UNIVERSITAT DE BARCELONA

Proposta de preservació de dades científiques en accés obert mitjançant tècniques d'anàlisi forense digital

Teodoro Wilderbeek López del Castillo

UNIVERSITAT DE BARCELONA

Facultat de Biblioteconomia i Documentació

Programa de doctorat:

Informació i Documentació en la societat del coneixement

Proposta de preservació de dades científiques en accés obert mitjançant tècniques d'anàlisi forense digital

Tesi doctoral

Teodoro Wilderbeek López del Castillo

Director de tesi: Dr. Miquel Térmens Graells

Barcelona, juny de 2017

Títol de la tesi:	Proposta de preservació de dades científiques en accés obert mitjançant tècniques d'anàlisi forense digital
Doctorand:	Teodoro Wilderbeek López del Castillo
Programa de doctorat:	"Informació i Documentació en la societat del coneixement" del Departament de Biblioteconomia, Documentació i Comunicació Audiovisual. Universitat de Barcelona
Director de la tesi:	Miquel Térmens Graells. Departament de Biblioteconomia, Documentació i Comunicació Audiovisual. Universitat de Barcelona

SUMARI

A la memòria del meu pare (1949-2016)

Agraïments

Acabar d'escriure aquesta tesi sens dubte ha estat una fita que no estava segur d'aconseguir quan el Dr. Térmens em va animar a matricular-me al programa de doctorat, un cop vaig finalitzar el Treball Final d'Estudis del Màster en Gestió de Continguts Digitals al 2013, on també vaig abordar la disciplina de l'anàlisi forense digital. Durant aquests últims quatre anys, per tant, he abordat la redacció de la tesi doctoral on he patit diverses dificultats per canvis en la meva vida personal i professional, però tots aquests esforços han donat bon fruit.

Gràcies en primer lloc al Dr. Miquel Térmens, que m'ha proporcionat un enorme suport amb els seus consells en les nombroses tutories que hem tingut. Sense els seus ànims i la seva fe en el meu treball aquesta tesi no existiria.

En l'elaboració del treball de recerca he tingut l'oportunitat de realitzar entrevistes a diversos encarregats/des de repositoris. Vull donar les gràcies per la seva atenció i amabilitat a Domingo Iglesias, Ignasi Labastida i Judit Casals (UB), Cristina Azorín (UAB), Ciro Llueca, Francesc March i Rosa Padrós (UOC), Antonio Juan Prieto, Jordi Prats i Anna Rovira (UPC) i Ricard de la Vega (CSUC).

Durant aquesta època com a doctorand el Departament ha organitzat diverses activitats formatives. Dono les gràcies a tots els professors que m'han ensenyat diverses tècniques i eines de recerca, amb una menció especial al Dr. Mario Pérez-Montoro per les seves frases d'ànim durant els moments en què ens hem trobat.

Vull agraïr finalment a la meva mare i als meus germans pel seu suport i la seva paciència, especialment en aquest últim any, on vaig fer l'esforç final per finalitzar el redactat i em preguntaven constantment pel desenvolupament de la tesi. En aquest respecte, el meu germà Francisco es troba actualment en el seu tercer any de doctorat i espera poder completar el seu treball de recerca l'any vinent.

1. Introducció

1.1 Presentació

Ja fa temps que les agències de finançament de la recerca són conscients de la importància d'una bona gestió de les dades de recerca per tal que aquesta es pugui compartir amb la societat. Els beneficis que dóna aquesta compartició són amplis: la transparència dels resultats científics i de la metodologia facilita la replicació i la verificació dels mateixos, facilita la participació d'altres investigadors, redueix costos perquè evita la duplicació de dades i ajuda a difondre la recerca entre la comunitat científica, entre altres avantatges. Les agències de finançament ja ho contemplen com una qüestió essencial; per exemple, tenim el cas de el NIH on la compartició de dades representa transformar els resultats de la recerca en coneixement, productes i procediments que puguin millorar la sanitat.

Per tal de poder compartir les dades de recerca en accés obert per a tothom, dites agències requereixen que els investigadors elaborin un pla de gestió de la recerca o *Data Management Plan*, un document on l'investigador descriu què farà amb les seves dades durant el projecte de recerca i després d'haver completat dita recerca. Un pas important és el de la preservació de les dades a llarg termini per així facilitar la seva compartició i aquí entren en joc els repositoris, ja que són els llocs més utilitzats per dipositar dades, encara que no els únics, atès que algunes publicacions científiques estan començant a requerir als investigadors que enviïn les dades científiques que hagin fet servir per arribar a les conclusions presentades als articles presentats.

Per tant, es planteja la qüestió de resoldre el repte de gestionar i preservar les dades científiques, les quals inclouen un nombre enorme de formats i de tipus de dades. Per exemple, hi ha dades que genera un investigador d'astrofísica que no tenen res a veure amb les que genera un investigador de ciències socials. Per altra banda, els fitxers multimèdia tenen unes necessitats de preservació molt diferents als fitxers de text o als tabulars. Gran part d'aquestes dificultats tècniques es poden resoldre amb l'aplicació de tècniques d'anàlisi forense digital.

L'anàlisi forense digital és una metodologia que es centra en l'ús de maquinari i programari per recollir, analitzar, interpretar i presentar informació de fonts digitals i donar garanties que la informació que s'ha recollit no s'ha alterat en el procés

(Kirschenbaum; Ovenden; Redwine, 2010). Aquesta tècnica ja s'utilitza en algunes biblioteques i arxius per a la preservació de dades, ja que assegura la integració i la no modificació de les mateixes.

1.2 Justificació i delimitació de l'estudi

Els nous requeriments de les agències de finançament a la recerca pel que respecta a la compartició de dades, així com una nova cultura d'accés obert a les dades amb suport governamental, exigeixen que els investigadors estiguin preparats per gestionar les seves dades i així fer-les aptes per a la seva reutilització i preservació a llarg termini. Aquesta recerca vol contribuir als passos que s'han produït en aquesta direcció, amb la presentació d'una proposta de solució que pot servir a qualsevol entitat on es preservin dades de recerca. En aquest respecte es tracta d'una continuació del treball realitzat en el treball final d'estudis de màster *Disseny d'una unitat d'anàlisi forense digital en una biblioteca* (Wilderbeek, 2013).

El context en què s'ha realitzat la tesi ha estat en primer lloc, per tal de tenir uns fonaments vàlids per a la nostra proposta de preservació digital, el de realitzar un estat de la qüestió en tres matèries: dades de recerca en accés obert, preservació digital i anàlisi forense digital. En segon lloc, un cop fet aquest estat de la qüestió, s'ha realitzat la recerca per tal d'assolir aquesta solució. A continuació es descriuen, capítol per capítol, els límits que s'han aplicat a cadascun dels temes i es justifiquen les decisions emprades.

1.2.1 Dades de recerca en accés obert

La intenció d'aquest capítol ha estat fer una panoràmica general d'un moviment relativament nou, l'accés obert de les dades de recerca fruit de les investigacions amb finançament públic. Un cop s'ha introduït el concepte i els passos que han fet diferents institucions per introduir la necessitat de l'*open data*, s'han analitzat les polítiques de les agències de finançament més importants a nivell europeu, nord-americà i espanyol per

tal d'exposar què necessita fer l'investigador per dipositar les seves dades de recerca. Aquest estudi s'ha limitat a les agències més rellevants i conscienciades amb el moviment *open data*, sense limitacions de disciplines o àrees de coneixement. És per aquesta raó que s'analitzen institucions tan dispars com el NIH, el NSF o Wellcome Trust.

En la mateixa línia de requeriments de les agències, s'ha fet una anàlisi de formats preferits i acceptats de les dades de recerca dins una mostra de quatre repositoris, dels quals dos es troben inclosos als dipòsits digitals que utilitzen els investigadors becats per les agències analitzades. La raó de fer aquesta anàlisi ha estat donar un estat de la qüestió dels formats de fitxer aptes per a la preservació, la qual cosa ha ajudat a la creació del model.

L'últim apartat ha tractat sobre el marc legal a tenir en compte quan es gestionen dades de recerca, atès que els drets d'autor són de gran importància dins la preservació digital, així com la privacitat i el dret a l'oblit.

1.2.2 Preservació digital

La preservació digital és un tema enormement ampli, així que dins aquesta tesi ens hem concentrat en tres qüestions: les possibilitats que tenen els investigadors per dipositar i compartir les seves dades de recerca, per tal d'oferir una panoràmica de les opcions més rellevants i així demostrar que el moviment de dades obertes de recerca no es tracta d'una qüestió marginal; l'estàndard de preservació OAIS i el seu estàndard d'auditoria i certificació corresponent, TDR, ja que el model de preservació que presentem en aquesta tesi es basarà en conceptes OAIS; i per últim, els esquemes de metadades que es faran servir en el nostre model: Dublin Core, DFXML, PREMIS i BagIt. Els motius de presentar aquests quatre és que són els més adequats i coherents, degut a les tècniques que farem servir per preservar les dades i a les característiques que presentarà el model.

1.2.3 Anàlisi forense digital

Hem volgut exposar la metodologia que presenta aquesta disciplina per tal de demostrar la seva adequació per preservar i garantir la integritat de les dades. Com a demostració pràctica d'aquesta metodologia, s'han presentat diversos casos d'ús de les tècniques forenses a institucions on s'han preservat amb èxit materials d'origen digital provinents de col·leccions i arxius privats d'escriptors i organitzacions no governamentals, sense ànim de ser una mostra exhaustiva, però sí suficientment representativa. Com aplicació directa d'aquestes tècniques, i aprofitant la disponibilitat de l'entorn forense BitCurator, s'han realitzat unes proves directament amb dades procedents del material de documentació que ha fet servir l'autor per a aquesta tesi. Aquestes proves han servit per definir el flux de treball del model de preservació.

1.2.4 Model de preservació

Dins les dades científiques podem trobar qualsevol tipus de disciplina: astronomia, biologia, enginyeria, matemàtica, etc. En moltes d'elles podem trobar formats altament especialitzats, els quals requereixen un alt nivell de coneixement de la matèria. S'ha decidit contemplar només les dades científiques en ciències socials i humanitats degut al millor coneixement de l'autor en aquestes disciplines que en d'altres com ciències humanes i així poder simplificar el procés d'investigació. S'han exclòs també totes aquelles disciplines amb presència de metodologia quantitativa (com és el cas de l'estadística), i contemplar disciplines amb metodologia qualitativa amb dades no estructurades. Les raons principals foren les enormes complicacions tècniques per crear un model de preservació amb dades estructurades i la manca de recursos disponibles.

El model de preservació que s'ha elaborat en aquesta tesi està basat en l'estàndard OAIS que descriu a un alt nivell com es preserven els continguts i dades digitals als repositoris (Reilly Jr.; Waltz, 2014, p. 110), en el qual s'han aplicat les tècniques d'anàlisi forense digital per preservar col·leccions de dades de recerca en un repositori programat sota DSpace. La raó per utilitzar DSpace és que es tracta del programari més utilitzat per crear repositoris en accés obert, tal i com es mostra a la Taula 1, en què s'ha fet una

recerca dins els directoris OpenDOAR[1] i ROAR[2]. A més, compta amb el suport d'una comunitat important d'usuaris (Lee; Stvilia, 2017).

Taula 1. Programaris més utilitzats en els repositoris d'accés obert (setembre de 2016)

Programari	Nombre d'ocurrències al directori OpenDOAR	Nombre d'ocurrències al directori ROAR
DSpace	1414	1696
Eprints	437	602
Fedora	51	52
Bepress	2	409
Digital Commons	158	-
CONTENTdm	56	10

Fonts: OpenDOAR. <http://www.opendoar.org>. [Consulta:05/09/2016]; ROAR. <http://roar.eprints.org> [Consulta:05/09/2016]

L'elaboració del model ha consistit en un flux de treball detallat amb totes les operacions necessàries on s'han utilitzat tècniques d'anàlisi forense digital i per aplicar-lo a nivell teòric a un repositori DSpace s'han elaborat una sèrie d'entrevistes a responsables institucionals i tècniques de diferents universitats catalanes.

1.3 Definicions

Es defineixen els conceptes clau de la terminologia relacionada amb les dades que es tractaran a la tesi i en últim lloc l'anàlisi forense digital, atès que és una disciplina nova dins les ciències de la informació.

1.3.1 Dades obertes

La iniciativa *open data* es troba dins una filosofia i pràctica que té com objectiu fer que es trobi disponible un cert nombre de dades de forma lliure a tothom, sense cap tipus de restricció. Es pot contemplar com "un movimiento que promueve la liberación de datos, generalmente no textuales y en formatos reutilizables como csv (*comma separated*

[1] < http://www.opendoar.org/index.html>. [Consulta: 05/09/2016]

[2] <http://roar.eprints.org>. [Consulta: 05/09/2016]

values), procedentes de organizaciones diversas" (Peset; Ferrer-Sapena; Subirats-Coll, 2011). La fundació CTIC considera l'*open data* com dades exposades en un format obert i estàndard, amb una estructura que permet la seva utilització dins serveis i aplicacions a qualsevol dispositiu electrònic (Ferrer-Sapena; Peset; Aleixandre-Benavent, 2011). Joel Gurin[3], assessor de la New York University, valora les dades obertes com dades públiques accessibles que poden ser utilitzades per persones, empreses i organitzacions per posar en marxa nous projectes, analitzar patrons i tendències i resoldre problemes complexos. Les dades han d'estar disponibles públicament per a tothom i han de tenir una llicència que permeti la seva reutilització. La definició més formal la trobem a l'Open Knowledge Foundation[4]: "Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share-alike".

En resum, les característiques més importants que han de complir les dades serien les següents:

- Disponibilitat i accés: les dades han d'estar disponibles de forma íntegra per un cost que no sigui superior al de la seva reproducció, preferentment mitjançant descàrrega per Internet. Les dades han d'estar disponibles en una forma convenient i modificable
- Reutilització i redistribució: les dades s'han de distribuir sota unes condicions que permetin la seva reutilització i la seva redistribució incloent la combinació amb altres conjunts de dades
- Participació de tothom: qualsevol ha de poder utilitzar, reutilitzar i redistribuir, sense discriminacions contra cap grup, persona o àrees d'activitat. Per tant, no es permetrien restriccions d'ús de les dades (per exemple, amb una limitació d'ús educatiu) o restriccions d'ús comercial

Existeixen molts tipus de dades obertes, però hi ha dos grups principals: les dades obertes de recerca o *open research data* i les dades obertes governamentals o *open*

[3] Gurin, Joel (2014, Apr. 15). "Big data and open data: what's what and why does it matter?". *The Guardian*. <http://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government/>. [Consulta: 17/05/2015]

[4] <https://okfn.org/>. [Consulta: 31/08/2015]

government data. És important diferenciar aquests conceptes, atès que aquesta tesi només inclourà el primer grup com a objecte d'estudi.

1.3.2 Dades de recerca "finals"

Tot i els múltiples tipus de dades que es poden trobar dins les disciplines científiques, hi ha un consens general entre els investigadors vers allò que s'entén com a "dades de recerca" (Guy; Donnelly; Molloy, 2013). La Monash University, al seu pla estratègic de gestió de dades[5] empra la següent definició: "Data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise a research project's observations, findings or outcomes, including primary materials and analysed data". La University of Bristol[6], per la seva banda, concreta una mica més:

> Data, or units of information which are created in the course of funded or unfunded research, and often arranged or formatted in a such a way as to make them suitable for communication, interpretation, and processing, perhaps by a computer. Examples of research data may include a spreadsheet of statistics, a series of email messages, a sound recording of an interview, a descriptive record of a rock specimen, or a collection of digital images. Research data does not include data generated in the course of personal activities, desktop or mailbox backups, or data produced by non-research activities such as University administration and teaching.

La NIH[7] fa una distinció important, ja que les dades que considera susceptibles de ser compartides amb la comunitat científica les anomena *final research data*:

> Recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Final research data do not include laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future

[5] Beitz, Anthony; Dharmawardena, Kheeran; Searle, Sam (2012). *Research data management strategy and strategic plan 2012–2015.* [Melbourne?]: Monash University. < https://goo.gl/ZFhpf6>. [Consulta: 31/05/2017].

[6] *Research data management glossary*. <http://vocab.bris.ac.uk/data/glossary/>. [Consulta: 30/08/2015].

[7] NIH (2004, Feb.16). *Frequently Asked Questions (FAQs) on data sharing*. <http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm#901>. [Consulta: 12/06/2016]

> research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens.

L'OECD (2007) es troba en la mateixa línia que la NIH:

> Factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research dataset constitutes a systematic, partial representation of the subject being investigated. This term does not cover the following: laboratory notebooks, preliminary analyses, and drafts of scientific papers, plans for future research, peer reviews, or personal communications with colleagues or physical objects (e.g. laboratory samples, strains of bacteria and test animals such as mice).

Així doncs, dins aquesta tesi només es consideraran dades de recerca "finals" aquelles que siguin necessàries per validar resultats científics i que per tant són aptes per ser compartides amb altres investigadors. Quedarien exclosos materials com notes, anàlisis preliminars, comunicacions personals o conjunts de dades parcials, que serien dades en "brut" i no processades.

1.3.3 Dades de recerca "en brut" i material de suport

Ara que queda establert que les dades de recerca finals són els *datasets* definitius, trobem el problema de com gestionar i preservar les dades de recerca "en brut", ja que es tracta habitualment d'un gran volum d'informació i de tipologies de fitxers, els quals corren un gran risc de pèrdua si no es prenen les mesures necessàries. S'inclouen en aquest grup les notes, anàlisis, vídeos, àudios i observacions que s'hagin fet al procés del projecte de recerca. Hi ha un gran risc de pèrdua d'aquests tipus de dades en el futur, ja que no sempre es requereix la seva preservació i/o dipòsit a un repositori/publicació científica, i per tant els científics estan perdent dades a un ritme alarmant, com per exemple a les disciplines de l'agricultura o de la biologia[8].

[8] Gibney, Elizabeth; Van Noorden, Richard (2013, Dec. 19). "Scientists losing data at a rapid rate". *Nature News*. <http://dx.doi.org/10.1038/nature.2013.14416>. [Consulta: 22/07/2014].

El material de suport també és rellevant, ja que aporten un context a la recerca i ajuda a entendre les dades. Aquest material pot prendre moltes formes (Ridge, 2014, p. 61), com comunicacions per correu electrònic, contractes de confidencialitat, programari especialitzat, diccionaris de dades o presentacions. De la mateixa manera que amb les dades en brut, aquest material no sempre es requereix per ser preservat, amb el subseqüent risc de pèrdua.

1.3.4 Compartició de dades de recerca

La compartició de dades o *data sharing* es defineix com un investigador que facilita l'accés a les seves dades mitjançant diverses vies com la seva publicació, difusió, o concessió d'accés públic per poder visualitzar, accedir o fer ús de les seves dades (Carlson, 2012). En aquesta línia, Lawrence et al. (2011) defineixen la publicació de dades com "To make data as permanently available as possible on the Internet", mentre que Borgman (2012) considera la compartició de dades com alliberar dades de recerca per tal de ser utilitzades per altri. Exemples de mètodes de compartició de dades són la inclusió de *datasets* a articles publicats, pujar dades a pàgines web institucionals o personals, dipositar *datasets* a repositoris o enviar dades en resposta de sol·licituds d'altres investigadors (Kim;Youngseek; Stanton, 2012; Wallis; Rolando; Borgman, 2013). Aquestes dades de recerca poden ser finals o en brut, tal com hem definit més amunt.

Els avantatges del *data sharing* inclouen (Tenopir et al., 2011) la verificació de resultats mitjançant la repetició d'anàlisi de les dades, poder aportar una nova interpretació o punt de vista a les dades existents, conservar la integritat de les dades amb una preservació a llarg termini i ben gestionada o optimitzar l'ús de recursos. Altres utilitats són que la disponibilitat de dades permet als investigadors la replicació i verificació de treballs de recerca, així com el seu desenvolupament (Andreoli-Versbach; Mueller-Langer, 2014) i que el l'investigador obté un reconeixement del seu treball gràcies a la citació de les seves dades (Dallmeier-Tiessen et al., 2014).

Un exemple dels seus avantatges fou el descobriment (Levelt Committee; Noort Committee; Drenth Committee (Eds.), 2012) del frau científic de Diederik Stapel, que

va falsificar dades a 55 articles publicats a revistes de psicologia i com a resultat d'això van concloure que les dades de recerca "must remain archived and be made available on request to other scientific practitioners". Un altre exemple seria que la compartició de dades minimitza la seva pèrdua i ajuda a la preservació (Vines et al., 2014), especialment si tenim en compte l'alta volatilitat de les dades a mesura que avança el temps[8].

Tot i els seus avantatges, la compartició de dades presenta diferents problemes com l'existència de les dades "fosques", definides per Heidorn (2008) com "not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost". Un exemple de dades fosques el trobem al Water Quality Field Station, una estació de recerca del Purdue Agronomy Center for Research and Education[9], on les dades que es generen no es trobaven accessibles, tot i els desitjos dels seus investigadors principals (Carlson; Stowell-Bracke, 2013). Les raons d'aquesta situació eren diverses: mancança de polítiques quant la gestió de dades, desconeixement entre els estudiants i investigadors de l'existència de repositoris que acceptin dades d'agronomia o falta de col·laboració amb les Purdue University Libraries per crear plans de preservació per a *datasets*. Per tant, es pot dir que no és suficient que l'investigador tingui la voluntat de compartir les seves dades, ja que necessita el suport de la institució o bé d'un tercer per facilitar el dipòsit.

1.3.5 Accés obert

En els últims temps, s'han fet grans esforços per facilitar l'accés obert als articles científics gràcies a la Iniciativa Budapest per a l'accés obert a l'any 2002[10], la Declaració de Bethesda sobre Publicació d'Accés Obert de 2003[11] i la Declaració de Berlin sobre accés obert del mateix any[12]. Actualment, els autors que vulguin publicar

[9] <https://ag.purdue.edu/agry/acre/Pages/default.aspx>. [Consulta: 08/12/2016]

[10] <http://www.budapestopenaccessinitiative.org/translations/spanish-translation>. [Consulta: 25/04/2015]

[11] <http://ictlogy.net/articles/bethesda_es.html>. [Consulta: 25/04/2015]

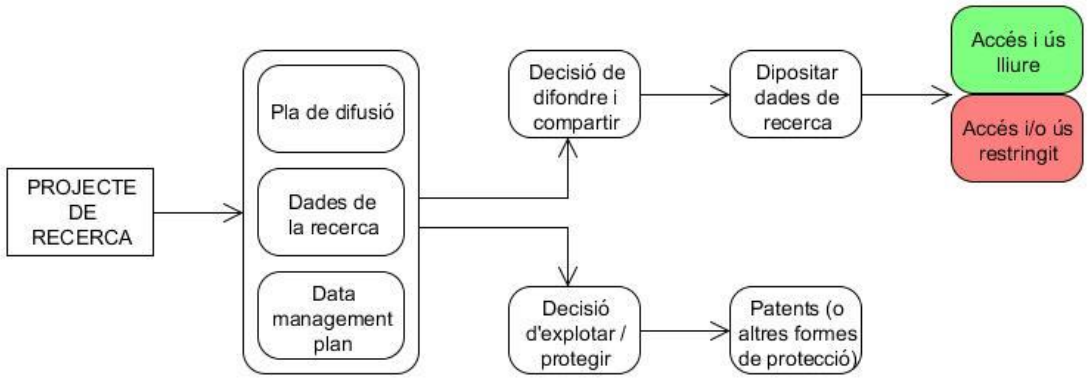
[12] <http://openaccess.mpg.de/67627/Berlin_sp.pdf>. [Consulta: 25/04/2015]

en accés lliure utilitzen dues vies principalment: les revistes en accés lliure (o "via daurada") i l'auto-arxiu (o "via verda").

Ara bé, una via addicional per difondre el coneixement que és especialment necessària per als científics és l'accés obert a les dades en format digital que es produeixen dins els processos de recerca. La Declaració de Berlin representà un gran pas en aquest aspecte, ja que considera com contribucions d'accés obert els resultats de la recerca científica, les dades en brut i les seves metadades, els materials originals, les representacions digitals de materials gràfics i el material multimèdia acadèmic, per la qual cosa no es limita a l'accés obert de la literatura científica com la Iniciativa Budapest i la Declaració de Bethesda.

Per tant, es defineix l'accés obert a les dades científiques com el dret d'accés i a la reutilització de les dades de recerca en format digital sota els termes i condicions estipulats a l'acord establert per l'agència de finançament (European Commission, 2016c). Les dades que estiguin obertament accessibles poden ser explotades, reproduïdes i difoses lliurament sense cap tipus de càrrega per a l'usuari. La Figura 1 mostra el procés que ha de travessar un projecte de recerca fins que aquest arriba a ser d'accés obert.

Figura 1. Accés obert de les dades de recerca en un context de difusió i publicació



Font: European Commission, 2012b. Traducció de l'autor

Un primer estudi de la Comissió Europea sobre l'accés obert a les dades de recerca en diverses disciplines (energia, ciències ambientals, ciències de la salut, tecnologies de la informació i la comunicació, infraestructures de recerca, ciència a la societat i ciències

socials i humanitats) va mostrar una bona predisposició a un mandat d'accés obert (European Commission, 2011b). Segons un segon estudi (European Commission, 2012a), l'accés obert rep una bona acollida a la Unió Europea, ja que un 90% dels agents implicats a la recerca (investigadors, agències de finançament o biblioteques) opinen que les dades de recerca que hagin resultat d'un finançament públic haurien d'estar disponibles en accés obert en Internet. En el cas d'un finançament mixt públic i privat, és un 72%.

1.3.6 Dataset

Encara que no existeix una definició formalitzada, hi ha un consens general dins la literatura científica i tècnica vers les característiques comunes que tenen els conjunts de dades o *datasets* (Renear; Sacchi; Wickett, 2010):

- Agrupació. Els *datasets* són conjunts de dades tractats de forma col·lectiva, com a unitat
- Contingut. Les dades dels *datasets* tenen un contingut comú, com per exemple resultats d'observacions
- Associació. Les associacions que es poden trobar dins les dades poden ser de diversos tipus, com el context en què les observacions s'han fet, una estructura comuna o el tema o matèria que tracten
- Propòsit. Tot conjunt de dades existeix per a un propòsit concret: contribuir a l'activitat científica

Així doncs, un *dataset* representa una col·lecció de dades reunides durant l'execució d'un projecte de recerca i poden comprendre diferents elements o tipus de dades, com text, fulls de càlcul, gràfics o imatges. El repositori Dryad utilitza el terme "paquet de dades" o *data package* per fer referència a un conjunt de fitxers de dades associat a una publicació, mentre que el repositori Figshare utilitza *dataset* per indicar dades i *fileset* per indicar un grup de fitxers múltiples que es poden citar com a un sol objecte (Assante et al., 2016).

1.3.7 Metadades i estàndards

La definició bàsica de metadades és "dades sobre dades". El sufix "meta" prové de la paraula grega que significa "juntament, amb, després de, a continuació". Per tant, les metadades es poden definir com "structured data about other data" (National Archives of Australia, 2010).

Les metadades s'utilitzen per descriure informació bàsica sobre el *dataset* que habiliten la seva indexació i recuperació. Els estàndards de metadades són expressions formals dels elements de metadades, amb especificacions sobre quina informació és l'apropiada per a certs tipus de dades i com s'han de documentar[13]. Alguns exemples són el Dublin Core Metadata Terms, Data Document Initiative (DDI) o Metadata Encoding and Transmission Standard (METS).

Com a mínim, s'han de proveir els camps suficients que permetin la citació d'un *dataset*, que serien:

- Creador. Productor principal de les dades o autors de la publicació on es troben les dades
- Títol. Nom pel qual es coneix el *dataset*
- Editor. Posseïdor del *dataset*
- Any de publicació. Any en què es van fer disponibles les dades
- Identificador. Identificador únic per a les dades

Les MIT Libraries[14] recomanen l'addició dels següents camps per a *datasets* de qualsevol disciplina:

- Dates. Dates clau associades amb les dades, com la de l'inici i de la fi del projecte, modificació, publicació, i període de temps cobert per les dades
- Matèria. Paraules clau o frases que descriguin la matèria o contingut de les dades
- Patrocinadors. Organitzacions o agències que han finançat la recerca

[13] *Metadata and Standards*. < https://goo.gl/qCi7fc>. [Consulta: 01/04/2016]

[14] *Documentation & metadata*. <https://libraries.mit.edu/data-management/store/documentation/>. [Consulta: 01/04/2016]

- Drets. Drets de propietat intel·lectual relacionats amb les dades
- Ubicació. Si les dades es refereixen a una ubicació física, s'ha de proveir informació sobre el lloc cobert
- Metodologia. Com es van generar les dades, incloent l'equipament o el programari que es va utilitzar, protocol per als experiments i altres

En funció del tipus de política de finançament que s'apliqui a l'investigador, aquest haurà d'utilitzar camps concrets, com els formats i el programari que s'han utilitzat al projecte de recerca.

Els registres de metadades poden ser recol·lectats mitjançant el protocol OAI-PMH[15], que habilita l'intercanvi de metadades entre proveïdors de dades (repositoris amb metadades estructurades) i proveïdors de serveis (sol·licituds de serveis que recol·lecten metadades).

1.3.8 Pla de gestió de dades

Diverses entitats, com la Rutgers University[16] o la University of Virginia[17] defineixen el DMP (*Data Management Plan*) o pla de gestió de dades de la següent manera: "Formal document that outlines what you will do with your data during and after you complete your research". Es tracta d'un document on l'investigador descriu què farà amb les seves dades durant el projecte de recerca i després d'haver completat dita recerca.

La preparació d'un DMP ofereix moltes avantatges com l'estalvi de temps, l'increment de l'eficiència dins la recerca i, molt important, compleix un requeriment que exigeixen moltes agències de finançament per a la recerca, com la NSF, agència federal dels EUA creada el 1950 amb l'objectiu, entre altres, "to promote the progress of science"[18]. En l'àmbit del Regne Unit, per tal de sol·licitar subvencions a la major part dels Research

[15] <https://www.openarchives.org/pmh/>. [Consulta: 01/04/2016]

[16] *Data management plans.* <https://data.rutgers.edu/Home/Plans>. [Consulta: 26/03/2016]

[17] University of Virginia Library Research Data Services (2016). *Data management components.* <http://data.library.virginia.edu/data-management/plan/>. [Consulta: 26/03/2016]

[18] *About the National Science Foundation*. <http://www.nsf.gov/about/>. [Consulta: 08/01/2014]

Councils és necessari fer un DMP[19]. Per poder desenvolupar-lo, s'ha d'incloure la següent descripció mínima de les dades:

- Context, descripció del projecte i propòsit de la recerca; metodologia
- Naturalesa de les dades, història de les dades, contingut i estructura, terminologia, programari, data de creació i dates de modificació, versions, responsables i participants
- Formats de fitxers, estructura i nomenclatura dels fitxers, sistema de emmagatzematge, procediment per a còpies de seguretat
- Aspectes legals, polítiques d'accés i seguretat

La NSF preveu que l'elaboració i l'avaluació dels DMP creixerà i canviarà amb el temps, gràcies a l'avaluació d'experts (o *peer review*), a directrius concretes per a diferents disciplines del coneixement (atès que el concepte d'intercanvi de dades és diferent a algunes disciplines i no es pot fer un mateix tipus de pla per a qualsevol tipus de projecte) i a què es dedicarà personal especialment per a l'avaluació dels DMP[20]. Un programari útil per a l'aplicació dels DMP és el DMPTool[21], desenvolupat per la California Digital Library, que permet generar-ne un pas per pas. El seu ús és gratuït i obert a tothom.

1.3.9 Repositori digital

Un repositori digital és un arxiu en línia en el qual els autors i els acadèmics poden dipositar el seu treball, amb la intenció de deixar-lo disponible de manera oberta en format digital (Pappalardo; Fitzgerald, 2007). Aquest terme també es pot referir a l'organització responsable del manteniment a llarg termini dels recursos digitals i de fer que aquests recursos es trobin disponibles al públic o a comunitats específiques

[19] *FAQ on Data Management Plans*. <http://www.dcc.ac.uk/resources/data-management-plans/faq-dmps>. [Consulta: 08/01/2014]

[20] Zhang, Tao (2013). *DMPTool: expert resources and support for data management planning.* <https://hubzero.org/resources/1062>. [Consulta: 30/08/2016]

[21] <https://dmp.cdlib.org/>. [Consulta: 27/03/2014]

d'usuaris[22]. Hi ha molts tipus diferents de repositoris, però es poden diferenciar quatre grans grups (Armbruster; Romary, 2010):

- Repositoris especialitzats en matèries. Habitualment s'estableixen per membres de la comunitat i s'utilitzen per fer autoarxiu
- Repositoris de recerca. Patrocinats per agències de recerca o organitzacions que obliguen a publicar els resultats, validen els resultats de la recerca
- Repositoris nacionals. Coordinats per administracions públiques i utilitzats per justificar inversions nacionals en recerca i permetre l'accés públic al coneixement
- Repositoris institucionals. Recullen les obres i materials de recerca dels diversos projectes i actuacions de la institució

1.3.10 Anàlisi forense digital

El concepte *digital forensics* (o en altres àmbits *computer forensics*) ha rebut moltes definicions. Kruse i Heiser (2001) ho van identificar com una pràctica aplicada que implica "preservation, identification, extraction, documentation, and interpretation of computer data". Palmer (2001) utilitzà una definició que serveix perfectament per justificar els objectius pràctics d'aquesta tesi i que és acceptada per Gengenbach (2012), Guarino (2013) i Reith, Carr i Gunsch (2002):

> The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.

L'anàlisi forense digital es centra en l'ús de maquinari i programari per recollir, analitzar, interpretar i presentar informació de fonts digitals i donar garanties que la informació que s'ha recollit no s'ha alterat en el procés. Per tant, aquest sistema pot

[22] *Attributes of a trusted digital repository.* <http://www.cni.org/wp-content/uploads/2013/05/Attributes01-RDale2001Ftf.pdf>. [Consulta: 04/05/2014]

facilitar al centre de recerca la preservació i assegurar la integritat de les dades, tal com ja s'ha demostrat a diverses institucions culturals (Gengenbach, 2012; Kirschenbaum et al., 2010; Lee et al., 2013), ja que la tècnica permet crear còpies autèntiques de les dades presents als suports, establir cadenes de custòdia fiables, localitzar i presentar informació contextual i identificar informació privada i sensible que s'hauria de filtrar, bloquejar o redactar (Lee, 2014).

1.4 Hipòtesi

La gestió de les dades de recerca és una realitat que paulatinament s'està fent obligatòria a nivell mundial, degut entre altres motius als requeriments d'agències de finançament, privades com públiques. Aquesta gestió requereix de forma forçosa preservar de forma íntegra les dades, facilitar un accés lliure, que no permeti alteracions que puguin afectar els resultats de la recerca i que protegeixi les dades confidencials dels subjectes investigats. Volem establir la següent hipòtesi:

Les tècniques d'anàlisi forense digital permeten preservar de manera eficaç dades de recerca de ciències socials i humanitats, tot tenint en compte:

- Els requeriments de les agències de finançament
- Les grans dimensions que poden tenir els *datasets*
- Aquests *datasets* són emmagatzemats en diferents suports físics
- La possible presència de dades confidencials

Aquesta hipòtesi es comprovarà a partir de la demostració dels següents paràmetres, que serviran per desenvolupar un model de preservació:

- Les agències de finançament requereixen el dipòsit de les dades de recerca i que aquestes siguin d'accés obert
- Les tècniques d'anàlisi forense digital permeten capturar les dades de recerca sense cap alteració en el seu contingut, fer-ne una anàlisi profunda i editar i/o bloquejar les dades confidencials

- Els repositoris poden donar accés obert a les dades de recerca de gran mida amb unes modificacions concretes

Un cop demostrats aquests paràmetres, s'elaborarà el nucli central de la nostra proposta: un flux de treball on s'empraran les tècniques d'anàlisi forense digital que permetrà preservar les dades de recerca en un repositori DSpace des del moment en què arriben a la institució fins al moment en què s'atorga l'accés final als usuaris.

1.5 Metodologia

La metodologia que s'ha utilitzat per tal de demostrar la hipòtesi que hem presentat al capítol 1.4 ha consistit en la realització de tres operacions principals:

- Estudi de la bibliografia disponible sobre la gestió i preservació de dades de recerca i sobre l'anàlisi forense digital
- Entrevistes personals amb responsables de repositoris
- Proves tècniques amb programari d'anàlisi forense digital

1.5.1 Anàlisi de la literatura

S'ha realitzat una anàlisi bibliogràfica amb l'objectiu de definir les necessitats existents quant a la preservació de dades de recerca, justificar quin estàndard de preservació utilitzarem com a base de la nostra proposta i presentar la metodologia de l'anàlisi forense digital i casos exitosos de preservació a biblioteques i arxius mitjançant aquesta tècnica. L'estudi bibliogràfic, per tant, ha pretès respondre a les següents qüestions:

- Quines són les polítiques de les agències de finançament pel que respecta a la gestió i la preservació de les dades de recerca?
- Quins són els formats digitals acceptats dins als repositoris i/o dipòsits de dades?
- Quin és el marc legal de les dades de recerca?
- Quines opcions té un investigador per dipositar les seves dades de recerca?

- Per què el model de referència OAIS és una bona opció com a base de la nostra proposta?
- Quins estàndards de metadades són vàlids per al nostre model?
- De quina manera s'han integrat les tècniques d'anàlisi forense digital als fluxos de treball de biblioteques i arxius?

Les fonts consultades han consistit en articles de revista, actes de congressos, informes, monografies i pàgines web. En aquest últim cas, les citacions s'han fet a peu de pàgina, mentre que amb la resta de fonts les citacions formen part de la bibliografia citada. En tots els casos de recursos accessibles via web s'ha afegit la data de consulta, mentre que en alguns casos en què s'ha consultat una adreça URL excessivament llarga s'ha utilitzat el servei Google URL Shortener[23] per escurçar-la i així poder utilitzar una citació més breu. Les fonts que s'han consultat inclouen:

- Investigadors i organitzacions especialitzats en la difusió i reutilització de dades de recerca en accés obert
- Agències de finançament de la recerca reconegudes internacionalment, com Horizon 2020, Wellcome Trust o el NIH
- Investigadors i organitzacions especialitzats en anàlisi forense digital aplicada a biblioteques i arxius
- Organitzacions especialitzades en el dipòsit de dades per a la preservació a llarg termini

1.5.2 Entrevistes a responsables tècnics i institucionals de repositoris

Per tal de poder contrastar les dades que s'han recollit de la bibliografia pel que pertoca a la preservació dins els repositoris, s'han realitzat sis entrevistes amb responsables de repositoris de diferents institucions amb seu a la província de Barcelona. Aquestes entrevistes no són representatives de tot el conjunt de repositoris, sinó que són una mostra que es limita a la província de Barcelona. Aquesta limitació geogràfica s'ha degut principalment a la manca de temps i recursos de l'autor per fer desplaçaments. No

[23] <https://goo.gl/>. [Consulta: 30/04/2017]

s'ha d'entendre, per tant, que els resultats siguin conclusius sinó més aviat orientatius; recordem que la nostra proposta s'ha elaborat per donar resposta a problemes de preservació de dades de recerca a qualsevol institució, no necessàriament institucions catalanes.

L'objectiu principal de l'entrevista era conèixer, per una banda, les opinions dels responsables sobre l'aplicació pràctica del model, i per altra, els processos aplicats als seus repositoris a la preservació de les dades de recerca. És per això que les entrevistes han consistit en un qüestionari obert, que es pot consultar a l'Annex C. A la Taula 2 presentem el llistat d'entrevistes que s'han realitzat on fem esment dels entrevistats i els seus càrrecs a la institució corresponent.

Taula 2. Llistat d'entrevistes

Institució	Responsable	Càrrec	Programari	Repositori	Data de l'entrevista
UB	Domingo Iglesias Sesma	Administrador informàtic	DSpace	Dipòsit Digital de la UB	9 de gener de 2017
UAB	Cristina Azorín Millaruelo	Administradora del repositori	Invenio	Dipòsit Digital de Documents de la UAB	30 de gener de 2017
UOC	Francesc March Mir	Administrador informàtic	DSpace	O2	1 de febrer de 2017
UOC	Rosa Padrós Cuxart	Administradora del repositori	DSpace	O2	1 de febrer de 2017
UOC	Ciro Llueca Fonollosa	Cap de biblioteca	DSpace	O2	1 de febrer de 2017
UB	Ignasi Labastida Juan	Cap de la unitat de la recerca del CRAI UB	DSpace	Dipòsit Digital de la UB	9 de gener de 2017
UB	Judit Casals Parladé	Cap de projectes del CRAI UB	DSpace	Dipòsit Digital de la UB	9 de gener de 2017
UPC	Antonio Juan Prieto Jiménez	Administrador informàtic	DSpace	UPCommons	14 de febrer de 2017
UPC	Jordi Prats Prat	Administrador del repositori	DSpace	UPCommons	14 de febrer de 2017
UPC	Anna Rovira Fernández	Cap d'Unitat dels Serveis Generals de Biblioteques	DSpace	UPCommons	14 de febrer de 2017
CSUC	Ricard de la Vega	Cap de Càlcul i Aplicacions	DSpace	Diversos	9 de març de 2017

Font: L'autor, a partir de les entrevistes realitzades

Per tal de realitzar les entrevistes, el Dr. Miquel Térmens es va posar en contacte amb cadascun dels responsables per concretar dia, hora i lloc de realització de l'entrevista. El desenvolupament de la mateixa va consistir en una petita presentació en Power Point vers la problemàtica que pretén resoldre aquesta tesi a tall d'introducció. A continuació es va requerir a cadascú dels entrevistats que signessin el seu consentiment per utilitzar les seves declaracions en l'elaboració d'aquesta tesi i finalment es van realitzar els qüestionaris, que van ser individuals o grupals segons el cas. En el cas d'una persona entrevistada, l'entrevista fou individual, mentre que va ser grupal en el cas de dues o més persones entrevistades.

1.5.3 Creació de la proposta de preservació

Per crear la proposta de preservació que permetés demostrar la hipòtesi presentada al capítol 1.4, es va fer una anàlisi del programari i maquinari disponibles per a l'anàlisi forense digital, així com un estudi de la terminologia de l'estàndard OAIS. Un cop escollits quin programari i maquinari utilitzaríem, es van realitzar una sèrie de proves tècniques per demostrar la viabilitat de la nostra proposta per a la preservació de dades de recerca. Per realitzar aquestes proves, el primer a fer fou adquirir un maquinari suficientment potent per complir amb els requeriments de l'entorn BitCurator i seguidament es va descarregar i instal·lar al maquinari l'entorn BitCurator mitjançant una memòria USB. De forma paral·lela, es van preparar tres suports físics per realitzar les proves de preservació.

Un cop completades aquestes proves i analitzats els resultats, es va passar a definir pas per pas els fluxos de treball de la nostra proposta. Per tal de tenir una base sòlida es va fer un estudi de casos d'ús d'anàlisi forense digital a biblioteques i a arxius, que presentem a la Taula 3.

Després d'haver definit els processos del nostre flux de treball, es van elaborar una sèrie de *scripts* programats en llenguatge Bash per tal d'executar les operacions de forma més

senzilla i eficient. Les fonts consultades per elaborar els *scripts* han estat els fòrums AskUbuntu[24] i la llista de correu BitCurator Users[25].

Taula 3. Llistat de casos d'ús estudiats d'anàlisi forense digital

Institució(ns)	Biblioteca/arxiu o projecte	Cas d'ús
British Library	Digital Lives	Desenvolupament de processos
University of Oxford	Bodleian Library	Preservació de l'arxiu de Barbara Castle
National Library of Australia	Prometheus	Desenvolupament d'un flux de treball automatitzat de preservació
Emory University	Manuscript, Archives, and Rare Book Library	Preservació de l'arxiu de Salman Rushdie, amb el desenvolupament d'un flux de treball bàsic
University of Hull	Hull University Archives	Preservació de l'arxiu de Stephen Gallagher, amb el desenvolupament d'un flux de treball detallat
Stanford University	Department of Special Collections and University Archives	Preservació de l'arxiu de l'organització Stop AIDS
University of Virginia	Albert and Shirley Small Special Collections Library	Desenvolupament d'un flux de treball detallat
Yale University	Beinecke Rare Book & Manuscript Library	Preservació de l'arxiu de George Whitmore, amb el desenvolupament d'un flux de treball detallat
University of North Carolina i Maryland Institute for Technology in the Humanities	BitCurator	Desenvolupament d'un programari adreçat a biblioteques i arxius
University of Montana	Maureen and Mike Mansfield Library	Preservació de l'arxiu de Patricia Goedicke

Font: L'autor, en funció dels casos estudiats al capítol 4.2

L'últim pas per la creació de la proposta ha consistit en estudiar el programari DSpace, on s'ha consultat diferent bibliografia publicada i pàgines web de repositoris creats amb aquest programari. Per tal de tenir una opinió millor formada sobre les possibilitats del programari, s'ha instal·lat un repositori DSpace de forma local a un maquinari propietat de l'autor amb sistema operatiu Windows 7 i s'han realitzat diferents ingestes de contingut. Per tal d'instal·lar DSpace, ha calgut la descàrrega i instal·lació addicionals dels següents programaris:

- Java Development Kit. Programari per a desenvolupadors del llenguatge Java

[24] <https://askubuntu.com/>. [Consulta: 30/04/2017]

[25] <https://groups.google.com/forum/#!forum/bitcurator-users>. [Consulta: 30/04/2017]

- Apache. Programari per instal·lar un servidor web
- PostgreSQL. Programari per a servidors de bases de dades relacionals
- TomCat. Contenidor de *servlets* que permet instal·lar un entorn de servidor web

Atès que la instal·lació de DSpace és una tasca que requereix de coneixements tècnics, s'han consultat dos fonts principals. La primera ha estat la wiki DuraSpace amb documentació del programari[26] i altra secundària per realitzar els passos concrets per instal·lar els servidors Apache i TomCat i de bases de dades PostgreSQL ha estat la Comunidad de Conocimiento Virtual[27], dedicada a la difusió de coneixement de ciències de la informació, on es detallen les instruccions concretes.

[26] <https://wiki.duraspace.org/display/DSDOC5x/DSpace+5.x+Documentation>. [Consulta: 01/05/2017]

[27] <http://conocimientovirtual.org/>. [Consulta: 01/05/2017]

2. Dades de recerca en accés obert

2.1 Orígens de l'*open research data*

El concepte d'accés obert a dades de recerca fou establert amb la formació del sistema World Data Center per preparar l'International Geophysical Year (National Research Council, 2008, p. 6), un projecte científic internacional que va estar actiu durant els anys 1957 i 1958 que tenia com objectiu fomentar l'intercanvi científic entre nacions que havia quedat interromput amb l'inici de la Guerra Freda. L'International Council of Science establí diversos World Data Centers l'any 1958 per arxivar i distribuir dades recollides dels programes observacionals de l'IGY.

No obstant, fou durant l'any 1995 quan va aparèixer per primera vegada el terme *open data* dins un context científic, amb la publicació de l'informe *Resolving conflicts arising from the privatization of environmental data* (National Research Council. Committee on Geophysical and Environmental Data, 2001) que fomentava un intercanvi total i obert d'informació científica: "In public-purpose environmental information systems a full and open data policy is optimal for collecting and synthesizing a wide range of observations, detecting scientific surprises, and avoiding or discovering processing or calibration errors". El mateix any, el GCDIS es posicionà a favor de les dades obertes en ciència amb l'informe *On the full and open exchange of scientific data*[28] on es declara el següent: "International programs for global change research and environmental monitoring crucially depend on the principle of full and open exchange (i.e., data and information are made available without restriction, on a non-discriminatory basis, for no more than the cost of reproduction and distribution [sic]". Aquesta última frase s'ha d'entendre en el context de 1995, on la xarxa d'Internet no tenia la capacitat actual de difusió d'informació.

Nou anys més tard, l'OECD va fer pública la seva *Declaration on access to research data from public funding*[29], on reconeix la importància de l'intercanvi internacional de dades, informació i coneixement per fer avançar la innovació i la recerca científiques i declara el seu compromís de treballar per establir formes d'accés per a les dades de recerca digitals provinents de finançament públic. El resultat final fou el document *OECD principles and guidelines for access to research data from public funding* (2007),

[28] <http://www.nap.edu/readingroom/books/exch/exch.html>. [Consulta: 17/05/2015]

[29] <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157>. [Consulta: 17/05/2015]

redactat com un seguit de recomanacions per als països membres per desenvolupar polítiques i bones pràctiques relacionades amb l'accessibilitat, ús i gestió de dades de recerca, que recull la següent declaració: "Sharing and open access to publicly funded research data not only helps to maximise the research potential of new digital technologies and networks, but provides greater returns from the public investment in research". Els principis són, resumits:

- Franquesa, entesa com accés en igualtat de condicions per a la comunitat de recerca internacional
- Transparència
- Conformitat legal
- Professionalisme
- Protecció de la propietat intel·lectual
- Interoperabilitat
- Qualitat i seguretat
- Eficiència
- Responsabilitat

Trenta defensors de l'accés obert es reuniren a Sebastopol (Califòrnia, EUA)[30] l'any 2007 amb l'objectiu de definir el concepte de dades públiques obertes, entre els quals es trobaven Tim O'Reilly que ha definit expressions com codi obert i Web 2.0 i Lawrence Lessig, fundador de les llicències Creative Commons. El resultat de la reunió fou la redacció dels vuit principis de les dades obertes[31]:

- Completes. Les dades públiques són aquelles que no estiguin subjectes a privacitat, seguretat o limitacions d'accés
- Primàries. Les dades es recullen de la font principal sense modificacions ni agregats
- Oportunes. Les dades estaran disponibles el més aviat possible per preservar el seu valor

[30] *Open Government Working Group*. <https://public.resource.org/open_government_meeting.html>. [Consulta: 02/05/2015]

[31] <http://opengovdata.org/>. [Consulta: 02/05/2015]

- Accessibles. Les dades estaran disponibles al major nombre d'usuaris possible, preferiblement mitjançant Internet
- Processades per ordinadors. Les dades estaran estructurades de forma que permetin un processat automatitzat
- Sense discriminacions. Les dades estaran disponibles per a tothom, sense requeriments de registres
- Sense propietari. Les dades estaran disponibles en un format sobre el qual cap entitat en té un control exclusiu
- Lliures de llicències. Les dades no estaran subjectes a cap dret d'autor, patent, marca registrada o regulació comercial. Sí que es permet l'aplicació d'una privacitat, una seguretat i unes restriccions d'accés que siguin raonables

Una altra entitat que promou l'accés obert a les dades és l'Open Knowledge Foundation. Fundada l'any 2004, té un grup de treball específic per a les dades obertes de recerca, l'Open Working Group on Open Data in Science[32], que actualment és una xarxa global d'investigadors, bibliotecaris i legisladors amb l'interès comú de dades obertes en ciència. Però, la definició de l'OKF per a les dades obertes (vegeu el capítol 1.3.1) es considerà massa àmplia i existia una necessitat de estendre aquesta definició amb una sèrie de principis específics al camp científic (Molloy, 2011).

Per tant, l'any 2009 tres membres de l'OKF (Rufus Pollock i Peter Murray-Rust de la University of Cambridge i Cameron Neylon del STFC) en col·laboració amb John Wilbanks de Creative Commons, van desenvolupar un primer esborrany dels Panton Principles[33] per publicar dades obertes de recerca que agafa com a base l'Open Definition i el protocol de Science Commons per implementar dades en accés obert[34]. Dit esborrany fou refinat per l'Open Working Group on Open Data in Science i foren publicades oficialment en febrer de 2010. La finalitat principal dels Principles és donar una sèrie de recomanacions per tal que les dades de recerca es puguin reutilitzar de forma òptima. En resum, els Principles recomanen:

[32] <http://science.okfn.org/about-us/>. [Consulta: 17/05/2015]

[33] <http://pantonprinciples.org/>. [Consulta: 17/05/2015]

[34] <http://sciencecommons.org/projects/publishing/open-access-data-protocol/>. [Consulta: 17/05/2015]

- Quan es publiquen dades, s'ha de fer una declaració explícita i sòlida dels desitjos de l'investigador
- S'ha d'utilitzar una llicència o renúncia de *copyright* apropiada per a dades
- Si es vol que les dades siguin utilitzades de forma efectiva per altres, han de ser obertes tal i com indica l'Open Definition; especialment, no han d'haver restriccions comercials ni altres de restrictives
- Es recomana especialment l'ús de llicències PDDL o CCZero, les quals garanteixen el compliment del protocol de Science Commons per implementar accés obert a dades i l'Open Definition (vegeu 2.4.2)

En aquesta mateixa línia, la Comissió Europea requerí també als seus Estats membres que defineixin polítiques clares per a la difusió i l'accés obert a les dades de recerca que resultin de recerca amb finançament públic i assegurar-se que dites dades es trobin accessibles, utilitzables i reutilitzables mitjançant infraestructures digitals (European Commission, 2012a).

Per últim, la Royal Society del Regne Unit també va mostrar la seva posició dins l'informe *Science as an open enterprise* (Royal Society, 2012), que recomana als científics que comuniquin les dades que recullen i els models que creen, per permetre l'accés lliure i obert a altres especialistes dins el mateix camp. Si les dades ho justifiquen, s'haurien de dipositar dins un repositori adequat i si és possible, s'hauria de possibilitar la comunicació de les dades a una audiència més àmplia, especialment dins àrees on la transparència és una prioritat.

2.2 Polítiques de les agències de finançament

Les agències de finançament de la recerca compleixen una funció vital per a la preservació de les dades, ja que en molts casos requereixen als investigadors que publiquin els resultats de la seva recerca en accés obert i que indiquin les condicions en què les dades es trobaran disponibles i accessibles a llarg termini. Això implica per força la preservació de les dades, bé dins un repositori o bé a un centre de dades.

La Unió Europea va fer el primer pas important en les polítiques d'accés obert a les dades l'any 2011 amb la publicació de l'Agenda digital europea[35], on es va insistir en la reutilització de les dades governamentals, patrimonials i científiques. Una comunicació de la Comissió Europea (2007) va ser cabdal en aquest sentit, ja que es va fer la declaració següent: "En principio, los datos de la investigación financiada íntegramente con fondos públicos deberían ser accesibles a todos, en consonancia con la Declaración Ministerial de 2004 sobre el Acceso a los Datos de la Investigación Obtenidos con Fondos Públicos de la OCDE". A més, aquesta comunicació també feia esments dels problemes que planteja la preservació de les dades a llarg termini.

En el cas dels EUA, el 22 de febrer de 2013 el govern del President Obama publicà un memoràndum que s'adreçava als departaments executius i agències de recerca federals amb pressupostos iguals o superiors al 100 milions de dòlars per tal que desenvolupessin un pla de suport a l'accés públic als resultats de la recerca finançada pels recursos governamentals[36]. Dins dels objectius d'aquest pla estava inclòs facilitar l'accés a les dades de recerca en format digital, assegurar que les agències desenvolupin els plans de gestió de dades, donar suport a la cooperació amb el sector privat per millorar l'accés a les dades i valorar la necessitat a llarg termini de la preservació de les dades de recerca dins les disciplines que controlin les agències. Aquests requeriments van augmentar amb la publicació d'un nou memoràndum, l'Open Data Policy[37], que obliga a les agències a "collect or create information in a way that supports downstream information processing and dissemination activities". A més, les agències han de construir o modernitzar els sistemes de tecnologies d'informació de manera que es maximitzi l'accessibilitat a la informació i es reforci la gestió de dades.

Analitzem a continuació les polítiques de les agències i/o institucions que més èmfasi han donat a la compartició i reutilització de dades de recerca (Eisenberg, 2006; Groves,

[35] *Digital Agenda: turning government data into gold*. <http://europa.eu/rapid/press-release_IP-11-1524_en.htm>. [Consulta: 26/03/2016].

[36] Holdren, John P. (comunicació, Feb. 22, 2013). *Increasing access to the results of federally funded scientific research.* Office of Science and Technology Policy. <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf>. [Consulta: 26/03/2017].

[37] Burwell, Sylvia M. *et al.* (comunicació, May 9, 2013). *Open data policy – Managing information as an asset.* Office of Management and Budget. <https://www.actiac.org/system/files/Open%20Data%20Policy%2C%20OMB%20m-13-13.pdf>. [Consulta: 26/03/2017].

2009; Higman; Pinfield, 2015; Mennes et al., 2013; Pitt; Tang, 2013; Tenopir et al., 2011, 2015; Wicherts; Bakker, 2012). S'aporta també un estat de la qüestió a Espanya.

2.2.1 Horizon 2020 (Comissió Europea)

L'any 2009, la Comissió Europea va encarregar a un grup d'experts l'elaboració d'un informe sobre l'evolució de les infraestructures per a dades científiques de cara l'any 2030. Una de les conclusions d'aquest informe final (High level Expert Group on Scientific Data, 2010) fou que era necessari donar un accés ampli a les dades científiques. Poc més tard, la Comissió Europea va donar suport explícit a l'accés a les dades científiques amb la previsió de la creació del programa de recerca Horizon 2020 en substitució de l'anterior programa FP7 (European Commission, 2011a). El següent pas fou la publicació d'una proposta de creació d'un projecte pilot (European Commission, 2012b), posant especial èmfasi en què la informació pagada amb diners públics no havia de ser pagada novament per al seu accés i ús. Finalment, Horizon 2020 s'aprovà i publicà oficialment (Unión Europea, 2013), el qual constitueix el programa de recerca i innovació més important de la Unió Europea, amb prop de 80.000 milions d'euros de pressupost [38] per al període 2014-2020, molt superior al de l'anterior programa, que era de 55.000 milions d'euros (European Commission, 2015).

A l'inici d'Horizon 2020 es va posar en marxa el programa de treball Open Research Data Pilot (European Commission, 2016c) de foment específic per a les dades de recerca. Durant el bienni 2014-2015 algunes àrees científiques ja van poder participar dins aquest projecte pilot, el qual va estar dotat amb un 20% del pressupost total d'Horizon 2020. Les àrees que van participar en el programa Open Research Data Pilot van ser:

- Tecnologies futures i emergents
- Infraestructures de recerca
- Tecnologies de la Informació i la Comunicació
- Energia segura, neta i eficient

[38] European Commission. *What is Horizon 2020?* <http://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020>. [Consulta: 29/03/2015]

- Acció del clima, medi ambient i eficiència dels recursos
- Societats inclusives i innovadores
- Ciència amb i per a la societat

No obstant, qualsevol projecte finançat per Horizon 2020 i que no es trobés dins aquestes àrees també hi podia participar de forma voluntària amb el mateix suport i seguiment. L'Open Research Data Pilot s'aplicà a dos tipus de dades:

- Totes aquelles dades, incloses les seves metadades, necessàries per validar els resultats presentats a les publicacions. Això no inclou tots els *datasets* resultants de la recerca, així que es tractaria de dades de recerca "finals"
- Altres dades, incloses les seves metadades, especificats en els plans de gestió de dades dels projectes de recerca

Els investigadors que van formar part del projecte pilot es trobaven obligats, per rebre el finançament corresponent, a complir l'article 29.3 del *grant agreement* (acord de subvenció) del programa Horizon 2020 (European Commission, 2016d, p. 222). Aquest article indica el següent, respecte a les dades de recerca:

- Dipositar-les en un repositori de dades de recerca i prendre mesures per tal que terceres persones puguin accedir, extreure, explotar, reproduir i distribuir –lliure de càrrec per a qualsevol usuari– el següent:
    - Totes aquelles dades, incloses metadades associades, necessàries per validar els resultats presentats a publicacions científiques tan aviat com sigui possible;
    - Altres dades, incloses les seves metadades, tal com quedi especificat i dins els terminis de lliurament establerts en els plans de gestió de dades dels projectes de recerca
- Proporcionar informació –mitjançant el repositori– de les eines i els instruments que estiguin a disposició dels beneficiaris i que siguin necessaris per validar els resultats (i si és possible, subministrar les mateixes eines i els instruments)

Aquestes condicions no canvien l'obligació de protegir els resultats de la recerca (art. 27 del *grant agreement*) ni les obligacions de confidencialitat, seguretat (art. 37) i de protecció de les dades personals (art. 39).

El programa pilot es va estendre l'any 2016 per cobrir totes les àrees temàtiques dins Horizon 2020[39], així que totes les dades de recerca dels projectes que rebin finançament del programa a partir de 2017 seran d'accés obert per defecte (European Commission, 2016a). No obstant, hi ha alguns casos en què els projectes de recerca poden ser alliberats d'aquest requeriment:

- Quan els resultats de la investigació seran explotats de forma comercial o industrial i han de ser protegits
- Quan és necessari protegir la confidencialitat per temes de seguretat
- Quan s'han de protegir dades personals
- Quan la participació al programa dificulta aconseguir l'objectiu principal de la recerca
- Quan el projecte no generarà ni recollirà cap dada de recerca
- Altres raons legítimes que s'hagin d'especificar

Per altra banda, a partir de 2016 els participants a l'Horizon 2020 han d'elaborar un DMP (European Commission, 2016b) que ha estat dissenyat segons els principis FAIR, que estipulen el següent sobre les dades de recerca[40]:

- Ser trobables:
    - F1. Les (meta)dades s'assignen a un identificador únic i persistent global
    - F2. Les dades es descriuen amb metadades riques
    - F3. Les metadades s'enregistren o s'indexen en un recurs que permeti fer recerques
    - F4. Les metadades especifiquen l'identificador de les dades que descriuen

[39] European Commission. *Open research data in Horizon 2020*. <http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit&pagemode=none>. [Consulta: 06/09/2016]

[40] *The FAIR Data Principles*. <https://www.force11.org/group/fairgroup/fairprinciples>. [Consulta: 22/12/2016]

- Ser accessibles:
    - A1. Les (meta)dades són recuperables mitjançant el seu identificador amb l'ús d'un protocol de comunicacions estandarditzat
        - A1.1. El protocol és obert, lliure i implementable de forma universal
        - A1.2. El protocol permet un procediment d'autenticació i autorització quan sigui necessari
    - A2. Les metadades són accessibles, fins i tot quan les dades ja no es troben disponibles
- Ser interoperables:
    - I1. Les (meta)dades utilitzen un llenguatge formal, accessible, compartit i aplicable de forma àmplia per a la representació del coneixement
    - I2. Les (meta)dades utilitzen vocabularis que segueixen els principis FAIR
    - I3. Les (meta)dades inclouen referències qualificades a altre (meta)dades
- Ser reutilitzables:
    - R1. Les (meta)dades tenen una pluralitat d'atributs acurats i exactes
        - R1.1. Les (meta)dades s'alliberen amb un llicència d'ús de dades clara i accessible
        - R1.2. Les (meta)dades s'associen amb la seva procedència
        - R1.3. Les (meta)dades satisfan estàndards comuns rellevants

Aquests principis es troben en evolució actualment per tal de superar les barreres presents a la recerca i la reutilització de dades (varietat de dades existents, diversitat de formats de dades, permisos per reutilitzar les dades o altres qüestions legals). Per tal de superar aquestes barreres, l'any 2014 es va organitzar un taller a Leiden, Països Baixos, que va començar a refinar i millorar els principis (Wilkinson et al., 2016). El resultat és un document en procés[41], que ha de servir per millorar els resultats d'implementació de les dades de recerca en accés obert.

La Comissió Europea no especifica l'ús de cap tipus d'estàndard de metadades, ni tampoc indica que s'hagin de depositar les dades en un repositori concret. Sí que facilita

[41] *FAIR principles living document* (2016, Jan 25). <http://datafairport.org/fair-principles-living-document-menu>. [Consulta: 22/12/2016]

algunes indicacions com els directoris de repositoris Registry of Research Data Repositories[42], el projecte OpenAIRE[43] de suport a les polítiques d'accés obert de la Comissió Europea o el repositori Zenodo que permet als investigadors el dipòsit tant de publicacions com de dades. En resum, els investigadors han de complir amb l'obligació de preservar les seves dades i metadades associades, sense haver d'utilitzar uns formats i/o estàndards concrets de dades i metadades; i han de fer el dipòsit en un repositori, però no es concreta cap, sinó que es deixa al criteri de l'investigador.

Durant el bienni 2014-2015, de 3.054 propostes de finançament que es van rebre a Horizon 2020, el 24,2% (442 de 1.824) de propostes a les àrees cobertes pel programa pilot va decidir no participar-hi. No obstant, es va produir una participació del 27,2% (334 de 1230) dins les àrees no cobertes (Ramjoué, 2015). Per tant, va haver-hi un total de 1.716 propostes de projectes que van optar per compartir les seves dades i amb la política actual de posar en accés obert totes les dades de recerca del programa el nombre augmentarà de forma significativa en els propers anys.

2.2.2 Research Councils UK (Regne Unit)

Dins l'àmbit britànic existeix un consorci estratègic de set consells de recerca, el Research Councils UK (RCUK)[44], la missió del qual és coordinar el finançament de la recerca dins les arts, les humanitats, la ciència i l'enginyeria. Els set Research Councils són l'Arts and Humanities Research Council (AHRC)[45], el Biotechnology and Biological Sciences Research Council (BBSRC)[46], l'Engineering and Physical Sciences Research Council (EPSRC)[47], l'Economic and Social Research Council (ESRC)[48], el Medical Research Council (MRC)[49], el Natural Environment Research Council

[42] <http://www.re3data.org/>. [Consulta: 01/03/2016]

[43] <https://www.openaire.eu/>. [Consulta: 01/03/2016]

[44] <http://www.rcuk.ac.uk/>. [Consulta: 21/04/2015]

[45] <http://www.ahrc.ac.uk/>. [Consulta: 06/03/2016]

[46] <http://www.bbsrc.ac.uk/>. [Consulta: 06/03/2016]

[47] <https://www.epsrc.ac.uk/>. [Consulta: 06/03/2016]

[48] <http://www.esrc.ac.uk/>. [Consulta: 06/03/2016]

[49] <http://www.mrc.ac.uk/>. [Consulta: 06/03/2016]

(NERC)[50] i el Science and Technology Facilities Council (STFC)[51]. Compten amb un pressupost anual de tres mil milions de lliures esterlines per a la recerca. El marc administratiu del RCUK dins el Regne Unit és el de Non-Departmental Public Body (organisme públic no ministerial), el qual és defineix com "body which has a role in the processes of national Government, but is not a Government Department or part of one, and which accordingly operates to a greater or lesser extent at arm's length from Ministers" (Great Britain. Cabinet Office, 2007). En el cas dels Councils, el Department for Business, Innovation and Skills (Ministeri de Empreses, Innovació i Qualificacions) té el control estatuari, amb el suport del Secretari d'Estat.

L'actual política de dades obertes del RCUK s'originà gràcies a un informe del Working Group on Expanding Access to Published Research Findings (2012), que va recomanar un programa d'acció per tal de permetre a més gent poder llegir i utilitzar les publicacions fruit de la recerca i fer progressos vers un entorn d'accés obert total. L'informe considerava els repositoris d'especial importància per preservar i facilitar l'accés a les dades de recerca. El govern britànic acceptà totes les recomanacions i va anunciar dins el seu *Open data white paper* (Great Britain. Cabinet Office, 2012) que implantaria mesures per desenvolupar polítiques sobre l'accés a les dades de recerca, amb col·laboració amb les universitats. Fruit d'aquestes accions, el RCUK va revisar la seva política d'accés obert (vigent des de 2005) i publicà la *RCUK policy on open access and supporting guidance*[52] en abril de 2013, que requereix la inclusió d'una declaració vers com es pot accedir als materials de recerca (com les dades).

Els set Councils van publicar l'any 2011 uns principis comuns vers la gestió i la compartició de dades de recerca, els *RCUK common principles on data policy*[53], que són els següents:

- Les dades de recerca amb finançament públic són un bé públic i han de trobar-se disponibles de forma oberta amb el menor nombre de restriccions possible

[50] <http://www.nerc.ac.uk>. [Consulta: 06/03/2016]

[51] <http://www.stfc.ac.uk/>. [Consulta: 06/03/2016]

[52] <http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/RCUKOpenAccessPolicy.pdf>. [Consulta: 21/04/2015]

[53] <http://www.rcuk.ac.uk/research/datapolicy/>. [Consulta: 05/03/2016]

- Les polítiques i plans de gestió de dades específiques de projectes haurien d'estar en conjunció amb estàndards rellevants i amb les millors pràctiques de la comunitat científica. Les dades amb un valor reconegut a llarg termini han de ser preservades, ser accessibles i útils per a la recerca futura
- S'han de registrar les metadades suficients per tal de permetre a altres investigadors entendre la recerca i permetre la reutilització de les dades. Els resultats publicats han d'incloure informació sobre com accedir a les dades de suport
- Degut als problemes legals, ètics i comercials que es poden crear degut a l'alliberament de dades de recerca, les polítiques i pràctiques de recerca han d'assegurar que no es segueixin procediments que facin malbé el procés de recerca
- Els investigadors que rebin finançament del RCUK poden tenir un període limitat d'ús privilegiat de les dades recollides per poder publicar els resultats de la seva recerca. La durada d'aquest període pot variar en funció de la disciplina de recerca
- Tots els usuaris de dades de recerca han de reconèixer les fonts de les seves dades i acatar els termes i condicions d'accés
- És adequat l'ús de fons públics per donar suport a la gestió i a la compartició de dades de recerca finançades de forma pública. Per tal de maximitzar el benefici de la recerca que es pot aconseguir de pressupostos limitats, els mecanismes per a aquestes activitats haurien de ser eficients i efectius en l'ús de fons públics

La missió comuna dels Research Councils és la promoció i el suport de la producció de la recerca en les seves disciplines i la formació de postgraduats. Actualment, la gran majoria ja inclouen dins les seves polítiques requeriments específics als investigadors i/o organitzacions de recerca per donar accés obert a les seves dades de recerca. Les analitzem a continuació.

Arts and Humanities Research Council

L'AHRC té la seva pròpia política d'accés obert[54] en consonància amb el RCUK. En el cas d'aquells plans de recerca on la producció de resultats o tecnologies digitals siguin una part essencial, es requereix l'elaboració d'un Technical Plan[55], que ha d'incloure les següents seccions:

- Secció 1. Resum de resultats i de tecnologies digitals
- Secció 2. Metodologia tècnica
    - Estàndards i formats
    - Programari i maquinari
    - Adquisició, procés, anàlisi i ús de dades
- Secció 3. Suport tècnic i experiència rellevant
- Secció 4. Preservació, sostenibilitat i ús
    - Preservar les teves dades
    - Assegurar l'accés i l'ús continuat a les teves dades digitals

Les dades de recerca s'han posar disponibles en un repositori en un terme màxim de tres anys després de la concessió de la subvenció, però no s'especifica cap, excepte en el cas de l'arqueologia, on l'AHRC disposa del servei ADS (Archaeology Data Service)[56]. En aquest últim cas, s'ha de fer esment de les seves pràctiques, ja que aquest dipòsit requereix als investigadors la inclusió de les següents metadades dins els seus *datasets*[57]:

- Títol. S'ha d'indicar el títol i títols alternatius del *dataset*
- Descripció. Un resum (màxim 200 o 300 paraules) del objectius principals del projecte amb un altre resum (màx. 200 o 300 paraules) del contingut del *dataset*
- Matèria. S'han de suggerir paraules clau per al contingut en matèries del conjunt de dades. Recomanen l'ús d'estàndards com el Monument Type Thesaurus[58] de

[54] <http://www.ahrc.ac.uk/about/policies/openaccess/>. [Consulta: 02/03/2016]

[55] <http://www.ahrc.ac.uk/funding/research/researchfundingguide/applicationguidance/technicalplan/>. [Consulta: 02/03/2016]

[56] <http://archaeologydataservice.ac.uk/. [Consulta: 02/03/2016]

[57] *What information is contained in the ADS catalogue?* <https://goo.gl/FK8whs>. [Consulta: 02/03/2016]

[58] <http://thesaurus.historicengland.org.uk/thesaurus.asp?thes_no=1>. [Consulta: 02/03/2016]

la Royal Commission on the Historical Monuments of England (RCHME) o el Archaeological Object Name Thesaurus[59]

- Cobertura. Noms contemporanis del país, regió, comtat, ciutat o municipi coberts dins la col·lecció de dades amb dates i períodes cronològics coberts
- Creadors. Detalls sobre els creadors, agències de finançament i altres grups o individus responsables de la col·lecció de dades
- Editor. Detalls sobre qualsevol organització que hagi publicat les dades
- Dates. Data de creació del *dataset* i de l'execució del projecte arqueològic
- Copyright. Nom del titular del *copyright* del dataset
- Relacions. Si la col·lecció de dades es deriva en part o totalment de fonts publicades o no publicades, bé en forma impresa o en digital, s'han de donar referències al material original
- Idioma. Idioma o idiomes del *dataset*
- Tipus de recurs. S'ha d'indicar si el *dataset* es descriu millor com a dades primàries, dades processades, una interpretació de dades o bé un informe final
- Format en què es guarden les dades (p. ex. Word, HTML, AutoCAD)

Biotechnology and Biological Sciences

Aquest Council requereix a tots els investigadors candidats a una subvenció l'enviament d'un DMP, segons indica la seva política de compartició de dades[60]. Aconsella la inclusió dels següents elements:

- Àrees de dades i tipus de dades. La BBSRC ha identificat un nombre d'àrees on la compartició de dades és important, com les dades que sorgeixen d'experiments de gran volum (p. ex. seqüenciació del genoma), dades de baix rendiment que provenen d'observacions en períodes prolongats de temps (p. ex. dades meteorològiques) i models que es generen amb enfocaments sistèmics (p. ex. models biològics)
- Estàndards i metadades. Els investigadors han de gestionar les seves dades utilitzant formats i metodologies acceptats, juntament amb metadades

[59] <https://goo.gl/auKw5J>. [Consulta: 31/05/2017]

[60] *BBSRC Data Sharing Policy: version 1.2 (March 2016 update)*. <http://www.bbsrc.ac.uk/datasharing>. [Consulta: 12/06/2016].

- Mètodes per a la compartició de dades. El BBSRC indica dues vies principals per compartir les dades: el dipòsit en un repositori o altre recurs d'ús comú o bé fer la compartició de manera directa, a petició d'altres investigadors; en aquest últim cas es requereix a l'investigador que conservi les seves dades durant un període de 10 anys després de la fi del projecte i l'ús de formats accessibles amb estàndards establerts. S'indiquen vies alternatives com compartir dades a comunitats tancades de recerca o combinar mètodes en *datasets* diferents
- Períodes de temps. El BBSRC espera que totes les dades, amb les seves metadades, es comparteixin en un temps no superior al temps de publicació dels resultats científics finals
- Ús secundari de les dades. Sempre que les dades es comparteixen mitjançant un recurs de tercers o una base de dades, els usuaris secundaris han de reconèixer la font de les dades. Si la compartició es fa de forma directa, seria apropiat fer el reconeixement a la font de les dades en funció del nivell d'ús i de col·laboració
- Dades propietàries. Es poden produir restriccions vers l'alliberament de dades si es produeix una col·laboració de finançament de recerca entre el BBSRC i un soci comercial. Aquest punt s'hauria de deixar clar en el DMP

Engineering and Physical Sciences

La política de dades de recerca de l'EPSRC[61] no exigeix l'elaboració d'un pla de dades als investigadors, però sí que presenta nou expectatives (amb clarificacions addicionals[62]) a les organitzacions que en reben finançament:

- S'han d'assegurar que els seus investigadors i estudiants de recerca siguin conscients de la regulació i les exempcions que es poden utilitzar si fos necessari per justificar la retenció de dades de recerca. Per tant, s'espera que la gestió de dades de recerca sigui una part del procés normal d'iniciació per als nous estudiants i investigadors

[61] *EPSRC policy framework on research data*. <https://www.epsrc.ac.uk/about/standards/researchdata/>. [Consulta: 03/03/2016]

[62] *Clarifications of EPSRC expectations on research data management*. <https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/>. [Consulta: 04/03/2016]

- Els articles científics publicats haurien d'incloure una descripció sobre com i en quines condicions es pot accedir a les dades de recerca de suport. S'espera la inclusió de citacions a les dades o a documentació de suport que descrigui les dades, les condicions d'accés i possibles restriccions, amb enllaços URL persistents com DOIs
- Cada organització científica tindrà polítiques específiques i processos associats per mantenir una conscienciació efectiva de les seves dades de recerca finançades públicament i de peticions de tercers per accedir-hi; tots els investigadors o estudiants de recerca finançats pel EPSRC han de complir amb les polítiques de recerca de l'organització en aquesta àrea o bé, si excepcionalment no és possible, donar una justificació. S'espera que l'organització faci registres dels accessos a dades de recerca que tingui en custòdia per tal de tenir evidències del seu impacte i així poder prendre decisions sobre la conservació futura de *datasets* específics
- Les dades de recerca finançades públicament que no es generin en format digital s'emmagatzemaran de forma que es faciliti la seva compartició si es dóna el cas de peticions d'accés a les dades. No s'espera que les organitzacions de recerca digitalitzin totes les seves dades de recerca, però tampoc s'espera que es rebutgin peticions d'accés perquè la forma no sigui apte per a la compartició
- Les organitzacions de recerca s'asseguraran que es publiquin metadades estructurades que descriguin les dades de recerca, normalment en un termini de dotze mesos des del moment en què es generen les dades, i distribuïdes de forma gratuïta a Internet. Les metadades han de ser suficients per permetre a altres entendre l'existència de les dades, per què, quan i com es van generar, i com s'hi pot accedir. Sempre que les metadades facin referència a dades de recerca en forma d'objecte digital, s'espera que les metadades incloguin un identificador robust, com DataCite. Les metadades estructurades no han de pertànyer a un estàndard específic, ja que encara no existeixen estàndards robustos de metadades que facilitin trobar dades de recerca
- Si l'accés de dades està restringit, les metadades han de donar un motiu i resumir les condicions que s'han de complir per aconseguir l'accés. Això pot incloure l'interès legítim de tercers dins organitzacions comercials en bloquejar dades confidencials, o bé informació sensible que pugui comprometre propietat

intel·lectual. No obstant, no s'ha d'impedir l'escrutini i la validació dels resultats científics publicats

- Les organitzacions de recerca s'asseguraran que les dades de recerca finançades per l'EPSRC es preserven de forma segura durant un mínim de deu anys des de la data en què el període d'accés de qualsevol investigador amb "accés privilegiat" hagi caducat o bé, si altres han accedit a les dades, des de la data en què l'accés a les dades fou sol·licitat per tercers. No s'espera, per tant, que es preservin les dades que no hagin resultat d'interès durant un període de 10 anys
- Les organitzacions de recerca s'asseguraran que la conservació efectiva de dades es fa durant el cicle complet de dades, segons les definicions de 'data curation' i 'data lifecycle' del Digital Curation Centre[63]. Sempre que les dades de recerca presentin un accés restringit, l'organització de recerca implementarà i gestionarà controls de seguretat i s'assegurarà que els processos de conservació de dades tinguin garanties de qualitat
- Les organitzacions de recerca s'asseguraran que hi ha recursos adequats per donar suport a la conservació de dades de recerca finançades públicament; aquests recursos estaran assignats dels seus fons públics, sempre que es rebin dels Research Councils com a suport directe o indirecte per a projectes específics o de Funding Councils d'educació superior

Economic and Social Research Council

La política de dades de recerca d'aquest Council[64] es compon de nou principis que coincideixen amb els principis comuns del RCUK. Les particularitats més importants quant a l'accés són:

- Les dades generals que donen suport als resultats publicats de la recerca seran accessibles al mateix temps que els resultats publicats
- Les dades de recerca que s'hagin produït gràcies a finançament de l'ESRC hauran d'estar dipositats en un repositori en un període de tres mesos després de la finalització de la concessió de subvencions

[63] *What is digital curation?* <http://www.dcc.ac.uk/digital-curation/what-digital-curation>. [Consulta: 04/03/2016]

[64] *ESRC Research Data Policy*. <http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/>. [Consulta: 13/03/2016]

- Els candidats a subvencions hauran d'incloure un DMP dins les seves sol·licituds, que ha d'especificar com gestionaran i prepararan les dades per a la seva compartició i reutilització
- Les dades de recerca hauran de tenir metadades i documentació per tal que usuaris secundaris puguin entendre les dades i així facilitar la seva reutilització. Es recomana en aquest cas l'ús d'estàndards com Data Documentation Initiative (DDI), SDMX o INSPIRE
- L'ESRC té el seu propi centre de dades per fer el dipòsit, l'UK Data Service. No obstant, l'investigador pot dipositar les seves dades en un altre repositori sempre que aquest permeti que les dades tinguin les condicions dels principis FAIR
- L'ESRC finança la preservació a llarg termini de totes les dades de recerca

Medical Research Council

La política de compartició de dades del MRC s'identifica amb els principis de l'OECD (2007) que declaren que les dades de recerca finançades públicament són un bé públic, produïts en l'interès públic i que haurien d'estar disponibles de forma oberta en la major mesura possible. Aquesta política també es coherent amb els principis comuns del RCUK.

El MRC no requereix als investigadors que indiquin com preservaran i compartiran les dades sinó que indiquin les seves previsions per fer-ho quan planegen i executen la seva recerca. Aquesta política es troba vigent al Council des de 2005[65], amb canvis menors inclosos l'any 2011[66]. Es tracta en els següents punts:

- Els investigadors haurien de fer que les dades de recerca finançades públicament pel MRC es trobin disponibles a la comunitat científica amb el menor nombre de restriccions possible
- Aquells que comparteixin les dades haurien de rebre el reconeixement total i apropiat per part dels finançadors, les seves institucions acadèmiques i altres agents que promocionin la recerca

[65] *MRC policy on research data sharing.* <http://www.mrc.ac.uk/research/research-policy-ethics/data-sharing/data-management-plans/>. [Consulta: 04/03/2016]

[66] *Data management plans*. <http://www.mrc.ac.uk/research/research-policy-ethics/data-sharing/data-management-plans/>. [Consulta: 04/03/2016]

- Els nous estudis que resultin de la compartició de dades han de complir els estàndards del MRC pel que fa a la qualitat científica, requeriments ètics i valor monetari. Haurien de donar també valor afegit al *dataset* original
- La recerca és més profitosa quan es produeix una col·laboració entre l'usuari nou i els creadors o conservadors de les dades originals
- Les dades que es generin de la recerca finançada pel MRC s'han de preservar de forma adequada durant tot el seu cicle de vida i alliberada amb les metadades apropiades. Això és responsabilitat dels custodis de les dades, que sovint són aquells individus o organitzacions que hagin rebut finançament del MRC per crear o recollir les dades
- S'ha de definir un període limitat d'ús exclusiu de dades per a la recerca primària en funció de la natura i valor de les dades i de com es generen i utilitzen
- La contribució a la recerca continuada a completar *datasets* no s'ha de comprometre degut a la compartició i anàlisi prematura o oportunista. La compartició sempre hauria de tenir en compte la millora del valor a llarg termini de les dades
- La política del MRC no pretén el rebuig de presentació d'aplicacions de patents abans de la publicació i reconeix que pot ser necessari retardar la publicació durant un període breu per tal de donar temps a la preparació de patents
- S'han de fer servir permisos per a tota aquella recerca mèdica que impliqui dades personals abans de compartir les dades
- Els investigadors, participants a la recerca i reguladors de la recerca s'han d'assegurar que dins el requeriments reguladors legals, es maximitzen les oportunitats per als nous usuaris
- Les polítiques d'accés i pràctiques per a les col·leccions de dades finançades pel MRC, noves i existents, han de ser transparents, equitatives, factibles, i que facilitin decisions clares i coherents amb la política de compartició de dades del MRC
- Tots els investigadors aspirants a rebre finançament del MRC han d'incloure un DMP a la seva sol·licitud

- Les dades de recerca han de ser de bona qualitat; han de tenir validesa a llarg termini; han d'estar ben documentades per tal que altres investigadors puguin accedir, entendre, utilitzar i afegir valor a les dades
- La informació de la recerca sobre persones s'ha de gestionar amb estàndards ètics de gran qualitat
- Els investigadors principals i institucions que rebin finançament són responsables de la planificació i l'execució de polítiques locals, i de sistemes i estàndards sobre com es gestionen les dades de recerca

Natural Environment Research Council

El NERC té la seva pròpia política de dades[67] que cobreix totes aquelles dades reunides o creades mitjançant la recerca, l'estudi o les activitats de control que siguin finançades totalment o en part pel NERC. També s'aplica a les dades gestionades pel NERC encara que aquest no sigui el finançador principal. És important indicar que la política s'aplica a dades mediambientals, definits pel NERC com ítems o registres individuals (tant digitals com analògics) que s'obtenen habitualment mitjançant el mesurament, l'observació o modelat del món natural. No s'aplicaria a aquells productes d'informació creats mitjançant l'addició de contribucions intel·lectuals que aporten valor afegit a les dades.

La política d'accés a les dades es concentra en els següents punts:

- Totes les dades de medi ambient custodiades pels centres de dades del NERC es trobaran disponibles de forma oberta a qualsevol persona o entitat que les demani
- Les úniques restriccions sobre l'accés s'aplicaran en els casos d'excepcions sobre difusió que s'indiquen en les Environmental Information Regulations[68] de 2004
- Per protegir el procés de recerca el NERC permetrà a aquells que rebin el seu finançament treballar durant un període de temps de forma exclusiva sobre les

[67] *NERC data policy*. <http://www.nerc.ac.uk/research/sites/data/policy/data-policy/>. [Consulta: 05/03/2016]

[68] Regulació d'accés a la informació mediambiental del Regne Unit

dades que recullin i així publicar els resultats. Aquest període serà normalment d'un màxim de dos anys des de la finalització de la recollida de dades

- Totes les dades custodiades pels centres de dades del NERC es distribuiran de forma gratuïta excepte en els casos que es produeixin peticions complexes on es pot demanar el càrrec dels costos de distribució
- Totes les dades mediambientals que distribueixin els centres de dades del NERC contindran una llicència de dades
- Tots aquells que utilitzin dades que subministri el NERC han de reconèixer la font de les dades

El NERC compta amb els seus propis dipòsits digitals, que cobreixen disciplines com les ciències marines, atmosfèriques o de la terra, entre d'altres[69]. La seva política indica el següent vers els seus centres de dades:

- NERC mantindrà els centres de dades per a la gestió i la difusió de dades mediambientals de valor a llarg termini que es generin mitjançant el finançament del NERC o d'aquelles dades dipositades per tercers
- Els centres de dades actuaran de forma imparcial amb tots els productors de dades, sense importar que es formin part o no del NERC
- NERC mantindrà criteris per identificar dades mediambientals de valor a llarg termini, concentrats en una llista de control[70]. Aquests criteris s'utilitzaran per informar de totes les decisions que faci el NERC per acceptar o disposar de les dades
- El NERC Data Discovery Service donarà informació sobre com es custodien les dades

Pel que fa als requeriments que el NERC demana als investigadors per rebre finançament, la seva política indica el següent:

- Totes les sol·licituds per rebre finançament han d'incloure un esborrany de DMP, que ha d'identificar quins són els *datasets* que es produeixen i que es consideren

[69] *Data centres*. <http://www.nerc.ac.uk/research/sites/data/>. [Consulta: 05/03/2016]

[70] *NERC data value checklist*. <http://www.nerc.ac.uk/research/sites/data/policy/data-value-checklist/>. [Consulta: 05/03/2016]

de valor a llarg termini, agafant com a base la llista de control del NERC. La sol·licitud també haurà d'identificar tots els recursos necessaris per implementar el DMP

- L'esborrany de DMP serà avaluat com a part del procés estàndard de concessió de subvencions. Totes les sol·licituds que passin el tall hauran d'elaborar un DMP detallat conjuntament amb el centre de de dades corresponent
- Tots els projectes finançats pel NERC han de treballar amb el centre de dades corresponent per implementar el DMP i s'han d'assegurar que les dades de valor a llarg termini s'envien al centre de dades en un format adequat i acompanyades per les metadades necessàries
- Les dades d'activitats finançades pel NERC s'enviaran als centres de dades de forma no exclusiva, sense prejudici de drets de propietat intel·lectual. Això es fa per tal que el NERC pugui gestionar i distribuir gratuïtament les dades de recerca finançades públicament
- Aquells que estiguin finançats pel NERC que no compleixin aquests requeriments s'arrisquen a no rebre subvencions o a no ser admesos per a futurs finançaments

Finalment, la seva política d'accés obert a les dades de recerca indica que:

- L'accés obert ajuda a assegurar la integritat, la transparència i la robustesa del procés de recerca
- Tots les publicacions de recerca que sorgeixin de finançament del NERC han d'incloure una declaració sobre com es pot accedir a les dades de suport i a qualsevol altre material rellevant per a la recerca
- Pel que fa a les publicacions de recerca produïdes per la plantilla del NERC, les dades de suport es trobaran disponibles als centres de dades del NERC

Science & Technology Facilities Council

La política[71] d'aquest Council s'aplica a totes les dades científiques produïdes com a resultat de finançament del STFC, però no a dades purament administratives ni a dades resultants de treballs executats pel STFC al servei d'altres organitzacions.

Els principis generals, pel que respecta a la gestió i preservació, indiquen el següent:

- Com a 'dades', el STFC inclou les dades científiques en brut que resultin d'experiments o observacions; les dades derivades que han estat subjectes a un procediment de reducció de dades i les dades publicades, aquelles que es mostren o bé són referides en una publicació científica
- Han d'existir DMPs per a totes les dades que es trobin dins l'abast de la política
- Es requereix l'elaboració d'un DMP per rebre finançament, amb una explicació de com es gestionaran les dades durant la vida del projecte i, on sigui apropiat, com es preservaran per a la seva reutilització en el futur
- Els DMPs haurien de seguir recomanacions nacionals i internacionals de bones pràctiques
- Les dades resultants de la recerca finançada públicament s'haurien de distribuir públicament després d'un període limitat, a menys que hi hagin raons específiques per evitar aquest fet
- Les dades publicades s'haurien de distribuir en un període de sis mesos després de la publicació
- El STFC buscarà assegurar la integritat de qualsevol dada i metadada que gestioni
- El STFC espera que les dades es gestionin mitjançant un repositori institucional, una universitat, un laboratori o un base de dades d'una matèria específica gestionada de forma independent
- Els DMPs haurien d'especificar quines dades es depositaran en un repositori, a on i durant quant de temps. Aquestes dades haurien d'estar acompanyats per metadades suficients per habilitar la reutilització. El STFC espera que les dades

[71] *STFC scientific data policy*. <http://www.stfc.ac.uk/stfc/cache/file/D0D76309-252B-4EEF-A7BFAF6271B8EC11.pdf>. [Consulta: 05/03/2016]

originals es retinguin durant un període prolongat, el qual seria un mínim de deu anys després de la fi del projecte

2.2.3 Wellcome Trust (Regne Unit)

La fundació privada Wellcome Trust[72] té la visió de millorar la salut mitjançant el suport a la recerca en ciència, les humanitats i les ciències socials. És l'agència de finançament privat més important del Regne Unit, amb un pressupost superior als 700 milions de lliures anuals per a sis àrees de recerca:

- Ciència biomèdica
- Innovació
- Internacional
- Compromís públic
- Humanitats mèdiques
- Societat i ètica

El Trust és conscient de la importància de la compartició de dades de recerca per avançar el coneixement i ajudar a verificar els resultats. Arran d'això, l'any 2007 publicà una política sobre gestió i compartició de dades, que fou revisada l'agost de 2010[73], coherent amb la posició de l'agència vers la publicació en accés obert dels resultats de la recerca finançada. Es pot resumir en els següents punts:

- El Trust demana a tots els investigadors que rebin el seu finançament que maximitzin la disponibilitat de les dades de recerca amb el menor nombre de restriccions possible
- Tots aquells que demanin finançament del Trust han de considerar com gestionaran i compartiran les seves dades a la proposta. Si els resultats esperats de la recerca resulten d'interès per a la comunitat científica, els candidats hauran d'incloure un pla de gestió de dades

[72] <http://www.wellcome.ac.uk>. [Consulta: 01/04/2015]

[73] *Policy on data management and sharing*. <http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm>. [Consulta: 01/04/2015]

- El Trust examinarà els plans de gestió de dades com a part integral de la decisió final de finançament i treballarà amb investigadors becaris per donar-los suport en maximitzar el valor a llarg termini dels conjunts de dades clau com a resultat de la seva recerca
- El Wellcome Trust demana a tots els usuaris de dades de recerca que reconeguin les fonts de les seves dades i que compleixin amb els termes i condicions en què van accedir a les dades originals
- El Trust fomentarà un entorn que permeti als investigadors maximitzar el valor de les dades de recerca

Pel que fa a l'accés a les dades, el Trust demana als investigadors que es faci el seu dipòsit a repositoris apropiats o bé a bases de dades[74]. En aquest aspecte, recomana repositoris i bases de dades de diverses disciplines, com ciències socials i humanitats, medicina o biologia. No especifica el temps mínim o màxim per dipositar les dades, però sí la dels articles publicats, que és de sis mesos[75]. La preservació és un aspecte que han de tenir en compte els investigadors durant la durada de la subvenció; han de considerar com es preservaran aquells *datasets* amb valor a llarg termini. No obstant, el Trust recomana un mínim de deu anys per a la conservació segura de totes les dades generades al llarg del projecte[76].

2.2.4 National Institutes of Health (Estats Units d'Amèrica)

Una de les agències federals que ha fet grans esforços per complir la directiva governamental és el NIH (National Institutes of Health), que és l'agència de recerca mèdica principal dels EUA i la font de finançament de recerca mèdica més gran del món[77], ja que subvenciona a més de 300.000 investigadors de més de 2.500 universitats i altres institucions de recerca. L'agència es compon de 27 instituts i centres, cadascun

[74] *Guidance for researchers: Developing a data management and sharing plan.* <http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/index.htm>. [Consulta: 06/03/2016]

[75] *Position statement in support of open and unrestricted access to published research.* <http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm>. [Consulta: 06/03/2016]

[76] *Guidelines on good research practice.* <http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002753.htm>. [Consulta: 06/03/2016]

[77] *About NIH.* <http://www.nih.gov/about/>. [Consulta: 04/10/2014]

amb una agenda de recerca específica, i la inversió total que dedica a la recerca és de prop de 30.000 milions de dòlars anuals[78].

La política general del NIH vers la compartició de dades de recerca es pot trobar a la seva NIH Data Sharing Policy[79], que fou publicada el 26 de febrer de 2003 i es va fer efectiva per a les propostes de recerca que es van rebre a partir de l'1 de octubre de 2003. Aquesta informació també es pot trobar dins el document general de requeriments de subvencions, la darrera versió del qual es publicà en novembre de 2015 (National Institutes of Health, 2015).

Segons el NIH, totes les dades s'haurien de compartir. La seva política es resumeix en la següent declaració: "Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data". Per tal de facilitar la compartició de dades, els investigadors que demanin una subvenció d'una quantitat igual o superior a 500.000 dòlars anuals han d'elaborar un pla per compartir les dades de recerca finals dels seus projectes (per ajudar a futures investigacions) o bé especificar els motius pels quals no es poden compartir les dades. El pla de gestió de dades ha de complir les següents condicions:

- Descripció del pla de treball i calendari per compartir les dades
- Format del *dataset* final
- Documentació sobre la metodologia emprada per recollir les dades
- Eines necessàries, si escau
- Mètode de compartició de les dades. El NIH suggereix alguns mètodes com el dipòsit a un arxiu de dades o bé a un repositori especialitzat, però no obliga a utilitzar cap en concret

És important destacar que en el cas de la recerca mèdica es poden presentar problemes de privacitat dels subjectes d'estudi, ja que encara que es bloquegin dades personals dels *datasets*, fer un estudi molt específic (com per exemple, afectats d'una malaltia molt minoritària) dins l'abast d'un àrea geogràfica reduïda pot fer fàcil identificar els

[78] *NIH Budget*. <http://www.nih.gov/about/budget.htm>. [Consulta: 03/04/2015]

[79] <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>. [Consulta: 06/03/2016]

subjectes utilitzats dins un projecte de recerca. No obstant això, hi ha opcions com requerir als investigadors que signin un compromís de confidencialitat i privacitat i fer que l'accés a les dades es faci dins un entorn controlat[80].

Al febrer de 2015 el NIH va publicar un informe amb una recopilació del seu plans actuals i futurs per assegurar l'accés a les publicacions i a les dades de recerca fruit del seu finançament[81]. Pel que fa a les dades, hi està treballant per facilitar el seu accés públic sense càrrec i a determinar quines dades haurien de ser les prioritàries per ser preservades a llarg termini, amb consultes a la comunitat científica. Un dels problemes de la política de 2003 es que no havia instruccions específiques vers la preservació a llarg termini, però aquest problema es pensa resoldre amb el desenvolupament de nous criteris per identificar de forma periòdica les necessitats que presenten els nous tipus de dades que apareixen degut a l'avanç de la tecnologia.

A més del NIH Data Sharing Policy, hi ha altres polítiques de compartició de dades dins el NIH i a instituts i centres que pertanyen al NIH. Es descriuen a continuació.

Dipòsit de coordenades atòmiques

Aquesta política, vigent des de l'any 1999[82], requereix a l'investigador que dipositi les coordenades atòmiques d'experiments de ressonància magnètica cristal·logràfica i nuclear de raigs X que hagin rebut finançament del NIH a la base de dades estructural adequada en el moment en què es lliuri un article de recerca que aporti conclusions d'aquestes dades. El repositori que s'indica en aquest cas és el Protein Data Bank[83].

[80] *Frequently asked questions. Data sharing.* <http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm>. [Consulta: 03/04/2015]

[81] Plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research. <http://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>. [Consulta: 03/04/2015].

[82] *NIH policy relating to deposition of atomic coordinates into structural databases*. <http://grants.nih.gov/grants/guide/notice-files/not99-010.html>. [Consulta: 07/03/2016]

[83] <www.rcsb.org>. [Consulta: 07/03/2016]

NHGRI: Alliberament de dades i ús de dades al consorci ENCODE

Aquesta política, amb data de darrera revisió de 22 de novembre de 2009[84], va ser desenvolupada al National Human Genome Research Institute (NHGRI), el qual dissenyà els projectes d'enciclopèdia d'elements d'ADN (ENCODE) i d'organisme model ENCODE (modENCODE) com a projectes de recursos de comunitat científica per accelerar l'accés i l'ús de les dades. Els productors d'ENCODE i modENCODE han d'alliberar les dades, tan aviat com hagin estat verificades i abans de la seva publicació, a bases de dades públiques apropiades com GenBank o els centres de coordinació de dades ENCODE/modENCODE i aquestes dades de pre-publicació estaran disponibles a tothom per al seu ús.

Compartició de dades genòmiques

En aquest cas, s'espera que es comparteixin les dades de recerca d'estudi del genoma que provinguin d'estudis finançats pel NIH mitjançant un repositori de dades amb accés públic. Per fer-ho, els investigadors han d'elaborar un pla de compartició de dades genòmiques, registrar els seus estudis a la base de dades de genotips i fenotips dbGaP[85] i enviar les dades a un repositori de dades designat pel NIH. La política actual[86] es troba vigent des de 2014 i substitueix una d'anterior, dissenyada per compartir dades d'estudis d'associació del genoma complet.

NHLBI: compartició d'assajos clínics i d'estudis epidemiològics

En aquest cas, aquesta política[87] es desenvolupa al National Heart, Lung and Blood Institute (NHLBI), que es dedica a la recerca, formació i educació per promoure la prevenció de malalties de cor, pulmó i sang. L'elaboració d'un pla de recerca es requereix per a aquells investigadors que demanin finançament per a projectes de

[84] *ENCODE consortia data release, data use, and publication policies*. <https://goo.gl/ijpEuT>. [Consulta: 07/03/2016]

[85] <http://www.ncbi.nlm.nih.gov/gap>. [Consulta: 08/03/2016]

[86] *NIH genomic data sharing policy*. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html>. [Consulta: 08/03/2016]

[87] *NHLBI Policy for Data Sharing from Clinical Trials and Epidemiological Studies*. <http://www.nhlbi.nih.gov/research/funding/human-subjects/data-sharing>. [Consulta: 08/03/2016]

recerca amb costos iguals o superiors a 500.000 dòlars. El repositori que es recomana a l'investigador és el dipòsit de dades de l'institut, el BioLINCC[88].

NIA: compartició de dades genòmiques de la malaltia d'Alzheimer

El National Institute on Aging (NIA)[89] és l'institut del NIH encarregat de l'estudi de l'envelliment i la manera d'augmentar els anys d'esperança de vida. La seva política, revisada per darrera vegada el 26 de febrer de 2015[90], és una revisió de la *NIA policies and guidance for sharing of resources and data from studies on the genetics of Alzheimer's disease*, amb data de darrera revisió en 24 de gener de 2008. Aquesta política s'aplica a totes aquelles dades derivades dels estudis finançats pel NIA del genoma de la malaltia d'Alzheimer. Per tal de complir amb la política, els investigadors han de compartir en primer lloc les mostres biològiques i dades fenotípiques associades al repositori NCRAD[91] i en el cas de les dades genòmiques, les han d'enviar al dipòsit de dades especialitzat en dades genòmiques de la malaltia d'Alzheimer NIAGADS[92] o altre lloc web aprovat pel NIA. Per tal de facilitar l'accés públic a *datasets* genòmics, el dipòsit NIAGADS funciona cooperativament amb la base de dades de genomes dbGaP.

ADNI: política de compartició de dades i de publicacions

La iniciativa Alzheimer's Disease Neuroimaging Initiative (ADNI)[93], creada sota el suport del NIA, és un esforç de recerca a nivell mundial que dóna suport a la recerca i el desenvolupament de tractaments que retardin o aturin la progressió de la malaltia d'Alzheimer. La seva política[94], revisada per darrera vegada el 15 d'abril de 2015, recomana l'accés obert a totes les dades sense informació personal i/o sensible d'ADNI a tots els investigadors que s'hi registrin i estiguin d'acord amb les condicions de

[88] <https://biolincc.nhlbi.nih.gov/home/>. [Consulta: 08/03/2016]

[89] <https://www.nia.nih.gov/>. [Consulta: 08/03/2016]

[90] *Policies and guidance for sharing of resource and data from studies on the genomics of Alzheimer's disease*. <https://www.nia.nih.gov/sites/default/files/nia_ad_gsp_fnl_2_posted_2_27_15.doc>. [Consulta: 08/03/2016]

[91] <https://ncrad.iu.edu/>. [Consulta: 08/03/2016]

[92] <https://www.niagads.org/>. [Consulta: 08/03/2016]

[93] <http://adni.loni.usc.edu/>. [Consulta: 08/03/2016]

[94] *Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Sharing and Publication Policy.* <http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_DSP_Policy.pdf>. [Consulta: 08/03/2016]

l'acord d'ús de dades. L'objectiu últim de l'ADNI és que aquestes dades es trobin disponibles a la comunitat científica en un cort període de temps.

NIAID/DMID: compartició i alliberament de dades

El National Institute of Allergy and Infectious Diseases (NIAID)[95] s'ocupa de donar suport a la recerca aplicada en malalties infecciones, immunològiques i al·lèrgiques, mentre que la Division of Microbiology and Infectious Diseases (DMID)[96] dóna suport a la recerca en control i prevenció de malalties causades per agents infecciosos humans, amb la excepció del VIH. Aquí trobem, més que una política, unes directrius[97] als investigadors per tal que elaborin plans de compartició de dades genòmiques i altres dades finançades pel NIAID. Aquestes dades s'haurien d'enviar el més aviat possible a bases de dades internacionals com dbGaP, el centre de recursos bioinformàtics del DMID o altres bases de dades designades i aprovades pel NIAID.

NIAID/DMID: compartició de dades al Systems Biology Program

Aquí novament trobem unes directrius[98], en aquest cas del Systems Biology Program (SBP)[99] del NIAID i el DMID, que dóna suport a la compartició i anàlisi de dades mitjançant la compartició de dades en brut no processades a investigadors centrals o bé realitzant les anàlisis de mostres a plataformes de diverses àrees. Aquell que estigui contractat pel SBP, ha de distribuir lliurement a la comunitat científica les dades de recerca, els protocols i els models estadístics i computacionals mitjançant els llocs web dels centres o altres bases de dades públic en un període de quatre setmanes després de la seva publicació o bé un any després d'haver estat generades.

[95] <https://www.niaid.nih.gov/about/Pages/default.aspx/>. [Consulta: 10/03/2016]

[96] <https://www.niaid.nih.gov/about/organization/dmid/Pages/default.aspx/>. [Consulta: 10/03/2016]

[97] *NIAID/Division of Microbiology and Infectious Diseases (DMID) Data Sharing and Release Guidelines.* <http://www.niaid.nih.gov/LabsAndResources/resources/dmid/Pages/data.aspx>. [Consulta: 10/03/2016]

[98] Data Sharing Guiding Principles for the Division of Microbiology and Infectious Diseases (DMID) Systems Biology Program. <http://www.niaid.nih.gov/labsandresources/resources/dmid/sb/pages/datareleaseguidelines.aspx>. [Consulta: 10/03/2016]

[99] <http://www.niaid.nih.gov/labsandresources/resources/dmid/sb/Pages/default.aspx>. [Consulta: 10/03/2016]

HIPC: pla de compartició de dades

L'Human Immunology Project Consortium (HIPC)[100] es va fundar l'any 2010 per la divisió del NIAID corresponent a al·lèrgies, immunologia i transmissió de malalties. Mitjançant aquest programa, s'estudien grups grans i ben caracteritzats utilitzant recursos de recerca centralitzats i una base de dades centralitzada. L'objectiu específic és compartir dades de forma àmplia i lliure per tal de promoure nova recerca i generar noves hipòtesis. Per tal d'aconseguir l'objectiu, s'ha desenvolupat una política de compartició de dades[101] que requereix als investigadors del consorci el dipòsit de les seves dades d'estudi amb les seves metadades en el sistema Immunology Database and Analysis Portal (ImmPort)[102].

NICHD: pla de compartició de dades per millorar la recerca del peix zebra

El National Institute of Child Health and Human Development (NICHD)[103] fou fundat l'any 1962 pel President John F. Kennedy amb l'objectiu d'estudiar el "complex process of human development from conception to old age". Com que l'ús del peix zebra com a organisme model (això vol dir que ha estat una espècie que ha estat estudiada abastament) és molt popular dins la recerca biomèdica i conductual, el NICHD va engegar una oferta de finançament [104] per als projectes que proposin recerca i desenvolupament de tecnologia per donar suport als diferents aspectes de la creació, identificació, detecció i caracterització de models de peix zebra de malalties humanes i de preservació de material genètic. Els participants han d'elaborar tres plans de compartició: plans de compartició de dades, compartició d'organismes model i pla de compartició de dades genòmiques. Per compartir les dades, es recomana contactar amb el Zebrafish International Resource Center (ZIRC)[105].

---

[100] <http://www.immuneprofiling.org/>. [Consulta: 10/03/2016]

[101] *HIPC Data Sharing Plan*. <http://www.immuneprofiling.org/hipc/page/showPage?pg=dataShare/>. [Consulta: 10/03/2016]

[102] <https://immport.niaid.nih.gov/immportWeb/home/home.do?loginType=full>. [Consulta: 10/03/2016]

[103] <https://www.nichd.nih.gov>. [Consulta: 10/03/2016]

[104] Development of Novel and Emerging Technologies to Support Zebrafish Models for Biomedical Research (R43/R44). <http://grants.nih.gov/grants/guide/pa-files/PA-15-086.html>. [Consulta: 12/03/2016

[105] <http://zebrafish.org/zirc/home/guide.php/>. [Consulta: 12/03/2016]

NIDA: política de compartició de dades

El National Institute on Drug Abuse (NIDA)[106] dóna suport a la recerca sobre la drogoaddicció i aplicar-la a la salut pública. La seva política, revisada en març de 2012[107], requereix que les dades de tots aquells estudis en genètica humana finançats pel NIDA estiguin disponibles per a la compartició, independentment dels costos directes, afiliació en el NIDA Genetics Consortium[108], o el tipus de dades genètiques que es generin. Els investigadors poden dipositar les seves dades al Consortium o bé al NIDA Center for Genetics Studies Repository[109].

NIDA: Política de compartició de dades de xarxes d'estudis clínics

El NIDA disposa de la xarxa d'estudis clínics CTN[110], un mitjà de cooperació entre el NIDA, investigadors, pacients participants i proveïdors de tractaments mèdics per tal de desenvolupar, validar, refinar i lliurar noves opcions de tractament per als pacients. La política del NIDA indica que aquells *datasets* que entrin en la tipologia d'estudis clínics, hauran d'estar disponibles un cop l'article final hagi estat acceptat per a la seva publicació o bé quan les dades estiguin bloquejades durant més de divuit mesos[111]. Per compartir les dades, s'ha d'utilitzar el web Data Share[112] de NIDA.

NIDDK: estudi TEDDY

El National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)[113] dóna suport a la recerca en la diabetis i altres malalties endocrines i metabòliques. Una de les seves activitats fou l'estudi TEDDY, que es va engegar per resoldre qüestions tècniques vers les dades de diabetis tipus 1. Segons indica la seva política, revisada el 30

[106] <https://www.drugabuse.gov/>. [Consulta: 12/03/2016]

[107] *NIDA Policy*. <https://goo.gl/0H6iak>. [Consulta: 12/03/2016]

[108] <https://goo.gl/YBU04E>. [Consulta: 12/03/2016]

[109] <https://nidagenetics.org/>. [Consulta: 12/03/2016]

[110] *About CTN*. <https://www.drugabuse.gov/about-nida/organization/cctn/ctn/about-ctn>. [Consulta: 12/03/2016]

[111] *Data Share Policy*. <https://datashare.nida.nih.gov/guidelines/data_share_policy>. [Consulta: 12/03/2016]

[112] <https://datashare.nida.nih.gov/index>. [Consulta: 12/03/2016]

[113] <http://www.niddk.nih.gov/Pages/default.aspx>. [Consulta: 12/03/2016]

d'octubre de 2012[114], tots aquells investigadors que van participar en aquest estudi van haver d'acceptar el reconeixement de l'estudi i compartir les dades de recerca mitjançant el repositori central del NIDKK[115].

NDA: política de compartició de dades

El NIH juntament amb el National Institute of Mental Health (NIMH) [116] han desenvolupat una federació de repositoris de dades, el NIMH Data Archive (NDA) , el qual està format pel National Database for Autism Research (NDAR) [117], el NIH Pediatric MRI Repository (PedsMRI) [118], el National Database for Clinical Trials Related to Mental Illness (NDCT) [119] i el Research Domain Criteria Database (RDoCdb)[120], amb l'objectiu d'emmagatzemar la col·lecció de dades de participants en estudis de recerca relacionats amb la salut mental. El seu document de terminis i condicions de compartició de dades[121], efectiva des del 7 de gener de 2015, requereix que totes les dades finançades pel NIH que impliquin a subjectes humans i que no continguin informació personal s'han d'enviar al NDA juntament amb la documentació necessària per habilitar l'ús eficient de les dades.

Compartició de dades mitjançant el FITBIR

El sistema informàtic Federal Interagency Traumatic Brain Injury Research (FITBIR)[122] és un recurs i repositori central per compartir dades que fou desenvolupat pel Departament de Defensa dels EUA i el NIH amb l'objectiu de promoure la col·laboració, la recerca i el coneixement avançat en caracterització, prevenció, diagnosi i tractament dels danys cerebrals traumàtics. Tots aquells estudis que es facin amb

[114] *Sample and Data Sharing Policy*. <https://www.niddkrepository.org/static/studies/teddy/teddy_sample_data_sharing_policy.pdf>. [Consulta: 12/03/2016]

[115] *NIDDK Central Repository*. <https://www.niddkrepository.org/home/>. [Consulta: 12/03/2016]

[116] <https://www.nimh.nih.gov/index.shtml>. [Consulta: 12/03/2016]

[117] <https://ndar.nih.gov>. [Consulta: 13/03/2016]

[118] <pediatricmri.nih.gov>. [Consulta: 13/03/2016]

[119] <ndct.nimh.nih.gov/>. [Consulta: 13/03/2016]

[120] <rdocdb.nimh.nih.gov/>. [Consulta: 13/03/2016]

[121] *NIMH Data Archive Data Sharing Terms and Conditions.* <https://ndar.nih.gov/ndarpublicweb/Documents/NDAR_data_sharing_language_fin.pdf>. [Consulta: 12/03/2016]

[122] <http://fitbir.nih.gov/>. [Consulta: 13/03/2016]

finançament del NIH i el Departament de Defensa han de dipositar les seves dades en el FITBIR, segons indica la seva política de compartició de dades vigent des del 27 de març de 2014[123]. Els investigadors han d'elaborar un pla de compartició de dades coherent amb la política del FITBIR i com a mínim han d'utilitzar l'estàndard de dades per a la recerca clínica TBI Common Data Elements[124].

2.2.5 National Science Foundation (Estats Units d'Amèrica)

Una altra agència federal important és la NSF, que té com a missió el suport per a tots els camps de la ciència i l'enginyeria i compta amb un pressupost anual superior als set mil milions de dòlars[125]. A partir de gener de 2011, la NSF va començar a requerir als sol·licitants de finançament la inclusió d'un pla de gestió de dades dins les propostes de recerca per tal de descriure "how the proposal will conform to NSF policy on the dissemination and sharing of research results"[126].

La política de requeriments de la NSF[127] respecte a les dades es resumeix en els següents punts:

- Les propostes han d'incloure un document suplementari de no més de dues pàgines amb l'etiqueta "Data Management Plan". Aquest suplement ha de descriure com s'ajusta la proposta a la política de la NSF vers la difusió i la compartició dels resultats de la recerca, a més dels plans d'arxiu de dades i la seva preservació[128]
- Els investigadors han de preparar i enviar per a la seva publicació tots els descobriments de la seva investigació sota el patrocini del NSF

[123] *FITBIR Data Sharing Policy*. <https://fitbir.nih.gov/assets/FITBIR_Data_Sharing_Policy.pdf>. [Consulta: 13/03/2016]

[124] <https://commondataelements.ninds.nih.gov/tbi.aspx#tab=Data_Standards>. [Consulta: 13/03/2016]

[125] *About the National Science Foundation*. <http://www.nsf.gov/about/>. [Consulta: 04/04/2015]

[126] National Science Foundation. *Dissemination and sharing of research results*. <https://www.nsf.gov/bfa/dias/policy/dmp.jsp>. [Consulta: 26/03/2017]

[127] *Award and Administration Guide Chapter VI.D.4. Dissemination and Sharing of Research Results*. <http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/aag_6.jsp#VID4>. [Consulta: 13/03/2016]

[128] *Grant Proposal Guide* II.c.2.j. <http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#IIC2j>. [Consulta: 04/04/2015]

- Els investigadors hauran de compartir les seves dades primàries, mostres, col·leccions físiques i altres materials de suport creats o reunits durant el seu treball sota el patrocini del NSF
- S'espera que els investigadors i persones subvencionades comparteixin programari i innovacions creats sota el patrocini o al menys distribueixin els productes de manera àmplia
- Els costos associats a les dades, com la seva documentació i preparació, seran susceptibles de finançament

La NSF no indica cap repositori en concret per dipositar les dades, ja que considera que la "gestió raonable de dades" s'ha de decidir per la comunitat d'interès mitjançant el procés de *peer review* i gestió de programes[129].

De la mateixa manera que el NIH, la NSF ha publicat un pla d'accés al resultats de la recerca finançada, per tal d'assolir els objectius plantejats al memoràndum de la Casa Blanca (National Science Foundation, 2015). L'abast d'aquest pla inclou les dades i els seus resultats associats i també aquells que entrin dins els plans de gestió de dades que hagin d'incloure els investigadors a les seves propostes de recerca.

Els plans de la NSF per facilitar l'accés a les dades inclouen la modificació del web de l'agència per tal de donar suport a la cerca de *datasets* i juntament amb altres agències federals, establir bones pràctiques per a la seva preservació i reutilització. Val a dir que aquests plans encara es troben en una fase inicial i s'espera que es compleixin en un termini de tres anys. Un altre projecte en funcionament finançat pel NSF és DataONE[130], un programa de compartició de dades biològiques i mediambientals que s'inicià l'any 2009 amb un pressupost de 20 milions de dòlars (Cohn, 2012).

[129] Data Management & Sharing Frequently Asked Questions (FAQs) - updated Nov. 30, 2010. <https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp>. [Consulta: 13/03/2016]

[130] <https://www.dataone.org>. [Consulta: 21/04/2016]

2.2.6 Plan Estatal de Investigación Científica y Técnica y de Innovación (Espanya)

L'article 149.1.15 de la Constitució espanyola de 1978[131] estableix que l'Estat espanyol té competència exclusiva vers el foment i la coordinació general de la recerca científica i tècnica. La política estatal es desenvolupa d'acord amb la Llei de la Ciència, la Tecnologia i la Innovació d'1 de juny de 2011[132], la qual va substituir la legislació prèvia de 1986, degut en bona part a la necessitat de consolidar l'activitat científica espanyola dins la Unió Europea. Per tal de complir amb els objectius d'aquesta llei es van definir dos instruments: per una banda, l'Estrategia Española de Ciencia y Tecnología y de Innovación 2013-2020[133] i el Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016[134].

L'Estrategia Española de Ciencia y Tecnología y de Innovación té quatre objectius generals: el reconeixement i la promoció del talent, el foment de la recerca científica i tècnica, potenciar el lideratge empresarial en R+D+I i la recerca orientada als reptes de la societat.

El Plan Estatal de Investigación Científica y Técnica y de Innovación té com a objectiu el desenvolupament i el finançament, per part de l'Administració estatal, de les actuacions compreses dins l'Estrategia. Per tal de complir aquest objectiu, el Plan es troba integrat per quatre Programas Estatales, que es subdivideixen en diferents Subprogrames de caràcter pluriennal.

Pel que respecta a les obligacions que té l'Administració estatal vers la preservació de dades de recerca, malauradament no són prou concretes. L'article 37 de la Llei de la Ciència indica el següent:

> 1. Los agentes públicos del Sistema Español de Ciencia, Tecnología e Innovación impulsarán el desarrollo de repositorios, propios o compartidos, de acceso abierto a las

[131] <http://www.boe.es/diario_boe/txt.php?id=BOE-A-1978-31229>. [Consulta: 28/03/2015]

[132] <http://www.boe.es/boe/dias/2011/06/02/pdfs/BOE-A-2011-9617.pdf>. [Consulta: 25/03/2014]

[133] Gobierno de España. Ministerio de Economía y Competitividad. *Estrategia Española de Ciencia y Tecnología y de Innovación 2013-2020.* <https://goo.gl/7cxN8V>. [Consulta: 11/09/2016]

[134] Gobierno de España. Ministerio de Economía y Competitividad. *Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016.* <https://goo.gl/4BSa1S>. [Consulta: 11/09/2016]

> publicaciones de su personal de investigación, y establecerán sistemas que permitan conectarlos con iniciativas similares de ámbito nacional e internacional.
>
> 2. El personal de investigación cuya actividad investigadora esté financiada mayoritariamente con fondos de los Presupuestos Generales del Estado hará pública una versión digital de la versión final de los contenidos que le hayan sido aceptados para publicación en publicaciones de investigación seriadas o periódicas, tan pronto como resulte posible, pero no más tarde de doce meses después de la fecha oficial de publicación.
>
> 3. La versión electrónica se hará pública en repositorios de acceso abierto reconocidos en el campo de conocimiento en el que se ha desarrollado la investigación, o en repositorios institucionales de acceso abierto.

Per altra banda, dins la convocatòria de concessió d'ajudes publicada el 30 de desembre de 2014 corresponents a la convocatòria Retos-Colaboración del Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad[135], declara al seu article 7 com a conceptes susceptibles d'ajuda:

> Cuando los resultados no sean susceptibles de protección de derechos de propiedad industrial o intelectual, de acuerdo con el artículo 37 de la Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación, las publicaciones científicas resultantes, total o parcialmente, de la financiación otorgada al amparo de la presente convocatoria deberán estar disponibles en acceso abierto. Para ello los autores podrán optar por publicar en revistas de acceso abierto o bien por autoarchivar en repositorios institucionales o temáticos de acceso abierto los trabajos científicos que hayan sido aceptados para su publicación en publicaciones seriadas o periódicas.

I a la convocatòria publicada en agost de 2014 corresponent al Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Subprograma Estatal de Generación de Conocimiento[136] s'indica al seu article 20 que entre les despeses susceptibles de finançament es troben les de publicació i difusió dels resultats de la recerca, inclosos aquells que es puguin derivar de la publicació en revistes d'accés obert.

[135] <http://www.boe.es/diario_boe/txt.php?id=BOE-A-2015-529>. [Consulta: 28/03/2015]

[136] <http://www.boe.es/boe/dias/2014/08/08/pdfs/BOE-A-2014-8601.pdf>. [Consulta: 28/03/2015]

Encara que la Llei de la Ciència no en faci un esment específic vers la preservació de dades de recerca, el que indica literalment són "contribuciones aceptadas por publicaciones de investigación" que no es trobin protegides pels drets de propietat intel·lectual. Certament, un gran part de les revistes de recerca només accepten articles científics, però hi ha d'altres que també requereixen el dipòsit de les dades fruit de la recerca, així que es pot dir que l'Estat espanyol contribueix al finançament de les dades obertes de recerca, però de moment és un fet força marginal. No obstant, podem esmenar dos casos on l'enviament de dades és obligatori segons els requeriments del Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad en la seva convocatòria d'ajudes per a l'any 2016[137]; el primer es troba en la gestió de les dades polars i oceanogràfiques en brut, on s'estipula que les dades polars s'han d'enviar al Centro Nacional de Datos Polares (Bermúdez; Barragán; Alonso, 2011) en un termini de tres mesos posteriors a la fi del projecte i que les dades oceanogràfiques s'han d'enviar segons un procediment que ha de publicar el Ministerio de Economía y Competitividad. El segon cas s'aplica a *datasets* que s'hagin generat a partir d'enquestes qualitatives en l'àmbit de les ciències socials, que s'han de dipositar en el Banco de Datos Específico de Estudios Sociales, un repositori del Centro de Investigaciones Sociológicas, en un terme màxim de dotze mesos.

2.2.7 Síntesi i conclusions

Un cop analitzades les diferents polítiques de finançament de la recerca, es presenta a la Taula 4 una síntesi de les mateixes.

[137] España. Ministerio de Economía y Competitividad. *Resolución de 8 de marzo de 2016, de la Secretaria de Estado de Investigación, Desarrollo e Innovación por la que se aprueba la convocatoria para el año 2016 del procedimiento de concesión de ayudas a proyectos de I+D+I correspondientes al Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016*. <https://goo.gl/lNmCZk>. [Consulta: 31/05/2017]

Taula 4. Polítiques de gestió de dades de recerca a les agències de finançament

Agència/política	Requeriment de pla de gestió de dades	Requeriment de dipòsit de dades	Repositoris/bases de dades
Horizon 2020	Sí, aquells que participin al programa Open Research Data Pilot	Sí, aquells que participin al programa Open Research Data Pilot	No s'especifica cap
AHRC	Sí, amb el nom "Technical Plan"	Sí, en un termini màxim de tres anys	Archaeology Data Service, en el cas de l'arqueologia
BBSRC	Sí	Sí, en un temps no superior al temps de publicació dels resultats científics finals	No indica cap, però recomana alguns com el BioModels Database
EPSRC	No	No. Són els investigadors i/o organitzacions els responsables de preservació de les dades	No s'especifica cap
ESRC	Sí	Sí, en un temps no superior al temps de publicació dels resultats científics finals	UK Data Service
MRC	Sí	No. Són els investigadors i/o organitzacions els responsables de preservació de les dades	No s'especifica cap
NERC	Sí	Sí, aquelles dades de valor a llarg termini	Centres de dades del NERC
STFC	Sí	Sí, en un termini màxim de sis mesos després de la publicació dels resultats científics finals	No s'especifica cap
Wellcome Trust	Sí, si les dades es consideren d'interès per a la comunitat	Sí, si les dades es consideren d'interès per a la comunitat	No s'especifica cap
NIH	Sí, per a propostes que demanin subvenció igual o superior a 500.000 dòlars anuals	Sí, en un temps no superior al temps de publicació dels resultats científics finals	No s'especifica cap
NIH: Dipòsit de coordenades atòmiques	No	Sí, en el moment en què es lliuri un article de recerca que tracti de les dades	Protein Data Bank

Agència/política	Requeriment de pla de gestió de dades	Requeriment de dipòsit de dades	Repositoris/bases de dades
NHGRI: Consorci ENCODE	No	Sí, un cop s'hagin verificat i abans de la seva publicació	ENCODE
NIH: Dades genòmiques	Sí	Sí, en un temps no superior al temps de publicació dels resultats científics finals	dbGaP (per al registre) Repositori designat pel NIH (per a dades)
NHLBI: Assajos clínics i estudis epidemiològics	Sí, per a projectes de recerca que demanin subvenció igual o superior a 500.000 dòlars	Sí, en un temps no superior al temps de publicació dels resultats científics finals	BioLINCC
NIA: dades genòmiques de la malaltia d'Alzheimer	Sí, per a mostres i dades fenotípiques i en el cas que l'investigador dipositi les dades en repositoris diferents dels especificats. Per a dades genòmiques, s'aplica la política del NIH per a aquest tipus de dades	Sí, en un temps no superior al temps de publicació dels resultats científics finals	NCRAD (mostres biològiques i dades fenotípiques) i NIAGADS (dades genòmiques)
NIA	No	L'investigador es compromet a compartir dades sense identificació personal un cop es registra a l'ADNI	ADNI
NIAID	Sí	Sí, en un temps no superior al temps de publicació dels resultats científics finals	dbGaP, centre de recursos bioinformàtics del DMID o altres bases de dades designades i aprovades pel NIAID
NIAID: Systems Biology Program	No	Sí, en un període de quatre setmanes després de la seva publicació o bé un any després d'haver estat generades.	SBP
NIAID: HIPC	Sí, amb el nom "Dataset Completion Plan"	Sí, en un temps màxim d'un any després de la seva generació o bé en un temps no superior al temps de publicació dels resultats científics finals	ImmPort

Agència/política	Requeriment de pla de gestió de dades	Requeriment de dipòsit de dades	Repositoris/bases de dades
NICHD: millora de la recerca del peix zebra	Requereix tres plans: plans de compartició de dades, compartició d'organismes model i pla de compartició de dades genòmiques	Sí, segons s'indica a la política de NIH: Dades genòmiques	ZIRC
NIDA	Sí	Sí, en un temps no superior al temps de publicació dels resultats científics finals	NIDA Genetics Consortium, NIDA Center for Genetic Studies Repository
NIDA: dades de xarxes d'estudis clínics	No	Sí, en un temps no superior al temps de publicació dels resultats científics finals o bé quan les dades estiguin bloquejades durant més de 18 mesos	Data Share
NIDDK: estudi TEDDY	No	Sí. El temps de dipòsit varia segons el tipus de dades	Repositori central del NIDDK
NIH i NIMH	No	Sí, juntament amb les publicacions i/o descobriments	Repositoris inclosos al NDA
NIH i Departament de Defensa	Sí	Sí, en períodes trimestrals	FITBIR
NSF	Sí	Sí, si les dades es consideren d'interès per a la comunitat	No s'indica cap
Plan Estatal de Investigación Científica y Técnica y de Innovación	No	Sí, en el cas de dades oceanogràfiques i polars i d'enquestes de l'àmbit de les ciències socials	Centro Nacional de Datos Polares, Banco de Datos Específico de Ciencias Sociales

Font: Elaboració pròpia a partir de les polítiques analitzades

Hi ha altres agències finançadores, com la Genome Canada[138] o la Gordon and Betty Moore Foundation[139], que també exigeixen la creació d'un pla de gestió de dades per rebre finançament, però considerem que aquesta mostra ja és prou representativa. Es pot concloure que la compartició de dades en el cas espanyol encara es troba en una fase

[138] *Data release and resource sharing*. <http://www.genomecanada.ca/sites/genomecanada/files/publications/datareleaseandresourcesharingpolicy.pdf>. [Consulta: 18/06/2016]

[139] Gordon and Betty Moore Foundation Grantee Resources (2008). *Data sharing philosophy*. <https://www.moore.org/docs/default-source/Grantee-Resources/data-sharing-philosophy.pdf>. [Consulta: 18/06/2016]

molt embrionària, ja que la Llei de la Ciència ni tan sols en fa un esment explícit i per tant l'única obligació de compartició de dades que té un investigador que rebi finançament del Plan Estatal de Investigación Científica y Técnica dependrà de si la publicació on apareguin els resultats de la recerca exigeix el dipòsit de les dades; un exemple és el cas de les revistes pediàtriques espanyoles (Aleixandre-Benavent et al., 2013), on cap d'elles fa esment d'aquesta possibilitat o de l'enviament de material suplementari. Tot el contrari que a les agències europees i americanes, on les dades de recerca en accés obert tenen una gran consideració i els investigadors han de plantejar la gestió de les seves dades des del principi del seu projecte. El suport de les administracions estatals és cabdal, ja que és evident que les directives de la Comissió Europea i de la Casa Blanca han donat un impuls definitiu per fer que la gestió de les dades sigui un element imprescindible dins qualsevol projecte important de recerca. Hom pot considerar que sigui qüestió de temps que aparegui una iniciativa semblant en l'àmbit espanyol.

2.3 Formats digitals de les dades de recerca

Dins les dades de recerca podem trobar una enorme varietat de formats de fitxer. Per una banda, hi ha multitud de disciplines amb formats molt especialitzats i per altra, bona part d'aquests formats són propietaris i requereixen de llicències d'ús per poder fer-los servir, la qual cosa dificulta les possibilitats de preservació a llarg termini. Les característiques ideals que haurien de complir els formats de fitxer per complir les tres característiques de preservació, reutilització i compartició serien les següents, tal com s'indica a les MIT Libraries,[140] les Stanford University Libraries[141] o a l'Australian National Data Service[142]:

- No propietaris
- Estàndard obert i documentat

[140] *File formats for long-term access.* <http://libraries.mit.edu/data-management/store/formats/>. [Consulta: 22/03/2016]

[141] *Best practices for file formats.* <https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats>. [Consulta: 22/03/2016]

[142] *File formats.* <http://ands.org.au/guides/file-formats>. [Consulta: 22/03/2016]

- Utilització habitual dins la comunitat científica
- Representació estàndard (ASCII, Unicode)
- Desencriptat
- No comprimit
- Sense pèrdua

2.3.1 Polítiques als repositoris

Per tal de donar una panoràmica dels requeriments de formats a l'*open research data*, s'ha fet un estudi de formats de fitxer acceptats tot fent servir una metodologia consistent en fer una selecció de repositoris que fan atenció especial a la preservació, reutilització i compartició, en analitzar els seus requeriments de formats de fitxer i finalment en fer una descripció de les seves característiques tècniques.

Recordem que un dels requeriments que han de complir els investigadors amb les agències de finançament és el dipòsit de dades a un repositori. Dins l'*open research data*, cada repositori té una política diferent pel que pertoca als estàndards acceptats i la documentació addicional que ha d'aportar l'investigador quan elabora el seu DMP. Per tal d'analitzar quins son els formats requerits, s'ha fet una selecció de repositoris que compleixin amb el següent:

- Atenció especial a la preservació
- Acceptació de dades de recerca
- Política clara respecte a formats a dipositar

Una sèrie de repositoris que compleixen amb aquests requisits són els que compten amb el Data Seal of Approval[143], un segell que fou desenvolupat per les organitzacions científiques neerlandeses KNAW [144] (Koninklijke Nederlandse Akademie van Wetenschappen o Real Acadèmia Neerlandesa de les Ciències i les Arts) i NWO[145] (Nederlandse Organisatie voor Wetenschappelijk Onderzoek o Organització

[143] <http://datasealofapproval.org/en/>. [Consulta: 24/03/2016]

[144] <https://www.knaw.nl/en/>. [Consulta: 24/03/2016]

[145] <http://www.nwo.nl/>. [Consulta: 25/04/2015]

Neerlandesa per a la Recerca Científica) l'any 2008. Aquest segell és una garantia del següent:

- Els productors de dades s'asseguren què les seves dades i materials associats estaran emmagatzemats de forma fiable i es podran reutilitzar
- Les agències de finançament tenen la confiança que les dades romandran disponibles per a la seva reutilització i que la seva inversió no es perdrà
- Els consumidors de dades poden avaluar i valorar els repositoris on es custodien les dades
- Es fa suport als repositoris de dades en l'arxiu i en la distribució eficient de dades

Per tal de tenir el segell, és necessari seguir un total de setze directrius, les quals es van elaborar d'acord amb directrius nacionals i internacionals per a l'arxiu de dades digitals com *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* (Directrius per a la creació d'una política institucional en preservació digital) desenvolupat per la xarxa alemanya de preservació digital nestor (2014); Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)[146], metodologia desenvolupada pel DCC i pel DPE[147] i *Trustworthy Repositories Audit and Certification (TRAC): criteria and checklist*, editat pel CRL i per l'OCLC (2007). Les directrius són les següents:

1. El productor de les dades diposita les dades de recerca en un repositori de dades amb la informació suficient per tal que altres puguin assolir la qualitat acadèmica i científica de les dades de recerca i complir amb les normes ètiques i disciplinàries
2. El productor de les dades facilita les dades de recerca en formats recomanats pel repositori de dades
3. El productor de les dades facilita les dades de recerca juntament amb les metadades sol·licitades pel repositori de dades
4. El repositori de dades té la missió explícita en l'àrea d'arxiu digital i el promulga

[146] <http://www.repositoryaudit.eu/>. [Consulta: 24/03/2016]

[147] <http://www.digitalpreservation.gov/series/edge/dpe.html/>. [Consulta: 24/03/2016]

5. El repositori de dades és diligent per assegurar que es compleixen les regulacions i els contractes legals
6. El repositori de dades aplica processos documentats i procediments per gestionar emmagatzematge de dades
7. El repositori de dades té un pla per a la preservació a llarg termini dels seus actius digitals
8. El fet d'arxiu té lloc en funció de fluxos de treball explícits al llarg del cicle de vida de les dades
9. El repositori de dades assumeix la responsabilitat dels productors de les dades per a l'accés i la disponibilitat dels objectes digitals
10. El repositori de dades habilita als usuaris per utilitzar les dades de recerca i fer-ne referències
11. El repositori de dades assegura la integritat dels objectes digitals i de les metadades
12. El repositori de dades assegura l'autenticitat dels objectes digitals i de les metadades
13. La infraestructura tècnica explícitament suporta les tasques i les funcions que es descriuen en estàndards d'arxiu acceptats internacionalment com OAIS
14. El consumidor de dades ha de complir amb regles d'accés que tingui el repositori de dades
15. El consumidor de dades estarà d'acord amb qualsevol codi de conducte que s'accepti de forma general en àmbits d'educació i recerca per a l'intercanvi i ús adequat del coneixement i la informació
16. El consumidor de dades respecta les llicències aplicables del repositori de dades quant a l'ús de les dades de recerca

A continuació es presenta una selecció de quatre repositoris (preferentment aquells que pertanyin a la disciplina de les ciències socials i les humanitats) que compten amb el segell Data Seal of Approval[148] i tenen polítiques específiques respecte als formats de fitxer. En la mesura del possible, s'ha procurat triar aquells amb la major varietat de formats per així tenir una mostra heterogènia.

[148] *List of repositories that have acquired the Data Seal of Approval*. <https://assessment.datasealofapproval.org/>. [Consulta: 24/03/2016]

3TU.Datacentrum

3TU.Datacentrum[149] és un repositori creat per les tres universitats tècniques dels Països Baixos (Delft, Eindhoven i Twente) i ofereix el servei de planificació de la recerca, realització de la recerca i dipòsit de dades als investigadors. La seva missió principal consisteix en assegurar l'accessibilitat de la recerca científica durant i després de la conclusió de la recerca per donar un impuls de qualitat a la recerca actual i futura.

Dins la seva política de formats, el repositori reconeix tres nivells de suport a la preservació:

- Suport de nivell 1: es prendran totes les mesures raonables per mantenir la usabilitat, les quals poden incloure la migració, normalització o conversió
- Suport de nivell 2: es prendran mesures limitades, com migració a altres formats per evitar l'obsolescència
- Suport de nivell 3: només es donarà accés al fitxer original

En el nivell 1, 3TU.Datacentrum garanteix que les dades de recerca romandran accessibles i que seran migrades o convertides a altres formats si fos necessari per evitar problemes d'obsolescència. Dins la Taula 5, es poden observar els diferents nivell de suports dels formats de dades acceptats al repositori.

Aquest repositori presenta categories per a presentacions i altres formats especialitzats com les dades de química, numèriques o aplicacions, categories que no existeixen a la resta de repositoris que formen part de l'estudi. Per altra banda, sorprèn que no s'hagi considerat un nivell 1 de preservació per als formats de vídeo, els quals són importants per als materials de suport a les dades.

[149] <http://datacentrum.3tu.nl/en/home/>. [Consulta: 25/03/2016]

Taula 5. Formats de dades requerits al 3TU.Datacentrum

Tipus de dades	Suport de nivell 1	Suport de nivell 2	Suport de nivell 3
Text	• PDF (PDF/1) (.pdf) • Text pla UFT-8 (Unicode) (.txt) • SGML amb DTD (.sgm, .sgml) • XML amb DTD (.xml)	• HTML (.htm, .html) • Postscript (.ps, .eps) • Richtext (.rtf) • OpenDocument Text (.odt)	• Microsoft Word • SGML sense DTD (.sgm, .sgml) • XML sense DTD (.xml) • Wordperfect (.wpd)
Imatge	• JPEG (.jpg) • TIFF (.tif)	• JPEG 2000 (.jp2) • PNG (.png)	• BMP/Bitmap (.bmp) • GIF (.gif) • Photoshop (.psd)
Presentació		• Opendocument Presentation (.odp) • OOXML (.pptx)	Microsoft Power Point (.ppt)
Vídeo		• AVI (.avi) • MPEG-1 (.mp1) • MPEG-2 (.mp2) • MPEG-4 (.mp4)	• Quicktime (.mov) • Windows Media Video (.wmv) • MPEG-2 Transport Stream (.ts, .tsv, .tsa)
Àudio	Waveform Audio file Format (WAVE) (.wav)	MPEG audio (.mp3)	• Real Audio (.ra, .rm, .ram) • Windows Media Audio (.wma)
Dades geoespacials	• ESRI Shapefile (.shp, .shx, .dbf; opcional: .prj, .sbx, .sbn) • TIFF amb georeferències (.tif, .tfw) • Google Earth (.kml+.xml) • Llenguatge de marques geogràfic (.gml)	• Format ESRI de bases de dades geogràfiques (.mdb) • Format d'intercanvi MapInfo (.mif) • Keyhole Mark-up Language (.kml, .kmz) • Adobe Illustrator (.ai) • Dades CAD (.dxf, .svg)	
Dades de química	• NMR • IR • Raman • UV • Espectrometria de masses • JCAMP (format pensat per a la compartició) • ChemDoodle		• Protein Data Bank (.pdb) • Química/x-xyz (.xyz)
Bases de dades	Fitxer pla delimitat amb DDL	Format dBASE (.dbf)	Microsoft Access (.dbf)

Tipus de dades	Suport de nivell 1	Suport de nivell 2	Suport de nivell 3
Fulls de càlcul / Dades tabulars quantitatives amb metadades mínimes	• Valors separats per comes (.csv) • Fitxer delimitat per tabulacions (.tab) • PDF/A (.pdf)	• dBASE (.dbf) • Opendocument Spreadsheet (.ods)	• Microsoft Excel (.xls, xlsx) • Microsoft Access (.mdb, .accdb)
Arxius		ZIP (.zip)	• GZIP (.gzip) • RAR (.rar) • TAR (.tar) • TGZ (.tar, .gz) • JAVA (.jar) • Matlab (.mat)
Aplicacions			• Javascript (.js) • Shockwave Flash (.swf) • Virtualbox (.vdi)
Dades numèriques	Netcdf (.nc, .cdf)		Hdf5 (.hdf, .h4, .hdf4, .h5, .hdf5, .he4, .he5)

Font: 3TU.Datacentrum (2015). <http://datacentrum.3tu.nl/fileadmin/editor_upload/File_formats/Digital_Preservation_Support_levels.pdf>. [Consulta: 25/03/2016]

Archaeology Data Service

Aquest repositori, fundat l'any 1996, dóna suport a la preservació digital i a la seva reutilització per a la recerca, aprenentatge i ensenyament de l'arqueologia i la història i tal com s'indica a l'apartat 2.2.2, és el repositori que utilitza l'AHRC per dipositar dades arqueològiques. Les seves directrius actuals es presenten a la Taula 6.

Dins aquest repositori sorprèn el fet que als formats preferits hi ha alguns de propietaris com els d'Autocad (.dwg) o els de Microsoft (Excel, Word i Access), ja que la reutilització dels fitxers és complica força amb aquests formats perquè es necessiten llicències de pagament per fer-los servir. Per altra banda, s'ha de destacar el requisit d'adjuntar documentació sobre els formats emprats, ja que és una qüestió rellevant com es veurà més endavant.

Taula 6. Formats de dades requerits a l'ADS

Tipus de dades	Formats de fitxer preferits	Formats de fitxer acceptats	Documentació (programari, versió, plataforma)
CAD (gràfics vectorials)	• AutoCAD (.dwg) • Drawing Interchange Format (.dxf) • Scalable Vector Graphics (.svg)		• .svg – Relació amb altres documents • .dwg, .dxf – Versió d'AutoCAD/DXF, significat de convencions (capes, colors, símbols, etc.), relació amb altres fitxers (bases de dades, biblioteques d'objectes, etc.)
Bases de dades	• Access (.mdb, .accdb) • OpenDocument Database (.odb) • Text delimitat	Dbase (.dbf)	• Un diccionari de dades, p. ex. una llista de totes les taules i els seus camps • Nombre de files per a cada taula • Per a text delimitat, delimitadors i qualificadors
GIS	• ESRI Shapefile (.shp + .sxh + dbf) • Imatge TIF amb georeferències • Llenguatge de marques de geografia (.gml)	• ESRI Grid • Format d'intercanvi MapInfo (.mif + .mid) • Estàndard de transferència de dades espacials • MOSS (.exp) • Vector product Format (.vpf)	• Quin és el propòsit del GIS? • Què representa cada capa? • Ús del sistema coordinat o arbitrari per a la quadrícula escollida • Mètode de captura • Font de les dades • Escala o resolució de la captura de dades • Escala o resolució a la qual s'han guardat les dades • Avaluació de la qualitat de les dades • Data de captura o adquisició
Imatges	TIFF versió 6 sense compressió (.tif)	• Format RAW (.raw) • Portable Network Graphics (.png) • Joint Photographic Expert Group (.jpg) • Graphics Interchange Format (GIF) • Bit-Mapped Graphics Format (.bmp) • PhotoCD (.pcd) • Photoshop (Adobe) (.psd)	• Títol • Fotògraf • Data • Ubicació (país, districte, etc.) • Declaració de drets d'autor • Paraules clau

Tipus de dades	Formats de fitxer preferits	Formats de fitxer acceptats	Documentació (programari, versió, plataforma)
Vídeo	• MPEG 1 i 2 (.mpg, .mpeg) • MPEG 4 (.mp4)	DivX (.divx, .avi)	• Nom i versió del còdec de vídeo, dimensió del vídeo en píxels, taxa de fotogrames (fps) i taxa de bits (*bitrate*) • Nom i versió del còdec d'àudio, amb freqüència de mostreig, taxa de bits i informació de canals • Durada (hores, minuts i segons) del fitxer i mida • Informació de drets d'autor si és aplicable • Títol i descripció breu per a cada vídeo
Fulls de càlcul	• CSV • Microsoft Excel (.xls, .xlsx) • Open Document Spreadsheet (.ods)	Lotus 1-2-3 (.123, .wk4, .wk3, .wk1, .wks)	• Propòsit i contingut del full de càlcul • Contingut de cada columna i de cada fila si no és evident • Tipus de dades i escala per a cada columna • Claus per a codis dins les dades • Documentació per a característiques addicionals que pugui contenir el full de càlcul, com fórmules, macros, gràfics, comentaris i qualsevol característica important que s'hagi de preservar
Estadístiques	Text delimitat	• SPSS (.sav, .por, .spo) • SAS (.sas7dbat, .sas) • Microsoft Excel (.xls, .xlsx) • Open Document Spreadsheet (.ods) • SYLK (.slk) • Microsoft Access (.mdb) • xBase (.dbf)	• Font(s) de les dades, metodologia de col·lecció • Propòsit de les dades • Detalls de taules i mostres (columnes) • Nombre de files • Tipus i escala de variables • Descripció completa de qualsevol codificació que s'hagi utilitzat • Anàlisis realitzades sobre les dades

Tipus de dades	Formats de fitxer preferits	Formats de fitxer acceptats	Documentació (programari, versió, plataforma)
Texts	• Microsoft Word (.doc, .docx) • OpenDocument Text (.odt)	• Rich Text Format (.rtf) • Microsoft Word (.docm) • OpenOffice.org 1.0 (.sxw) • .txt • HTML, XHTML, XML, SGML	• DOC: programari, versió i plataforma • SXW, RTF, ODT: programari i versió • TXT: codificació de text • HTML, XHTML: programari utilitzat en la creació • XML: codificació de text, DTD o esquema • SGML: codificació de text
Realitat virtual	• X3D • VRML • Java3D • QTVR		• Fitxers de dades originals que formen el model (fitxers d'imatge, models CAD) si estan disponibles • Una renderització en vídeo del món original en realitat virtual per preservar el *look and feel*
Geofísica	• Dades xyz en brut (.txt, .csv) • Imatges renderitzades (.tif)	Dades en brut (.dat, .rep)	• Dades en brut: ubicació de l'estudi, condicions i instruments • Imatges: detalls del procés de les dades i interpretació de les dades
Àudio	• Broadcast Wave Format (.bwf) • Waveform Audio (.wav) • Audio Interchange Format (.aif)	Au de Sun Microsystems (.au)	• Taxa de bits (kbps) i freqüència de mostreig (KHz) si és aplicable. • Còdec utilitzat (on sigui apropiat) i documentació del procés de conversió • Longitud de la gravació (minuts i segons) • Informació de drets d'autor (en especial per a història oral) • Transcripcions d'entrevistes, etc. (on sigui apropiat)

Font: ADS (2014). *Guidelines for depositors: version 2.0 September 2014*. <http://archaeologydataservice.ac.uk/advice/FileFormatTable>. [Consulta: 25/03/2016]

UK Data Archive

L'UK Data Archive[150] és l'arxiu digital en ciències socials i humanitats més gran del Regne Unit. Fundat l'any 1967 amb finançament del Social Science Research Council, actualment proporciona un servei de suport a la recerca mitjançant l'UK Data Service[151].

[150] <http://www.data-archive.ac.uk>. [Consulta: 25/03/2016]

[151] <https://www.ukdataservice.ac.uk/>. [Consulta: 25/03/2016]

Aquest és el repositori que utilitzen els investigadors subvencionats per l'ESRC per dipositar les seves dades, a més de ser l'arxiu de referència per a moltes universitats com la University of Bristol [152], l'Aberystwyth University [153], la University of Glasgow[154], la University of British Columbia[155], i la University of Delaware[156]. La Taula 7 mostra els formats de fitxer acceptats al centre.

A més d'indicar els formats preferits, el repositori aporta directrius sobre com organitzar les dades, fer-ne control de qualitat, mantenir un control de versions i fer transcripcions d'enregistraments sonors, amb models i exemples.

Taula 7. Formats de dades requerits a l'UK Data Archive

Tipus de dades	Formats de fitxer acceptats per a la compartició, reutilització i preservació	Altres formats de fitxer acceptats
Dades tabulars quantitatives amb metadades extensives *(Dataset* amb etiquetes variables, etiquetes de codi, i valors de pèrdua definits, a més de la matriu de dades)	• Format portable SPSS (.por) • Fitxers de text delimitat (SPSS, Stata, SAS, etc.) que contenen informació de metadades • Fitxers de text estructurat o de marques que contenen informació de metadades, p. ex. fitxers DDI XML	• Formats propietaris de paquets estadístics p. ex. SPSS (.sav), Stata (.dta) • MS Access (.mdb, .accdb)
Dades tabulars quantitatives amb metadades mínimes (Matriu de dades amb o sense encapçalaments de columna o noms de variables, però sense altres metadades o etiquetes)	• Fitxers de valors separats per comes (CSV) (.csv) • Fitxers delimitats per tabulacions (.tab) • Inclou text delimitat o conjunt de caràcters amb declaracions de definició de dades SQL on sigui aplicable	• Text delimitat de conjunt de caràcters; només caràcters que no estiguin present a les dades s'haurien d'utilitzar com a delimitadors (.txt) • Formats propietaris utilitzats abastament, p. ex. MS Excel (.xls, .xlsx), MS Access (.mdb, .accdb), dBASE (.dbf) • OpenDocument (.ods)

[152] University of Bristol Research Data Service (2013). *An introduction to managing research data*. <https://zenodo.org/record/28547/files/Introduction-to-research-data-management-for-researchers.pdf>. [Consulta: 12/06/2016].

[153] *Orgainsing* [sic] *your data*. <https://www.aber.ac.uk/en/research/good-practice/data-management/how/organising/>. [Consulta: 12/06/2016]

[154] *Choosing file formats*. <http://www.gla.ac.uk/services/datamanagement/creatingyourdata/choosingfileformats/>. [Consulta: 28/03/2016]

[155] *Research data management dataguide* [sic]. <http://researchdata.library.ubc.ca/files/2015/10/RDM_DataGuide_V03.1_20151020.pdf>. [Consulta: 28/03/2016]

[156] *Research data management: file formats*. <http://guides.lib.udel.edu/c.php?g=371489&p=2511404>. [Consulta: 28/03/2016]

Tipus de dades	Formats de fitxer acceptats per a la compartició, reutilització i preservació	Altres formats de fitxer acceptats
Dades geoespacials (Dades de vectors i *raster*)	• ESRI Shapefile • (essencials: .shp, .shx, .dbf ; opcionals: .prj, .sbx, .sbn) • TIFF amb georeferències (.tif, .tfw) • Dades CAD (.dwg) • Dades d'atributs tabulars GIS	• Format ESRI de bases de dades geogràfiques (.mdb) • Format d'intercanvi MapInfo (.mif) per a dades vectorials • Llenguatge de marques per a dades geogràfiques (.kml) • Adobe Illustrator (.ai), dades CAD (.dxf o .svg) • Formats binaris de paquets GIS i CAD
Dades qualitatives (Text)	• Text XML (eXtensible Mark-up Language) segons un DTD (Document Type Definition) apropiat o esquema (.xml) • Rich Text Format (.rtf) • Dades de text pla, ASCII (.txt)	• Llenguatge de marques Hypertext (HTML) (.html) • Formats propietaris utilitzats abastament, p. ex. MS Word (.doc, .docx) • Alguns formats propietaris o específics de programari, p. ex. NUD*IST, NVivo i ATLAS.ti
Imatge digital	• TIFF versió 6 sense compressió (.tif)	• JPEG (.jpeg, .jpg) però només si és creat en aquest format • TIFF (altres versions) (.tif, .tiff) • Adobe Portable Document Format (PDF/A, PDF) (.pdf) • Format d'imatge RAW (.raw) • Photoshop (.psd)
Àudio digital	• FLAC sense pèrdua (.flac)	• MPEG-1 Audio Layer 3 (.mp3) però només si és creat en aquest format • Audio Interchange File Format (AIFF) (.aif) • Waveform Audio Format (WAV) (.wav)
Vídeo digital	• MPEG-4 (.mp4) • motion JPEG 2000 (.jp2)	
Documentació i *scripts*	• Rich Text Format (.rtf) • PDF/A o PDF (.pdf) • OpenDocument Text (.odt)	• Text pla (.txt) • Alguns formats propietaris utilitzats abastament, p. ex. MS Word (.doc, .docx) o MS Excel (.xls, .xlsx) • Text XML (eXtensible Mark-up Language) segons un DTD (Document Type Definition) apropiat o esquema (.xml)

Font: UK Data Archive. *File formats table*. <http://data-archive.ac.uk/create-manage/format/formats-table>. [Consulta: 25/03/2016]

DANS

DANS (Data Archiving and Networked Services)[157] és una institució neerlandesa del KNAW i del NWO fundat l'any 2005 que té com a missió la promoció de l'accés sostenible a les dades digitals de recerca. Per tal d'assegurar l'ús continuat d'aquests recursos l'arxiu segueix una política de preservació activa amb l'objectiu d'assegurar l'autenticitat, confiabilitat i integritat lògica de tots els recursos que es confien al seu càrrec mentre que proveu de formats adequats per a la recerca a llarg termini. La comunitat de l'arxiu consisteix en acadèmics en Humanitats i Ciències Socials. Per a l'arxiu de dades de recerca, s'utilitzen tres vies: el sistema d'arxiu EASY[158] (que compta amb el segell Data Seal of Approval), la xarxa DataverseNL[159] i el portal acadèmic NARCIS[160] (Dijk; Doorn, 2014). DANS té dues tipologies de formats, els quals es mostren en la seva totalitat a la Taula 8:

- Formats preferits són formats de fitxer que DANS considera que oferiran les millors garanties a llarg termini quant a usabilitat, accessibilitat i sostenibilitat. Dipositar dades de recerca en formats preferits sempre estarà acceptat per DANS
- Formats acceptats són formats de fitxer que s'utilitzen de forma àmplia a més dels formats preferits, i que són raonablement usables, accessibles i robustos a llarg termini. DANS afavoreix l'ús de formats preferits, però els formats acceptats també es permetran en la major part dels casos

[157] <http://www.dans.knaw.nl/en>. [Consulta: 25/03/2016]

[158] <https://easy.dans.knaw.nl/ui/home>. [Consulta: 25/03/2016]

[159] <https://dataverse.nl/dvn/>. [Consulta: 25/03/2016]

[160] <http://www.narcis.nl/>. [Consulta: 25/03/2016]

Taula 8. Formats requerits de dades al DANS

Tipus	Formats preferits	Formats acceptats
Documents de text	PDF/A (.pdf)	• ODT (.odt) • MS Word (.doc, .docx) • RTF (.rtf) • PDF (.pdf)
Text pla	Unicode (.txt)	No Unicode (.txt)
Llenguatge de marques	• XML (.xml) • HTML (.html; .xhtml) (Sempre que sigui vàlid i complet) • Si fos necessari: Fitxers relacionats: .css; .xslt; .js; .es	SGML (.sgml)
Fulls de càlcul	• ODS (.ods) • CSV (.csv)	• MS Excel (.xls, .xlsx) • PDF/A (.pdf) • OOXML (.docx, .docm)
Bases de dades	• SQL (.sql) • SIARD (.siard) • Taules DB (.csv)	• MS Access (.mdb, .accdb) (v. 2000 o posterior) • dBASE (.dbf) (v.7 o posterior) • HDF5 (.hdf5, he5, .h5)
Dades estadístiques	• SPSS portable (.por) • SPSS (.sav) • Stata (.dta) • DDI (.xml) • Dades (.csv) + configuració (.txt)	• SAS (.7bdat; .sd2; .tpt) • R (pendent d'estudi)
Imatges rasteritzades	• JPEG (.jpg; .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2)	• DICOM (.dcm)
Imatges vectorials	SVG (.svg)	• Illustrator (.ai) • EPS (.eps)
Àudio	• WAVE; BWF (.wav) • FLAC (.flac)	• AIFF (.aif, .aiff) • MP3 (.mp3) • AAC (.aac, .m4a)
Vídeo	• MPEG-2 (.mpg; . mpeg,...) • MPEG-4 H.264 (.mp4) • AVI sense pèrdua (.avi) • QuickTime (.mov)	MKV (.mkv)
CAD	AutoCAD DXF v. R12 (.dxf)	AutoCAD, altres versions (.dwg, dxf)
GIS	• GML (.gml) • MIF/MID (.mif/.mid)	• ESRI Shapefiles (.shp i fitxers relacionats) • MapInfo (.tab i fitxers relacionats) • KML (.kml)
Imatges (georeferència)	GeoTIFF (.tif, .tiff)	TIFF World File (.tfw i .tif)
GIS rasteritzat	ASCII GRID (.asc, .txt)	• ESRI Grid (.grd i fitxers relacionats)

Tipus	Formats preferits	Formats acceptats
3D	• WaveFront Object (.obj) • X3D (.x3D)	• COLLADA (.dae) • Autodesk FBX (.fbx)
RDF	Estàndards W3C	
Computer Assisted Qualitative Data Analysis (CAQDAS)	Formats utilitzats en aplicació, processats en funció de cada tipus de fitxer individual	• Formats d'exportació de l'aplicació (paquet de còpia ATLAS.TI; projecte d'exportació NVIVO; etc.) • QuDEX

Font: DANS (2015). *Preferred formats: September 2015, version 3.0.* <www.dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf>. [Consulta: 25/03/2016]

2.3.2 Archivematica

Un projecte on s'ha fet una feina extensiva de normalització de formats de fitxers adients per a la preservació és el programari de codi obert Archivematica[161], que té com a base conceptual un model OAIS totalment funcional mitjançant l'ús de microserveis, cadascun dels quals és equivalent a un paquet d'informació OAIS (vegeu el capítol 3.2.1): el Paquet d'Informació d'Enviament, el Paquet d'Informació d'Arxiu i el Paquet d'Informació de Difusió (Van Garderen, 2010). La primera versió que es publicà fou la 0.6 alpha en maig de 2010, sent la versió estable més recent la 1.5 amb data de juny de 2016. El codi d'Archivematica es publica sota una llicència GNU i la documentació es publica sota una llicència Creative Commons.

Pel que fa als formats de fitxer, Archivematica gestiona una base de dades, la FPR (Format Policy Registry), la qual és un registre per a les seves polítiques per a la normalització, extracció i identificació de formats. El criteri per seleccionar els formats és que aquests han de ser lliures de restriccions de llicències i patents, han de tenir especificacions d'accés lliure i han de ser utilitzats abastament i/o avalats per repositoris importants (Van Garderen; Mumma, 2013). La Taula 9 descriu les polítiques actuals.

[161] <https://www.archivematica.org/es/>. [Consulta: 26/10/2016]

Taula 9. Polítiques de formats de fitxer a Archivematica

Tipus	Formats de fitxer	Formats de preservació	Formats d'accés	Eina de normalització
Àudio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Correu electrònic	PST	MBOX	MBOX	readpst
Office Open XML	DOCX, PPTX, XLSX	Format original	Format original	Per definir
Text pla	TXT	Format original	Format original	Cap
Portable Document Format	PDF	PDF/A	Format original	Ghostscript
Presentacions	PPT	Format original	PDF	Per definir
Imatges rasteritzades	BMP, GIF, JPG, JP2, PCT, PNG, PSD, TIFF, TGA	TIFF sense comprimir	JPEG	ImageMagick
Fitxers *raw* de càmera	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	Format original	JPEG	ImageMagick/ UFRaw
Fulls de càlcul	XLS	Format original	Format original	Cap
Imatges vectorials	AI, EPS, SVG	SVG	PDF	Inkscape
Vídeo	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV	FFV1/LPCM en contenidor MKV	MP4	FFmpeg
Fitxers Word	DOC, WPD, RTF	Format original	Format original	Per definir

Font: Archivematica (2016, Aug. 22). *Format policies*. <https://wiki.archivematica.org/Format_policies>. [Consulta: 26/10/2016]

Val a dir que a les primeres versions d'Archivematica sí que es normalitzaven els fitxers Word (Microsoft Word o Word Perfect) a PDF o formats Open Office, però les proves van mostrar que els resultats eren massa inconsistents, amb pèrdues significatives d'informació de format.

2.3.3 Síntesi de formats preferits

Un cop feta aquesta panoràmica de tipologies de formats, s'ha elaborat una simplificació de tipologies de formats, ja que s'han eliminat formats massa especialitzats com les dades d'estructures químiques, les bases de dades o els models de dades. Per tant, del total de formats presentats s'han seleccionat aquells de tipus textual, visual i audiovisual. Els resultats es mostren a la Taula 10.

Taula 10. Formats preferits de preservació als repositoris estudiats i a Archivematica

Tipus	3TU.Datacentrum	ADS	UK Data Archive	DANS	Archivematica
Text	• PDF • Text pla	• MS Word • Open Document Text	• Text pla • Rich Text Format • PDF • PDF/A, • HTML • OpenDocument Text	• PDF • Text pla	• PDF/A • Text pla
Text amb llenguatge de marques	• SGML amb DTD (.sgm, .sgml) • XML amb DTD (.xml)	• No s'indiquen, però sí que hi ha d'acceptats: • SGML • XML	XML amb DTD	• XML • HTML	No s'indiquen
Fulls de càlcul i dades tabulars	• Valors separats per comes CSV • Fitxer delimitat per tabulacions • PDF/A	• Valors separats per comes CSV • MS Excel • OpenDocument Spreadsheet	• Valors separats per comes CSV • Fitxer delimitat per tabulacions	• OpenDocument Spreadsheet • Valors separats per comes CSV	• XLS
Dades estadístiques	No s'indiquen	Text delimitat	• Format portable SPSS • Text delimitat amb fitxer de configuració • Text estructurat	• Format portable SPSS • SPSS • Stata • DDI • Text delimitat amb fitxer de configuració	No s'indiquen
Imatges rasteritzades	• JPEG • TIFF	• TIFF versió 6 sense compressió	• TIFF versió 6 sense compressió	• JPEG • TIFF • PNG • JPEG 2000	TIFF sense compressió
Gràfics vectorials	No s'indiquen	• AutoCAD • Drawing Interchange Format • SVG	No s'indiquen	• SVG • AutoCAD	SVG

Tipus	3TU.Datacentrum	ADS	UK Data Archive	DANS	Archivematica
Dades geoespacials (GIS)	• ESRI Shapefile • TIFF amb georeferències • Google Earth • Llenguatge de marques geogràfic	• ESRI Shapefile • TIFF amb georeferències • Llenguatge de marques geogràfic	• ESRI Shapefile • TIFF amb georeferències • Dades CAD • Dades amb atributs GIS tabulars	• MIF/MID • TIFF amb georeferències • ASCII GRID • Llenguatge de marques geogràfic	No s'indiquen
Gràfics 3D	No s'indiquen	• X3D • VRML • Java3D • QTVR	No s'indiquen	• WaveFront Object • X3D	No s'indiquen
Àudio	WAVE	• BWF • WAVE • AIFF	FLAC	• WAVE • BWF • FLAC (.flac)	WAVE
Vídeo	No hi ha formats preferits, però sí acceptats: • AVI • MPEG-1 • MPEG-2 • MPEG-4	• MPEG-1 • MPEG-2 • MPEG-4	• MPEG-4 • Motion JPEG 2000	• MPEG-2 • MPEG-4 H.264 • AVI sense pèrdua • QuickTime	FFV1/LPCM en contenidor MKV

Font: Elaboració pròpia a partir de les taules 5 a la 9 (ambdues incloses)

A continuació es descriuen les característiques dels formats que tenen més preferència als repositoris. Per tal de fer la selecció de formats, s'han triat aquells presents a dos o més repositoris i que siguin especialment rellevants per a la preservació a llarg termini. En els casos que ha estat necessari, s'ha aportat una definició de la tipologia de dades.

Text

El format PDF d'Adobe és un estàndard obert des de 2008 i publicat com a norma ISO 32000-1[162]. Una de les seves extensions és el PDF/A, un format especialitzat per a la preservació digital de documents electrònics [163] i publicat com a estàndard ISO

[162] *ISO 32000-1:2008. Document management -- Portable document format -- Part 1: PDF 1.7.* <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=51502>. [Consulta: 28/03/2016]

[163] *PDF/A facts: an introduction to the standard.* <http://www.pdfa.org/2013/02/pdfa-facts/>. [Consulta: 28/03/2016]

19005[164][165]. Encara que només un repositori especifica aquesta extensió com a format preferit, val la pena tenir-lo en compte per les seves característiques.

En el cas del text pla, el més habitual és que els fitxers de text pla tinguin l'extensió TXT, els quals es poden obrir fàcilment amb diverses aplicacions. Els fitxers de text poden utilitzar diversos conjunts de caràcters, com ASCII i Unicode. Al repositori DANS es considera que l'ús del conjunt de caràcters Unicode amb la utilització de les codificacions Byte Order Mark i UTF garanteix que tots els caràcters es representen de forma correcta a qualsevol entorn informàtic.

Finalment, l'OpenDocument Text és un format obert i estàndard, publicat com a ISO/IEC 26300[166][167][168] i desenvolupat l'any 2005 pel consorci sense ànim de lucre OASIS[169]. Al ser un format obert, permet l'accés a llarg termini de les dades sense barreres tècniques o legals.

Els fitxers de text es poden utilitzar també amb llenguatges de marques, els quals són una forma de codificar un document que, juntament amb el text, incorpora etiquetes o marques amb informació sobre l'estructura del text o la seva presentació. Els més adequats per a la preservació són XML i SGML.

El llenguatge de marques SGML és un estàndard desenvolupat per la norma ISO 8879[170] originalment l'any 1986, amb revisions fetes els anys 1996 i 1998. En el cas de

---

[164] *ISO 19005-1:2005. Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1*).
<http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920>. [Consulta: 28/03/2016]

[165] *ISO 19005-2:2011. Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2).*
<http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50655>. [Consulta: 28/03/2016]

[166] *ISO/IEC 26300-1:2015. Information technology -- Open Document Format for Office Applications (OpenDocument) v1.2 -- Part 1: OpenDocument Schema*.
<http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=66363>. [Consulta: 28/03/2016]

[167] *ISO/IEC 26300-2:2015. Information technology -- Open Document Format for Office Applications (OpenDocument) v1.2 -- Part 2: Recalculated Formula (OpenFormula) Format.*
<http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=66375>. [Consulta: 28/03/2016]

[168] *ISO/IEC 26300-3:2015. Information technology -- Open Document Format for Office Applications (OpenDocument) v1.2 -- Part 3: Packages.*
<http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=66376>. [Consulta: 28/03/2016]

[169] <https://www.oasis-open.org/>. [Consulta: 28/03/2016]

[170] *ISO 8879:1986. Information processing -- Text and office systems -- Standard Generalized Markup Language (SGML).* <www.iso.org/iso/catalogue_detail.htm?csnumber=16387>. [Consulta: 28/03/2016]

XML, es tracta d'un estàndard obert, definit pel World Wide Web Consortium (W3C)[171] i creat l'any 1996 pel XML Working Group[172]. El motiu de la seva creació fou poder utilitzar una versió simplificada de SGML dissenyada pel seu ús en pàgines web.

Un tipus de llenguatge de marques específic per a dades geogràfiques és el GML (Geography Markup Language), el qual és una sintaxi XML adaptada per expressar característiques geogràfiques i definida pel Open Geospatial Consortium (OGC)[173] durant l'any 2000. Com a estàndard obert, té la norma ISO 19136[174].

Fulls de càlcul i dades tabulars

Un full de càlcul pot tenir un format, com l'ús de colors a les cel·les. S'ha de tenir en compte també l'estructura del full de càlcul: poden haver-hi cel·les amb càlculs fets amb valors d'altres cel·les. En aquest aspecte, OpenDocument Spreadsheet és un format obert i robust desenvolupat per OASIS i es recomana per a l'emmagatzematge permanent de fulls de càlcul amb càlculs i/o altres propietats estructurals.

Si és possible reduir el full de càlcul a una taula pla de files i columnes, una bona opció és crear un fitxer de text CSV, que no tenen format, fórmules ni enllaços a recursos externs. Altra opció és utilitzar fitxers delimitats per tabulacions que tampoc tenen format.

SPSS (Statistical Package for the Social Sciences) és un format propietari desenvolupat originalment per SPSS Inc. l'any 1968 i que actualment és propietat de la companyia IBM. Tot i ser un format propietari, es considera adequat per a la preservació perquè s'utilitza de forma freqüent i s'espera que el format romandrà accessible en el futur.

[171] <http://www.w3.org/>. [Consulta: 28/03/2016]

[172] *The birth of XML*. Recuperat del web Internet Archive. <https://web.archive.org/web/20120418030904/http://java.sun.com/xml/birth_of_xml.html>. [Consulta: 28/03/2016]

[173] <http://www.opengeospatial.org/>. [Consulta: 30/03/2016]

[174] *ISO 19136:2007. Geographic information -- Geography Markup Language (GML).* <http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32554>. [Consulta: 30/03/2016]

Imatges rasteritzades

TIFF és un format per a imatges creat originalment per la companyia Aldus l'any 1986, però es troba controlat avui en dia per Adobe. La versió 6.0, publicada l'any 1992[175] és la que recomanen els repositoris, ja que es considera el millor format per a l'arxiu d'imatges[176]. Una extensió de la versió 6.0, TIFF/IT, ja és un estàndard amb la norma ISO 12639[177].

JPEG és un mètode de compressió amb pèrdua per a imatges digitals desenvolupat l'any 1992 pel grup homònim. Com a estàndard, té la norma ISO/IEC 10918-1[178]. El seu ús dins els conjunts de dades és complementari amb TIFF, ja que ocupa molt menys espai. DANS, per exemple, recomana l'ús de TIFF per a l'arxiu d'imatges i utilitzar JPEG per a la publicació.

Gràfics vectorials

SVG és un format de fitxer basat en XML per a gràfics vectorials dinàmics. Es tracta d'un estàndard obert desenvolupat pel W3C l'any 1999. Per altra banda, AutoCAD és un format propietari de la companyia Autodesk llançat inicialment l'any 1982 i s'ha convertit en l'estàndard *de facto* per a l'intercanvi de gràfics CAD en dues dimensions degut al llarg temps que porta al mercat.

Àudio

Quan es vol preservar àudio a llarg termini s'han d'utilitzar formats sense pèrdua, que tinguin la millor qualitat sense pèrdua de dades. El format sense compressió WAVE, desenvolupat per Microsoft i IBM l'any 1991, s'utilitza abastament per a àudio sense comprimir, però té la limitació de 4 GB que corresponen a 6,8 hores d'àudio. Una extensió de WAVE, el BWF, fou creada per l'European Broadcasting Union durant

[175] *TIFF: revision 6.0 final*. <http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf>. [Consulta: 28/03/2016].

[176] *Tagged Image File Format (TIFF)*. <http://ec.europa.eu/ipg/standards/image/tiff/>. [Consulta: 21/04/2016]

[177] *ISO 12639:1998. Graphic technology -- Prepress digital data exchange -- Tag image file format for image technology (TIFF/IT).* <http://www.iso.org/iso/catalogue_detail?csnumber=2181>. [Consulta: 28/03/2016]

[178] *ISO/IEC 10918-1:1994. Information technology -- Digital compression and coding of continuous-tone still images: Requirements and guidelines.*
<http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=18902>. [Consulta: 28/03/2016]

l'any 1997 i permet l'ús de fitxers més grans de 4 GB mitjançant l'especificació RF64[179].

Una alternativa és FLAC, un format de compressió sense pèrdua, obert i lliure i també és el nom d'un còdec d'àudio sense pèrdua[180]. Fou desenvolupat l'any 2001 per Josh Coalson i la Xiph.Org Foundation[181] i presenta l'avantatge de comprimir àudio entre un 40 i un 50%, la qual cosa permet un estalvi important en espai d'emmagatzematge, sense pèrdues en les dades. Com que FLAC és un format sense pèrdua, és possible restaurar les dades originals d'un suport en àudio si aquest es perd o queda danyat.

Vídeo

En el cas del vídeo digital, és necessari fer la compressió degut al seu gran pes i per motius d'eficiència. El grup MPEG fou l'encarregat de desenvolupar el format de compressió MPEG-2, el qual és un estàndard obert segons la norma ISO/IEC 13818[182], desenvolupat per a DVD i televisió digital. No està optimitzat per a *bitrates* baixos (inferiors a 1 Mbit/s) però ofereix una qualitat superior a *bitrates* superiors (superiors a 3 Mbit/s) si el comparem amb MPEG-1.

MPEG-4 és un estàndard obert desenvolupat de l'estàndard de vídeo MPEG amb ISO/IEC 14496[183]. Les millores que es van realitzar foren especialment per facilitar l'ús d'equipament audiovisual a Internet mitjançant una millor compressió amb el còdec H.264, el qual fou desenvolupat pel Video Coding Experts Group (VCEG) juntament amb el grup MPEG.

[179] *MBWF/RF64: an extended file format for audio*. <https://tech.ebu.ch/docs/tech/tech3306-2009.pdf>. [Consulta: 29/03/2016]

[180] *What is FLAC?* <https://xiph.org/flac/>. [Consulta: 29/03/2016]

[181] <https://xiph.org/>. [Consulta: 29/03/2016]

[182] *ISO/IEC 13818-1:2000. Information technology -- Generic coding of moving pictures and associated audio information: Systems.* <http://www.iso.org/iso/catalogue_detail?csnumber=31537>. [Consulta: 29/03/2016]

[183] *MPEG-4.* <http://mpeg.chiariglione.org/standards/mpeg-4>. [Consulta: 29/03/2016]

S'ha de tenir en compte que MPEG-2 i MPEG-4 són formats patentats, però existeixen diversos còdecs de codi obert que permeten el seu ús mitjançant llicències com la GNU LGPL[184].

Dades geoespacials

Els sistemes d'informació geogràfica o GIS utilitzen diversos tipus de format per a l'intercanvi de dades geogràfiques. El format ESRI Shapefile és propietat de la companyia Esri[185], i serveix per emmagatzemar la localització d'elements geogràfics i els atributs associats. Com que es tracta d'un format multiarxiu, requereix d'un mínim de tres fitxers:

- .shp: fitxer amb entitats geomètriques dels objectes
- .shx: fitxer amb l'índex de les entitats geomètriques
- .dbf: base de dades en format dBASE on es guarda la informació dels atributs dels objectes

Tot i ser un format propietari, compta amb especificacions obertes per a la interoperabilitat de dades entre ESRI i altres productes informàtics. A més, es troba suportat abastament per sistemes d'informació geogràfics i per programari de codi obert[186].

Finalment, el GeoTIFF[187] o TIFF amb georeferències és un estàndard de metadades obert i de domini públic, creat pel Dr. Niles Ritter quan treballava al Jet Propulsion Laboratory de la NASA[188], que permet incrustar informació geogràfica dins un fitxer TIFF. Això pot incloure sistemes de coordenades, el·lipsoides, datums i tot allò necessari per tal que la imatge pugui posicionar-se de forma automàtica en un sistema de referència espacial. La imatge TIFF ha d'estar acompanyada amb un fitxer de text

[184] *Llicència Pública General Reduïda de GNU.* <https://www.gnu.org/copyleft/lesser.html>. [Consulta: 29/03/2016]

[185] <http://www.esri.com/>. [Consulta: 30/03/2016]

[186] *ESRI Shapefile*. <http://www.digitalpreservation.gov/formats/fdd/fdd000280.shtml>. [Consulta: 30/03/2016]

[187] <http://trac.osgeo.org/geotiff/>. [Consulta: 30/03/2016]

[188] *Who owns the GeoTIFF format?*
<http://www.remotesensing.org/geotiff/faq.html#Who%20owns%20GeoTIFF%20Format?> [Consulta: 30/03/2016]

pla amb extensió .tfw, el qual conté la mida en píxels dels eixos de coordenades X i Y, informació rotacional i informació de les coordenades horitzontal i vertical.

Gràfics 3D

El format més adequat per a la preservació de gràfics en 3D seria l'X3D, un format estàndard obert[189] amb les ISO/IECs 19775[190], 19776[191] i 19777[192]. Fou desenvolupat pel Web 3D Consortium[193] com a sistema per a emmagatzematge, recuperació i reproducció de gràfics en temps real. Compta amb diverses característiques, com integració amb XML o interoperabilitat amb el web.

2.4 Marc legal de les dades de recerca

Les dades de recerca obertes presenten diversos problemes quant al seu marc legal, ja que s'han de considerar els drets legals de propietat intel·lectual dins Espanya, així com les llicències que permeten el seu ús (que les dades siguin obertes no significa que els seus autors renunciïn als seus drets), i els drets que tenen els subjectes d'estudi a la seva privacitat i al seu dret de ser anònims dins la xarxa.

2.4.1 Propietat intel·lectual

La Llei de Propietat Intel·lectual que es troba vigent avui en dia (España, 1996) té com a principis bàsics el reconeixement de drets d'autor o drets morals (que protegeixen el reconeixement d'autoria de l'obra, la seva integritat, la forma de divulgació i si cal, el

[189] *What is X3D*. <http://www.web3d.org/x3d/what-x3d>. [Consulta: 30/03/2016]

[190] *ISO/IEC 19775-1:2013. Information technology -- Computer graphics, image processing and environmental data representation -- Extensible 3D (X3D) -- Part 1: Architecture and base component.* <http://www.iso.org/iso/catalogue_detail?csnumber=60760>. [Consulta: 30/03/2016]

[191] *ISO/IEC 19776-1:2015. Information technology -- Computer graphics, image processing and environmental data representation -- Extensible 3D (X3D) encodings -- Part 1: Extensible Markup Language (XML) encoding.* <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=60502>. [Consulta: 30/03/2016]

[192] *ISO/IEC 19777-1:2006. Information technology -- Computer graphics and image processing -- Extensible 3D (X3D) language bindings -- Part 1: ECMAScrip.* <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=33915>. [Consulta: 30/03/2016]

[193] <http://www.web3d.org/>. [Consulta: 30/03/2016]

dret de l'autor a retirar l'obra o modificar-la) que permeten fer ús dels drets patrimonials o d'explotació (en anglès s'utilitza el terme *copyright*). Els drets patrimonials són quatre: reproducció, distribució, comunicació pública i transformació.

Quant a les dades de recerca, l'article 12 indica el següent:

> 1. También son objeto de propiedad intelectual, en los términos del Libro I de la presente Ley, las colecciones de obras ajenas, de datos o de otros elementos independientes como las antologías y las bases de datos que por la selección o disposición de sus contenidos constituyan creaciones intelectuales, sin perjuicio, en su caso, de los derechos que pudieran subsistir sobre dichos contenidos. La protección reconocida en el presente artículo a estas colecciones se refiere únicamente a su estructura en cuanto forma de expresión de la selección o disposición de sus contenidos, no siendo extensiva a éstos.
> 2. A efectos de la presente Ley, y sin perjuicio de lo dispuesto en el apartado anterior, se consideran bases de datos las colecciones de obras, de datos, o de otros elementos independientes dispuestos de manera sistemática o metódica y accesibles individualmente por medios electrónicos o de otra forma.
> 3. La protección reconocida a las bases de datos en virtud del presente artículo no se aplicará a los programas de ordenador utilizados en la fabricación o en el funcionamiento de bases de datos accesibles por medios electrónicos.

Per tant, podem veure que les col·leccions de dades o *datasets* es troben protegides per la LPI sempre i quan sigui constitueixin una obra original, entesa com que la selecció i disposició dels continguts formen l'obra, o bé suposin una inversió substancial (Bercovitz Rodríguez-Cano, 2006, p. 262). Aquest últim concepte és l'anomenat dret *sui generis*, que està regulat a la Llei 5/98 (España, 1998), que protegeix la inversió en temps, diners i esforç per assolir els continguts de la col·lecció de dades.

Les dades científiques en accés obert es difonen en funció de les condicions establertes per les agències de finançament, però s'ha de tenir en compte que la preservació implica fer algunes operacions sobre les dades originals com el *refreshing* o la migració, la qual cosa possiblement requereixi l'exercici del dret de transformació. En el primer cas, el *refreshing* només implica un canvi de suport. Això no és un acte de transformació, ja que no hi ha cap aportació original que doni com a resultat una obra diferent a l'anterior.

Però, en el cas que el conjunt de dades sigui interoperable i permeti la creació d'una nova obra (com un nou producte científic) ens trobem clarament amb un acte de transformació. Així doncs, per tal de complir la llei el titular dels drets haurà de fer una cessió explícita dels seus drets de transformació quan dipositi les seves dades, si és que no es fa esment a l'acord de finançament amb l'agència.

Pel que respecta al dret de reproducció, la reforma de la LPI de 2006 va incloure una novetat: les biblioteques poden fer reproduccions sempre que es facin sense finalitat lucrativa. En el cas de la preservació dels *datasets*, això vol dir que es poden fer còpies per motius de conservació (no es fa esment al suport ni al nombre de còpies) si no actuen com a obstacle a l'explotació normal de l'obra o causen un perjudici als interessos legítims de l'autor (Fernández Molina, 2010).

La reforma inclou un nou límit al dret de comunicació pública; en el cas dels *datasets*, això implica que els centres poden comunicar-los públicament sense autorització de l'autor a efectes de recerca si es fa mitjançant una xarxa tancada i interna als terminals de la biblioteca (no obstant, hi ha una obligació de remunerar a l'autor). Per tant, si no es compta amb la cessió d'aquest dret per part de l'autor, la biblioteca o el centre només podrà fer-ne difusió dins les seves instal·lacions.

2.4.2 Llicències

Les llicències d'ús són cabdals per poder compartir les dades i permetre la seva reutilització de forma legal. L'Open Data Institute ho deixa ben clar amb aquesta declaració: "Open data has to have a licence that says it is open data. Without a licence, the data can't be reused" [194]. Existeixen diverses llicències d'ús actualment, però "arguably the most important standard open data licence offered to date is the Creative Common [sic] (CC) licence" (Khayyat; Bannister, 2015). Utilitza els mateixos principis que el *copyleft* que permeten exercir drets d'autor que permetin la lliure distribució de còpies i versions modificades, sempre i quan els mateixos drets es preservin a les versions modificades. A més, es una llicència que es considera per a molts com la forma predeterminada per defecte de protegir la seva obra quan la deixen disponible en la

[194] Open Data Institute. *What makes data open?* <https://theodi.org/guides/what-open-data>. [Consulta: 11/09/2016]

forma que ells desitgen (Murray-Rust, 2008). Per tal d'implementar una llicència CC[195] l'investigador primer ha d'escollir una o vàries d'aquestes quatre opcions:

- Reconeixement (Attribution). En qualsevol explotació de l'obra autoritzada per la llicència caldrà reconèixer l'autoria
- No Comercial (Non commercial): L'explotació de l'obra queda limitada a usos no comercials
- Sense Obres Derivades (No Derivate Works): L'autorització per explotar l'obra no inclou la transformació per crear una obra derivada
- Compartir Igual (Share alike): L'explotació autoritzada inclou la creació d'obres derivades sempre que mantinguin la mateixa llicència en ser divulgades

Un cop s'hagi triat una o més d'aquestes opcions, es generaran sis opcions de llicències:

- Reconeixement (by): Es permet qualsevol explotació de l'obra, incloent-hi una finalitat comercial, així com la creació d'obres derivades, la distribució de les quals també està permesa sense cap restricció
- Reconeixement – No Comercial (by-nc): Es permet la generació d'obres derivades sempre que no se'n faci un ús comercial. Tampoc es pot utilitzar l'obra original amb finalitats comercials
- Reconeixement – No Comercial – Compartir Igual (by-nc-sa): No es permet un ús comercial de l'obra original ni de les possibles obres derivades, la distribució de les quals s'ha de fer amb una llicència igual a la que regula l'obra original
- Reconeixement – No Comercial – Sense Obra Derivada (by-nc-nd): No es permet un ús comercial de l'obra original ni la generació d'obres derivades
- Reconeixement – Compartir Igual (by-sa): Es permet l'ús comercial de l'obra i de les possibles obres derivades, la distribució de les quals s'ha de fer amb una llicència igual a la que regula l'obra original
- Reconeixement – Sense Obra Derivada (by-nd): Es permet l'ús comercial de l'obra però no la generació d'obres derivades

[195] Creative Commons Catalunya. *Llicències.* <http://cat.creativecommons.org/?page_id=184>. [Consulta: 11/09/2016]

Una llicència addicional és la CCZero, que no presenta cap restricció ni requeriment, ni tan sols l'atribució de la font. Per tant, és una llicència que deixa l'obra en el domini públic i que permet a l'autor fer una renúncia dels seus drets sempre i quan la llei ho permeti.

Un exemple de llicència de dades amb CC el trobem al servei de l'Ajuntament de Barcelona OpenDataBCN[196], que fa servir una llicència CC - Reconeixement (CC-BY 3.0), excepte en els casos de dades on hi ha participació de tercers, on s'utilitza la llicència Reconeixement - Sense Obra Derivada (CC BY-ND 3.0).

L'OKF, dins la seva política de suport a la difusió del coneixement obert, va crear l'any 2009 un conjunt de llicències pensades expressament per a les dades obertes, Open Data Commons[197], el qual permet que les dades i bases de dades siguin obertes de forma legal amb certes condicions. Els tipus de llicències disponibles són tres:

- Public Domain Dedication and License (PDDL). Permet posar en domini públic dades i bases de dades, renunciant a tots els drets
- Attribution License (ODC-By). Requereix atribuir l'ús públic de la base de dades, o obres produïdes des de la base de dades
- Open Database License (ODC-ODbL). Requereix atribuir l'ús públic de la base de dades, o obres produïdes des de la base de dades, compartir obres derivades sota la mateixa llicència ODbL i redistribuir les obres derivades de forma oberta

Les llicències, per tant, que permeten compartir i reutilitzar dades de recerca de forma més efectiva serien la CCZero i la PDDL, ja que asseguren el compliment amb els protocols del Science Commons i amb la definició de dades obertes de l'OKF, tal com s'indica als Principis Panton (Molloy, 2011).

[196] <http://opendata.bcn.cat>. [Consulta: 11/09/2016]

[197] <http://opendatacommons.org/>. [Consulta: 11/09/2016]

2.4.3 Privacitat

La privacitat és una qüestió que tot investigador ha de tenir en compte a l'hora de gestionar les seves dades de recerca. L'article 18 de la Constitució espanyola estableix el següent, respecte a la intimitat personal (España, 1978):

> 1. Se garantiza el derecho al honor, a la intimidad personal y familiar y a la propia imagen.
> 2. El domicilio es inviolable. Ninguna entrada o registro podrá hacerse en él sin consentimiento del titular o resolución judicial, salvo en caso de flagrante delito.
> 3. Se garantiza el secreto de las comunicaciones y, en especial, de las postales, telegráficas y telefónicas, salvo resolución judicial.
> 4. La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos.

En cas que l'investigador hagi fet estudis que requereixen de dades personals, haurà de fer els esforços necessaris per tal de respectar el dret de respecte de l'honor i a la pròpia imatge. Per tal de regular aquest drets, es va promulgar la Llei Orgànica de Protecció de Dades de Caràcter Personal (España, 1999) que té com objecte "garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar". L'investigador ha de tenir presents els següents conceptes, tal i com s'indiquen al Reglament de desenvolupament de la LOPD (España, 2008):

- Dades de caràcter personal. Qualsevol informació pertanyent a persones físiques identificades o identificables
    - Dades públiques. Dades personals que són conegudes per un nombre important de persones sense que el titular de les dades necessàriament ho sàpiga
    - Dades privades: Dades personals que tenen regulades i taxades les situacions o circumstàncies en les quals les persones estan obligades a donar-les, com l'adreça, el número de compte corrent, etc.
    - Dades sensibles. Dades personals que es refereixen a característiques morals o físiques, com l'origen racial, conviccions religioses, vida sexual, condemnes criminals, etc.

- Consentiment. Qualsevol manifestació de la voluntat lliure, inequívoca, específica i informada, mitjançant la qual l'interessat dóna el seu consentiment per al tractament de les seves dades. El consentiment implica:
    - Que per obtenir les dades s'ha de demanar permís
    - Que s'ha d'informar l'ús que es farà de les dades
    - Que no es poden fer servir les dades per a usos diferents dels que es van informar

L'investigador ha de saber diferenciar correctament entre els tipus de dades personals, ja que (especialment en els estudis clínics) facilitar certes dades com l'origen racial i la ubicació geogràfica poden ser suficients per identificar una persona. Quant al consentiment, no serà vàlid si les dades es recullen de forma fraudulenta i tenint en compte que hi ha dades especialment protegides com la ideologia, la religió o l'origen racial. No és una qüestió baladina, ja que sovint els participants no entenen completament la terminologia que s'utilitza als estudis, com "compartició de dades" o "dades obertes" i es poden produir casos de destrucció de dades confidencials si la llei ho requereix (Childs et al., 2014). En aquest respecte, els titulars de les dades poden exercir, si ho consideren necessari, els seus drets ARCO mitjançant l'Agencia Española de Protección de Datos [198] en l'àmbit geogràfic espanyol i l'Autoritat Catalana de Protecció de Dades[199] en l'àmbit geogràfic català.

2.4.4 Dret a l'oblit

És un fet incontestable que les generacions actuals processen de forma pública la seva privada dins les xarxes socials. Es poden trobar fàcilment vídeos, imatges i comentaris amb noms i cognoms. Però, de la mateixa manera que un individu ha donat consentiment per publicar aquests continguts, també té dret a revocar aquest consentiment i exercir el seu dret de cancel·lació, si pensa que pot ser perjudicat per continguts que hagi publicat en el passat.

[198] <http://www.agpd.es/portalwebAGPD/index-ides-idphp.php/>. [Consulta: 11/09/2016]

[199] <http://www.apd.cat/ca/>. [Consulta: 11/09/2016]

En els últims anys ha hagut un augment de demandes per exercir aquest dret, com mostra la memòria de l'any 2010 de l'AEPD, en què es va registrar un augment del 56% de reclamacions, respecte el 2009, per eliminar dades a Internet (Simón Castellano, 2012). Recentment, el Tribunal de Justícia de la Unió Europea va sentenciar que el dret a l'oblit davant el motor de cerca Google és una aplicació dels drets d'oposició i cancel·lació dins l'entorn d'Internet. A nivell legislatiu, el Reglament General de Protecció de Dades es publicà en maig de 2016 al Diari Oficial de la Unió Europea (Unión Europea, 2016) i ja incorpora el dret de supressió o dret a l'oblit (art. 17), que indica el següent:

> 1. El interesado tendrá derecho a obtener sin dilación indebida del responsable del tratamiento la supresión de los datos personales que le conciernan, el cual estará obligado a suprimir sin dilación indebida los datos personales cuando concurra alguna de las circunstancias siguientes:
> a) los datos personales ya no sean necesarios en relación con los fines para los que fueron recogidos o tratados de otro modo;
> b) el interesado retire el consentimiento en que se basa el tratamiento de conformidad con el artículo 6, apartado 1, letra a), o el artículo 9, apartado 2, letra a), y este no se base en otro fundamento jurídico;
> c) el interesado se oponga al tratamiento con arreglo al artículo 21, apartado 1, y no prevalezcan otros motivos legítimos para el tratamiento, o el interesado se oponga al tratamiento con arreglo al artículo 21, apartado 2;
> d) los datos personales hayan sido tratados ilícitamente;
> e) los datos personales deban suprimirse para el cumplimiento de una obligación legal establecida en el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento;
> f) los datos personales se hayan obtenido en relación con la oferta de servicios de la sociedad de la información mencionados en el artículo 8, apartado 1.
>
> 2. Cuando haya hecho públicos los datos personales y esté obligado, en virtud de lo dispuesto en el apartado 1, a suprimir dichos datos, el responsable del tratamiento, teniendo en cuenta la tecnología disponible y el coste de su aplicación, adoptará medidas razonables, incluidas medidas técnicas, con miras a informar a los responsables que estén tratando los datos personales de la solicitud del interesado de supresión de cualquier enlace a esos datos personales, o cualquier copia o réplica de los mismos.

> 3. Los apartados 1 y 2 no se aplicarán cuando el tratamiento sea necesario:
> a) para ejercer el derecho a la libertad de expresión e información;
> b) para el cumplimiento de una obligación legal que requiera el tratamiento de datos impuesta por el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento, o para el cumplimiento de una misión realizada en interés público o en el ejercicio de poderes públicos conferidos al responsable;
> c) por razones de interés público en el ámbito de la salud pública de conformidad con el artículo 9, apartado 2, letras h) e i), y apartado 3;
> d) con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos, de conformidad con el artículo 89, apartado 1, en la medida en que el derecho indicado en el apartado 1 pudiera hacer imposible u obstaculizar gravemente el logro de los objetivos de dicho tratamiento, o
> e) para la formulación, el ejercicio o la defensa de reclamaciones.

L'investigador haurà de tenir en compte especialment aquest article quan dipositi les seves dades a un repositori, perquè l'interessat tindrà dret a què siguin esborrades les seves dades personals un cop aquestes hagin servit a la finalitat de la recerca, i sempre i quan aquesta recerca no quedi compromesa per l'esborrat de les dades. La millor opció seria que l'investigador tingui en compte això des del principi del seu projecte, per així saber quines són les dades que s'han d'esborrar o bloquejar degut a la informació personal que es trobi a les mateixes.

3. Preservació digital

La preservació digital és un camp enormement ampli, que inclou diverses aproximacions. Quan es projecta la preservació d'una col·lecció de documents o dades, hom ha de respondre diverses qüestions, com el pressupost i infraestructura necessaris, tenir personal format i qualificat per executar les tasques, la metodologia emprada, els estàndards a utilitzar, el temps que cal invertir o l'objectiu cabdal de la preservació (es prioritza la qualitat de la preservació o la quantitat d'objectes preservats?). Aquesta tesi no pretén analitzar tots els aspectes presents dins la preservació digital, sinó que es limita a explicar quines són les possibilitats tècniques més adients per assolir un model de preservació digital per a dades científiques de ciències socials i humanitats.

Així doncs, dins el present capítol, fem un estat de la qüestió quant a les possibilitats que tenen els investigadors per dipositar les dades a diversos tipus de repositoris; s'analitzen dos estàndards de preservació: el model de referència OAIS i el sistema d'auditoria TDR i per últim es fa un estudi dels esquemes de metadades que es faran servir a la nostra proposta per tal de demostrar la seva pertinència.

3.1 Dipòsits de dades

L'investigador disposa de diversos mètodes per compartir les dades i fer que aquestes es trobin disponibles a llarg termini. Una és l'ingrés de dades a repositoris, on podem trobar tres categories, especialitzats per disciplines, generals o institucionals (Nina-Alcocer; Blasco-Gil; Peset, 2013), altra és la publicació de dades juntament amb l'article de recerca en forma d'*enriched* o *enhanced publication* (García-García; López-Borrull; Peset, 2015) i una tercera seria guardar-les al web curricular de l'autor (Kousha; Thelwall, 2014) o del departament/projecte de recerca. El model basat en revistes presenta problemes com la preservació a llarg termini, ja que és un material complementari i no l'objectiu principal de la plataforma (els articles científics), però aquesta és una opció que agrada a una gran part dels investigadors degut al seu desig d'enllaçar les seves dades i les publicacions (Smit, 2011).

Com ja s'ha vist, l'opció que s'utilitza dins les polítiques de les agències de finançament per emmagatzemar i compartir els *datasets* es troba als repositoris de dades. Els repositoris institucionals presenten l'avantatge que són interoperables, tenen un

responsable institucional que garanteix la seva qualitat i continuïtat i compten amb personal ben format (Nina-Alcocer et al., 2013). Però, s'han d'enfrontar amb diferents barreres: la percepció general per part dels agents implicats (agències de finançament, biblioteques, centres de recerca, investigadors individuals) que els repositoris necessiten desenvolupar-se per tal de permetre aquesta interoperabilitat o la manca d'un imperatiu legal que faci que les dades de recerca ingressin obligatòriament a un repositori (European Commission, 2012c). Un altre problema dels repositoris institucionals és que a moltes universitats i centres d'investigació no hi han repositoris de dades; una possible solució a curt termini seria realitzar acords amb algun d'existent, ja sigui específic o d'ús general (Hernández-Pérez; García-Moreno, 2013). Una síntesi dels avantatges i desavantatges de les possibilitats existents les podem trobar a la Taula 11.

Taula 11. Tipologies de dipòsits de dades i els seus avantatges i desavantatges

Tipus de dipòsit	Avantatges	Desavantatges
Repositori de dades específic per a una disciplina, centre de dades o base de dades científica	Ofereix coneixement i experiència en gestió de dades que assegura la col·lecció es preserva i s'utilitza correctament	Requereix planificació avançada per complir amb estàndards de metadades i documentació
Repositori de dades d'ús general (p. ex., Dryad, Figshare, Zenodo)	Ofereix funcionalitats de recerca, navegació i visualització	Requereix escrutini de condicions per tal d'assegurar la consistència amb les polítiques del finançador, la publicació o la institució quant a la recuperació de costos, drets d'autor o preservació a llarg termini
Repositori de dades institucional	Accepta qualsevol tipus de dades, en especial si no hi ha altre dipòsit adient i garanteix els requeriments de la política de preservació a llarg termini	No compta amb els mateixos recursos que els repositoris específics o d'ús general
Servei de material de suport per a publicacions	Compleix amb els requeriments de la publicació o de l'editorial	Pot presentar costos i no és probable que ofereixi una solució a llarg termini o funcionalitats de repositori de dades
Pàgina web de departament, de projecte o personal	Permet funcionalitats adaptades a la col·lecció de dades i/o als usuaris de dades i a la xarxa d'investigadors	Hi ha menys possibilitats que la col·lecció de dades estigui visible als nous usuaris i de mantenir un accés a llarg termini de la col·lecció

Font: Whyte, 2015

En un principi, les opcions més favorables serien els repositoris de dades específics i els d'ús general, ja que donen les millors opcions de accés i especialment de preservació.

Presenten les millors opcions de visibilitat i de preservació de les col·leccions de dades i els seus inconvenients no són pas insalvables.

S'analitzen a continuació cinc serveis de dipòsits de dades. Com que aquesta tesi està orientada a la creació d'un model de preservació optimitzat per a un centre, s'han descartat les publicacions científiques que accepten *datasets* com a material complementari als articles i pàgines web.

3.1.1 ICPSR

L'Interuniversity Consortium for Political and Social Research [200] és un consorci internacional de més de 700 institucions acadèmiques amb seu a l'Institute for Social Research a la University of Michigan, que dóna serveis de formació en accés a les dades, preservació i mètodes d'anàlisi per a la comunitat de recerca en ciències socials. El seu arxiu especialitzat en ciències socials es fundà l'any 1962 i actualment està considerat com el més gran del món en la seva especialitat (Lyle et al., 2013) amb una col·lecció superior als 500.000 ítems digitals al seu repositori.

Per dipositar dades, els investigadors tenen dues opcions:

- Repositori ICPSR[201], que facilita un servei de curació de dades; les dades seran revisades per evitar problemes tècnics i distribuïdes en diversos formats. Si el dipositant paga per aquest servei, les dades es distribuiran de forma immediata. En cas contrari, les dades només estaran disponibles per als membres de l'ICPSR
- Repositori openICPSR[202], que permet l'accés lliure als usuaris per una tarifa de dipòsit de 600 dòlars (per als membres d'ICPSR no hi ha cap cost). Les dades es troben disponibles als usuaris de forma immediata sense cap tipus de curació; les dades seran accessibles tal i com arriben originalment

[200] <http://www.icpsr.umich.edu/icpsrweb/landing.jsp>. [Consulta: 26/04/2015]

[201] <http://www.icpsr.umich.edu/index.html>. [Consulta: 25/06/2016]

[202] <https://www.openicpsr.org/>. [Consulta: 25/06/2016]

El repositori facilita instruccions concretes sobre com preparar els conjunts de dades a dipositar al repositori; això inclou els formats a utilitzar, el nivell de confidencialitat de les dades o una descripció de l'estudi de recerca (ICPSR, 2012). Dins les tipologies de *datasets*, també accepten dades en brut relacionades amb articles, els anomenats *replication datasets*, que permeten a altres investigadors replicar les anàlisis presentades a publicacions.

La política de preservació[203] de l'ICPSR inclou el compliment amb els requeriments dels estàndards de preservació digital OAIS (Vardigan; Whiteman, 2007) i TDR, a més de l'afegit de l'adquisició del DSA l'any 2010 (Vardigan; Lyle, 2014). Per altra banda, el consorci és conscient dels reptes que ha d'afrontar dins la preservació a llarg termini: el canvi tecnològic, els nous continguts digitals i la formació i conscienciació del seu personal i de la comunitat científica quant a la importància de la preservació. Dins els processos que està adoptant, s'inclouen la migració de formats i l'autenticitat i integritat dels actius digitals.

3.1.2 Dryad

Una solució interessant d'arxiu de dades dins les disciplines científiques i mèdiques la proporciona l'organització sense ànim de lucre Dryad[204], que es concentra en un repositori digital homònim escrit en Java i desenvolupat sota el programari de codi obert DSpace pel Massachusetts Institute of Technology i Hewlett-Packard, que permet l'accés, reutilització i citació de les dades científiques de forma oberta. Encara que Dryad inicialment es va dissenyar com a un repositori per a dades ecològiques i biològiques, actualment és un repositori general per a dades científiques (Akers; Green, 2014) on s'inclouen disciplines com la medicina i les ciències socials (Krause et al., 2015) i compromès amb la preservació a llarg termini de l'arxiu de dades (Mannheimer et al., 2014). La comunitat es troba oberta a investigadors, societats científiques, revistes, editors, biblioteques, acadèmies i agències de recerca. Dryad permet l'arxiu de dades associades a qualsevol article publicat a més de vuitanta revistes científiques i/o

[203] *ICPSR Digital Preservation Policy Framework* (2007, Apr.). Last revised June 2012. <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/dpp-framework.html>. [Consulta: 01/06/2016]

[204] <http://datadryad.org>. [Consulta: 06/04/2015]

mèdiques, sense restricció de formats. A diferència de Dataverse, no es demana una preparació de les dades; Dryad es coordina amb les publicacions per integrar els articles i les dades al procés d'ingesta. A més, un tècnic (*digital curator*) de l'organització comprova tots els objectes digitals per verificar qualsevol problema que es pugui presentar. S'accepta qualsevol tipus de format per a les dades, però també es demana a l'investigador l'ús de formats preferents com Open Office o text pla per a text, PDF o JPEG per a imatge, FLAC o AIFF per a àudio, o AVI per a vídeo. Per altra banda, totes les dades i metadades reben un identificador DOI que facilita la seva citació i reutilització. Els estàndards de metadades emprats a Dryad són DC, per a la descripció del recurs; DDI, per a la gestió i administració del recurs; EML, per a recursos de la disciplina ecològica; i PREMIS, per a la preservació del recurs (Greenberg et al., 2009).

Una dada important és que els investigadors poden publicar les seves dades encara que les dades del seu article es publiquin dins una revista que no estigui patrocinada a Dryad. En aquest cas, és necessari fer el pagament de noranta o vuitanta dòlars, segons el cas. Pel que fa als drets d'autor, Dryad utilitza una llicència de domini públic CCZero[205] per tal de reduir els impediments legals i tècnics a la reutilització de dades.

Quan l'investigador ha fet l'enviament de dades, Dryad verifica els fitxers, comprova les possibles restriccions de drets d'autors i problemes que es puguin presentar amb dades sensibles. Amb les metadades, es completa i corregeix tota la informació, com la publicació associada o paraules clau i es registra l'identificador DOI per al paquet de dades, que es preserva a llarg termini dins la xarxa DataONE, la qual pot fer còpies en diferents formats per facilitar la preservació en alguns casos. S'ha de destacar que les dades només seran accessibles de forma oberta un cop la publicació associada es trobi disponible, sempre i quan no s'hagi seleccionat una opció diferent. Les metadades associades es comparteixen amb serveis d'indexació per tal de facilitar la recerca dels *datasets*.

[205] <http://creativecommons.org/publicdomain/zero/1.0/>. [Consulta: 06/04/2015]

3.1.3 Dataverse

Una solució que s'està aplicant dins les ciències socials és el programari en codi obert per a repositoris Dataverse[206], un projecte per compartir, preservar, citar, explorar i analitzar dades de recerca, que facilita la reutilització de dades, la seva replicació, a més que en tot moment els editors i distribuïdors de dades, juntament amb les institucions finançadores, reben el crèdit apropiat. Aquest projecte es va començar a desenvolupar l'any 2006 al Institute for Quantitative Social Science (IQSS) de la Harvard University gràcies al treball previ durant el període 1999-2006 al projecte Virtual Data Center (Crosas, 2011). Qualsevol institució pot descarregar de forma lliure el programari, consistent en una aplicació web, i no necessita que s'instal·li cap programari especialitzat addicional (King, 2007).

El funcionament de Dataverse és en xarxa; un cop l'investigador té les seves dades llestes juntament amb tota la documentació i les metadades descriptives que en permetin la seva reutilització, haurà de descarregar l'aplicació del web del projecte, seguidament haurà de seleccionar un repositori de Dataverse per depositar el *dataset*, i finalment les dades estaran disponibles en forma d'un *study*, un contenidor que inclou la informació catalogràfica, arxius de dades i arxius complementaris.

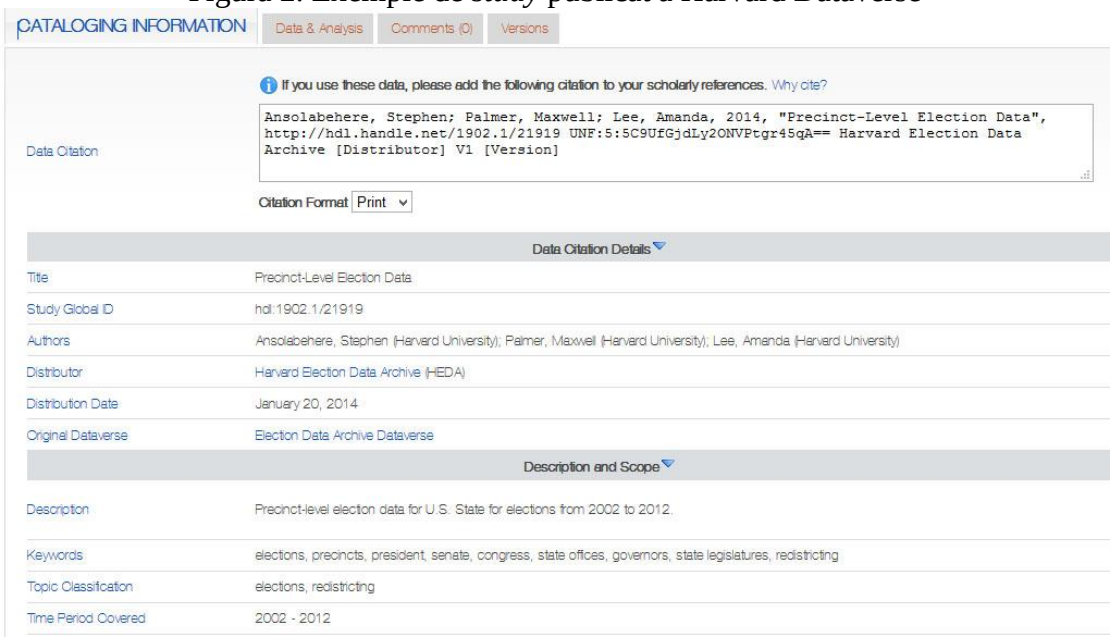
Existeixen repositoris Dataverse tant a universitats com a entitats governamentals. Actualment, els més destacats de la comunitat són els següents:

- Harvard Dataverse
- Odum Institute Dataverse
- DANS - Dutch Dataverse
- Fudan University Dataverse
- University of Alberta Libraries Dataverse
- Scholars Portal Dataverse
- Abacus Dataverse
- HeiDATA Dataverse
- UiT Open Research Data

[206] <http://dataverse.org/>. [Consulta: 05/04/2015]

Agafem per exemple el Harvard Dataverse que accepta dades de qualsevol disciplina. L'investigador ha de crear un compte, descriure el *dataset* per rebre una citació formal (amb una adreça URL persistent), pujar les dades, el codi i la documentació i finalment, publicar el data set. Gràcies a aquest sistema, qualsevol investigador pot consultar, descarregar i citar, si escau, de forma senzilla qualsevol dada que necessiti, ja que es troben tots els recursos que ha utilitzat l'investigador en un sol enllaç; això inclou *working papers*, el programari i les dades necessàries per replicar l'experiment.

Figura 2. Exemple de *study* publicat a Harvard Dataverse



Font: Harvard Dataverse. <https://dataverse.harvard.edu>. [Consulta: 05/04/2015]

L'autor té l'avantatge de no haver de renunciar a la propietat de les dades i com que de tota manera ha de planificar la gestió de les seves dades per aconseguir finançament, no és un pas complicat haver de preparar-les per a la seva ingesta a Dataverse. La citació generada inclou: autor(s), data de distribució, títol, identificador *handle* i URL, empremta numèrica universal (UNF) que serveix per identificar i verificar de forma unívoca el paquet de dades i altres camps opcionals com distribuïdor i versions. La URL presentada a Dataverse correspon al sistema Handle, que és el mateix que utilitza el sistema DOI per tenir referències permanents als objectes digitals i l'empremta UNF ajuda a verificar de forma permanent que les dades no han patit cap canvi respecte a les dades originals proporcionades per l'autor (Crosas, 2011). Per preservar les dades, el sistema de replicació LOCKSS en fa còpies múltiples, converteix els *datasets* a un

format d'arxiu que permet que es puguin convertir fàcilment a altres formats, exporta la informació catalogràfica de descripció de les dades a formats estàndards de metadades com Dublin Core o DDI (Altman; Crosas, 2013) i interopera amb altres sistemes mitjançant protocols com OAI-PMH per buscar i trobar dades d'altres aplicacions web (Crosas, 2012).

3.1.4 Figshare

El repositori figshare[207] és un projecte creat l'any 2011 amb el suport de Digital Science-Macmillan Publishers Co. (Nina-Alcocer et al., 2013) que permet als investigadors publicar i compartir fàcilment les seves dades de recerca de forma oberta mitjançant llicències Creative Commons, excepte en el cas dels *datasets*, que utilitzen la CCZero, la qual allibera l'obra dels totes les restriccions de propietat intel·lectual arreu del món[208]. Els investigadors no tenen cap restricció de format de dades, les quals poden incloure figures, *datasets*, vídeo, imatges, articles (els *pre-prints* també s'accepten), pòsters i grups de fitxers. Aquesta plataforma ja ha estat descrita com un Dropbox per a la recerca acadèmica, que a més de permetre la compartició de dades científiques, també dóna suport a la col·laboració entre investigadors (Leach-Murray, 2016).

El servei inclou l'obtenció d'un DOI, integració amb ORCID i curació de dades que verifica el seu contingut acadèmic. S'ha de destacar que el servei no només està adreçat a investigadors sinó també als editors i a les institucions acadèmiques que emmagatzemen dades. Un compte de figshare gratuït permet a l'usuari fins a 1 GB per a dades d'ús privat, però sense límits en el cas de les dades obertes; l'usuari també té accés a una interfície de programació oberta (*open API*) i espais de col·laboració de fins a cinc investigadors. Un compte de pagament, per altra banda, té un cost de 149 dòlars anuals que permet un emmagatzematge de 20 GB de dades d'ús privat amb un límit d'1 GB de mida de fitxer, tres espais de col·laboració de fins a vint investigadors i espai públic il·limitat. Una altra possibilitat és el compte institucional de figshare, que permet

[207] <http://figshare.com/>. [Consulta: 26/04/2015]

[208] *CC0*. <https://creativecommons.org/choose/zero/>. [Consulta: 26/04/2015]

a la universitat o institució un espai privat de per exemple 10 TB[209]. Les dades s'allotgen al servei Amazon Web Services[210], mentre que els fitxers i *datasets* públics es preserven mitjançant l'arxiu CLOCKSS[211].

Un recent estudi (Thelwall; Kousha, 2016) sobre l'impacte de figshare en la compartició de dades científiques va aportar, entre d'altres, els següents resultats, després d'analitzar 2.753 pàgines web d'aquesta plataforma:

- El tipus de recurs que més usuaris comparteixen dins la plataforma són els *datasets,* amb 1.283 usuaris, que representen un 47% del total. Per altra banda, és un dels que menys es consulten
- Les disciplines a les que pertanyen els investigadors que més comparteixen *datasets* són, per aquest ordre: informació i documentació, política científica i biologia
- Les disciplines a les que pertanyen els investigadors que més consulten *datasets* són, per aquest ordre: informació i documentació, informàtica aplicada i bioinformàtica

Aquestes dades són prou rellevants, ja que indiquen què figshare té prou èxit com a eina per a compartir *datasets,* encara que no es limita a aquest tipus de recurs. De fet, el recurs més visualitzat són els articles, amb un nombre màxim de consultes de 34.182, mentre que el corresponent als *datasets* només presenta un màxim de 7.336 consultes. No obstant això, és una eina molt útil per a tot aquell investigador que vulgui compartir el seu treball científic i preservar-lo, ja que figshare s'ha unit recentment al Digital Preservation Network[212], un servei que garanteix als seus membres la preservació a llarg termini dels seus actius digitals mitjançant el dipòsit al seu arxiu DPN, a més que assegura la propietat i la gestió dels continguts per part de figshare[213].

---

[209] Enis, Matt (2013, Sep. 11). "Figshare debuts repository platform for institutions". *The Digital Shift.* <http://www.thedigitalshift.com/2013/09/publishing/figshare-debuts-repository-platform-for-institution>. [Consulta: 29/05/2016]

[210] https://aws.amazon.com/es>. [Consulta: 29/05/2016]

[211] <https://clockss.org/clockss/Home/>. [Consulta: 29/05/2016]

[212] <http://dpn.org/>. [Consulta: 29/05/2016]

[213] Hyndman, Alan (2016, Jan. 20). *Figshare joins the Digital Preservation Network (DPN) to ensure survival, ownership and management of research data into the future*. <https://goo.gl/uopqpg>. [Consulta: 29/05/2016]

3.1.5 Zenodo

Un servei pensat per a la integració de l'entorn de treball de l'investigador és Zenodo[214], un repositori de dades d'àmbit general creat l'any 2013 mitjançant Invenio, un programari gratuït de creació de repositoris i desenvolupat conjuntament pel CERN (Nina-Alcocer et al., 2013) i OpenAIRE amb suport de la Comissió Europea, que permet pujar qualsevol tipus de fitxer mitjançant l'eina Dropbox, amb un sistema de metadades automàtic que es comunica amb serveis en línia com Mendeley, ORCID, CrossRef i OpenAIRE. S'ha de destacar que es permeten tipus de llicències diferents de Creative Commons, ja que cada investigador és lliure de decidir la llicència de dret d'autor que aplicarà als seus fitxers. Per altra banda, s'inclou la integració amb OpenAIRE, que permet informar a les agències de finançament de la Comissió Europea sobre els resultats de la recerca.

El servei inclou la possibilitat de crear repositoris propis dins aquest entorn i l'adquisició d'un identificador DOI per facilitar la citació de les dades i la seva curació, per tal de verificar que es trobin relacionades amb una recerca. S'accepten dades de qualsevol disciplina, sense restriccions de registre o d'accés, però amb un límit de 2 GB per fitxer i sense limitacions de mida total quant a col·leccions de dades dins les comunitats, que es guarden als centres de dades del CERN. Aquest límit de 2 GB, però, no implica que Zenodo no accepti fitxers més grans, ja que la infraestructura actual s'ha provat fins als 10 GB. Si es donés el cas de haver d'enviar fitxers més grans, l'investigador ha de contactar amb el servei, el qual a diferència de figshare, no ofereix comptes institucionals.

La preservació de les dades està garantida per a tot el temps en actiu del repositori, el qual està establert pel CERN per un mínim de 20 anys. En el cas que el servei hagués de tancar les seves operacions, es garanteix la migració de tot el contingut a repositoris adequats i la no afectació de les citacions i enllaços a recursos de Zenodo, ja que tots compten amb un DOI.

[214] <http://zenodo.org/>. [Consulta: 31/05/2016]

3.1.6 Edinburgh DataShare

A la University of Edinburgh trobem un exemple de repositori de dades institucional, l'Edinburgh DataShare[215], el qual s'utilitza com a dipòsit per a *datasets* d'origen multidisciplinari produïts a aquesta universitat. Està programat amb el paquet de programari de codi obert DSpace (Westra et al., 2010) i fou desenvolupat entre els anys 2007 i 2009 (Macdonald; Martinez-Uribe, 2010), amb 61 *datasets* emmagatzemats en data de març de 2013 (Pampel et al., 2013). A més, està considerat com a un repositori de confiança, ja que compta amb el DSA des de l'any 2015[216].

Per dipositar un *dataset*, l'investigador ha de registrar-se a Edinburgh DataShare si es tracta d'un membre de la University of Edinburgh. En cas contrari, ha de registrar-se en un compte EASE Friend. El següent pas és seleccionar una col·lecció de dades adient a la disciplina del *dataset*, afegir una descripció i enviar-lo juntament amb la documentació que faci que les dades siguin reutilitzables. S'especifica un límit de mida de fitxer de 9 GB, però és possible enviar fitxers més grans si es contacta amb l'administració del repositori. La llicència que recomana el repositori és la Creative Commons Attribution 4.0 International[217], però hi ha altres opcions que pot escollir l'investigador. Un cop l'investigador ha comprovat que totes les dades de l'enviament estiguin correctes, es fa l'enviament i ha d'esperar l'aprovació de l'administrador. Un cop s'hagi aprovat, es publicarà un DOI per a l'ítem.

La política de preservació del repositori[218] indica el següent:

- Els ítems es preservaran de forma indefinida, amb migració a nous formats de fitxers si fos necessari, però no es donen garanties per accedir a formats de fitxer poc habituals
- Es fan còpies de seguretat de forma regular

[215] <http://datashare.is.ed.ac.uk/>. [Consulta: 05/06/2016]

[216] Macdonald, Stuart (2015, Oct. 28). *Edinburgh DataShare receives 'Data Seal of Approval'*. <http://datablog.is.ed.ac.uk/2015/10/28/edinburgh-datashare-receives-data-seal-of-approval/>. [Consulta: 05/06/2016]

[217] <https://creativecommons.org/licenses/by/4.0/>. [Consulta: 05/06/2016]

[218] DataShare repository (2015, July 31). *Preservation policy.* <https://goo.gl/2jpdVj>. [Consulta: 05/06/2016]

- El *bit stream* original es conservarà per a tots els ítems, a més de fer-ho per als nous formats
- Si el propietari dels drets d'autor ho demana, es poden treure els ítems del repositori (les raons vàlides per això poden ser la violació dels drets d'autor, seguretat nacional o recerca fraudulenta). En aquests casos, els ítems retirats no s'esborraran, sinó que es retiraran de l'accés públic
- Els identificadors i/o URLs dels ítems retirats es retenen de forma indefinida, per tal d'evitar enllaços trencats. S'inclouran notes explicant les raons de la retirada i un enllaç a la versió més actual, si és aplicable
- Si es requereix, es poden incloure llistes d'errates i de correccions al registre original
- Si fos necessari, es pot dipositar una versió actualitzada. La versió anterior es retiraria de l'accés públic i s'indicaran enllaços entre les versions anteriors i posteriors, amb indicacions clares de quina és la versió més recent
- Els ítems inclouran un *checksum* per facilitar la detecció d'alteracions
- Si el repositori tanqués, la base de dades es transferirà a un altre arxiu

3.2 Estàndards de preservació digital

Existeixen dues grans categories d'estàndards de preservació digital. Dins la primera, trobem serveis d'infraestructura de sistemes que donen suport a un repositori de confiança. A la segona categoria s'inclouen formats de fitxer d'estàndard obert i que no depenen d'una tecnologia concreta (Dollar; Ashley, 2013, p. 289).

Per tal d'escollir els estàndards més adients de preservació digital, s'han seleccionat aquells que estan reconeguts com a estàndards ISO i en la mesura del possible, aquells amb casos d'estudi documentats. Encara que els estàndards ISO són coneguts per ser de compliment "voluntari" més que "obligatori", són una bona guia per a les organitzacions (Yoon, 2014) i també són acceptats per un consens internacional (Giménez Chornet, 2014).

Dins la categoria d'estàndards de suport per a repositoris de confiança, un estàndard de referència per a la preservació de dades d'arxiu a llarg termini és el model OAIS (Downs; Chen, 2010a; Shaw; Corns; McAuley, 2009), ja que defineix els atributs de sistemes i organitzacions compromeses en la preservació digital (Dearborn; Barton; Harmeyer, 2014; Steinhart; Dietrich; Green, 2009). Per altra banda, l'auditoria i certificació de repositoris és un estàndard reconegut (Johnston, 2012; Reilly Jr.; Waltz, 2014).

3.2.1 Model de referència OAIS

El model de referència OAIS (Open Archival Information System) fou creat pel Consultative Committee for Space Data Systems (CCSDS), el qual fou fundat l'any 1982 per les agències espacials més importants del món, com la NASA, per tal d'aportar un espai de discussió per a problemes comuns dins el desenvolupament de sistemes de dades espacials (Shaw et al., 2009). S'ha de posar èmfasi que es va dissenyar com a un model de referència per donar suport a estàndards formals per a la preservació a llarg termini de dades ciència espacial i actius d'informació, i no com un model d'implementació (Smallwood, 2013, p. 247); la prova d'això és que OAIS actualment és la *lingua franca* de la preservació digital, ja que la major part d'arxius i repositoris digitals seriosos l'utilitzen de referència per a la preservació i l'accés dels seus actius digitals (Lee; Tibbo, 2007). De fet, una estratègia de preservació a llarg termini que compleix amb el model OAIS ofereix els millors mitjans disponibles per preservar el patrimoni digital de qualsevol organització, ja sigui privada o pública (Dollar; Ashley, 2013, p. 291).

L'any 2002 OAIS fou aprovat com a estàndard internacional ISO 14721, amb una revisió i actualització que fou publicada l'any 2012 com a ISO estàndard 14721:2012[219]. La definició que rep el model és "Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community" (CCSDS, 2012, p. 1-1). Això vol dir que un repositori d'arxiu de tipus OAIS ha de complir dos

[219] *ISO 14721:2012 - Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model*. <http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284>. [Consulta: 18/04/2016]

funcions principals: una és preservar informació i l'altra és facilitar l'accés a la informació arxivada de manera consistent amb les necessitats dels usuaris de l'arxiu o "comunitat designada" (Lavoie, 2014). Addicionalment, les responsabilitats d'un arxiu OAIS són (Schumann; Recker, 2013):

- Negociar per i acceptar informació adequada de productors d'informació
- Obtenir control suficient de la informació amb l'objectiu de complir amb els objectius de preservació a llarg termini
- Determinar l'abast de la comunitat d'usuaris de l'arxiu
- Assegurar que la informació preservada es pot entendre de forma independent en el sentit que la informació es pot entendre per part d'usuaris sense l'assistència del productor de la informació
- Seguir polítiques i procediments documentats per assegurar que la informació es preserva contra totes les contingències raonables, i que no s'han produït supressions
- Fer que la informació preservada estigui disponible per a la comunitat d'usuaris, i habilitar la difusió de còpies autenticades de la informació preservada en la seva forma original, o en una forma que es pugui rastrejar

La Taula 12 presenta una selecció dels components més importants dins el model OAIS i el seu significat.

Taula 12. Terminologia selecta d'OAIS

Terminologia OAIS	Descripció
Entorn OAIS	
Gestors	Responsables de la formulació, revisió i si és necessari, reforçar el marc de polítiques a alt nivell que governin les activitats de l'OAIS
Productors	Individus, organitzacions o sistemes que transfereixen informació a l'OAIS per a la preservació a llarg termini
Consumidors i Comunitat Designada	Individus, organitzacions o sistemes que consumeixen, o utilitzen, la informació preservada a l'OAIS
Entitats funcionals	
Ingesta	Conjunt de processos responsables d'acceptar informació enviada pels Productors i de preparar-la per a la seva inclusió a l'arxiu
Emmagatzematge d'arxiu	Part del sistema d'arxiu que gestiona l'emmagatzematge a llarg termini i el manteniment dels materials digitals que es confien a l'OAIS
Gestió de dades	Manteniment de bases de dades de metadades descriptives que identifiquen i descriuen la informació arxivada per donar suport a les eines de cerca
Planificació de la preservació	Recomanacions i plans de preservació per tal d'assegurar que la informació romaní accessible, comprensible i usable
Accés	Gestiona els processos i serveis mitjançant els quals els Consumidors localitzen, sol·liciten i reben els objectes dins OAIS
Administració	Gestiona les operacions diàries de l'OAIS, així com coordinar les activitats de les altres cinc entitats funcionals. Estableix estàndards i polítiques de l'arxiu
Serveis comuns	Serveis de suport com serveis de sistema operatiu, serveis de xarxa o serveis de seguretat
Paquets d'informació	Encapsulament dels objectes digitals
Paquet d'Informació d'Enviament	Versió del paquet d'informació que es transfereix del Productor a l'OAIS quan la informació s'ingesta dins l'arxiu
Paquet d'Informació d'Arxiu	Versió del paquet d'informació que es guarda i es preserva per part de l'OAIS
Paquet d'Informació de Difusió	Versió del paquet d'informació que s'entrega al Consumidor en resposta a una petició d'accés

Fonts: CCSDS, 2012, p. 1-8-1-16; Lavoie, 2014; Smallwood, 2013, p. 90

Es descriuen a continuació com funcionen aquestes entitats i quines són les seves tasques. També s'expliquen amb més profunditat els Paquets d'Informació SIP, DIP i AIP (Lavoie, 2014; Smallwood, 2013, p. 90). A més, es mostren les relacions entre les sis entitats funcionals i els Paquets d'Informació dins la Figura 3.

- Ingesta. Conjunt de processos responsables d'acceptar informació enviada pels Productors i de preparar-la per a la seva inclusió a l'arxiu. Dins de les funcions

d'ingesta s'inclouen la recepció d'informació transferida a l'OAIS per un Productor, la validació que la informació que es rep no es troba corrompuda i està completa, la transformació de la informació enviada en una forma apropiada per a l'emmagatzematge i gestió dins el sistema d'arxiu, extracció i/o creació de metadades descriptives per donar suport a la cerca dins OAIS i eines de recuperació i transferència de la informació enviada i les seves metadades associades a l'arxiu. En resum, la funció d'ingesta serveix com a interfície externa de l'OAIS amb els Productors, gestionant el procés d'acceptar la custòdia de la informació enviada i preparar-la per la retenció a l'arxiu

- Emmagatzematge de l'arxiu. Es tracta de la part del sistema d'arxiu que gestiona l'emmagatzematge a llarg termini i el manteniment dels materials digitals que es confien a l'OAIS. És responsable d'assegurar que el contingut arxivat resideix en formes apropiades d'emmagatzematge i que els bits que comprenen la informació preservada resten complets i gestionables a llarg termini. Per fer això, l'Emmagatzematge de l'arxiu executa procediments com *refreshing* o migracions. També implementa mecanismes de seguretat, com procediments de comprovació d'errors, avaluar els resultats dels processos de preservació, o polítiques de prevenció de desastres per evitar resultats catastròfics de pèrdua de dades. L'Emmagatzematge també recupera objectes dels sistemes d'emmagatzematge de l'OAIS per donar suport a peticions d'accés per part dels Consumidors.
- Gestió de dades. La funció de Gestió de dades fa el manteniment de bases de dades de metadades descriptives que identifiquen i descriuen la informació arxivada per donar suport a les eines de cerca; també gestiona les dades administratives que donen suport a les operacions del sistema intern de l'OAIS, com les dades de rendiment del sistema o les estadístiques del sistema. Les funcions primàries inclouen el manteniment de les bases de dades dels quals és responsable; realitzar preguntes sobre aquestes bases de dades i generar informes en resposta a peticions d'altres entitats funcionals dins l'OAIS; conduir actualitzacions a les bases de dades sempre que arriba nova informació, o quan la informació existent es modifica o s'esborra. La funció de Gestió de dades dóna suport a la cerca i recuperació del contingut arxivat de l'OAIS, i a l'administració de les operacions internes de l'OAIS

- Planificació de la preservació. La funció Planificació de la preservació és responsable de mapar l'estratègia de preservació d'OAIS, així com recomanar revisions apropiades per a aquesta estratègia en resposta a canvis que es puguin produir dins l'entorn OAIS. Monitoritza l'entorn extern per a canvis i riscos que puguin afectar la capacitat d'OAIS per preservar i mantenir l'accés a la informació que custodia. També desenvolupa recomanacions per actualitzar les polítiques i procediments d'OAIS que acomodin aquests canvis. La funció també representa una seguretat contra l'usuari i l'entorn tecnològic en constant evolució. Detecta canvis o riscos que impactin la capacitat d'OAIS per afrontar les seves responsabilitats, dissenya estratègies per afrontar-los, i assisteix en la implementació d'aquestes estratègies dins el sistema d'arxiu
- Accés. La funció d'Accés gestiona els processos i serveis mitjançant els quals els Consumidors localitzen, sol·liciten i reben els objectes dins OAIS. També coordina la recuperació i l'entrega de contingut que s'hagi demanat i és responsable d'implementar mecanismes de control i de seguretat associats al contingut arxivat. La funció d'Accés representa la interfície d'OAIS amb els seus Consumidors (i Comunitat Designada); és el mecanisme primari mitjançant el qual l'OAIS compleix amb la seva responsabilitat de fer la informació disponible a la comunitat
- Administració. Aquesta funció és responsable de gestionar les operacions diàries de l'OAIS, així com coordinar les activitats de les altres cinc entitats funcionals. Altres responsabilitats inclouen interactuar amb els Productors (p.e. negociar *Submission Agreements* o Acords d'Enviament), Consumidors (p.e. proveir suport de servei al client), i Gestió (p.e. implementar i mantenir polítiques i estàndards d'arxiu). La funció d'Administració també és responsable de supervisar l'operació d'arxiu i de sistemes d'accés, monitoritzar el rendiment del sistema, i coordinar actualitzacions al sistema si escauen. L'Administració funciona com un centre d'activitat central per a les interaccions internes i externes; comunica directament amb els altres cinc serveis d'alt nivell, així com els agents externs d'OAIS: Productors, Consumidors i Gestió
- Serveis comuns. Hi ha una setena entitat funcional: Serveis comuns. Aquests serveis són freqüents dins l'arxiu, i inclouen (entre d'altres), serveis de sistema

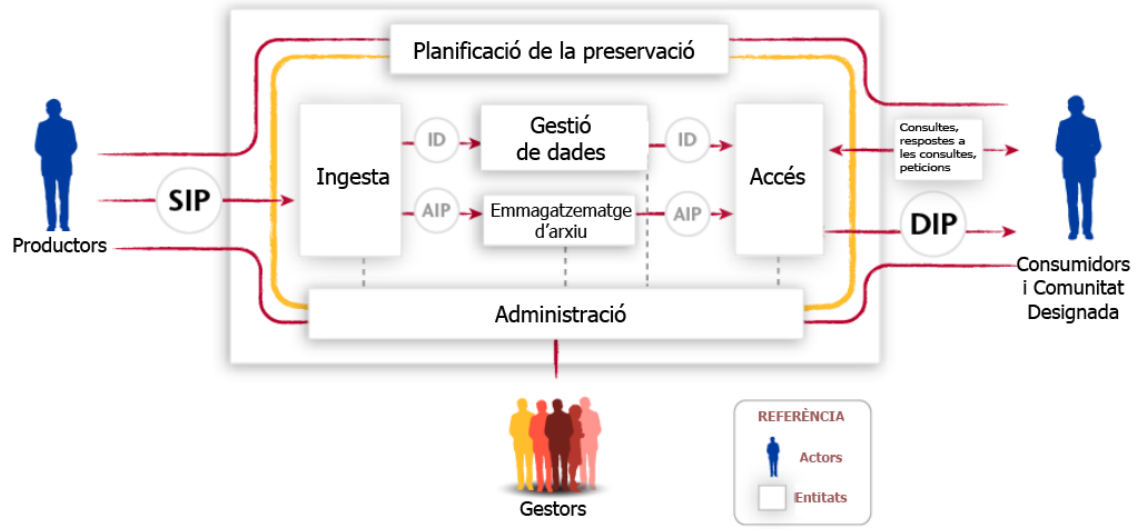
operatiu, serveis de xarxa i serveis de seguretat. Els Serveis Comuns són la columna vertebral de qualsevol arxiu de tipus OAIS

- Paquets d'Informació. Dins cada Paquet s'inclouen el contingut de l'objecte digital (una seqüència de bits) i informació de representació que habilita la visualització d'un objecte en informació usable juntament amb Informació de Descripció de Preservació (*Preservation Description Information*) com la procedència i el context
    - Paquet d'Informació d'Enviament (*Submission Information Package*). SIP és la versió del paquet d'informació que es transfereix del Productor a l'OAIS quan la informació s'ingesta dins l'arxiu. La forma exacta del SIP pot ser el resultat d'un acord negociat entre el Productor i l'OAIS, o es pot construir a mida. El concepte SIP dóna èmfasi al fet que la informació podria no preservar-se en la forma exacta en la qual és enviada per part del Productor. Per exemple, l'objecte preservat podria ser una incorporació de contingut proveït en múltiples SIPs; o bé el Productor podria proveir la informació en un format no suportat per l'OAIS, i per tant necessitaria migració a un altre format abans d'incloure'l a l'arxiu. També podria donar-se el cas que les metadades que el Productor subministra estan incompletes o són inadequades, i s'han d'augmentar durant el procés d'ingesta
    - Paquet d'Informació d'Arxiu (*Archival Information Package*). AIP és la versió del paquet d'informació que es guarda i es preserva per part de l'OAIS. L'AIP consisteix en la informació que és el focus de preservació, acompanyat per un conjunt complet de metadades suficients per donar suport als serveis d'accés i de preservació. La informació arxivada i les seves metadades associades representen un paquet lògic dins el sistema d'arxiu: no obstant, no hi ha requeriments que s'hagi de mantenir qualsevol forma d'associació física, com la incrustació de metadades dins el mateix objecte d'informació i guardar l'objecte combinat com un sol flux de bits. El model de referència defineix dues especialitzacions de l'AIP: Unitat d'Informació d'Arxiu (*Archival Information Unit*) i Col·lecció d'Informació d'Arxiu (*Archival Information Collection*). Un AIU guarda el contingut i les metadades per un sol objecte 'atomístic'

(com una pel·lícula o un llibre), mentre que un AIC consisteix en múltiples AIUs que s'han agrupat en una col·lecció

- Paquet d'Informació de Difusió (*Dissemination Information Package*). DIP és la versió del paquet d'informació que s'entrega al Consumidor en resposta a una petició d'accés. El concepte de DIP dóna èmfasi al fet que el paquet d'informació que es distribueix de l'OAIS al Consumidor pot diferir en forma o contingut al que resideix en l'arxiu. Les diferències entre DIP i AIP poden incloure el format del contingut (p. ex., un fitxer d'imatge es podria convertir de TIFF a JPEG abans de la seva distribució); la quantitat de contingut (un DIP pot correspondre a un AIP, múltiples AIPs o una part d'un AIP), i la quantitat de metadades que subministren juntament amb el contingut (és probable que el DIP no contingui el conjunt complet de metadades associades amb un objecte digital arxivat, ja que gran part és de poc interès per al Consumidor). L'AIP és el focus de preservació: és la variant de paquet d'informació el qual l'OAIS es compromet a perpetuar a llarg termini

Figura 3. Entitats funcionals a OAIS



Font: CCSDS, 2012, p. 4-1. Traducció de l'autor

Per tal d'entendre millor el funcionament del model OAIS, veurem a continuació alguns exemples d'aplicacions reals.

Arxiu de dades GESIS

Aquest arxiu fou fundat l'any 1960 a la Universitat de Colònia com a arxiu central de recerca social empírica (Zentralarchiv für empirische Sozialforschung), el primer arxiu de dades en ciències socials a Europa. L'any 1986 va entra a formar part del GESIS (Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen) – Institut Leibniz per a les Ciències Socials [220], la institució més gran de recerca en ciències socials d'Alemanya. Actualment, la seva col·lecció compta amb més de 500.000 fitxers provinents d'un nombre superior a 5.100 estudis. El 8 de maig de 2014 va rebre el DSA, el qual garanteix que és un servei de confiança i ja ha començat a fer passos per executar un mapatge al model de referència OAIS (Schumann; Recker, 2013), amb els objectius següents:

- Aconseguir una visió de conjunt més estructurada dels fluxos de treball i dels processos de preservació i de difusió
- Identificar i tancar buits en els processos i fluxos de treball
- Introduir la terminologia i els conceptes OAIS per donar suport a la comunicació dins l'arxiu i amb altres organitzacions

Es va començar amb l'elaboració d'una figura dels fluxos de treball existents a l'arxiu per poder analitzar la seva adequació a OAIS. La Figura 4 conté les funcions principals que s'executen quan les dades s'adquireixen, es dipositen, s'arxiven i es difonen.

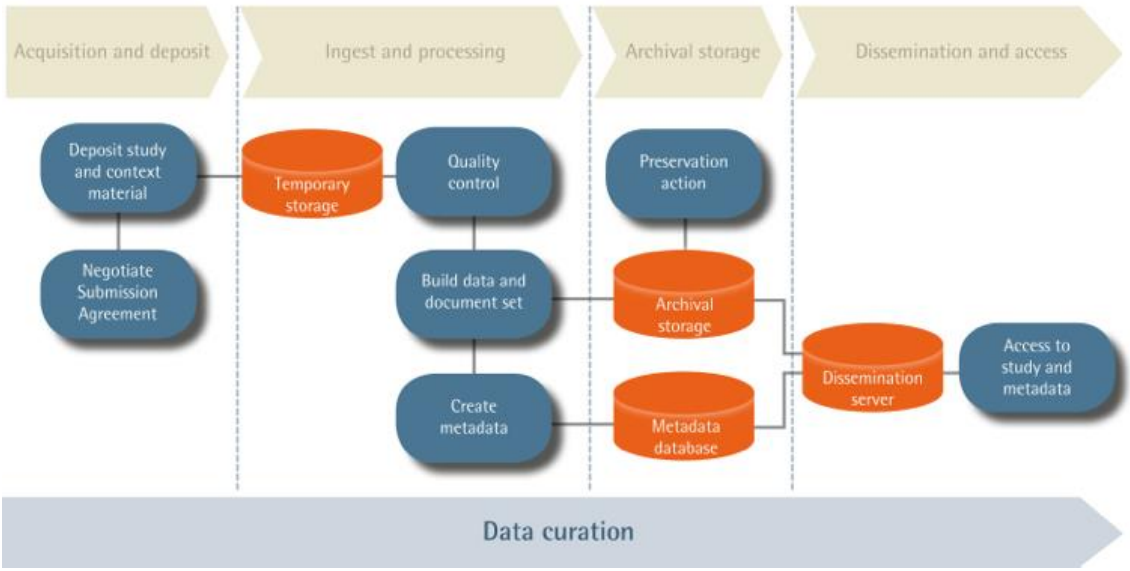
Un dels primers descobriments va ser que el procés d'ingesta d'OAIS no es trobava mapada correctament dins els fluxos de treball, ja que diverses funcions dels processos de preingesta i d'ingesta es trobaven dins altres entitats funcionals d'OAIS, com el d'Administració, el qual és el responsable de negociar l'Acord d'Enviament amb el Productor. Això va fer que es modifiquessin algunes funcions del procés d'ingesta per acomodar-les al model OAIS.

Els primers beneficis de l'aplicació del model foren la comunicació, l'auto-reflexió i l'avaluació de processos. En el primer cas, el vocabulari de termes OAIS ajuda al personal de l'arxiu a evitar ambigüitats dins les seves tasques. En el segon cas, el model

[220] <http://www.gesis.org/home/>. [Consulta: 23/04/2016]

ajuda a l'anàlisi de les rutines i els procediments habituals, ja que permet una millor entesa dels fluxos de treball, els processos de comunicació de suport i els papers i les responsabilitats dels agents implicats. En últim lloc, el mapat va ajudar a avaluar els processos per així detectar problemes i posar-hi solució.

Figura 4. Fluxos de treball a l'arxiu de dades GESIS



Font: Schumann; Recker, 2013

SHARE IT

El projecte Spatial Heritage & Archaeological Research Environment IT[221] va ser creat per l'Heritage Council (Irlanda) l'any 2008 amb l'objectiu d'investigar dades arqueològiques a Irlanda i desenvolupar una eina WebGIS per a l'exploració en recerques futures. Hi ha un cas de preservació d'un estudi de magnetometria que ens permet entendre com gestionen els paquets d'informació (Shaw et al., 2009). En primer lloc, el productor presenta el SIP, el qual inclou:

- Els fitxers de dades en brut en format GeoPlot
- L'informe detallat de l'estudi en format PDF, creat com a requeriment de llicència
- Fitxers addicionals que inclouen informació de georeferències per a les dades de l'estudi i imatges JPEG

[221] <http://www.share-it.ie/index.html/>. [Consulta: 23/04/2016]

Aquestes dades necessitaven tenir una estructura, i per aquesta raó es va generar un AIP, en el qual les dades es van migrar en formats apropiats per a l'arxiu a llarg termini. Un component important és la generació de metadades estructurades en un fitxer XML que es trobi d'acord amb un esquema designat. Per tal de crear aquest fitxer, es va extreure la informació de l'informe en PDF. Després de fer aquests passos, ja es pot generar un DIP a petició dels interessats en el format requerit.

3.2.2 TDR: auditoria i certificació de repositoris

L'any 2002, un informe del RLG i l'OCLC (2002) aportà els primers passos per descriure un marc d'atributs i responsabilitats per als repositoris digitals de confiança o TDR (*Trusted Digital Repository*), on es definí el concepte de TDR com un repositori que té polítiques, estàndards i infraestructura tecnològica que proporcionen el marc per a la preservació digital i amb un sistema de confiança, com un sistema de programari i de maquinari que ha de seguir certes regles. L'any 2003 es va constituir un equip de treball entre el RLG i el NARA per assolir una certificació específica per a repositoris digitals (Yoon, 2014), el qual va esbrinar que les institucions sovint declaraven que complien amb els requisits del model OAIS per recalcar el nivell de confiança dels seus repositoris, encara que no existia en aquell moment una definició establerta de què volia dir complir amb el model OAIS (RLG; NARA, 2005, p. 1). Així doncs, l'equip de treball va començar a construir criteris per mesurar el nivell de confiança d'un repositori amb mètriques que es van posar a prova l'any 2005 mitjançant el Certification of Digital Archives Project[222] del CRL, un projecte que va realitzar auditories a tres arxius digitals: la Koninklijke Bibliotheek (Biblioteca nacional dels Països Baixos), Portico [223] i l'ICPSR, i també a un sistema d'arxiu, LOCKSS. Per altra banda, la xarxa nestor també va fer esforços en programes d'auditoria i certificació, els quals van culminar amb la publicació del *Catalogue of criteria for Trusted Digital Repositories* l'any 2006 (nestor, 2006).

La combinació d'aquests treballs per realitzar programes d'auditoria i certificació van causar la publicació de *Trustworthy Repositories Audit and Certification: criteria and*

[222] <http://www.crl.edu/reports/certification-digital-archives-project>. [Consulta: 25/04/2016]

[223] <http://www.portico.org/digital-preservation/>. [Consulta: 26/04/2016]

checklist (TRAC) l'any 2007 (CRL; OCLC, 2007), que fou la base del document de recomanacions *Audit and Certification of Trustworthy Digital Repositories: recommended practice CCSDS 652.0-M-1* (CCSDS, 2011), i que es va formalitzar més tard com a estàndard ISO 16363:2012[224]. El document de recomanacions presenta una llista de criteris que han de complir els repositoris per poder ser TDR. El total de criteris són 105, que cobreixen tres àrees (Downs; Chen, 2010b; Houghton, 2015; Johnston, 2012):

- Infraestructura de l'organització. Inclou governabilitat, estructura i viabilitat de l'organització, personal, comptabilitat, sostenibilitat financera i problemes legals
- Gestió d'objectes digitals. Inclou adquisició i ingesta de contingut, planificació i procediments de preservació, gestió i accés a la informació
- Gestió d'infraestructura i de riscos de seguretat. Inclou infraestructura tècnica i problemes legals

Veurem a continuació alguns exemples d'auditories i/o avaluacions tot utilitzant l'estàndard ISO 16363. Com que la seva publicació encara és prou recent, no hi ha molts casos d'estudi.

Deakin Research Online

DRO [225] és el repositori de recerca de la Deakin University, universitat pública australiana ubicada a la ciutat de Victoria. El repositori es constituí l'any 2007 amb l'objectiu de facilitar el dipòsit de publicacions de recerca i fa preservació dels resultats produïts pels científics de la universitat. Es basa en programari Fez/Fedora i compta amb suport per a JHOVE i PREMIS.

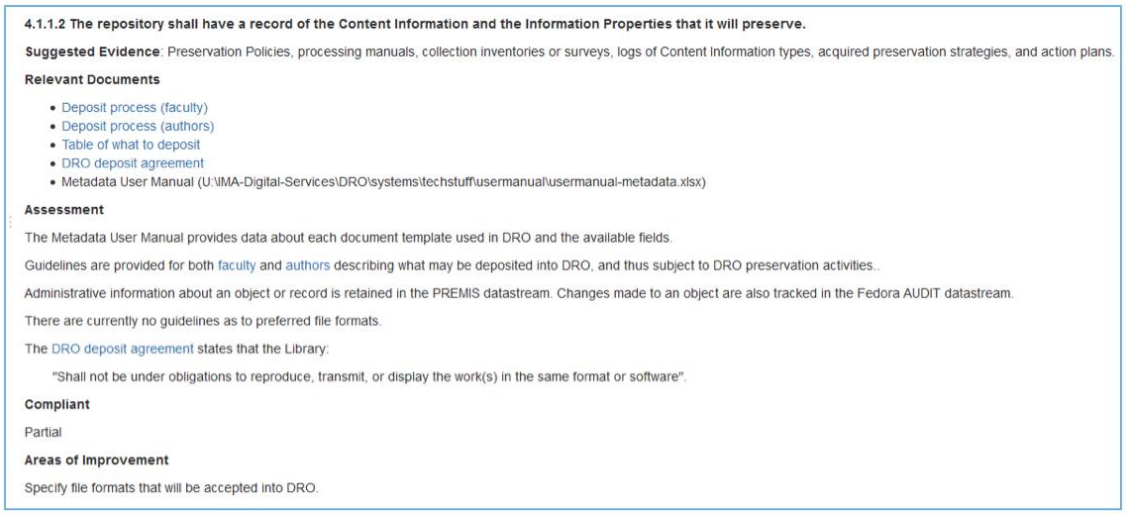
En aquest cas, es va fer una autovaloració (Houghton, 2015) durant l'any 2013, on el primer pas fou la creació d'una wiki per documentar els processos. Encara que el web del CRL inclou una plantilla de full de càlcul per realitzar aquesta tasca, alguns dels criteris requerien respostes massa extenses que no encaixaven bé en un full de càlcul.

[224] *ISO 16363:2012 - Space data and information transfer systems -- Audit and certification of trustworthy digital repositories*. <http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510>. [Consulta: 25/04/2016]

[225] <http://dro.deakin.edu.au/>. [Consulta: 27/04/2016]

Un cop es completà la wiki, es va fer una recerca de documentació que pogués donar evidències del nivell de compliment dels criteris. En la mesura del possible, aquests documents es van enllaçar a la wiki. Aquesta col·lecció preliminar de documentació es va contrastar amb cada criteri de compliment, amb l'addició de nova documentació si era necessària i també es van fer reunions amb el personal per tal de clarificar els procediments i els fluxos de treball. Llavors es va fer la valoració de compliments de criteris, amb una identificació d'àrees de millora. Els resultats es van documentar a la wiki amb una anotació del nivell de compliment de cada criteri en tres categories: "compliment total", "compliment parcial" i "no hi ha compliment". Del total de 105 criteris, DRO compleix totalment amb 67, parcialment amb 32, no compleix amb 15 i un criteri no és aplicable. La Figura 5 mostra un exemple de redacció final d'una anàlisi de compliment d'un criteri, amb l'anotació del mateix, les fonts consultades, la valoració, el nivell de compliment i les àrees de millora.

Figura 5. Exemple de documentació d'autovaloració a DRO



4.1.1.2 The repository shall have a record of the Content Information and the Information Properties that it will preserve.

Suggested Evidence: Preservation Policies, processing manuals, collection inventories or surveys, logs of Content Information types, acquired preservation strategies, and action plans.

Relevant Documents

- Deposit process (faculty)
- Deposit process (authors)
- Table of what to deposit
- DRO deposit agreement
- Metadata User Manual (U:\IMA-Digital-Services\DRO\systems\techstuff\usermanual\usermanual-metadata.xlsx)

Assessment

The Metadata User Manual provides data about each document template used in DRO and the available fields.

Guidelines are provided for both faculty and authors describing what may be deposited into DRO, and thus subject to DRO preservation activities..

Administrative information about an object or record is retained in the PREMIS datastream. Changes made to an object are also tracked in the Fedora AUDIT datastream.

There are currently no guidelines as to preferred file formats.

The DRO deposit agreement states that the Library:

"Shall not be under obligations to reproduce, transmit, or display the work(s) in the same format or software".

Compliant

Partial

Areas of Improvement

Specify file formats that will be accepted into DRO.

Font: Houghton, 2015

Una de les conclusions més interessants de la valoració és que DRO fa tasques de preservació fetes a mida, però ja estan treballant per millorar aquesta àrea, amb la incorporació de millores trobades a l'autovaloració al pla estratègic de preservació de Deakin. Per altra banda, es fa una sèrie de recomanacions als gestors de repositoris que vulguin fer una valoració amb la norma ISO 16363:

- És molt més econòmic fer una autovaloració que una auditoria externa

- El suport de l'alta direcció és necessari, ja que la preservació digital és una qüestió que s'ha de tractar a llarg termini
- L'encarregat de fer la valoració ha d'estar familiaritzat amb les polítiques i procediments de l'organització i del repositori
- En el cas de no tenir temps ni recursos per fer una valoració ISO 16363, una alternativa és la valoració segons NDSA Levels of Digital Preservation[226]
- S'ha de configurar una wiki de forma específica per fer l'autovaloració
- S'han de tenir en compte el temps i recursos disponibles per fer l'autovaloració
- Decidir el nivell de detall que tindrà l'autovaloració
- Utilitzar el coneixement de l'organització quan es reuneix documentació
- Conèixer els criteris abans de començar l'autovaloració per així estalviar temps quan es reuneix documentació que pot servir per a diferents criteris
- Pot ser que alguns criteris no siguin aplicables
- Després d'acabar la valoració, s'hauran de millorar els problemes detectats. No serveix de res fer una valoració si no es fan passos addicionals de millores
- Programar les autovaloracions de forma periòdica
- No assumir que si el programari del repositori compleix amb OAIS, també ho fa el propi repositori

SEDAC

El Socioeconomic Data and Applications Center (SEDAC)[227], gestionat pel Center for International Earth Science Information Network (CIESIN) [228] de la Columbia University, compta amb un arxiu iniciat l'any 2004 per preservar dades científiques i informació relacionada amb la recerca difosa pel SEDAC per al seu ús i accés en el futur. Aquest arxiu es troba gestionat conjuntament pel SEDAC i les Columbia University Libraries[229]. Diverses comunitats d'usuaris tenen el SEDAC com un referent

[226] Owens, Trevor. *NDSA Levels of Digital Preservation: Release Candidate One.* <http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/>. [Consulta: 27/04/2016]

[227] <http://sedac.ciesin.columbia.edu/>. [Consulta: 01/05/2016]

[228] <http://www.ciesin.org/>. [Consulta: 01/05/2016]

[229] Downs, Robert R.; Chen, Robert S. (2008, Oct. 5-8). "Creating a Trustworthy Digital Repository for a long-term archive of interdisciplinary data: a case study" [presentació de Power Point]. *CODATA*. Kyiv, Ukraine. <http://ciesin.columbia.edu/documents/CreatingTrustworthyDgtlReposityPrsntn.pdf>. [Consulta: 01/05/2016].

per l'accés a dades científiques i serveis que donin suport a la recerca i l'educació (Downs; Chen, 2004).

Per tal de millorar les capacitats i els serveis que ofereixen a les comunitats interessades en els seus serveis, SEDAC demanà una auditoria independent per avaluar el nivell de compliment del seu arxiu amb la norma ISO 16363. Encara que ja van realitzar diverses autovaloracions, com la que es va fer segons els criteris del TRAC (Downs; Chen, 2010a, 2010b), un test independent oferia l'oportunitat d'identificar formes de millorar les seves estratègies, polítiques i pràctiques o aconseguir garanties externes per a comunitats de proveïdors de dades, col·laboradors, usuaris i patrocinadors, a més del reconeixement intrínsec del compliment oficial amb un estàndard internacional[230].

El SEDAC, per tant, va haver de preparar la visita dels auditors, la qual es realitzà l'any 2012, amb una identificació de l'abast i de les limitacions de l'auditoria, ja que per raons de seguretat, l'accés a diversos documents financers i confidencials va estar restringit. Es va contestar un qüestionari amb una descripció de com l'arxiu complia els requeriments de l'estàndard ISO, la qual cosa ajudà a millorar els seus processos de gestió. Finalment, set auditors externs, membres del Primary Trustworthy Digital Repository Authorization Board (PTAB), van fer una visita de dos dies, durant la qual es van realitzar introduccions, una descripció de l'auditoria, inspeccions de documents i de les instal·lacions, observacions de realització de tasques, entrevistes a membres del personal, verificació de registres d'activitats per comprovar el compliment de polítiques i procediments i reunions informatives[231]. Els resultats de l'auditoria es presenten a la Taula 13, on s'especifiquen les àrees que necessiten millores.

[230] Downs, Robert R.; Chen, Robert S. (2011, Dec. 9). "Audit of a scientific data center for certification as a Trustworthy Digital Repository: a case study" [presentació de Power Point]. *AGU Meeting*. San Francisco, CA. Recuperat del web Internet Archive. <https://goo.gl/0XwnK4>. [Consulta: 26/04/2016].

[231] Downs, Robert R.; Chen, Robert S. (2012, June 6). "Improving the trustworthiness of an interdisciplinary scientific data archive" [presentació de Power Point]. *IASSIST*. Washington, DC. <http://www.iassistdata.org/conferences/2012/presentation/3329>. [Consulta: 01/05/2016].

Taula 13. Resultats de l'auditoria ISO 16363 a l'arxiu SEDAC

Mètriques que necessiten millores	Àrees pendents de rebre millores
3.1 Governabilitat i viabilitat de l'organització	Declaració de la missió i de les polítiques: s'ha de donar èmfasi al compromís per continuar la custòdia i la preservació de les dades científiques i dels serveis Plans per transferir dades, operacions, responsabilitats, i autoritat a una altra entitat en el cas que es produeixi un esdeveniment imprevist Plans de preservació per incloure detalls de nous procediments en el moment en què s'adoptin
3.2 Estructura i personal de l'organització	Formació en la custòdia de dades s'ha de completar per part del nou personal i de forma periòdica pel personal experimental, el qual inclou estàndards i termes OAIS
3.3 Responsabilitat de procediments i marc de polítiques de preservació	Processos per definir la comunitat designada per a cada AIP durant el desenvolupament de les dades i la planificació de de difusió de dades
4.1 Ingesta: adquisició de contingut	Procediments per enregistrar les activitats d'inventari, verificació i manteniment realitzades sobre objectes i col·leccions
4.2 Ingesta: Creació de l'AIP	Procediments per testar i millorar l'entesa de cada AIP per a la comunitat designada Procediments per enregistrar la procedència d'activitats completades durant el desenvolupament i difusió de les dades
4.3 Planificació de la preservació	Procediment per identificar, enregistrar, i mantenir informació sobre dependències de programari per a cada fitxer que es rep
4.4 Preservació de l'AIP	Procediments per verificar la integritat d'objectes i fitxers digitals
5.1 Gestió de riscos de la infraestructura tècnica	Plans de gestió de riscos per incloure un registre de riscos de l'organització que contingui una agenda de reducció de riscos Procediments per separar còpies de circulació d'AIPs de còpies d'arxiu

Font: Downs, Robert R.; Chen, Robert S. (2013, Sept. 16). *Independent evaluation of a scientific data center for compliance with the ISO 16363 requirements for Audit and Certification of Trustworthy Digital Repositories*. Pòster presentat al RDA Second Plenary Meeting. Washington, DC: RDA. <http://hdl.handle.net/10022/AC:P:21793>. [Consulta: 26/04/2016]

Totes aquestes recomanacions han conduit a la creació d'un pla de millora, l'adopció de formats especialitzats com l'especificació BagIt (vegeu el capítol 3.3.4), el programari DROID[232] de reconeixement de formats de fitxer, modificacions als fluxos de treball i una valoració de la comunitat d'usuaris.

[232] National Archives. *Download DROID: file format identification tool*. <https://goo.gl/t1S17O>. [Consulta: 20/11/2016]

3.3 Metadades

L'ús de metadades és cabdal per a la construcció del model de preservació. No només és important fer servir les metadades apropiades; també s'han de decidir els estàndards que permetin la recuperació ràpida dels registres de *datasets* per part de l'usuari, aquells adients per gestionar la preservació, aquells que facin una descripció apropiada dels conjunts de dades i els que donin informació vers els drets d'autor dels registres. Les metadades es poden classificar en quatre grups principals, segons Dappert i Enders (2010):

- Metadades descriptives, que descriuen el contingut intel·lectual del recurs i s'utilitza per a la indexació, recuperació i identificació mitjançant propietats com autor o títol
- Metadades estructurals, que capturen les relacions estructurals físiques, com quina imatge s'ha incrustat dins un web, així com relacions estructurals lògiques, com quina pàgina va a continuació dins un llibre digitalitzat. En resum, es tracta de dades sobre els contenidors de dades
- Metadades tècniques, que descriuen les característiques d'un fitxer digital, com la resolució, la mida en píxels o la mida en bytes
- Metadades administratives, les quals inclouen la informació de procedència de l'objecte digital i quines accions de preservació s'han executat sobre el mateix, així com drets d'autor i informació de permisos que especifiqui quin és el tipus d'accés

Altres autors, com Smallwood (2013, p. 273), separen les metadades administratives de les de preservació. Dins aquesta tesi, considerarem metadades de preservació com a part de les metadades administratives per tal de no separar la història de l'objecte digital (qui ha donat el suport original o qui l'ha custodiat) i els drets d'autor associats dels actes de preservació (creació d'imatges de disc o escanejat de virus).

Per tal de decidir els estàndards més apropiats, és important concretar diferents punts com: Quins usuaris utilitzaran les metadades? Quins camps de metadades són els importants per a ells? S'haurien d'utilitzar estàndards especialitzats? Aquesta elecció seria inapropiada, atès que existeixen uns 250 estàndards de metadades per diferents

tipus de dades (Dallmeier-Tiessen et al., 2014), com Darwin Core, Ecological Markup Language o Discovery Interchange Format (Wright et al., 2013). Per tant, el més pràctic i convenient serà definir els requeriments que hauran de complir els estàndards per al nostre model orientat a dades de ciències socials i humanitats, que serien:

- Descriure de forma adequada els conjunts de dades. És necessari permetre la recuperació efectiva per paraules clau, matèria, autor, etc.
- Recuperar dades tècniques bàsiques com la mida del fitxer o el tipus de fitxer
- Gestionar la preservació. És necessari documentar tots els actes de preservació que s'hagin produït, com creació d'imatges forenses, generació d'informes, anàlisi del sistema de fitxers, verificació, etc.
- Acreditar els drets d'autor. Això inclou l'investigador o investigadors del projecte de recerca, però també als directors/tutors de projecte, finançadors i altres participants

Degut a les necessitats del nostre model, pensat per dades científiques de les disciplines de ciències socials i d'humanitats, s'ha decidit que els estàndards a utilitzar seran de metadades no estructurades, ja que es farà la preservació de contenidors de dades, on el contingut dels quals no es farà una estructuració més enllà de la seva descripció, informació tècnica i de preservació. Per tant, no s'ha contemplat l'estudi de metadades estructurades, atès que no es necessitaran etiquetes d'estructuració com capítols, parts, número de pàgina o nombre d'imatges. Alguns autors com Bengtson (2012) consideren les metadades estructurals com informació intrínseca preservada a un fitxer, però nosaltres considerarem això com metadades tècniques.

Un estudi recent sobre l'ús de metadades en repositoris destacats en ciències socials i humanitats (Gómez; Méndez; Hernández-Pérez, 2016) demostra la gran acceptació de DC com a esquema de metadades descriptiu. Dels sis repositoris analitzats, tots sis fan ús de DC (entre d'altres esquemes). Les raons són el seu gran nivell de normalització en el cas de Dublin Core Simple i la seva capacitat d'interoperabilitat OAI-PMH entre repositoris. En quan a metadades tècniques, és cabdal utilitzar un esquema que ens permeti la recuperació d'informació tècnica relacionada amb els processos d'anàlisi forense, com la informació tècnica del suport original, generació de valors *hash*,

verificació del procés de captura, etc. Amb les metadades administratives, hem d'utilitzar un que guardi informació relacionada amb la preservació, com el dia i la hora en què s'ha fet la captura de la imatge forense.

Tenint en compte això, s'han analitzat quatre estàndards, que són els que utilitzarà el nostre model: Dublin Core, per a metadades descriptives; DFXML, per a metadades tècniques; i PREMIS i BagIt, per a metadades administratives.

3.3.1 Dublin Core

L'esquema de metadades Dublin Core fou creat l'any 1995 a partir d'un taller organitzat per l'OCLC l'any 1995. "Dublin" fa referència al lloc on es va aquesta reunió, la ciutat de Dublín a Ohio (EUA), i "Core" fa referència al fet que DC és un conjunt d'elements bàsics, però amb possibilitats d'expansió (Kurtz, 2010). Aquests elements bàsics van ser definits pel DCMI[233], han rebut una àmplia difusió com a part de l'OAI-PMH i han estat ratificats com a estàndards ISO 15836:2009[234] i ANSI/NISO Z39.85-2012[235], a més de l'estàndard d'Internet IETF RFC 5013[236]. El total d'elements són quinze, tots opcionals i repetibles, els quals van rebre amb posterioritat el nom de Dublin Core Simple.

L'any 2000 el DCMI va aprovar una llista de qualificadors recomanats que van rebre el nom de Dublin Core Qualificat. Aquesta nova llista incorpora dos tipus de qualificadors: refinaments d'elements i esquemes de codificació (Sugimoto; Baker; Weibel, 2002). El primer redueix el significat d'un element associat, mentre que el segon especifica un nom de un vocabulari o un nom d'esquema de codificació de dades per codificar un valor d'un element. Per exemple, per indicar el format d'un suport (imatge, vídeo, so, etc.) es pot utilitzar un refinament amb el nom "Medium" i utilitzar l'esquema de codificació IMT, el qual és una llista normalitzada de descriptors.

---

[233] <http://dublincore.org/>. [Consulta: 15/08/2016]

[234] *ISO 15836:2009 - Information and documentation -- The Dublin Core metadata element set*. <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52142>. [Consulta: 15/08/2016]

[235] *ANSI/NISO Z39.85-2012 - The Dublin Core metadata element set.* <http://www.niso.org/apps/group_public/project/details.php?project_id=105>. [Consulta: 15/08/2016]

[236] *The Dublin Core metadata element set*. <http://www.ietf.org/rfc/rfc5013.txt>. [Consulta: 15/08/2016]

Taula 14. Qualificadors recomanats pel DCMI

Element	Refinament d'element	Esquema de codificació
title	alternative	
subject		LCSH, MeSH, DDC, LCC, UDC
description	Table of Contents, Abstract	
date	Created, Valid, Available, Issued, Modified	DCMI Period, W3C-DTF
type		DCMIType
format	Extent	
	Medium	IMT
identifier		URI
source		URI
language		ISO 639-2, RFC 1766, RFC 3066
relationship	Is Version Of, Has Version, Is Replaced By, Replaces, Is Required By, Requires, Is Part Of, Has Part, Is Referenced By, References, Is Format Of, Has Format, Conforms To	URI
coverage	Spatial	DCMI Point, ISO 3166, DCMI Box, TGN
	Temporal	DCMI Period, W3C-DTF

Font: Sugimoto; Baker; Weibel, 2002

L'any 2012, DCMI va unificar tots els seus termes en un vocabulari, els DCMI Metadata Terms, que inclouen els quinze termes del Dublin Core Simple i els qualificadors. A la Taula 15 es troben tots els termes, amb els quinze elements bàsics en cursiva.

Taula 15. DCMI Metadata Terms

Terme	Descripció	Comentari
abstract	Resum del recurs	
accessioned	Data en què el recurs es adquirit	
accessRights	Informació sobre qui pot accedir al recurs o indicació del seu nivell de seguretat	
accrualMethod	Mètode pel qual els ítems s'afegeixen a la col·lecció	
accrualPeriodicity	Freqüència amb la qual els ítems s'afegeixen a la col·lecció	
accrualPolicy	Política que regeix l'addició d'ítems a la col·lecció	
alternative	Nom alternatiu del recurs	
audience	Entitat a la qual es destina el recurs	

Terme	Descripció	Comentari
available	Data o rang de dates en què el recurs es va fer disponible o es farà disponible	
bibliographicCitation	Referència bibliogràfica per al recurs	
conformsTo	Estàndard establert al qual s'ajusta el recurs	
contributor	Entitat responsable de fer contribucions al contingut del recurs	
coverage	Extensió o abast del contingut del recurs	
created	Data de creació del recurs	
creator	Entitat responsable principal de crear el contingut del recurs	
date	Data d'un esdeveniment en el cicle vital del recurs	Es recomana l'ús d'un esquema de codificació com el W3C-DTF
dateAccepted	Data d'acceptació del recurs	Un exemple pot ser la data d'acceptació d'un article per part d'una publicació en sèrie
dateCopyrighted	Data de *copyright*	
dateSubmitted	Data de presentació del recurs	
description	Descripció del contingut del recurs	
educationLevel	Nivell educatiu de l'audiència	
extent	Mida o duració del recurs	
format	Format del fitxer, suport físic, o dimensions del recurs	
hasFormat	Recurs relacionat que en essència és el mateix que el recurs descrit preexistent, però en altre format	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
hasPart	Recurs relacionat que s'inclou o bé de forma física o bé de forma lògica en el recurs descrit	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
hasVersion	Recurs relacionat que és una versió, edició, o adaptació del recurs descrit	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
identifier	Referència unívoca del recurs dins un context	
instructionalMethod	Un procés, utilitzat per generar coneixement, aptituds i habilitats, que el recurs descrit ha dissenyat per donar suport	S'inclouen formes de presentar materials d'instrucció o conduir activitats d'instrucció
isFormatOf	Recurs relacionat que en essència és el mateix que el recurs descrit, però en altre format	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
isPartOf	Recurs relacionat en què el recurs descrit s'inclou de forma física o lògica	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI

Terme	Descripció	Comentari
isReferencedby	Recurs relacionat que fa referència, cita o apunta al recurs descrit	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
isReplacedBy	Recurs relacionat que suplanta o substitueix el recurs descrit	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
isRequiredBy	Recurs relacionat que requereix que el recurs descrit doni suport a la seva funció, distribució o coherència	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
issued	Data de publicació formal del recurs	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
isVersionOf	Recurs relacionat del qual el recurs descrit és una versió, edició o adaptació	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
language	Idioma del contingut intel·lectual del recurs	
license	Document legal que dona permís oficial per fer quelcom amb el recurs	
mediator	Entitat que fa d'intermediari per donar accés al recurs al destinatari	En un context educatiu, un mediador pot ser un professor o un dels pares
medium	Suport material o físic del recurs	
mimetype	Identificador MIME de tipus de fitxer	
modified	Data en què s'han fet canvis al recurs	
provenance	Declaració de canvis en la custòdia i propietat del recurs des de la seva creació que siguin significants per a la seva autenticitat, integritat i interpretació	La declaració pot incloure una descripció de canvis que s'hagin produït amb els dipositaris
publisher	Entitat responsable de fer que el recurs estigui disponible	
references	Recurs relacionat que és referenciat, citat o apuntat pel recurs descrit	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
relation	Recurs relacionat	Es recomana identificar el recurs relacionat mitjançant una cadena conforme un sistema d'identificació formal
replaces	Recurs relacionat que és suplantat o substituït pel recurs descrit	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI
requires	Recurs relacionat que es requerit pel recurs descrit per donar suport a la seva funció, distribució o coherència	El terme s'ha d'utilitzar amb valors no literals tal com es defineixen al Model Abstracte DCMI

Terme	Descripció	Comentari
rights	Informació sobre drets executats sobre el recurs, com drets de propietat intel·lectual	
rightsHolder	Persona o organització propietari o gestor dels drets del recurs	
source	Recurs relacionat del qual es deriva el recurs descrit	El recurs descrit es pot derivar en tot o en part del recurs relacionat. Es recomana identificar el recurs relacionat mitjançant una cadena conforme un sistema d'identificació formal
spatial	Característiques d'espai del recurs	
sponsorship	Agències de finançament que donen suport al recurs	
subject	Tema del recurs	
tableOfContents	Llista de subunitats del recurs	
temporal	Característiques temporals del recurs	
title	Nom donat al recurs	
type	Natura o gènere del recurs	
valid	Data o rang de dates de validesa del recurs	

Font: DCMI Usage Board (2012). <http://dublincore.org/documents/dcmi-terms>. [Consulta: 16/08/2016]; Kurtz, 2010.

El Model Abstracte DCMI proporciona un modelo de informació independent de qualsevol sintaxi particular de codificació[237], que es va crear degut a la necessitat de fer interoperables la gran varietat d'aplicacions que es van fer amb DC a principis dels 2000. La primera recomanació del model es formulà l'any 2005, amb una revisió feta l'any 2007[238], que van generar quatre especificacions: *Expressing Dublin Core metadata using the Resource Description Framework (RDF)*[239], *Expressing Dublin Core metadata using HTML/XHTML meta and link elements*[240], *Expressing Dublin Core Description Sets using XML (DC-DS-XML)*[241] i *Expressing Dublin Core metadata using the DC-Text format*[242].

[237] Powell et al. (2007). *DCMI Abstract Model*. <http://dublincore.org/documents/abstract-model/>. [Consulta: 16/08/2016]

[238] DCMI_MediaWiki (2011). *Glossary/DCMI Abstract Model*. <http://wiki.dublincore.org/index.php/Glossary/DCMI_Abstract_Model>. [Consulta: 16/08/2016]

[239] <http://dublincore.org/documents/2008/01/14/dc-rdf/>. [Consulta: 16/08/2016]

[240] <http://dublincore.org/documents/2008/08/04/dc-html/>. [Consulta: 16/08/2016]

[241] <http://dublincore.org/documents/2008/09/01/dc-ds-xml/>. [Consulta: 16/08/2016]

[242] <http://dublincore.org/documents/2007/12/03/dc-text/>. [Consulta: 16/08/2016]

El Model Abstracte va permetre tenir suficients fonaments per a la creació de Perfils d'Aplicació l'any 2008 segons el model del Singapore Framework[243], el qual és un document o conjunt de documents que especifica i descriu les metadades que s'utilitzen en una aplicació en particular[244]. Per tal d'aconseguir això, un Perfil:

- Descriu què es el que vol aconseguir una comunitat amb la seva aplicació (Requeriments Funcionals)
- Caracteritza els tipus de coses descrites per les metadades i les seves relacions (Model de Domini)
- Numera els termes de metadades que seran utilitzats per al seu ús (Perfil de Conjunt de Descripció i Directrius d'Ús)
- Defineix la sintaxi mecànica que serà utilitzada per codificar les dades (Directrius de Sintaxi i Formats de Dades)

Aquests Perfils d'Aplicació s'utilitzen generalment dins comunitats amb necessitats d'informació específiques que no poden assolir els DCMI Metadata Terms (Vogel, 2014). Un exemple d'ús de Perfil d'Aplicació el trobem al repositori Dryad, on els seus Requeriments Funcionals dins el projecte DRIADE (nom inicial de Dryad) tenien com a objectiu principal la preservació de dades publicades dins el camp de la biologia evolutiva (Dube; Carrier; Greenberg, 2007).

Per tal d'il·lustrar millor l'ús de les etiquetes amb qualificadors, s'ha elaborat un fitxer XML mitjançant l'eina en línia Dublin Code Generator[245] amb metadades descriptives de l'article "Creación de unidades de análisis forense digital en bibliotecas" escrit per l'autor i el Dr. Miquel Térmens. Les etiquetes bàsiques que no necessiten més explicació són les de títol i autor, mentre que per matèria s'ha utilitzat el vocabulari controlat LCSH, però també és possible utilitzar d'altres com el MeSH o l'AAT. S'ha descrit el contingut de l'article breument, sense utilitzar el resum publicat originalment ja que això correspondria a l'etiqueta <dc:abstract>. L'etiqueta d'editor és EPI, ja que és

[243] Nilsson, Mikael; Baker, Thomas; Johnston, Pete (2008).*The Singapore Framework for Dublin Core Application Profiles*. <http://dublincore.org/documents/singapore-framework>. [Consulta: 16/08/2016].

[244] Coyle, Karen; Baker, Thomas (2009). *Guidelines for Dublin Core Application Profiles*. <http://dublincore.org/documents/profile-guidelines>. [Consulta: 16/08/2016]

[245] <http://www.dublincoregenerator.com/>. [Consulta: 21/08/2016]

l'editorial on es publicà originalment l'article. Dins <dc:contributor> s'ha fet esment del finançament del Plan Nacional I+D+i amb el seu codi per poder elaborar l'article. Les diferents dates estan elaborades amb l'esquema W3C-DTF, les quals són de tres tipus: data de creació de l'article, data d'acceptació definitiva de l'article i data de publicació de l'article. L'esquema utilitzat per a indicar el tipus de recurs fou DCMIType, mentre que per indicar el format (un fitxer PDF) s'utilitzà l'esquema IMT. El tipus d'identificador unívoc va ser el DOI amb esquema de codificació URI i la font original de l'article, l'enllaç web on es pot trobar l'article. Per indicar l'idioma en què es va escriure l'article, s'emprà l'esquema ISO 639-2 i per especificar la relació de l'article com "es part de" s'ha usat l'ISSN de la publicació en sèrie. Per acabar, s'ha indicat també l'audiència objectiva de l'article i els posseïdors i/o gestors dels drets.

Figura 6. Exemple d'etiquetes Dublin Core amb qualificadors

```
<?xml version="1.0" encoding="UTF-8"?>

<metadata
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
>

xmlns:dcterms="http://purl.org/dc/terms/">
<dc:title>Creación de unidades de análisis forense en bibliotecas</dc:title>
<dc:creator>Wilderbeek, Theo</dc:creator>
<dc:creator>Térmens, Miquel</dc:creator>
<dc:subject xsi:type="dcterms:LCSH">Digital preservation</dc:subject>
<dc:description>Article sobre com es poden crear laboratoris d'anàlisi forense
digital a biblioteques, amb l'exposició de diversos exemples i mostres de
maquinari i programari.</dc:description>
<dc:publisher>EPI</dc:publisher>
<dc:contributor>El acceso abierto (open access) a la ciencia en España. 2012-2014. Plan Nacional I+D+i,
código CSO2011-29503-C02-01</dc:contributor>
<dcterms:created xsi:type="dcterms:W3CDTF">2014-06-06</dcterms:created>
<dcterms:dateAccepted xsi:type="dcterms:W3CDTF">2014-10-08</dcterms:dateAccepted>
<dcterms:issued xsi:type="dcterms:W3CDTF">2015-01</dcterms:issued>
<dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
<dc:format xsi:type="dcterms:IMT">application/pdf</dc:format>
<dc:identifier xsi:type="dcterms:URI">doi:10.3145/epi.2015.ene.06</dc:identifier>
<dc:source>http://www.elprofesionaldelainformacion.com/contenidos/2015/ene/index.html</dc:source>
<dc:language xsi:type="dcterms:ISO639-2">spa</dc:language>
<dcterms:isPartOf xsi:type="dcterms:URI">urn:issn:1386-6710</dcterms:isPartOf>
<dc:rights>http://www.elprofesionaldelainformacion.com/copyright.html</dc:rights>
<dcterms:audience>Profesionales de la información y la documentación</dcterms:audience>
<dcterms:rightsholder>El Profesional de la Información</dcterms:rightsholder>
<dcterms:rightsholder>Theo Wilderbeek</dcterms:rightsholder>
<dcterms:rightsholder>Miquel Térmens</dcterms:rightsholder>

</metadata>
```

Font: L'autor

Es pot concloure que Dublin Core representa una bona opció per ser utilitzat dins el nostre model per indexar les metadades a nivell descriptiu. És adaptable, extensible i el seu és ampli dins la comunitat de preservació digital. A més, DSpace està configurat per ser utilitzat amb Dublin Core, així que el seu ús no ha de representar excessius problemes tècnics.

3.3.2 DFXML

Digital Forensics XML és un llenguatge XML que s'utilitza per a l'automatització del processat de l'anàlisi forense digital, que compta amb un esquema[246] que permet validar un document DFXML. Fins ara, el seu desenvolupament ha estat limitat degut a la manca de finançament, però ja permet fer tasques molt interessants, com representar el nombre de fitxers, la seva localització física dins la imatge forense i els seus valors *hash*. També permet documentar la procedència, com l'entorn de programari que s'ha utilitzat per crear el fitxer DFXML, el suport original a partir del qual es va crear la imatge forense o el seu sistema de fitxers.

La raó per la qual es va crear el DFXML fou la necessitat de normalitzar la funcionalitat de diverses eines forenses, confinades a sistemes molt específics i que dificultaven poder comparar resultats produïts per diferents eines i algoritmes (Garfinkel, 2012). Per solucionar això, DFXML va ser dissenyat com un llenguatge que descriu processos típics forenses (como la generació de valors *hash*), anàlisi forense (com la localització de fitxers en un disc dur) i metadades (com noms de fitxer i dates i hores de creació).

Un dels programes que produeixen fitxers DFXML és el programari lliure fiwalk que s'inclou en el paquet d'eines forenses SleuthKit (Garfinkel, 2009a). Els documents XML que genera estan formats per quatre parts: informació sobre les eines utilitzades per crear el fitxer, informació sobre la imatge de disc, informació sobre les particions i informació sobre cada fitxer. A continuació analitzem cadascuna d'aquestes parts, amb informació més detallada de les etiquetes a la Taula 16.

Per tal d'il·lustrar millor els exemples, s'ha utilitzat un fitxer DFXML creat pel propi autor. Com es pot veure a la Figura 7, es poden utilitzar etiquetes DCMI per anotar objectes DFXML com el tipus de fitxer (imatge de disc) o un resum del mateix. Pel que fa a la creació del fitxer, es va fer mitjançant fiwalk en la versió 4.2.0 de SleuthKit, compilat en GCC versió 5.4, amb biblioteques AFFLIB i libewf, les quals estan dissenyades per recuperar informació d'imatges forenses. Finalment, es pot consultar

[246] DFXML Working Group (2016). *XML Schema for Digital Forensics XML*. <https://github.com/dfxml-working-group/dfxml_schema/>. [Consulta: 20/08/2016]

també el comanament que s'utilitzà per crear la imatge i la data i hora local en què es va començà a crear la imatge.

Pel que respecta a la imatge de disc i la partició, a la Figura 8 es poden veure les etiquetes creades per identificar la imatge de disc, i a continuació informació sobre la partició, que identifica el sistema de fitxers utilitzat originalment (FAT16), i altres dades tècniques com la mida dels sectors de la partició i el nombre de blocs. Altres etiquetes que no s'han utilitzat en aquest exemple, però que poden ser útils, són el número de sèrie del suport original, i el nom del model del suport.

Figura 7. Etiquetes DFXML amb informació sobre com s'ha creat la imatge forense

```
<?xml version='1.0' encoding='UTF-8'?>
<dfxml
  xmlns='http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML'
  xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  version='1.0'>
  <metadata>
    <dc:type>Disk Image</dc:type>
    <dc:abstract>Memòria USB d'1 GB del fabricant Toshiba</dc:abstract>
  </metadata>
  <creator version='1.0'>
    <program>fiwalk</program>
    <version>4.2.0</version>
    <build_environment>
      <compiler>GCC 5.4</compiler>
      <library name="afflib" version="3.7.7"/>
      <library name="libewf" version="20140608"/>
    </build_environment>
    <execution_environment>
      <command_line>fiwalk -f -X /media/sf_SharedBitCurator/ToshibaBitCuratoraff.xml
      /media/sf_SharedBitCurator/ToshibaBitCuratoraff.aff</command_line>
      <start_time>2016-08-06T14:56:33Z</start_time>
    </execution_environment>
  </creator>
  .....
</dfxml>
```

Font: L'autor, a partir d'una imatge forense AFF

Figura 8. Etiquetes DFXML amb informació sobre la imatge de disc i la partició original

```
  <source>
    <image_filename>/media/sf_SharedBitCurator/ToshibaBitCuratoraff.aff</image_filename>
  </source>
<!-- fs start: 32256 -->
  <volume offset='32256'>
    <partition_offset>32256</partition_offset>
    <sector_size>512</sector_size>
    <block_size>16384</block_size>
    <ftype>4</ftype>
    <ftype_str>fat16</ftype_str>
    <block_count>2004929</block_count>
    <first_block>0</first_block>
    <last_block>2004928</last_block>
    .....
  </volume>
<!-- end of volume -->
```

Font: L'autor, a partir d'una imatge forense AFF

Finalment, a la Figura 9 apareix un exemple d'etiquetes emprades per a cada fitxer present al suport original. En aquest exemple s'ha utilitzat el document Word en què s'està redactant la tesi de l'autor, on s'especifica el nom de fitxer, la partició on està ubicat, un número identificador unívoc, si es tracta d'un fitxer ("r" a name_type) o un directori ("d" a name_type), la mida en bytes, les dades de creació, modificació i d'accés i els valors *hash* MD5 i SHA1. També permet identificar el tipus de fitxer mitjançant la biblioteca libmagic.

Figura 9. Etiquetes DFXML amb informació de fitxer

```
<fileobject>
  <parent_object>
    <inode>2</inode>
  </parent_object>
  <filename>tesi_esborrany.docx</filename>
  <partition>1</partition>
  <id>1224</id>
  <name_type>r</name_type>
  <filesize>7742399</filesize>
  <unalloc>1</unalloc>
  <used>1</used>
  <inode>514</inode>
  <meta_type>1</meta_type>
  <mode>511</mode>
  <nlink>0</nlink>
  <uid>0</uid>
  <gid>0</gid>
  <mtime prec="2">2016-08-04T13:23:02</mtime>
  <atime prec="86400">2016-08-04T04:00:00</atime>
  <crtime prec="2">2016-08-04T13:23:00</crtime>
  <libmagic>XML 1.0 document, ASCII text, with very long lines,
  with CRLF line terminators (Microsoft Word 2007+) </libmagic>
  <byte_runs>
   <byte_run file_offset='0' fs_offset='1957888' img_offset='1990144' len='16384'/>
  </byte_runs>
  <hashdigest type='md5'>a127bd1d34005db250099d19e93a25d0</hashdigest>
  <hashdigest type='sha1'>59adc3b54d252595c2e36d5f3cf357a6d6b523a5</hashdigest>
</fileobject>
```

Font: L'autor, a partir d'una imatge forense AFF

Taula 16. Etiquetes DFXML

Nom d'etiqueta	Descripció
<alloc>	"1" implica que el fitxer es va trobar en un estat assignat. "0" indica espai no assignat, com seria el cas d'un fitxer esborrat
<allocated_only>	Indica que aquest volum només conté *fileobjects* per a fitxers assignats i no per a fitxers recuperats o esborrats
<arch>	Especifica l'arquitectura del processador que s'ha utilitzat per capturar la imatge forense
<atime>	Hora en què es va accedir per últim cop al fitxer
<bkup_time>	Hora en què es va fer la còpia de seguretat del fitxer
<block_count>	Llista del nombre total de blocs en el volum de destinació
<block_size>	Mida en bytes d'un bloc individual de dades en un volum, tal com es defineix al sistema de fitxers
<build_environment>	Informació sobre l'entorn que s'ha utilitzat per generar el fitxer XML
<byte_run>	Etiqueta que s'utilitza per descriure execucions seqüencials de bytes que componen un fitxer
<byte_runs>	Element pare de <byte_run> que descriu la localització específica de fragments de fitxer en un volum. L'etiqueta <byte_runs> mapa els bytes lògics del fitxer a una localització física dins la imatge de disc
<command_line>	Línia de comanaments utilitzat per executar el programa per capturar la imatge forense
<compilation_date>	Data en què es va compilar el programa
<compiler>	Informació sobre el compilador
<compressed>	Valor booleà "0" o "1"
<creator>	Documentació sobre el programa i l'entorn en què s'ha fet la captura i/o l'anàlisi del disc
<crtime>	Hora en què el fitxer es va crear
<ctime>	Hora en què el fitxer de metadades es va modificar per darrera vegada
<dfxml>	Element arrel que marca l'inici i el final del fitxer de metadades DFXML
<dtime>	Hora en què el fitxer es va enregistrar com a esborrat
<error>	Cadena que descriu un error trobat al processar un fitxer
<execution_environment>	Informació sobre el sistema en què es van capturar les dades forenses
<filename>	Nom de fitxer
<fileobject>	Informació del fitxer i les seves metadades, amb informació de la mida i tipus de fitxer, valors *hash* i informació de procedència
<filesize>	Mida del fitxer en bytes
<first_block>	Adreça del primer bloc del sistema de fitxers, en bytes
<ftype>	Identificador numèric que representa el sistema de fitxers present a la partició del volum
<ftype_str>	Cadena de text corresponent al sistema de fitxers que es representa a <ftype>
<gid>	Identificador de grup d'usuaris del fitxer que s'utilitza pe representar interaccions amb documents compartits. Si no està present, el <gid> dóna un valor de "0"
<hashdigest>	Valor *hash* criptogràfic

Nom d'etiqueta	Descripció
<host>	Informació sobre el nom de *host* en què s'ha executat el programa per capturar la imatge
<id>	Identificador únic per al fitxer assignat per fiwalk o DFXML
<image_filename>	Conté el nom de fitxer complet, amb la seva ruta original, de la imatge forense que s'ha creat per a l'anàlisi
<inode>	L'inode és una estructura de dades que gestiona informació sobre fitxers en un sistema de fitxers Unix. El sistema de fitxers assigna un número identificador inode unívoc per a cada fitxer al sistema de fitxers
<last_block>	Número del darrer bloc dins el volum. En els casos en què el primer bloc és "0" el darrer bloc serà un menys que el nombre total de blocs dins el volum
<libmagic>	Valor corresponent al tipus de fitxer identificat per libmagic, biblioteca que llegeix valors en bytes de capçaleres de fitxers
<library>	Biblioteques de programari utilitzades pel programa que realitza l'anàlisi forense
<link_target>	Fitxer al qual fa referència un enllaç
<maxrss>	Mesura d'ús de memòria per part del sistema durant el procés de captura
<meta_type>	Codificació numèrica del tipus de fitxer
<metadata>	Informació de la capçalera que defineix les metadades al document DFXML
<mode>	Identificació de tipus de permisos per poder obrir el fitxer
<mtime>	Hora en què les dades del fitxer van ser modificades per darrera vegada
<name_type>	Representació del valor de tipus de fitxer per a una entrada en una estructura de directoris. Generalment, els valors seran "r" per a un fitxer i "d" per a un directori
<nlink>	Nombre d'enllaços a un inode específic
<orphan>	Un fitxer sense una estructura de metadades de referència
<os_release>	Número de versió del sistema operatiu utilitzat per capturar dades forenses
<os_sysname>	Nom del sistema operatiu utilitzat per capturar dades forenses
<os_version>	Informació completa del sistema operatiu utilitzat per capturar dades forenses
<partition>	Partició en què es troba el fitxer. Si només hi ha una partició, el valor serà "1"
<partition_offset>	Localització d'inici de la partició dins la imatge de disc, mesurada en bytes. Quan el valor és 0, vol dir que la primera partició comença a l'inici de l'espai de disc disponible
<program>	Nom del programa que genera el fitxer XML
<rusage>	Durada en l'execució d'ús del processador
<sector_size>	Mida en bytes d'un sector de disc del volum
<seq>	Número incremental per a entrades a sistemes de fitxers NTFS
<source>	Font de les dades forenses. Pot contenir informació contextual sobre el volum objectiu de l'anàlisi forense, com dispositiu i metadades d'adquisició, i l'estructura de les dades al volum objectiu
<start_time>	Data i hora en què s'executà el programa de captura de dades forenses

Nom d'etiqueta	Descripció
<uid>	Número identificador de l'usuari si està present. Si no ho està, retorna el valor "0"
<unalloc>	"1" implica que el fitxer s'ha trobat sense assignació
<unused>	Indica si l'estructura de metadades del fitxer no s'ha utilitzat mai
<used>	Indica si l'estructura de metadades del fitxer s'ha utilitzat al menys una vegada
<username>	Nom de l'usuari sota el qual s'ha executat el programa
<version>	Número de versió del programa d'anàlisi forense
<volume>	Informació sobre fitxers individuals a un fitxer DFXML

Fonts: Garfinkel, 2012; DFXML Tag Library: v4. <https://goo.gl/BTSTvs>. [Consulta: 20/08/2016]

Es pot concloure que DFXML és una bona solució per documentar les metadades tècniques d'un suport digital, ja que dóna informació molt acurada sobre les operacions que es realitzen per crear la imatge forense i l'anàlisi posterior, així com la possibilitat d'utilitzar com a complement Dublin Core per fer descripcions del suport original.

3.3.3 PREMIS

La norma ISO 14721 especifica que les metadades de preservació associades amb les activitats d'Emmagatzematge d'Arxiu, com el *refreshing* de suports o la normalització/migració de formats, s'ha de capturar i guardar dins el PDI. Per tal d'assolir aquesta meta, l'any 2003 es formà un grup de treball entre l'OCLC i el RLG, el qual tenia com objectius principals definir un conjunt d'elements de metadades de preservació que es poguessin aplicar a la comunitat de preservació digital i crear un diccionari de dades que donés suport al conjunt de metadades. El resultat final fou la publicació, dos anys més tard, del diccionari de dades PREMIS (PREMIS Working Group, 2005, p. vii). Aquest diccionari defineix un conjunt de metadades de preservació que es poden aplicar abastament a repositoris de preservació digital i es troba acompanyat d'un esquema XML que facilita la seva adopció de manera directa (Ariza López et al., 2012). Després d'un procés de revisió, es publicà la versió 2 del diccionari tres anys més tard (PREMIS Editorial Committee, 2008) i recentment, la versió 3 (PREMIS Editorial Committee, 2015). El diccionari de dades inclou conceptes com la procedència (qui ha estat el propietari de l'objecte digital?), l'autenticitat (és l'objecte digital allò que pretén ser?), l'activitat de preservació (què s'ha fet per preservar l'objecte digital?), l'entorn tècnic (què és necessita per representar i utilitzar l'objecte

digital?) i la gestió de drets (quins són els drets de propietat intel·lectual que s'han de tenir en compte?) (Johnston, 2012).

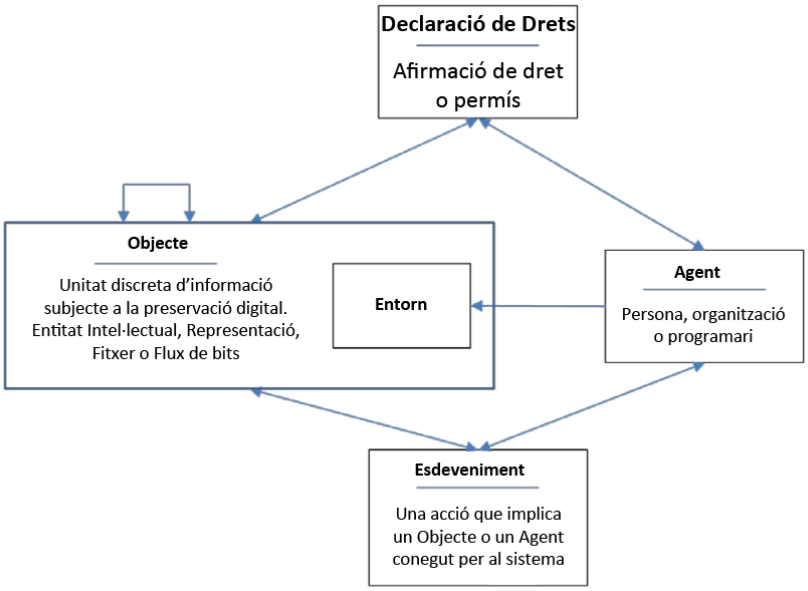
Actualment, PREMIS ja es considera l'estàndard *de facto* per a metadades de preservació (Lavoie, 2014) i el més influent dins la comunitat de preservació digital (Donaldson; Yakel, 2013). Prova de la seva importància és la concessió de importants premis, com el Digital Preservation Award[247] i el Preservation Publication Award[248].

Una de les bases de PREMIS és que fou dissenyat per poder ser utilitzat en organitzacions que utilitzin l'estàndard de preservació OAIS, encara que existeixen algunes petites diferències en la terminologia (PREMIS Editorial Committee, 2015, p. 2). Aquestes diferències habitualment es refereixen en què la informació que ha de conèixer un repositori de preservació ("unitat semàntica" en terminologia de PREMIS) requereix ser més específica que la que donen les definicions d'OAIS. Cada unitat semàntica definida al model de dades PREMIS es mapa a una entitat que s'organitza dins el model de dades. Per tant, un unitat semàntica es pot entendre com una propietat d'una entitat. Inicialment, el model contemplava cinc entitats d'alt nivell (Lee; Stvilia, 2014), però amb la publicació de la versió 3 actualment consta de quatre, ja que l'Entitat Intel·lectual (definida més avall) ara forma part de l'Entitat Objecte. La Figura 10 mostra les relacions d'aquestes entitats dins el model.

[247] Digital Preservation Coalition (2005, Nov. 22). *International team wins the 2005 Digital Preservation Award*. <http://www.dpconline.org/newsroom/not-so-new/110-awards-2005>. [Consulta: 03/07/2016]

[248] Digital Preservation Coalition (2006, Aug. 25). *PREMIS Data Dictionary wins 2006 Society of American Archivists' Preservation Publication Award*. <http://www.oclc.org/research/news/2006/08-25b.html>. [Consulta: 03/07/2016]

Figura 10. Model de dades PREMIS



Font: PREMIS Editorial Committee, 2015, p. 6. Traducció de l'autor

Dins les relacions de la Figura 10, quan la fletxa és bidireccional cada tipus d'entitat conté una entitat semàntica que permet enllaçar-la a l'altra. Per exemple, l'entitat Drets inclou una unitat semàntica que enregistra informació sobre la relació amb un Agent, i l'entitat Agent inclou una unitat semàntica que enregistra informació sobre Drets associats. Les entitats i altres termes presents en el model de dades es defineixen a continuació:

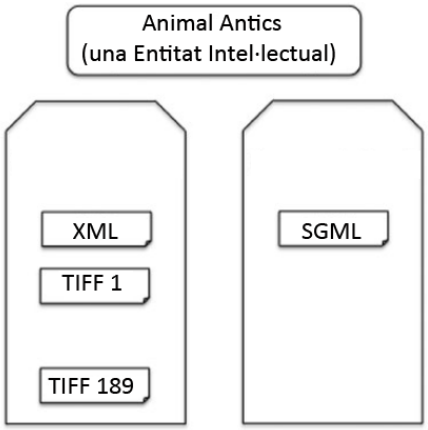
- Entitat Objecte. Agrupa informació sobre un objecte digital gestionat per un repositori de preservació i descriu les característiques que siguin rellevants per a la gestió de la preservació. L'entitat inclou quatre subcategories:
    - Entitat Intel·lectual. Creació intel·lectual o artística que es considera rellevant per a una comunitat designada dins el context de la preservació digital. Una Entitat Intel·lectual pot incloure altre Entitats Intel·lectuals com un web, el qual pot incloure imatges. Una Entitat Intel·lectual també pot tenir una o més Representacions digitals o no digitals
    - Representació. Conjunt de fitxers, incloent metadades estructurals, necessaris per a una interpretació d'una Entitat Intel·lectual. Per exemple, la Figura 11 mostra que la monografia *Animal Antics* es pot representar amb una imatge TIF per a cadascuna de les 189 pàgines amb un fitxer XML que representa les metadades estructurals i això seria una

Representació; però es pot donar una segona representació amb la creació d'un sol fitxer de text en format SGML que inclou tot el contingut en text del llibre. El repositori, per tant, enregistraria metadades de dos objectes de representació i 191 objectes de fitxer

  - Fitxer. Seqüència de bytes amb un nom i un ordre que reconeix un sistema operatiu. Un Fitxer pot ser de zero o més bytes i té un format de Fitxer, permisos d'accés i característiques de sistemes de Fitxer com mida en bytes i data de modificació
  - Flux de bits. Dades adjacents o no adjacents dins un fitxer que tenen unes propietats comuns per a propòsits de preservació

- Entitat Esdeveniment. Agrupa informació vers una acció que afecta un o més Objectes. Les metadades sobre un Esdeveniment s'haurien d'enregistrar i emmagatzemar separadament de l'objecte digital. En funció de la importància de l'Esdeveniment, quedarà enregistrat dins el repositori. Per exemple, totes aquelles accions que modifiquin objectes s'han d'enregistrar sempre
- Entitat Agent. Agrupa informació sobre atributs o característiques d'Agents, que inclouen persones, organitzacions o programari, associats amb la gestió de Drets i accions de preservació presents a la vida de l'objecte de dades. La informació d'Agent serveix per identificar un Agent de forma inequívoca de la resta d'entitats Agent.
- Entitat Drets. Dins el model de dades, totes les declaracions de drets i permisos legals s'utilitzen com a construccions que poden ser descrits com a una entitat de Drets. Els drets són legitimacions concedides pels Agents mitjançant drets de propietat intel·lectual, mentre que els permisos serien poders o privilegis concedits mitjançant un acord entre el titular dels drets i altre grup o grups.
- Entorn. Tecnologia (programari o maquinari) que dóna suport a un Objecte Digital d'alguna manera (com representació o execució)

Figura 11. Exemple d'Entitat Intel·lectual amb *Animal Antics*



Fonts: Lee; Stvilia, 2014; PREMIS Editorial Committee, 2008. Traducció de l'autor

Un exemple d'aplicació de metadades PREMIS el trobem a la biblioteca de la University of North Carolina, on utilitzen aquest estàndard al seu Carolina Digital Repository[249]. Algunes de les funcions en què es troba implementat es mostren a la Figura 12; com a valor d'identificació del PID de l'entitat Objecte s'empra un UUID; un altre UUID com a valor d'identificació del PID de l'entitat Esdeveniment, el qual és del tipus validació en què s'ha fet un escanejat de virus; i la resta d'etiquetes identifiquen l'agent vinculat que ha fet l'escanejat, que ha estat l'administrador amb un programari sense identificar.

[249] <https://cdr.lib.unc.edu/>. [Consulta: 04/07/2016]

Figura 12. Exemple de metadades PREMIS al CDR

```
<premis xmlns="info:lc/xmlns/premis-v2" version="2.0">
  <object xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:type="representation">
    <objectIdentifier>
      <objectIdentifierType>PID</objectIdentifierType>
      <objectIdentifierValue>uuid:c0a8beae-b6a3-4012-8ef0-00b25d19eaed</objectIdentifierValue>
    </objectIdentifier>
  </object>
  <event>
    <eventIdentifier>
      <eventIdentifierType>URN</eventIdentifierType>
      <eventIdentifierValue>urn:uuid:b0d7b307-89a5-484a-ba0d-23c849ca32b5</eventIdentifierValue>
    </eventIdentifier>
    <eventType>
      http://id.loc.gov/vocabulary/preservationEvents/validation
    </eventType>
    <eventDateTime>2011-04-14T00:00:00</eventDateTime>
    <eventDetail>virus scan</eventDetail>
    <linkingAgentIdentifier>
      <linkingAgentIdentifierType>Name</linkingAgentIdentifierType>
      <linkingAgentIdentifierValue>admin:ADMINISTRATORS</linkingAgentIdentifierValue>
      <linkingAgentRole>Initiator</linkingAgentRole>
    </linkingAgentIdentifier>
    <linkingAgentIdentifier>
      <linkingAgentIdentifierType>Name</linkingAgentIdentifierType>
      <linkingAgentIdentifierValue>n/a</linkingAgentIdentifierValue>
      <linkingAgentRole>Software</linkingAgentRole>
    </linkingAgentIdentifier>
  </event>
```

Font: PREMIS Implementation Registry (2014, May 29). <https://www.loc.gov/standards/premis/registry/index.php>. [Consulta: 04/07/2016]

Així doncs, podem concloure que PREMIS és molt eficaç per gestionar metadades de preservació, atès que documenta adequadament totes les accions i transformacions per els quals ha passat un fitxer de dades, a més de donar informació molt precisa sobre els drets d'autor i identifica tots els agents implicats durant els processos. Com que es tracta d'un estàndard pensat per a la preservació, el seu ús requereix un estàndard de metadades addicional per a casos d'ús com indexació de citació del *dataset* o etiquetes adequades com la cobertura geogràfica i temporal de l'estudi, la metodologia emprada, el finançament rebut, etc.

3.3.4 BagIt

L'especificació BagIt és un format d'empaquetat de fitxers per a la creació de contenidors estàndards anomenats *bags*, els quals s'utilitzen per emmagatzemar i transferir contingut digital. Un *bag* consisteix en una 'càrrega' (el contingut) i unes 'etiquetes', que són fitxers de metadades que documenten l'emmagatzematge i la transferència del paquet. Aquesta especificació fou creada degut als problemes tècnics que va patir l'any 2007 la California Digital Library per transferir diversos TB de contingut (majorment dades de pàgines web) a la Library of Congress. Treballant

conjuntament amb John Kunze de la California Digital Library i amb Andy Boyko, Justin Littman, Liz Madden i Brian Vargas de la LC es va produir una primera versió que es va anomenar inicialment 'LC Package Specification' i actualment té el nom BagIt (Johnston, 2010). L'especificació es troba disponible públicament i la seva darrera revisió és la 0.97, publicada el 21 d'octubre de 2016[250]. Un *bag* ha de contenir com a mínim:

- Un directori de dades que inclou el contingut digital que s'ha de preservar
- Un fitxer de manifest que llista els fitxers presents en el directori de dades, així com els seus valors *hash*
- Un fitxer 'bagit.txt' que identifica el directori com a *bag*, la versió de l'especificació BagIt que s'ha emprat i la codificació de caràcters que s'ha utilitzat per etiquetar els fitxers.

Les metadades BagIt es troben dins el fitxer 'bag-info.txt', que es tracta d'un fitxer d'etiquetes que conté elements de metadades que descriuen el *bag* i la càrrega en bytes. Tots aquests elements són opcionals i repetibles i es detallen a la Taula 17.

A tall d'exemple, la Figura 13 mostra com seria un fitxer 'bag-info.txt' amb tots els camps emplenats. L'especificació estipula que cada metadada ha de consistir d'una etiqueta, dos punts i un valor, amb una separació opcional d'un espai en blanc. S'ha d'esmenar que l'etiquetes 'Payload-Oxum' i 'Bagging-Date' són generades de manera automàtica, mentre que la resta les pot editar la institució tot creant un perfil personalitzat de metadades.

[250] Kunze, J. et al. *The BagIt File Packaging Format (V0.97).* <https://tools.ietf.org/html/draft-kunze-bagit-14/>. [Consulta: 23/10/2016]

Taula 17. Camps de metadades BagIt

Etiqueta	Descripció
Source-Organization	L'organització que transfereix el contingut
Organization-Address	Adreça postal de l'organització
Contact-Name	Persona de l'organització responsable de la transferència del contingut
Contact-Phone	Número de telèfon (amb format internacional) de la persona o càrrec responsable
Contact-Email	Correu electrònic de la persona o càrrec responsable
External-Description	Descripció breu dels continguts i la procedència
Bagging-Date	Data (AAAA-MM-DD) en què el contingut s'ha preparat per al lliurament
External-Identifier	Identificador per al *bag*, subministrat pel remitent
Bag-Size	Mida del *bag* que es transfereix, seguit d'una abreviació com MB, GB o TB
Payload-Oxum	Signatura en nombre de bytes i en nombre de fitxers del *bag*
Bag-Group-Identifier	Identificador per al conjunt de *bags* subministrat pel remitent. Aquest identificador ha de ser únic en tot el contingut del remitent, i s'ha de reconèixer que pertany a un esquema únic
Bag-Count	Dos números separats per "of", en particular, "N of T", on T és el nombre total de *bags* en un grup de *bags* i N és el número ordinal dins el grup. P. ex.: 1 of 2
Internal-Sender-Identifier	Identificador altern subministrat pel remitent per al contingut i/o el *bag*
Internal-Sender-Description	Descripció dels continguts del *bag* basada en estàndards interns

Font: Kunze, J. et al. *The BagIt file packaging format (V0.97).* <https://tools.ietf.org/html/draft-kunze-bagit-14>. [Consulta: 23/10/2016]

Figura 13. Etiquetes BagIt

```
Source-Organization: Spengler University
Organization-Address: 1400 Elm St., Cupertino, California, 95014
Contact-Name: Edna Janssen
Contact-Phone: +1 408-555-1212
Contact-Email: ej@spengler.edu
External-Description: Uncompressed greyscale TIFF images from the Yoshimuri papers colle...
Bagging-Date: 2008-01-15
External-Identifier: spengler_yoshimuri_001
Bag-Size: 260 GB
Payload-Oxum: 279164409832.1198
Bag-Group-Identifier: spengler_yoshimuri
Bag-Count: 1 of 15
Internal-Sender-Identifier: /storage/images/yoshimuri
Internal-Sender-Description: Uncompressed greyscale TIFFs created from microfilm and are...
```

Font: Kunze, J. et al. *The BagIt file packaging format (V0.97).* <https://tools.ietf.org/html/draft-kunze-bagit-14>. [Consulta: 23/10/2016]

En el cas d'aquesta tesi, BagIt és una molt bona opció per crear empaquetats de fitxers addicionals a la imatge forense (com registres de metadades) que com veurem més endavant, també s'hauran de preservar i preparar per a la ingesta al repositori. La raó d'escollir aquesta especificació d'empaquetat és que s'utilitza per transferir grans volums de dades; per exemple, la LC ha transferit més de 80 TB mitjançant BagIt

(Minor et al., 2010). A més, s'ha demostrat que és un mitjà molt eficaç per ingestar dades, ja que a la xarxa de preservació digital Chronopolis[251] s'anima als proveïdors a crear col·leccions de dades de fins a 5 TB abans de subdividir-les en *bags* més petits (Minor et al., 2009). Per altra banda, alguns repositoris institucionals ja fan ús de BagIt, com el de l'Emory University, on s'han ingestat 436 imatges de disc (on s'inclouen formats forenses com EnCase), que sumen un total d'1,2 TB[252]. En aquest cas, però, l'Emory utilitza un repositori Fedora. El mateix cas el trobem en el Stanford Digital Repository (que també utilitza Fedora), on BagIt és el format principal de transferència per al contingut que s'hi diposita (Cramer; Kott, 2010). Per altra banda, la xarxa DataONE de repositoris de dades fa servir BagIt per al transport de paquets de dades de repositoris als usuaris finals[253]. Finalment, un exemple de repositori DSpace que utilitza BagIt és Dryad, que fa ús del mòdul BagIt Handshaking[254] per compartir paquets de dades amb altres repositoris. Actualment, aquest mòdul s'utilitza per compartir dades NeXus entre Dryad i TreeBASE, un repositori d'informació filogenètica.

[251] <http://libraries.ucsd.edu/chronopolis/>. [Consulta: 31/03/2017]

[252] Koeser, Rebecca Sutton; Roke, Elizabeth Russey; Waugh, Dorothy (2017). *Sipping from a bag: ingesting disk images with BagIt [*pòster]. <https://pid.emory.edu/ark:/25593/rpff8>. [Consulta: 02/04/2017]

[253] DataONE (2014). *DataONE archictecture, version 1.2*. <https://releases.dataone.org/online/api-documentation-v1.2.0/design/DataPackage.html>. [Consulta: 13/04/2017]

[254] <http://wiki.datadryad.org/BagIt_Handshaking>. [Consulta: 11/05/2017]

4. Anàlisi forense digital

Com ja s'ha explicat (vegeu el capítol 1.3.10), l'anàlisi forense digital és un conjunt de tècniques que permeten la preservació (entre altres operacions) de fonts digitals amb l'objectiu de reconstruir fets que hagin estat producte d'activitats criminals. L'obra de Donn Parker, *Crime by computer*, escrita l'any 1976, és "perhaps the first description of the use of digital information to investigate and prosecute crimes committed with the assistance of a computer" (Pollitt, 2010). Dins l'àmbit privat, hi ha grans corporacions que utilitzen els seus propis grups de recerca forense per investigar problemes interns; dins l'àmbit públic, els governs utilitzen recursos forenses per a la intel·ligència militar (Dietrich; Adelstein, 2015).

Pel que fa a la seva aplicació a la preservació digital, un dels primers estudis el podem trobar a un informe de Ross i Gow (1999), on es discutien els avenços en la recuperació de dades i l'anàlisi forense digital i la seva rellevància potencial. Poc més tard, es van publicar alguns articles sobre mètodes per recuperar dades amb eines forenses, com el presentat per la National Library of Australia on explicaven com recuperar dades de disquets[255] o sobre les possibilitats d'ús d'eines forenses en el patrimoni cultural (Duranti, 2009; Garfinkel; Cox, 2009). Per altra banda, un informe del projecte FIDO de 2011 va investigar i documentar l'ús de diverses eines d'anàlisi forense digital per donar suport a la conservació i la preservació d'informació digital dipositada en sistemes informàtics i suports digitals (Knight, 2012), i segons un informe de Jeremy Leighton John (2012), existeixen diverses vies d'explotar el seu potencial dins el camp de la preservació digital, com la varietat d'eines de codi obert que permeten fins i tot a institucions petites l'ús de les tècniques, la possibilitat de recuperar informació de tecnologies obsoletes, la recuperació d'informació encriptada amb la col·laboració del dipositant, donar autenticitat i valor històric a objectes digitals o la recerca continua de les tècniques forenses com a conseqüència inevitable de l'aparició de noves tecnologies.

Un reconeixement de la importància de l'anàlisi forense el va fer la UNESCO i la UBC l'any 2012 amb el document *UNESCO/UBC Vancouver Declaration*, on instaven a la secretaria de la UNESCO a "encourage engagement of cultural heritage professionals knowledgeable about digital forensics concepts, methods and tools in order to ensure

[255] Woodyard, Deborah (2001, Jan. 16-18). "Data recovery and providing access to digital manuscripts". *Information Online 2001 Conference*. Sidney, Australia. Recuperat del web Internet Archive. <https://goo.gl/DdxGMg>. [Consulta: 23/03/2017]

capture and reliable preservation of authentic, contextualized and meaningful information, and appropriate mediation of access to the information"[256].

Un altre exemple el tenim amb la publicació de la National Agenda for Digital Stewardship 2014[257], que va tenir com a objectiu "offer inspiration and guidance and suggest potential directions and areas of inquiry for research and future work in digital stewardship". L'agenda va incloure el següent paràgraf:

> As more digital materials are selected for long-term digital preservation, the need to integrate digital forensics tools into production workflows for collections becomes increasingly important. This will require identifying the boundaries between technical infrastructure development and organizational policies, and where there is tension that creates issues for providing access or pursuing work that reduces tension whether it be new or refined policies or services and tools development. Integration of these tools can build on exploratory work using digital forensics. Tools currently under development can be leveraged and workflows can be implemented. Aside from the need for tools and workflow developments, there are also important opportunities for organizations to share resources in order to tackle these issues. In this respect, pioneering new organizational models for centers of stewardship, such as SWAT sites, can help to support the development of centers of excellence that help to scale up this kind of activity.
>
> There is a clear need to move the basic research in digital forensics tools from research to implementation in production workflows for organizations. This would require investment in scaling up tools and creating collaborative models for sharing resources to make this work possible. The digital preservation community would also benefit from a shared space for exchanging knowledge around how forensics tools are being integrated into production preservation activities.

Dins d'aquesta tesi tractarem aquesta disciplina en tres vessants: en primer lloc, analitzarem la metodologia que permet preservar objectes digitals en consonància amb el nostre model; en segon lloc, es presentaran casos d'ús en què s'han preservat amb èxit col·leccions patrimonials amb tècniques forenses o bé casos on es presentin

[256] *UNESCO/UBC Vancouver Declaration: the memory of the world in the digital age: digitization and preservation*. <https://goo.gl/uUsrqA>. [Consulta: 23/03/2017]

[257] <http://www.digitalpreservation.gov/documents/2014NationalAgenda.pdf>. [Consulta: 23/03/2017]

innovacions per a l'aplicació de les tècniques a biblioteques i/o arxius i en tercer lloc, es faran diverses proves de programari forense per preservar dades.

4.1 Metodologia dins l'anàlisi forense digital

Els processos que utilitzen els cossos i forces de seguretat com el Departament de Justícia dels EUA per tal de preservar proves inclouen la identificació de l'incident, preparar les eines, assegurar i documentar l'escena criminal, recollir les proves en suport informàtic, empaquetar, transportar i guardar de forma segura les proves, fer la inspecció inicial, fer la imatge forense, examinar-la i fer l'informe final (Ashcroft; Daniels; Hart, 2004; Ayers; Brothers; Jansen, 2014). Aquesta metodologia es basa en la ciència forense preexistent, la qual té com a mètode bàsic el principi d'intercanvi del Dr. Edmond Locard (1877-1966), que sosté que el responsable d'un crim deixarà alguna cosa enrere dins l'escena que el delatarà més endavant. Aquest principi es pot resumir amb la frase "Every contact leaves a trace" (Byard et al., 2016; Crispino, 2008), que posteriorment fou refinat pel Dr. Paul L. Kirk (1902-1970) amb la següent declaració (Chisum; Turvey, 2006, p. 30-31):

> Wherever he steps, whatever he touches, whatever he leaves, even unconsciously, will serve as a silent witness against him. Not only his fingerprints or his footprints, but his hair, the fibers from his clothes, the glass he breaks, the tool mark he leaves, the paint he scratches, the blood or semen he deposits or collects. All of these and more, bear mute witness against him. This is evidence that does not forget.

Aquesta formulació es podria resumir en la següent frase, segons Crispino (2008): "Any object of our universe is unique". Queda clar que una escena del crim pot contenir una o diverses proves que identificaran el criminal i per tant, és cabdal que un grup de tècnics forenses s'ocupin de segellar l'escena i recollir possibles proves sense incórrer en cap tipus de contaminació, a més d'enregistrar en vídeo i prendre fotografies de l'escena. Dins els principis arxivístics, això és equivalent al principi de l'ordre original, que s'entén amb més utilitat dins el context més ampli de la cadena de custòdia, definida per Pearce-Moses (2005, p. 67) com "succession of offices or persons who have held

materials from the moment they were created". Aquest principi de mantenir una custòdia responsable mitjançant control, documentació i justificació de tots els estats en què les dades es preserven i les seves possibles transformacions permetrà assegurar la seva autenticitat, integritat i compliment normatiu (Lee, 2012a), a més de documentar la seva procedència, definida per Nesmith (1999, p. 146) com "social and technical processes of the records' inscription, transmission, contextualization, and interpretation which account for its existence, characteristics, and continuing history". Ja s'han utilitzat aquests mètodes per preservar materials nascuts digitals, com veurem més endavant al capítol 4.2.

A continuació tractarem els passos més concrets i específics per tal de preservar a llarg termini grans col·leccions de dades amb mètodes d'anàlisi forense digital, els quals parteixen d'escrits de John (2012), Kirschenbaum et al. (2010), Knight (2012) i de Redwine et al. (2013), a més del treball realitzat per James Baker[258].

4.1.1 Preparatius inicials

El primer pas es fer la recepció dels suports físics o bé gestionar l'arribada en línia dels objectes digitals com a Paquet d'Informació d'Enviament. En el cas de suports físics, s'ha de fer un inventari de tots els materials, fer un examen exhaustiu, valorar el seu estat i assignar un número identificador a cada ítem físic amb informació digital. Si es donés el cas que hi ha suports amb informació inaccessible, ja sigui pel seu mal estat, per incompatibilitats de format o per oxidació, es retornarien al donant. Un cop controlat i inventariat tot el maquinari, amb una atenció especial al que es trobi en mal estat, s'haurien de separar i classificar el suports. En funció del centre, es pot fer segons el tipus de suport o segons el sistema de fitxers. Els discs de programari d'instal·lació s'haurien de deixar apart per tal de ser emmagatzemats però no per ser capturats.

Llavors s'ha de preparar una estació de treball, preferiblement una unitat FRED amb el programari especialitzat FTK, o bé una estació amb l'entorn BitCurator, el qual expliquem amb més detall al capítol 4.2.6. S'hauran de bloquejar contra escriptura amb

[258] *Processing workflow for digital media*. <https://github.com/drjwbaker/bitcurator-workflow>. [Consulta: 20/07/2016]

write blockers els suports en què sigui possible que s'alteri el seu contingut (com és el cas dels discs durs), excepte aquells en què ja es troben bloquejats d'origen contra escriptura (com els CD-ROMs i DVD-ROMs). Si fos necessari, s'han de connectar unitats de lectura externes si l'estació de treball no les té d'internes per a suports com discs òptics; en el cas de disquets es pot utilitzar maquinari especialitzat com KryoFlux[259] o FC5025[260], els quals inclouen programari especialitzat per a la creació d'imatges de disquet. Per tal de facilitar la transferència posterior de paquets d'informació, es pot configurar una memòria USB com a unitat d'emmagatzematge temporal i finalment s'han d'ajustar les opcions d'energia per evitar que l'equip entri en hibernació o en estat de suspensió.

4.1.2 Captura forense

Arribats a aquest punt, ja entren en joc les eines forenses amb l'adquisició física (creació d'una imatge forense) o lògica (captura del conjunt de fitxers d'un disc dur). S'han de crear valors *hash* per a cada objecte digital creat i de forma opcional es pot repetir el procés amb un altre equip independent per tal de comprovar que s'obté el mateix valor exacte. En el cas de produir-se errors en la captura, s'utilitzaria un programari especialitzat però si els errors persisteixen, es deixarà el suport a un costat i s'anotarà aquest error dins els fitxers de metadades.

4.1.3 Examen i anàlisi de contingut

A continuació, s'avaluen i s'inspeccionen les imatges forenses i els objectes digitals amb comprovacions de presència de virus i identificació de sistemes de fitxers i de formats presents. El sistema BitCurator permet crear un directori i una llista de fitxers per a la imatge forense, a més de poder navegar dins la imatge i fer captures selectives. Un pas important aquí és l'examinació molt curosa del contingut de les imatges forenses per tal de bloquejar informació sensible amb la funció de cerca del programari forense i crear les metadades que utilitzarà el centre per fer l'arxiu definitiu, amb especial atenció

[259] <https://kryoflux.com/>. [Consulta: 25/03/2017]

[260] *FC5025 USB 5.25" floppy controller - Device Side Data*. <http://www.deviceside.com/fc5025.html>. [Consulta: 25/03/2017]

als drets d'autor implicats. De forma opcional es pot fer servir l'emulació, una de les eines més valorades per a la preservació del *look and feel* original, amb experiències exitoses com la de l'Emory University (vegeu el capítol 4.2.4) o la realitzada a la New York Public Library amb disquets de 5 ¼ polzades de la col·lecció Timothy Leary Papers amb els emuladors DOSBox i WinUAE (Dietrich et al., 2016), per accedir al contingut. Seguidament s'examinen i s'extreuen les metadades per tal de seleccionar els camps més adequats a un catàleg. A continuació s'exporten còpies d'objectes digitals per tal de transferir-les a l'emmagatzematge intern del centre i si es considerés necessari es farien passos de conversió o migració per tal de tenir objectes digitals interoperables i aptes per a la preservació a llarg termini.

Un cop fet el pas de l'anàlisi i la visualització, es poden crear versions dissenyades per a futures consultes d'usuaris (Paquets d'Informació de Difusió), amb accés restringit o un redactat d'informació sobre les condicions d'accés si és necessari.

4.1.4 Emmagatzematge digital

En aquest pas s'ha de preparar el paquet de dades juntament amb les seves metadades de processat forense per tal de transferir-los al sistema d'emmagatzematge intern del centre, el qual seria, en terminologia OAIS, el Paquet d'Informació d'Arxiu. A continuació, es fa la ingesta definitiva de Paquets d'Informació de Difusió al sistema de repositori digital del centre. Tot depenent dels acords previs amb el donant o dipositant, es retornaran o es retindran els suports originals amb dades i les imatges forenses creades originalment.

4.1.5 Accés i consulta dels recursos

S'han de fer comprovacions per tal d'assegurar que es compleix la política de preservació del centre quant a l'accés dels materials digitals o si aquesta no existeix, s'haurà de desenvolupar. L'accés es farà en funció d'aquesta política: lliure descàrrega, consulta en sala o bé accés institucional o en línia.

4.2 Casos d'ús de tècniques forenses a biblioteques i arxius

A continuació es descriuen les experiències més rellevants en què s'han utilitzat eines d'anàlisi forense digital per preservar objectes digitals en institucions rellevants dins els camps de la recerca i la cultura i projectes de creació de fluxos de treball a biblioteques i arxius amb mètodes i tecnologies forenses. Primerament, s'exposa el primer intent seriós de proporcionar un flux de treball dins les institucions culturals per preservar arxius personals en forma digital, el qual va estar representat pel projecte Digital Lives de la British Library; en segon lloc, es presenten els passos que va haver de fer la Bodleian Library per tal de gestionar els arxius híbrids dins les seves col·leccions, on es van utilitzar eines com la migració, i que van representar canvis profunds dins la institució; en tercer lloc, s'expliquen els passos que va fer la National Library of Australia per tal de tenir un flux de treball que va facilitar en gran mesura la preservació i catalogació de materials nascuts digitals; en quart lloc, s'exposa un cas d'emulació d'arxiu digital dins l'Emory University, el qual va suposar un gran repte tècnic i que va requerir utilitzar tècniques forenses per preservar informació sensible; en cinquè lloc, es descriu el projecte AIMS, una associació entre quatre universitats nord-americanes per dissenyar procediments d'accés i processat de materials nascuts digitals, en els quals l'anàlisi forense digital jugà un paper important; i en últim lloc, parlem del projecte BitCurator, que es va dissenyar per integrar eines forenses digitals als fluxos de treball de biblioteques i arxius.

4.2.1 British Library

Dins les operacions que compleix la British Library com a biblioteca nacional del Regne Unit, es troba l'acceptació d'arxius personals els quals inclouen cartes, diaris, fotografies i també suports informàtics com cintes magnètiques, disquets, discs òptics (com CD-ROMs) i discs durs tant interns com externs. L'any 2000 la British Library va adoptar el terme eMANUSCRIPTS (eMSS) per fer referència a l'equivalent digital de la documentació personal (John, 2008). Per tal de capturar, replicar i retenir la informació digital dels eMSS, es va posar en marxa de forma interna el projecte Digital

Manuscripts en el qual es van estudiar tres tecnologies clau: programari i maquinari forense, maquinari de sistemes antics ja obsolets i tècniques i perspectives informàtiques que considerin l'evolució tecnològica en el futur, que planteja reptes com la preservació permanent de dades d'ADN, la qual presenta la mateixa informació durant centenars de milions d'anys (Wong; Wong; Foote, 2003). Un concepte interessant que es va plantejar fou el de *scriptorium* digital, un lloc on copiar i recuperar materials nascuts digitals de suports obsolets com disquets, targetes perforades o cassets[261] mitjançant quatre etapes clau, les quals han d'incloure requeriments de preservació digital a llarg termini: l'adquisició de materials originals, el replicat de les dades, l'avaluació i la descripció dels materials i l'accés a l'usuari final.

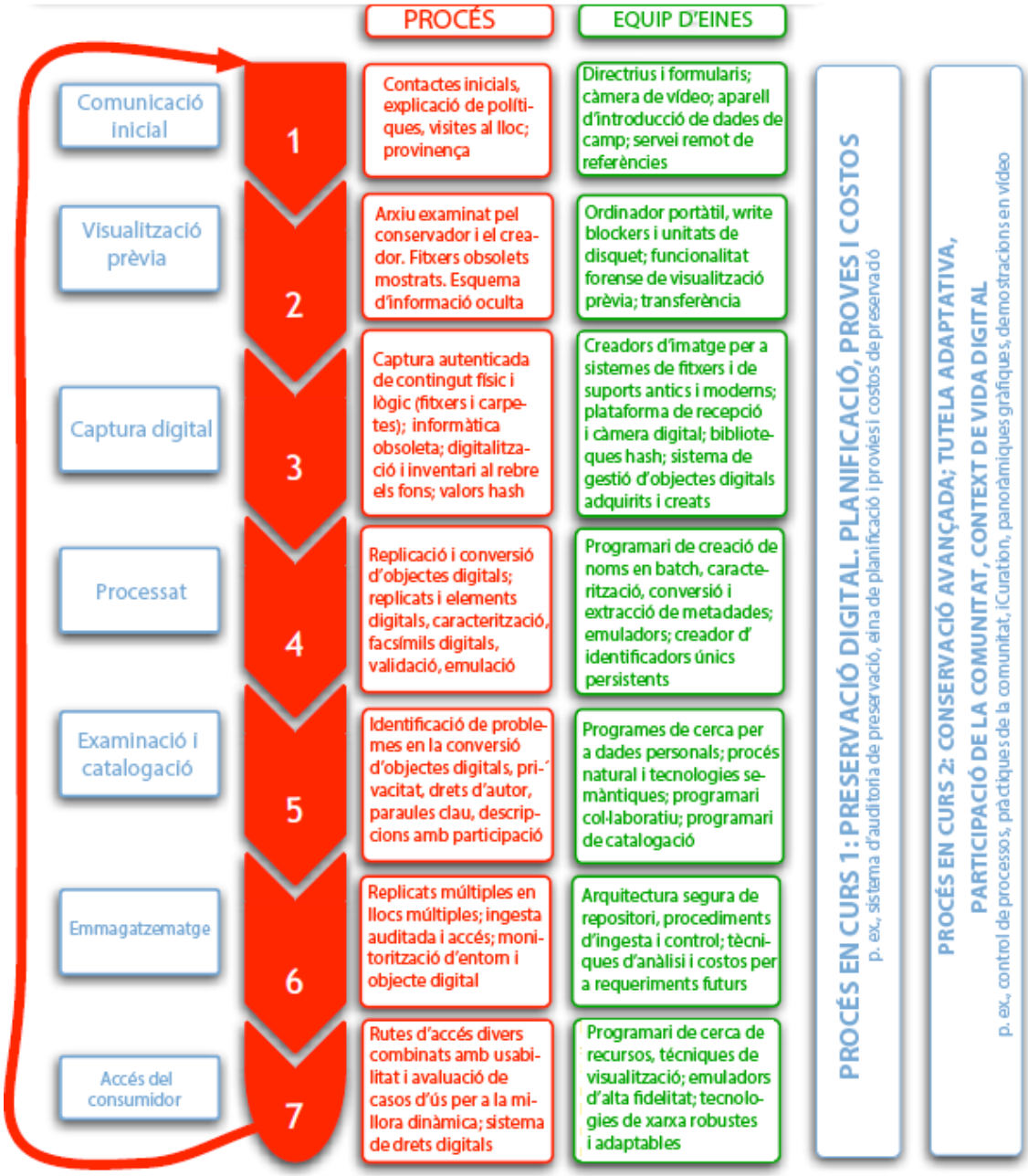
Per tal d'entendre millor els arxius personals digitals i el seu potencial dins la recerca al segle XXI, l'any 2008 s'inicià el projecte Digital Lives amb finançament de l'AHRC (John, 2009; Williams et al., 2008), dins el qual va tenir una implicació important el projecte Planets[262], un programa de recerca de la British Library cofinançat per la Unió Europea i desenvolupat entre els anys 2006 i 2010 amb l'objectiu de crear serveis i eines de preservació digital que assegurin un accés a llarg termini a actius digitals culturals i científics (Farquhar; Hockx-Yu, 2007).

Les conclusions preliminars de Digital Lives més interessants quant al contingut d'aquesta tesi són que tothom hauria de gestionar el seu arxiu digital personal i conservar una còpia en un altre lloc, preferiblement un repositori (John et al., 2010). La Figura 14 mostra una il·lustració de l'escenari ideal, sense ànim de servir com a flux de treball formal. Dins la Taula 18, expliquem amb més detall els processos i les eines.

[261] John, Jeremy Leighton (2006, Apr. 24-25). "Digital Manuscripts; capture & context" [presentació de Power Point]. *Digital Curation Centre Conference*. Newcastle. Recuperat del web Internet Archive. <https://goo.gl/gUFNHK>. [Consulta: 23/07/2016]

[262] <http://www.planets-project.eu/>. [Consulta: 16/07/2016]

Figura 14. Processos d'arxiu digital i eines necessàries dins el projecte Digital Lives



Font: John et al., 2010, p. 189. Traducció de l'autor

Taula 18. Explicació de processos i eines al projecte Digital Lives

Processos	Eines
Pas 1: Comunicació inicial	
L'equip de conservació inicia el contacte o bé rep una petició de recerca	Pàgina web amb detalls de contacte, juntament amb desenvolupament de la col·lecció i polítiques d'assessorament
Pas 2: Visualització prèvia	
Examen de suports moderns i retirats, així com d'obsolets	S'utilitzen ordinadors portàtils amb *write blockers* i amb lectors apropiats per a suports obsolets
Establir quins són els requeriments dels creadors originals i confirmar permisos	Formularis, acords de dipòsit i llicències
Transferència segura de suports i maquinari al repositori	Transport, amb empaquetat de preservació i conservació o en determinades circumstàncies, amb servei de missatgeria de confiança
Pas 3: Captura digital	
Adquisició mitjançant tècniques forenses amb *write blockers* i creació de valors *hash*. Comprovació de presència de virus	Per a fitxers d'imatge oberts, hi ha opcions de formats com Forensic Toolkit Imager i Advanced Forensic Format. Per a l'obtenció de valors *hash*, s'ha de repetir el procés amb un segon equip de programari i de *write blockers* per comprovar que s'obtenen els mateixos valors
Pas 4: Processat	
Exportació de fitxers replicats del fitxer imatge autenticada	Programari forense i altres sistemes amb eines de creació d'imatges
Descriure i identificar formats de fitxer	El programari DROID permet identificar formats amb el sistema Pronom
Extracció de metadades	Serveis com Pronom i programari forense; JHOVE, DROID i Planets
Creació de facsímils digitals; conversió de formats en una forma interoperable; creació de noms de fitxer en *batch*	El projecte Planets permetrà fer la conversió amb Interoperability Framework; Irfanview es pot utilitzar per a noms de fitxer en *batch*
Validació de facsímils digitals; conversió i sostenibilitat en el temps	Mitjançant Planets i JHOVE
Descripció i validació de la duplicació o la conversió	Mitjançant Planets
Accés directe al fitxer mitjançant el fitxer imatge	Edició del fitxer imatge o encriptació selectiva compatible amb sistemes d'emulació
Pas 5: Examinació i catalogació	
Examen de l'arxiu per comprovar problemes de privacitat i de drets d'autor; catalogació d'objectes digitals, compilats amb metadades	Ordinador per examinar aïllat i protegit de forma segura. Archivist's Toolkit és una eina de codi obert per catalogar
Pas 6: Emmagatzematge	
Transferir objectes i metadades digitals	Molts repositoris utilitzen el seu propi sistema o associats amb altres repositoris. L'estàndard comú per a la interoperabilitat és OAI-PMH
Pas 7: Accés	
Informació dels objectes digitals disponibles per a tothom en línia	Sistemes de codi obert com Drupal i Joomla o específics de repositori com DSpace, Fedora i Greenstone. L'emulació també permet l'accés

Processos	Eines
Procés en curs 1: Preservació digital	
Gestió sistemàtica i control amb auditories i cadenes de custòdia	Plato i DRAMBORA tenen eines per a la planificació i l'auditoria. El model de referència OAIS i les metadades PREMIS són estàndards establerts
Procés en curs 2: Conservació avançada	
Addició de valor als objectes amb informació contextual: fotografies, imatges 3D, demostracions en vídeo	Programari d'edició de vídeo; càmera digital amb programari de modelat; il·luminació per a fotografia
Participació de la comunitat per tal que aporti nova informació contextual	HubZero ofereix un sistema de codi obert a la col·laboració en línia. WikiGenes permet als col·laboracions atribuïbles a autors individuals
Tutela adaptativa degut als canvis continus que produeixen les noves tecnologies i models de negoci	Informes de control tecnològic

Font: John et al., 2010, p. 190-198

Encara que el projecte no especifica en detall alguna experiència de preservació concreta, el seu treball va ser dels primers en fer un estudi seriós de l'aplicació de la tecnologia forense dins la preservació digital adreçada al patrimoni cultural. Malauradament, no hi ha notícies del projecte a hores d'ara i tot indica que el projecte ja es va tancar, atès que el seu web ja no es troba operatiu i que la darrera versió arxivada té data d'abril de 2012[263]. Per altra banda, el projecte Planets es va tancar l'any 2010, però va deixar tres eines molt interessants: Plato[264], una eina de planificació que ajuda a prendre decisions, especialment per documentar cada pas del procés (Strodl et al., 2007); Testbed, un entorn controlat d'objectes digitals que permet executar experiments que posin a prova fluxos de treball i eines de preservació (Eitken et al., 2008) i Interoperability Framework, una plataforma basada en Java que permet als usuaris descobrir i executar eines de preservació digital com la migració[265].

[263] Segons s'indica al resultats de la recerca a l'UK Web Archive. <http://www.webarchive.org.uk/wayback/archive/20120423115222*/http://www.bl.uk/digital-lives/about.html/>. [Consulta: 23/07/2016]

[264] <http://www.ifs.tuwien.ac.at/dp/plato/intro/>. [Consulta: 16/07/2016]

[265] *Interoperability Framework: the glue that holds Planets together*. <http://www.planets-project.eu/docs/newsletters/Planetarium8_December09.pdf/>. [Consulta: 16/07/2016]

4.2.2 Bodleian Library

La University of Oxford compta amb una xarxa de biblioteques universitàries per als seus usuaris, les Bodleian Libraries, entre les quals es troba la biblioteca de recerca principal de la universitat, la Bodleian Library. És una de les sis biblioteques responsables de rebre el dipòsit legal al Regne Unit, juntament amb la British Library, la Cambridge University Library, la National Library of Scotland, la National Library of Wales i la Library of Trinity College a Dublín[266].

De la mateixa manera que a la British Library, la Bodleian va començar a plantejar-se el repte dels canvis tècnics que presenten els arxius personals digitals, amb la diferència que es van considerar com un tot amb el nom "arxiu híbrid", el qual conté tant documents en suport paper com en suport digital (Thomas, 2011). El primer pas en aquest aspecte començà l'any 2005 amb el projecte Paradigm[267], el qual va explorar els problemes culturals, legals i tècnics involucrats en la preservació a llarg termini dels arxius digitals (Thomas; Martin, 2006). El projecte, finançat sota el programa del JISC Supporting Institutional Digital Preservation and Asset Management[268], finalitzà l'any 2007 i els seus resultats finals es van sintetitzar en un quadern de treball, el *Workbook on digital private papers*[269], on es tracta l'adquisició, processat i accés de documents personals en format digital en una varietat de formats. Però, el *Workbook* no contemplava les tècniques d'anàlisi forense digital en profunditat.

El següent projecte de la Bodleian, futureArch[270], es va posar en marxa en 2008 amb finançament de l'Andrew W. Mellon Foundation en aquest cas (Thomas, 2011) i finalitzà l'any 2012. El seu objectiu era el de crear fonaments sòlids per al futur repositori BEAM on es conservarien manuscrits digitals a la biblioteca. Aquests fonaments van incloure ajustos als següents activitats i serveis de preservació de la biblioteca:

---

[266] *Agency for the Legal Deposit Libraries*. <http://www.legaldeposit.org.uk/index.html/>. [Consulta: 16/07/2016]

[267] <http://www.paradigm.ac.uk/>. [Consulta: 17/07/2016]

[268] *Digital preservation and asset management*. Recuperat del web Internet Archive. <https://goo.gl/M2Mzl8>. [Consulta: 17/07/2016]

[269] <http://www.paradigm.ac.uk/workbook/index.html>. [Consulta: 17/07/2016]

[270] <http://www.bodleian.ox.ac.uk/beam/about/projects/futurearch-project>. [Consulta: 17/07/2016]

- Política. El projecte va permetre l'inici dels canvis d'organització, estratègics, estructurals i financers necessaris per establir serveis com el repositori BEAM
- Canvi cultural. Es van desenvolupar polítiques, processos, sistemes i formació dissenyats per ajudar als conservadors a aplicar les seves tècniques a les col·leccions híbrides
- Infraestructura i sistemes. El projecte desenvolupà processos de captura, preservació i difusió per a molts tipus de materials nascuts digitals dins el nou repositori BEAM
- Col·leccions digitals. El projecte recuperà registres ja existents de col·leccions digitals i els agrupà en un entorn gestionat que va permetre l'accés als investigadors. També va donar suport a la biblioteca quant l'adquisició i cura de nous arxius híbrids
- Serveis a l'usuari. El projecte desenvolupà un sistema prototip per tal que l'usuari tingui accés als arxius digitals

Dins aquest projecte es van utilitzar tècniques d'anàlisi forense digital, degut a la seva capacitat de mantenir l'autenticitat de les dades originals i que permeten gestionar grans quantitats de dades heterogènies. La Taula 19 descriu alguns passos del flux de treball que s'aplica habitualment als arxius híbrids.

Un dels casos més interessants dins els processos de captura fou la recuperació d'una part dels materials nascuts digitals de l'arxiu de Barbara Castle, consistents en dades de text escrites originalment en el programari LocoScript de Locomative Software dissenyat per a la sèrie d'ordinadors personals Amstrad PCW, la qual va ser distribuïda entre els anys 1985 i 1998[271]. Els primers models utilitzaven com a mitjà de emmagatzematge disquets de 3 polzades, un suport molt poc comú i que va plantejar un gran problema, ja que els documents de text es trobaven guardats en tres disquets d'aquest tipus.

La solució va arribar després de consultar abastament la comunitat d'usuaris d'Amstrad i localitzar un Amstrad PCW en funcionament, programari específic i manuals. Es connectà un cable de comunicacions entre l'Amstrad i un ordinador portàtil amb sistema

[271] <https://en.wikipedia.org/wiki/Amstrad_PCW>. [Consulta: 17/07/2016]

operatiu SUSE Linux en el qual s'instal·là una màquina virtual amb el sistema operatiu Windows 95, el qual executà el programari de migració que convertí el text en format Locomotive a format ASCII o RTF, de forma que els documents ja es poden consultar en sistemes moderns (Thomas, 2011).

Taula 19. Processos i eines al repositori BEAM (Bodleian Library)

Processos	Eines
Separació	
Es separen els suports digitals de la documentació en paper; incorporació al repositori	Repositori BEAM; base de dades de col·leccions
Captura	
Assignació d'un número d'inventari a cada ítem amb fotografies	Càmera digital
Creació d'imatges de disc amb metadades sobre el procés	*Write blockers*, estacions de treball forense FRED, programari forense FTK Imager
Verificació de la imatge de disc	Programari de verificació de *hash*
Empaquetat del material i ingesta amb metadades bàsiques de col·lecció i accés	Repositori de preservació
Treball de camp	
Creació selectiva d'imatges amb control i supervisió del donant	*Write blocker* extern, programari de creació d'imatges
Anàlisi	
Visió en conjunt dels formats presents a l'arxiu, identificació de problemes, preparació de l'arxiu per al catalogador	Programari forense FTK

Font: Kirschenbaum et al., 2010, p. 36-37

Actualment, el repositori BEAM[272] ja s'utilitza com el servei de confiança de la biblioteca per a materials d'arxiu digitals. Entre les col·leccions en què estan treballant, es troba l'arxiu del Partit Conservador del Regne Unit, l'arxiu literari de la impremta Clutag, l'arxiu de la política del Partit Laborista Barbara Castle o l'arxiu de la política del Partit Liberal Demòcrata Emma Nicholson. Encara que l'accés a BEAM no és obert a hores d'ara al públic general, s'estan fent treballs en aquest aspecte[273]. Un altre projecte relacionat amb la preservació en què estan treballant en col·laboració amb les

[272] <http://www.bodleian.ox.ac.uk/beam>. [Consulta: 17/07/2016]

[273] Cliff, Peter; Gittens, Renhart (2009, Apr. 23). *Same goal, new challenge: an introduction to the futureArch project & BEAM*. Digital Repositories Workshop: Tools and Infrastructure. University of Oxford. <http://goo.gl/aSeGKQ>. [Consulta: 23/07/2016]

Stanford University Libraries és el projecte ePADD[274], el qual consisteix en un paquet de codi obert que dóna suport als processos d'arxiu dins els fitxers de correu electrònic

4.2.3 National Library of Australia

A l'igual que la British Library, la NLA va haver d'estudiar com preservar els materials nascuts digitals que ingressaven a les seves col·leccions. Fruit d'això es van desenvolupar processos com fer una catalogació més acurada, donar prioritat a materials amb risc de pèrdues (com els disquets) i establir un procediment per valorar, documentar i copiar materials nascuts digitals. Aquests processos van durar entre set i vuit anys. Però, es van adonar que els seus esforços consumien massa temps ja que no havia un flux de treball concret per als materials digitals, dels quals es feia la preservació cas per cas sense cap tipus d'automatització. Per tant, l'any 2004 es van deixar de banda, però el problema de la preservació de materials amb risc encara estava pendent de solució (Elford et al., 2008).

La resposta fou el desenvolupament intern a la NLA del Digital Preservation Workflow Project, el qual tenia com a objectiu la creació d'un sistema de flux de treball apropiat per a la preservació, emmagatzematge i accés a llarg termini dels objectes digitals adquirits a la biblioteca, ja sigui mitjançant suports físics o mitjançant fitxers discrets. Els processos concrets que havia d'assolir el projecte foren:

- Capturar dades dels suports físics o d'altres fonts digitals de forma acurada
- Generar i capturar metadades relacionades
- Analitzar, processar i validar els objectes digitals
- Enviar les dades i metadades capturades al sistema d'emmagatzematge de la biblioteca
- Permetre accés limitat als materials
- Donar atenció especial a materials amb dificultats de captura o rèplica, amb avisos a personal especialitzat si fos necessari

[274] <http://library.stanford.edu/projects/epadd>. [Consulta: 17/07/2016]

Per tal d'assegurar l'accés al contingut digital, el projecte va produir l'aplicació Prometheus[275], la qual permet un procés escalable i semi-automatitzat per transferir dades des de suports físics a l'emmagatzematge digital de preservació (Del Pozo; Elford; Pearson, 2009b). El procés implica l'ús d'estacions de treball portàtils i personalitzables que es poden connectar a altres ordinadors del personal de la NLA. Aquestes estacions de treball tenen instal·lades unitats de lectura de CD-ROMs, DVD-ROMs i disquets de 3 ½ polzades, així com ranures USB. Un cop connectada l'estació de treball, el personal utilitza l'eina Prometheus per als processos de creació de la imatge de disc, calcular valors *hash*, identificar formats de fitxer i extreure metadades, els quals executa mitjançant diferents aplicacions de codi obert. La Figura 15 mostra el flux de treball que s'utilitza amb l'aplicació, que expliquem a continuació.

En primer lloc, la biblioteca té la seva política on es seleccionen els ítems susceptibles de ser preservats dins Prometheus. Un cop connectat a l'aplicació, l'usuari pot veure directament quins són aquests ítems, o si aquests no existeixen, s'hauran de crear. En aquest moment s'arriba al procés de catalogar i descriure els materials, que poden incloure l'addició de metadades de requeriments del sistema si es tracta d'un CD-ROM o informació de drets d'autor. Un cop acabat aquest procés, s'ha d'afegir la imatge del suport concret i replicar-la. Un cop creada la imatge de disc, es copien els continguts a l'àrea de treball.

Dins la creació d'imatges s'utilitzen dues eines. Per una banda, tenim cdrdao[276], un programari pensat per crear imatges de CD-ROMs i de Video-CDs. En el cas dels DVD-ROMs, els discs durs, memòries USB i disquets, s'utilitza dd[277]. Per crear valors *hash* de les imatges, es fa servir Jacksum[278], que dóna suport a 58 algoritmes diferents.

Si tots els processos s'han realitzat sense contratemps, els sistema munta la imatge i desempaqueta els sistemes de fitxers i els fitxers de la imatge de disc i fa un registre de les estructures de directoris. A continuació fa una anàlisi de fitxers i enregistra metadades tècniques mitjançant DROID per a la identificació de formats de fitxer i

[275] <http://prometheus-digi.sourceforge.net/>. [Consulta: 23/07/2016]

[276] <http://cdrdao.sourceforge.net/>. [Consulta: 23/07/2016]

[277] <http://www.chrysocome.net/dd>. [Consulta: 23/07/2016]

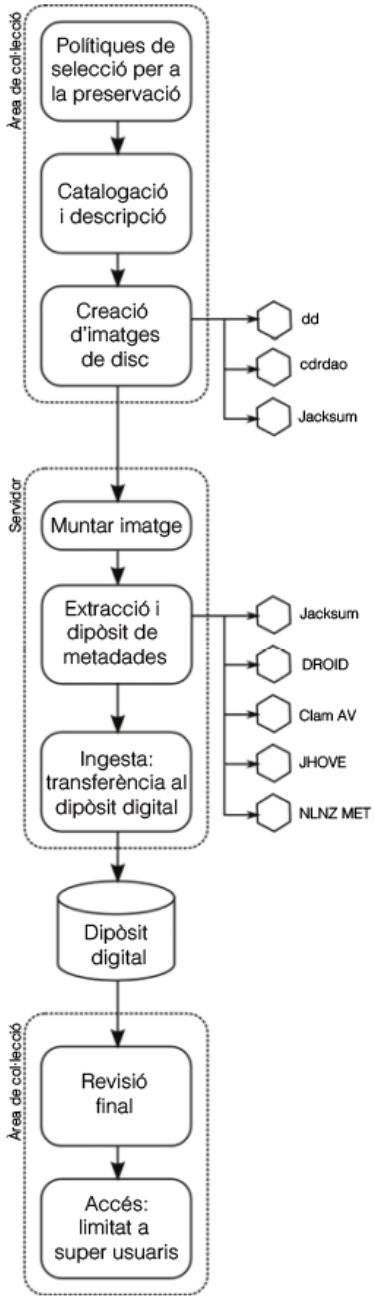
[278] <http://jacksum.net/en/index.html/>. [Consulta: 23/07/2016]

JHOVE i NLNZ MET per a l'extracció de metadades, mentre que ClamAV s'encarrega de comprovar la presència de virus. Mentre es fa aquest procés d'anàlisi, l'usuari s'encarrega d'afegir notes addicionals com incompatibilitats de la imatge amb altres sistemes i pot anar revisant la llista de fitxers capturats amb creació opcional d'informes en format PDF. Un cop hagi acabat completament el procés d'anàlisi, el treball estarà llest per a la ingesta al dipòsit de preservació. Quan el treball hagi estat transferit, aquest serà revisat exhaustivament per al seu accés final, el qual està reservat a usuaris d'un nivell alt de privilegi.

Encara que Elford et al. (2008) reconeixen que encara hi ha molt treball per fer, com la creació de directrius per refinar els fluxos de treball quant el descartat de materials digitals, la formació del personal o reconèixer materials amb dificultats de preservació, així com fer avenços en la recerca d'estratègies per reconèixer i superar l'obsolescència de formats de fitxers, Prometheus ha suposat un gran pas per automatitzar processos per rescatar materials nascuts digitals. Encara que l'aplicació sigui d'ús exclusiu per al personal de la NLA, presenta un flux de treball on no es requereix un programari propietari, ja que tot és de codi obert, i l'ús d'una estació de treball portàtil i optimitzada per a suports "fugitius" en paraules de Forstrom (2009), com els disquets i els CD-ROMs i DVD-ROMs, és una opció interessant i econòmica per a centres amb pressupostos baixos. Quant els suports fugitius, el Digital Preservation Workflow Project desenvolupà una iniciativa per tal d'identificar diversos tipus de suports i les seves dependències associades (Del Pozo; Elford; Pearson, 2009a). Aquesta iniciativa, amb la col·laboració de la IASA, la National Film & Sound Archives of Australia, el Powerhouse Museum i els Archives New Zealand, es va convertir en la Mediapedia[279], on es poden identificar ràpidament diferents suports com bobines de cinta magnètica, microformes o targetes intel·ligents. Gràcies a aquesta web de referència, l'usuari pot identificar el risc d'obsolescència d'un suport de forma directa, però també es tracta d'un servei que es pot reutilitzar dins altres sistemes per conèixer els seus requeriments d'ús i el context històric.

[279] <http://mediapedia.nla.gov.au/home.php/>. [Consulta: 24/07/2016]

Figura 15. Flux de treball amb Prometheus



Font: Del Pozo et al., 2009b. Traducció de l'autor

4.2.4 Emory University

L'any 2004 l'escriptor britànic Salman Rushdie començà a donar conferències en les Richard Ellmann Lectures in Modern Literature de l'Emory University, un cicle semestral de conferències en què un escriptor o crític rellevant ofereix tres ponències i una lectura pública. Fruit d'aquesta relació, l'any 2006 la universitat va adquirir els documents de Rushdie i els va integrar dins la MARBL, secció component de les Emory

University Libraries. Llavors es va presentar un repte important, ja que l'adquisició va incloure, a més de material d'arxiu tradicional com diaris, correspondència i manuscrits, maquinari entre el qual havien ordinadors complets. Encara que el centre ja havia rebut altres col·leccions amb disquets, CDs i DVDs, mai havia rebut ordinadors. Específicament, es van rebre tres ordinadors portàtils i un de sobretaula que pertanyien a la línia Macintosh, un disc dur portàtil Firewire i diversos discos els quals contenien en la seva major part fitxers de programari. Carroll et al. (2011) van fer un estudi molt detallat del procés i el presentem a continuació.

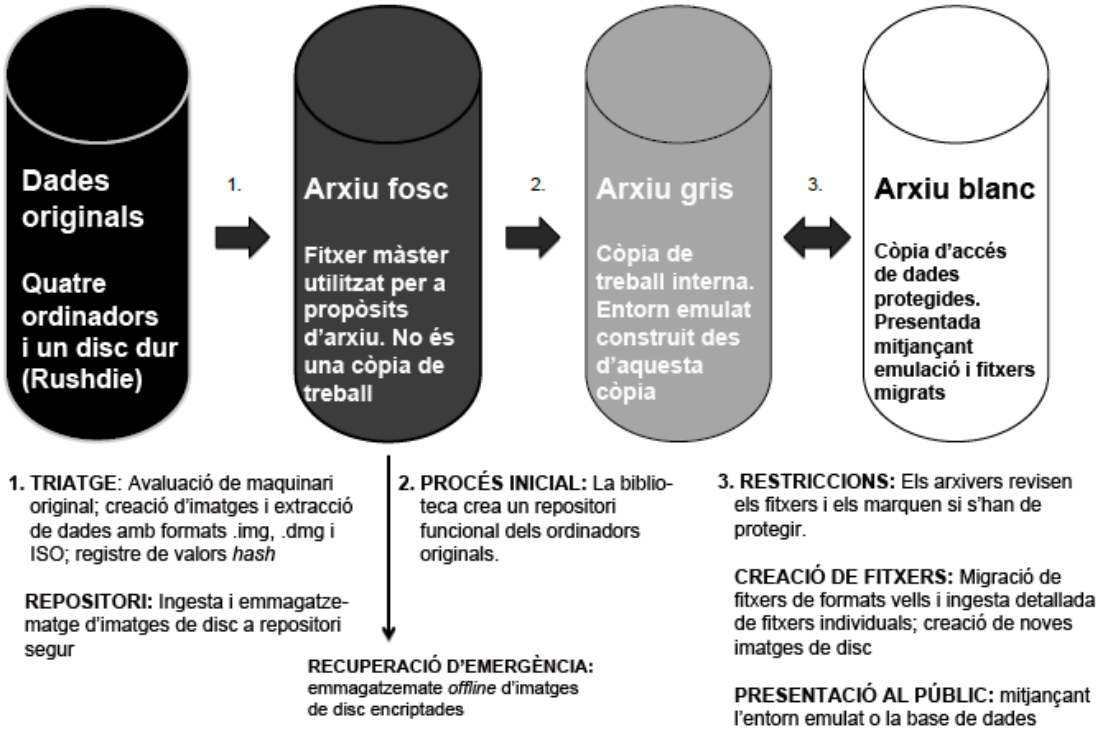
El repte que representava preservar el material nascut digital i fer-lo accessible a investigadors va provocar la creació d'un grup de treball, el BoDAR, que va decidir fer un tractament d'arxiu híbrid de col·lecció, buscar un equilibri entre les necessitats del donant i els de l'investigador i oferir una experiència de recerca genuïna. Un dels passos importants dins el procés del material nascut digital fou la privacitat; després de diverses negociacions amb Rushdie, es va acordar que tota la documentació relacionada amb la família de Rushdie es devia bloquejar fins la mort d'un parent específic o bé setanta anys després de la data d'adquisició. També es va especificar que tots els diaris escrits a partir de 1989 es devien bloquejar degut a què Rushdie estava començant a escriure una autobiografia després de la fàtua que es va proclamar aquell any. Finalment, es va decidir bloquejar tota la correspondència amb una petita selecció oberta a investigadors.

El següent pas consistí en preparar els materials nascuts digitals per al procés d'arxiu. Un cop s'inventarià el maquinari i es calculà l'espai en bytes necessari, que fou un total de 18 GB (Loftus, 2010), es va passar a replicar les dades. La Figura 16 mostra el flux de treball que es va utilitzar en aquest cas particular, on el primer cilindre representa les dades i el maquinari originals, els quals queden sense alterar. En el segon cilindre es troba l'arxiu "fosc", on es troba la còpia de les dades originals. Aquest "màster" només es troba accessible a personal del grup de treball i no es treballa sobre ell directament, sinó que s'utilitza per extreure còpies cap a l'arxiu "gris", que representa la còpia de treball amb el qual el personal pot revisar el material i restringir-ne l'accés si és necessari. Finalment, l'arxiu "blanc" representa el fitxers ja processats i disponibles als investigadors. Això inclou fitxers migrats a formats d'ús comú com el PDF.

El procés utilitzat per recuperar les dades dins el pas 1 fou la creació d'imatges de cada disc dur present, un total de cinc. A continuació es van extreure els fitxers de cada imatge i es van calcular valors *hash* de cada fitxer, a més de generar metadades sobre els tipus de fitxer en un full de càlcul. El nombre final de fitxers fou de 11.350 amb un total de 12.205 MB.

Arribats a aquest punt, l'equip de treball es plantejà com donar accés als materials nascuts digitals, amb el compromís de preparar dues interfícies: una base de dades per poder fer cerques i un entorn emulat. Amb aquesta estratègia en ment, s'inicià el pas 2 amb l'ús d'un repositori intern per processar les còpies de treball. Gràcies a l'ús dels valors *hash*, es van detectar i filtrar els fitxers duplicats i es va crear un màster de fitxers únics. A continuació, es van avaluar tots els tipus de fitxer i la seva nomenclatura i es van classificar per categories com correspondència, fotografies, diaris, etc. Aquesta informació s'incorporà al full de càlcul de metadades.

Figura 16. Flux de treball de preservació de l'arxiu de Salman Rushdie



Font: Carroll et al., 2011. Traducció de l'autor

Un cop completades les còpies de treball, començà el pas 3, amb la primera prova d'execució en un entorn emulat. Aquesta prova es va realitzar mitjançant el programari

de codi obert SheepSaver[280], el qual permet emular l'entorn Macintosh a sistemes operatius Linux. La prova consistí en, per una banda, avaluar si els continguts de cada fitxer efectivament corresponien a les valoracions fetes al pas 2 i, per altra banda, determinar si un fitxer contenia informació restringida o bé si requeria redactar-la. Finalment, es va assignar un "veredicte" a cada fitxer:

- Tal qual: el fitxer pot ser alliberat sense cap canvi o alteració dins l'entorn emulat i dins la base de dades
- Redactat: el fitxer necessita una redacció per al seu accés; no es trobarà disponible per a l'emulació, però sí dins la base de dades
- Restringit: el fitxer estarà restringit i no estarà disponible en cap entorn
- Només virtual: els fitxers només es trobaran a l'entorn emulat i no a la base de dades

A continuació es van fer dues operacions més. La primera fou la migració de documents de text a PDF per tal d'incorporar-los a la base de dades. Això va incloure formats complicats, com missatges de correu electrònic del programari Eudora i faxos. En aquest últim cas, va ser necessari instal·lar una versió de l'aplicació de faxos, exportar els fitxers com a TIFFs individuals i fer una conversió massiva de TIFFs a PDF. En el cas dels correus electrònics, s'utilitzà un programari basat en Python per exportar els missatges a xml CERP[281]. La segona operació fou crear noves imatges de disc amb l'esborrat de tots els fitxers que haguessin estat creats abans del seu ús per part d'en Rushdie i també d'aquells fitxers que va generar Rushdie com a usuari del sistema operatiu. Després es van carregar tots aquells fitxers marcats com "tal qual" i "només virtual" dins l'entorn; per tal d'assegurar l'autenticitat dels fitxers es van verificar tots els valors *hash*.

El 26 de febrer de 2010 va ser la data en què es va permetre l'accés a l'entorn emulat i a la base de dades dins la sala de lectura del MARBL[282]. Els usuaris poden explorar

[280] <http://sheepshaver.cebix.net/>. [Consulta: 23/07/2016]

[281] Projecte per preservar correus electrònics històrics. Més informació a <http://siarchives.si.edu/cerp/>. [Consulta: 23/07/2016]

[282] *Salman Rushdie Archive to open Feb. 26*. <http://shared.web.emory.edu/emory/news/releases/2010/01/salman-rushdie-archive-to-open-at-emory.html>. [Consulta: 23/07/2016]

l'estructura de directoris original, poden modificar fitxers, esborrar-los i fins i tot poden jugar als jocs preinstal·lats. Encara que dins l'entorn sembla que es guardin els canvis, la imatge refresca i reinicia la imatge original cada cop que torna a arrencar l'entorn, sense que es guardin els canvis ni les modificacions. Quant a la base de dades, es poden fer recerques per paraules clau, navegar per sèries o subsèries i navegar per carpetes i directoris. Aquest accés dual es va fer pensant en les necessitats dels investigadors. La base de dades permet la recerca d'informació (amb el requeriment de migrar dades) i l'entorn emulat permet donar a conèixer el context en què va treballar Rushdie i com va impactar el mitjà tecnològic dins la seva producció literària.

Gràcies a l'experiència adquirida, el MARBL ja té establert un laboratori per a materials nascuts digitals, amb procediments per a creació d'imatges de disc, revisió de material, filtrat de fitxers duplicats, creació de valors *hash* i avaluació general de col·leccions digitals. Els investigadors ja poden fer consultes de materials com els arxius de l'escriptora Alice Walker, la poetessa Lucille Clifton, el poeta i dramaturg Turner Cassity, l'artista Mildred Thompson, el poeta i crític literari Nathaniel Mackey i l'entrevistador Matt Schaffer[283]. Segons s'indica a les notes de processament de les col·leccions, ja s'utilitza programari forense com FTK Imager per a la creació d'imatges i l'entorn BitCurator per detectar informació sensible i fitxers duplicats.

4.2.5 Projecte AIMS

El projecte AIMS fou una associació entre quatre universitats, la University of Hull, la Stanford University, la University of Virginia i la Yale University, amb finançament de l'Andrew W. Mellon Foundation, que tenia l'objectiu de construir un marc per preservar i gestionar material nascut digital. La tasca principal del projecte fou la de fer recomanacions de millors pràctiques mitjançant l'ús d'eines i fluxos de treball en diferents escenaris, però no va ser possible degut a la diversitat d'eines i de factors institucionals com la tecnologia i el personal disponibles. Així doncs, l'esquema AIMS fou desenvolupat per definir bones pràctiques quant a tasques d'arxiu per a la

[283] Rose Library - Emory University (2016). *Digital archives access.* <http://rose.library.emory.edu/using/reading-room/digital-archives-access.html>. [Consulta: 23/07/2016]

preservació (AIMS Work Group, 2012, p. II), les quals estan dividides en quatre funcions principals:

- Desenvolupament de la col·lecció. Accions i polítiques d'una institució per adquirir materials
- Adquisició. Funció nuclear d'arxius on una institució arxivística assumeix la custòdia d'un grup de documents d'un donant i documenta la seva transferència en un registre
- Organització i descripció. Processos que realitza una institució per establir control intel·lectual del material un cop s'ha assegurat el control físic mitjançant l'adquisició
- Recerca i accés. Sistemes i fluxos de treball que fan que el material i les seves metadades associades estiguin disponibles als usuaris

El projecte començà en octubre de 2009 i finalitzà en octubre de 2012, del qual es pot consultar tota la seva documentació en el blog Born Digital Archives[284]. Els resultats més interessants inclouen la preservació de tretze col·leccions de figures literàries i un programari de codi obert, el repositori Hydra[285], que permet difondre i preservar aquestes col·leccions. Entre els anys 2011 i 2015, parts d'aquestes col·leccions es van trobar accessibles públicament mitjançant Hypatia, una demostració d'aplicació de programari Hydra, però a hores d'ara, els continguts de aquestes col·leccions es guarden a les seves institucions respectives.

Seguidament s'exposen els casos particulars de cadascuna de les universitats implicades en el projecte, amb una especial atenció als fluxos de treball que es van utilitzar.

University of Hull

La University of Hull és una universitat situada al comtat anglès de Yorkshire i es fundà l'any 1927[286]. Les seves col·leccions de manuscrits i arxius es guarden dins els

[284] <http://born-digital-archives.blogspot.com.es/>. [Consulta: 07/08/2016]

[285] <https://projecthydra.org/>. [Consulta: 07/08/2016]

[286] *About us*. <http://beta.www.hull.ac.uk/Choose-Hull/University-and-region/About-us/About-us.aspx>. [Consulta: 07/08/2016]

University of Hull Archives[287], amb més de 750.000 documents. Abans de la seva entrada dins el projecte AIMS, els Archives no tenien cap procediment per arxivar i processar materials nascuts digitals. En una primera fase, es decidí fer un inventari de suports digitals mitjançant el seu sistema de gestió de col·leccions d'arxiu, on es descobrí un disquet de 5 ¼ polzades. En aquest cas es va encarregar la recuperació del contingut a una empresa externa i es van desenvolupar procediments per extreure el suport dels documents en paper i guardar-lo en un entorn adequat, a més de fotografiar-lo per tal de capturar la informació de l'etiqueta original del disquet.

En una segona fase, els Archives van decidir gestionar directament els suports fugitius i amb aquest objectiu van configurar una estació de treball que comptava amb ports USB i també amb unitats de lectura de disquets i de CD-ROMs. Posteriorment es va afegir una unitat de lectura de discs Zip per tal de poder accedir a la major part possible de suports. Pel que respecta al programari, s'instal·laren FTK Imager, DROID i Karen's Directory Printer. Per tal de protegir la integritat de les dades, l'estació de treball operava de forma autònoma i no estava connectada a la xarxa, així que fou necessari descarregar el programari en una altre ordinador, guardar-lo en una memòria USB i instal·lar-lo posteriorment mitjançant aquesta memòria. Aquesta estació de treball va rebre el nom de HAROLD i el centre ja ha encarregat un nou maquinari per gestionar grans volums de material[288], ja que, en data de març de 2012, el centre ja custodiava set col·leccions nascudes digitals, amb un total de 46.576 fitxers, que sumaven més de 100 GB[289].

El primer cas on es preservà una col·lecció sencera de materials nascuts digitals fou el dels arxius del guionista i escriptor Stephen Gallagher, que l'any 2010 va donar a la universitat un disc dur extern amb 14.320 fitxers amb una mida de 13,6 GB i 39 discs Amstrad amb 300 fitxers. Entre els anys 2012 i 2013 envià material addicional, que afegit a l'anterior ja sumava 39,2 GB[290]. Els passos concrets per preservar aquesta

[287] <http://www.hull.ac.uk/arc/index.html/>. [Consulta: 07/08/2016]

[288] Hull History Centre. *Work in progress.* <https://goo.gl/ughozl>. [Consulta: 07/08/2016]

[289] Wilson, Samuel (2012, Mar. 6). *Born-Digital archives at Hull: early steps & early lessons* [presentació de Power Point]. <http://hullhistorycentre.org.uk/discover/pdf/Archives%20and%20Society.pdf>. [Consulta: 07/08/2016]

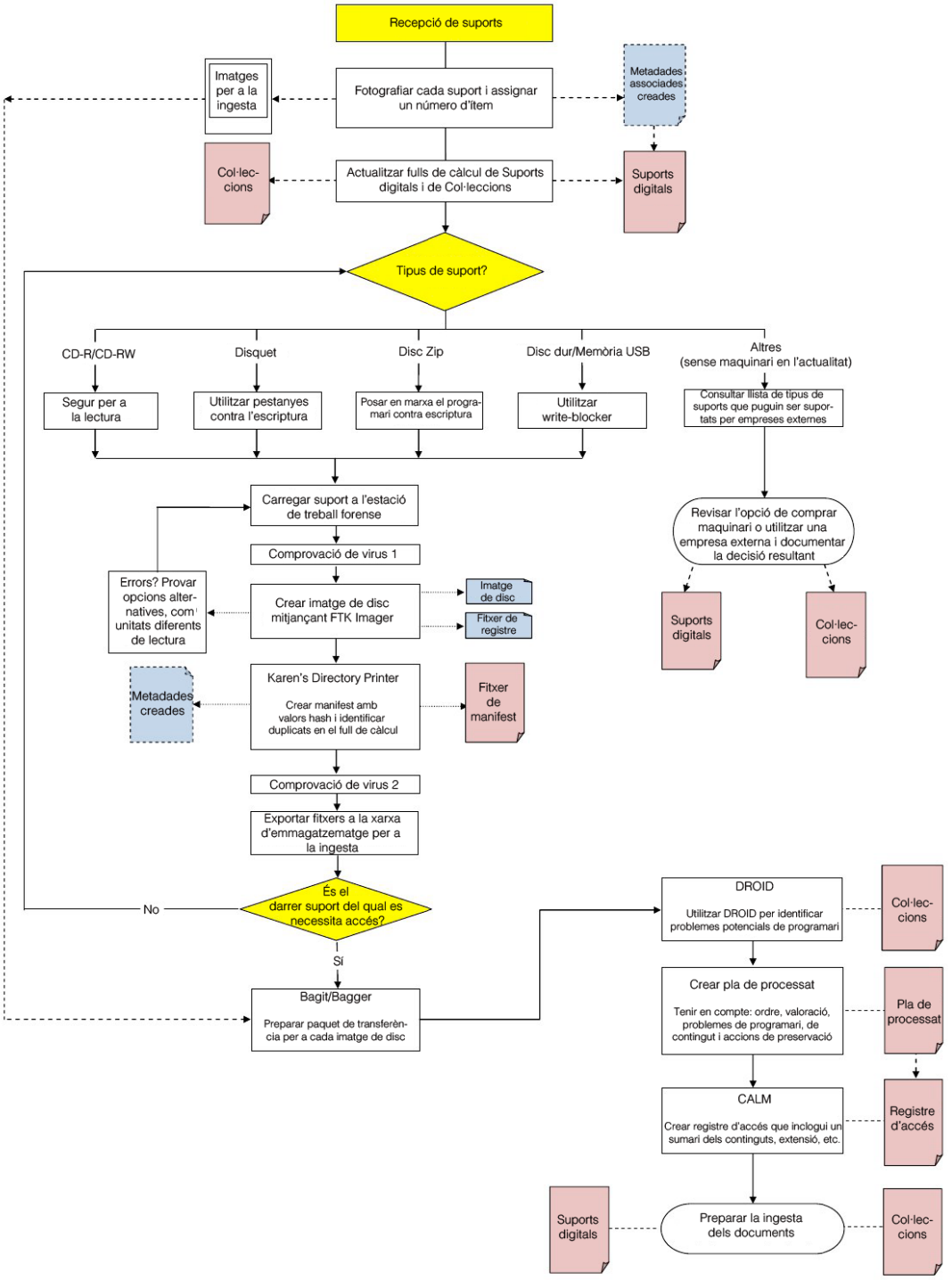
[290] Wilson, Stephen (2013, Nov. 15). *Stephen Gallagher: lessons learnt (so far) from a hybrid literary collection* [presentació de Power Point]. <http://glam-archives.org.uk/wp-content/uploads/2013/12/Simon-Wilson.pdf/>. [Consulta: 07/08/2016]

col·lecció van originar nous procediments a la institució amb la creació de dos fluxos de treball, un per a la separació de suports digitals dins els arxius en paper i un altre per a l'accés de materials nascuts digitals. El segon és el que més s'adequa als objectius d'aquesta tesi i per tant el mostrem a la Figura 17.

El primer que es fa quan es reben els materials es fer una fotografia de l'ítem (com ja s'ha indicat abans) i documentar el màxim possible la seva etiqueta original mitjançant una plantilla transparent feta a mida amb disseny de claqueta. Aquesta informació s'emmagatzema en forma de metadades i tant la foto com les metadades es guarden al repositori intern de la universitat. Seguidament s'editen dos fulls de càlcul; un és el de Col·leccions, on es controla el contingut dels ítems amb detalls com el seu nom, el valor *hash*, la seva mida, etc.; i l'altra és el de Suports digitals, on es controlen els tipus de suports presents a les col·leccions. En funció del tipus d'ítem, s'hauran de posar en marxa diferents tipus de maquinari per bloquejar-lo contra l'escriptura. Si es tracta d'un suport amb format no contemplat per l'entitat, es plantejarà utilitzar una empresa externa per accedir al contingut.

Un cop bloquejat el suport, s'utilitza l'estació HAROLD per fer-ne un escanejat antivirus, es crea una imatge lògica de disc, el qual contindrà només els fitxers que els dipositants han acordat amb l'entitat, i el seu fitxer de registre corresponent mitjançant FTK Imager. Seguidament, Karen's Directory Printer genera un manifest del contingut del suport, del qual es poden extreure metadades de forma separada. Aquest manifest inclou data de creació, nom del fitxer, valors *hash*, etc. A continuació es fa un nou escanejat antivirus i s'exporten els fitxers desitjats a l'espai d'emmagatzematge intern. Per fer-ho, la University of Hull utilitza el format BagIt com a paquet que integra tant la imatge de disc com les metadades. Mitjançant DROID, s'identifiquen acuradament els formats de fitxer presents a la imatge, els quals s'anoten al full de càlcul Col·leccions, es crea el document Pla de processat, el qual reuneix la informació recollida als fulls de càlcul i aporta documentació sobre com catalogar el material, la valoració global de l'ítem o ítems, accions necessàries per a la preservació, etc. Finalment, s'ha de crear un registre al programari de gestió de col·leccions d'arxiu CALM amb descripcions acurades de la col·lecció i ingestar els fulls de càlculs amb els nous canvis.

Figura 17. Flux de treball d'accés a materials nascuts digitals a la University of Hull



Font: AIMS Work Group, 2012, p. 109. Traducció de l'autor

Algunes de les particularitats que es van presentar al tractament de la col·lecció de Stephen Gallagher i que s'anotaren al Pla de processat fou la impossibilitat de recuperar el contingut dels disquets Amstrad de 3 polzades degut a la manca de maquinari necessari, per una banda, i els passos per migrar fitxers del format FinalDraft a PDF, per

l'altra. A més, es va acordar el bloqueig d'informació sensible com una carpeta de pàgines web i material personal de l'autor.

Actualment, la University of Hull continua gestionant materials digitals, però encara no es permet l'accés a les seves col·leccions. S'estan fent passos en aquesta direcció i el centre espera poder oferir a curt termini l'accés restringit a part del seu catàleg dins les seves instal·lacions i també fer ús del repositori Hydra per crear una interfície web d'accés.

Stanford University

El primer cas en què la Stanford University, universitat nord-americana situada a Palo Alto, Califòrnia, va realitzar tasques per capturar i processar materials nascuts digitals d'un arxiu híbrid es va presentar entre els anys 2011 i 2012, durant els quals el seu Department of Special Collections and University Archives, secció que pertany a la xarxa d'universitats SUL (Stanford University Libraries), va realitzar aquestes tasques amb els arxius híbrids de l'organització no governamental Stop AIDS que va rebre l'any 2005. El projecte, finançat pel NHPRC, va permetre crear nous procediments de treball per generar imatges forenses de disquets, discs Zip i CDs, suports que estaven presents a l'arxiu híbrid (Wilsey et al., 2013).

Anteriorment a l'inici del projecte, la universitat ja havia fet passos previs de gran importància per processar aquests tipus de materials, atès que ja s'havia incorporat al projecte AIMS i havia posat en marxa el Born-Digital Program[291] per preservar i donar accés a materials amb risc d'obsolescència i tenia preparada una estació de treball forense per poder treballar amb els materials. Encara que la Stanford University ha presentat més casos de preservació d'arxius híbrids, com els del paleontòleg Stephen Jay Gould, el poeta Robert Creeley i l'impressor Peter Koch, Wilsey et al. (2013) van descriure amb detall el flux de treball que es va utilitzar per ingestar els materials de Stop AIDS al seu repositori i els descrivim a continuació.

El primer pas tracta dels preparatius previs, en què es van fer reunions amb el donant per analitzar el contingut digital de la col·lecció, el qual comptava amb 210 suports

[291] <https://library.stanford.edu/spc/more-about-us/born-digital-program>. [Consulta: 14/08/2016]

físics a l'inici del projecte de preservació. Un dels passos més importants fou la identificació de la informació sensible, molt important en aquest cas, ja que l'arxiu contenia informació de participants dins el projecte Stop AIDS com el seu nom, informació de contacte, data de naixement i la seva orientació sexual. Gràcies a la bona infraestructura del programa, el donant va acordar la transferència addicional d'un disc dur amb 836 GB. Una altra informació valuosa que proporcionà el donant fou el programari utilitzat als suports de la col·lecció i el sistema de fitxers per formatar-los.

El segon pas consistí en l'adquisició, la qual té diverses facetes dins l'esquema AIMS. Per una banda, s'ha d'assegurar que hi ha un control físic adequat sobre els fitxers i per una altra, que s'han fet imatges forenses dels suports amb controls antivirus. En aquest cas, primer es van separar els 210 suports físics dins les més de 300 caixes de material de la col·lecció. Es va etiquetar cada ítem amb un número identificador normalitzat i es va catalogar cada ítem dins un full de càlcul Excel per tal de localitzar-lo dins el procés d'adquisició. Dins aquest full de càlcul es va enregistrar el número identificador, el tipus de suport, el fabricant, una transcripció de les metadades presents a les etiquetes originals o la capsa, el tipus de sistema de fitxers, si s'havia comprovat la presència de virus, si s'havia realitzat la imatge forense, les dades de creació i/o modificació dels fitxers, la data de creació de la imatge forense, el personal que va crear la imatge i el tipus d'ordinador que es va utilitzar per a la captura. A continuació, es va fer una fotografia de cada ítem per tal d'enregistrar la informació original present a la seva etiqueta o capsa. Per comprovar la presència de virus, s'utilitzà el programari gratuït Sophos[292] i es crearen les imatges forenses amb FTK Imager. El maquinari involucrat fou una estació forense FRED, un ordinador portàtil Macintosh, una unitat externa forense de disquets de 3 ½ polzades del fabricant Digital Intelligence, una unitat externa de disquets de 3 ½ polzades del fabricant Fujitsu i una unitat externa de discs Zip del fabricant iOmega.

Quant al procés de creació d'imatges forenses, es van presentar dificultats serioses per capturar els CDs, ja que d'un total de 218, només es van poder capturar 18. Les possibles raons serien la baixa qualitat dels discs i el maquinari utilitzat, ja que els continguts no es van gravar de forma professional sinó als equips personals dels

[292] <https://www.sophos.com/en-us.aspx>. [Consulta: 14/08/2016]

membres de l'ONG. A diferència dels disquets, no es van presentar problemes d'incompatibilitats de sistemes de fitxers, atès que tots els CDs utilitzaven el mateix sistema de format acordat per la norma ISO 9660. Quant el disc dur, es va crear una imatge lògica degut a què els fitxers ja havien estat seleccionats pel donant. Aquest procés automatitzat de creació d'imatge va requerir unes 45 hores; 24 hores per a la creació i 21 hores més per a la seva verificació, mentre que el procés manual de configuració només va requerir cinc minuts. Els resultats finals de l'adquisició (sense comptar el disc dur afegit *a posteriori*) foren un total de 22 imatges de discs Zip, 109 de disquets i 18 de CDs que van sumar 29.423 fitxers únics. Les imatges de disc, juntament amb les fotografies dels suports, van ser ingestats al Stanford Digital Repository per a la seva preservació.

El tercer pas, organització i descripció, va arrencar amb la identificació d'informació sensible per poder filtrar-los i així tenir menys fitxers que processar. Mitjançant l'eina de cerca Live Search (que fa cerques en tot el contingut de les dades) del programari FTK es van poder filtrar fitxers amb números de seguretat socials i números de targetes de crèdit gràcies a l'ús de patrons i expressions regulars. Index Search (que fa cerques dins l'índex de paraules), per altra banda, va permetre filtrar contingut com nom, data de naixement o orientació sexual. Finalment, es van identificar 1.816 fitxers amb informació restringida, i la resta, 27.607 fitxers (5.925 MB) i les seves metadades tècniques es van ingestar al SDR juntament amb les imatges de disc i les fotografies dels suports.

Per tal de permetre el quart pas, la recerca i l'accés, es va decidir donar accés als investigadors mitjançant la sala de lectura de col·leccions especials dins un servidor de xarxa intern. Els fitxers s'exportarien mitjançant FTK en funció del tipus de fitxer i es guardarien a les seves carpetes corresponents dins a una carpeta de col·leccions de la unitat de xarxa. Els investigadors poden navegar dins l'estructura de carpetes i si desitgen fer recerques més avançades, poden utilitzar FTK demanant una cita prèvia amb l'arxiver digital del centre.

Com a nota negativa, s'ha de dir que la universitat encara no té un flux de treball realment definit per processar materials nascuts digitals, ja que Wilsey et al. (2013) reconeixen que necessiten refinar els seus protocols i les seves estratègies. A hores d'ara,

les SUL estan preparant un nou laboratori forense amb la futura adquisició d'una segona estació FRED portàtil que permetrà capturar material nascut digital *in situ* sense haver d'esperar la cessió del donant, mentre que l'altra estació FRED s'instal·larà al Department of Special Collections[293].

University of Virginia

La University of Virginia és una de les universitats més antigues dels EUA, ja que va ser fundada l'any 1819 per Thomas Jefferson, autor de la Declaració d'Independència. Dins la xarxa de biblioteques de la universitat, l'Albert and Shirley Small Special Collections Library[294] és el repositori principal de materials d'arxiu i manuscrits, amb col·leccions que arriben als 16 milions d'objectes, on podem trobar mapes, fotografies i llibres antics.

Gràcies a la participació de la UVA al projecte AIMS, es començà a desenvolupar un programa per gestionar els materials nascuts digitals dins les seves col·leccions, mitjançat l'elaboració d'un inventari, creació de fluxos de treball, i l'exploració d'opcions per facilitar-ne l'accés en el futur. Bradley Daigle, que va ser el Director de Serveis de Conservació Digital entre els anys 2009 i 2015[295], explicà els processos de gestió dels materials a una entrevista amb Gengenbach (2012) i els resumim a continuació.

El contingut nascut digital s'identifica un cop s'ha rebut una col·lecció, o bé s'identifica dins una col·lecció ja existent. Els suports individuals reben identificadors per tal de relacionar-los a una col·lecció específica i les metadades descriptives i tècniques es recullen en funció del tipus, format i condició dels suports, les quals s'enregistren en una base de dades. El suport físic es fotografia i es connecta a una estació forense FRED.

FRED es troba equipat amb el programari forense FTK, que permet fer diverses operacions com creació d'imatges forenses, vista prèvia de carpetes i fitxers, muntar la imatge per visualitzar el contingut, exportar fitxers i carpetes, recuperar fitxers esborrats

[293] Stanford University Libraries (2016?). *Digital Forensics Lab.* <http://library.stanford.edu/digital-forensics-stanford-university-libraries/first-draft-our-forensic-workflow>. [Consulta: 14/08/2016]

[294] <http://small.library.virginia.edu/>. [Consulta: 28/08/2016]

[295] Perfil professional disponible a LinkedIn. <https://www.linkedin.com/in/bradleydaigle>. [Consulta: 28/08/2016]

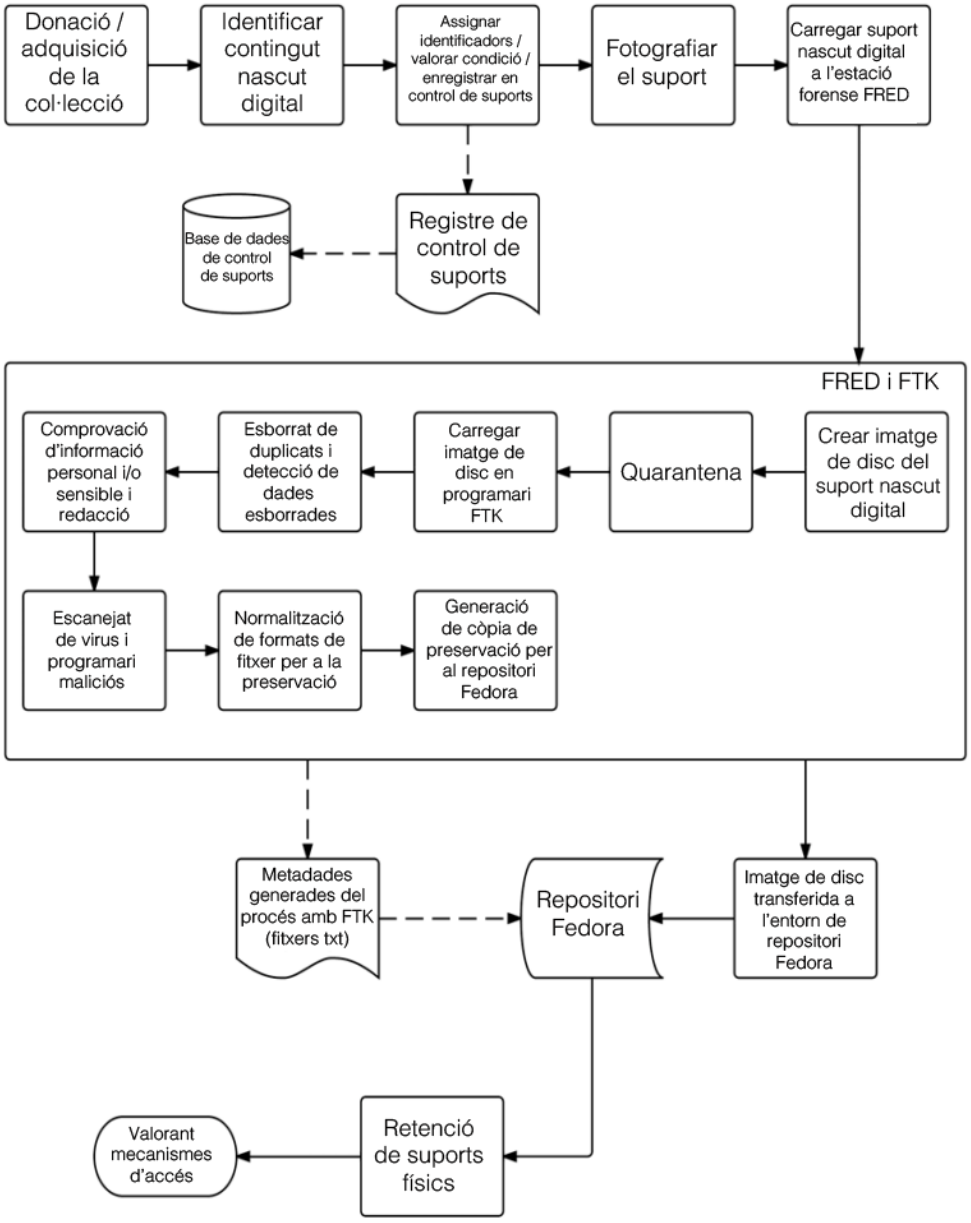
dins la imatge i crear valors *hash* en cadena per comprovar la integritat dels continguts. Un cop fets els passos de creació d'imatge, la qual ha de passar un procés de quarantena per tal d'assegurar que no hi ha cap programari maliciós, es carrega la imatge de disc per tal de localitzar i eliminar fitxers duplicats, així com detectar dades esborrades que poden existir a la imatge. Fet això, es fa la recerca d'informació personal i/o sensible que pot requerir redacció i es fa un escanejat de virus per verificar que efectivament no hi ha cap programari maliciós. Finalment, els formats de fitxer passen per un procés de normalització a estàndards de preservació i FTK s'utilitza per generar una còpia de preservació de la imatge forense.

La imatge de disc de preservació, juntament amb totes les metadades que s'han generat durant el procés amb FRED i FTK, s'envien a un repositori de preservació basat en Fedora. Acabat aquest procés, el suport original es conserva dins les instal·lacions de la biblioteca. A hores d'ara, la UVA encara no ha desenvolupat cap política d'accés als suports originals o les imatges forenses. La Figura 18 mostra el flux de treball descrit.

Entre les col·leccions ja processades, podem nombrar el cas de 83 discs (CDs i disquets) que es troben als arxius del escriptor i crític Alan Cheuse (AIMS Work Group, 2012, p. 99-100). El pla de processat de la UVA especifica que es van fer imatges forenses amb FTK, amb restriccions vers fitxers que contenien informació confidencial. En el cas de treballs comercials ja comercialitzats, no es fa creació d'imatge, atès que segons la política de l'entitat no necessiten ser preservats. No es va fer cap tipus de migració de formats obsolets, encara que no es descarta fer-ho en el futur. Un cop fetes les imatges forenses, es transfereixen al sistema de preservació de l'entitat, juntament amb les seves metadades. Els suports físics es retenen dins les col·leccions originals d'arxiu, les quals es poden consultar gràcies a l'eina d'ajuda[296] que utilitza la institució segons l'estàndard EAD.

[296] Special Collections, University of Virginia Library. *A guide to the additional papers of Alan Cheuse, 1971-1998*. <http://ead.lib.virginia.edu/vivaxtf/view?docId=uva-sc/viu03663.xml>. [Consulta: 28/08/2016]

Figura 18. Flux de treball d'accés a materials nascuts digitals a la UVA



Font: Gengenbach, 2012, p. 67. Traducció de l'autor

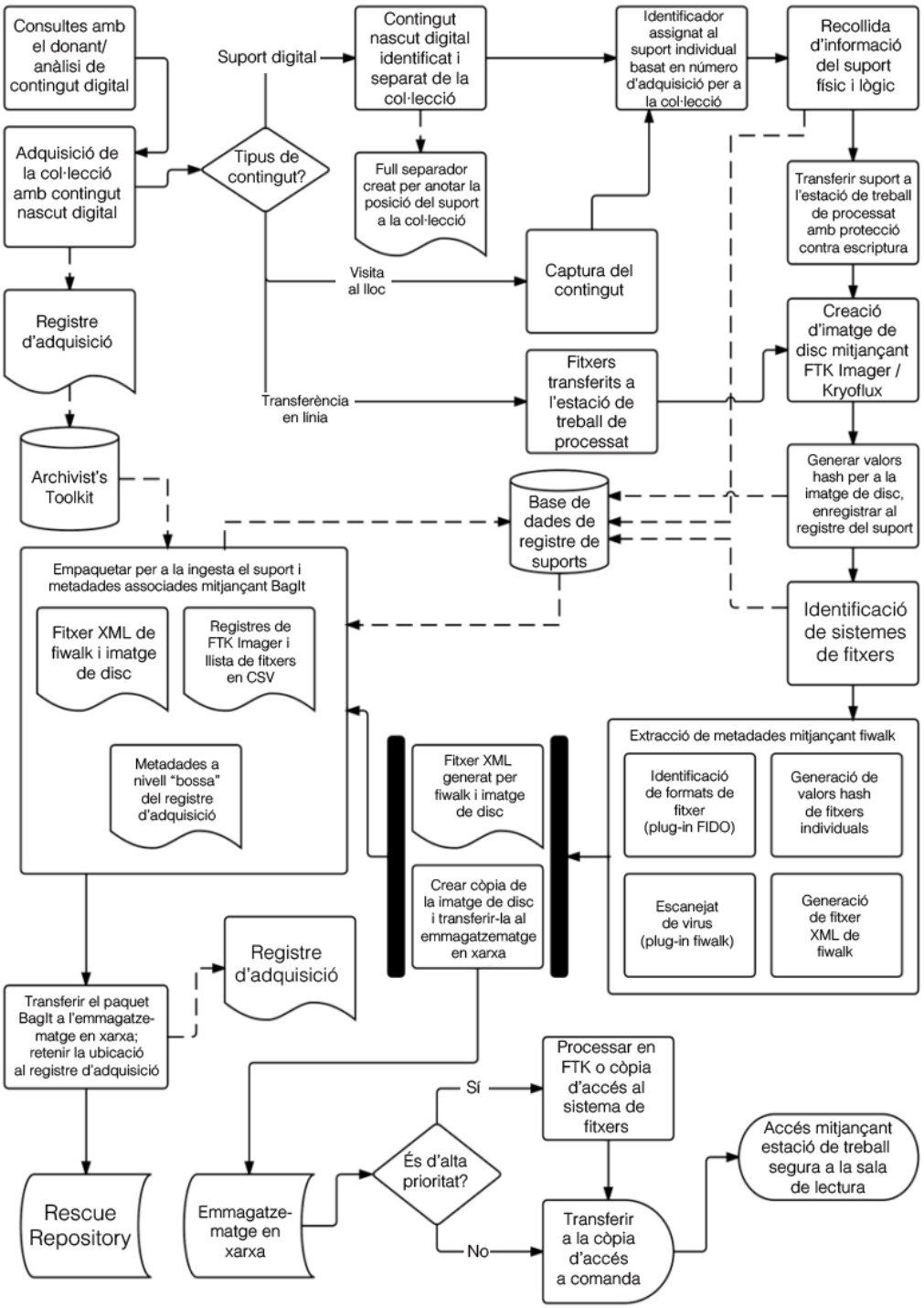
Yale University

La Beinecke Rare Book & Manuscript Library, fundada l'any 1963, és el repositori principal de la Yale University per a documents literaris i per a manuscrits i llibres antics en els camps de la literatura, teologia, història i ciències naturals[297]. D'igual forma que altres institucions ja comentades amb anterioritat, la Beinecke va començar en els últims anys a rebre material nascut digital dins les seves col·leccions. L'arxiver

[297] *Collections. Beinecke Rare Book & Manuscript Library*. <http://beinecke.library.yale.edu/collections>. [Consulta: 01/09/2016]

Michael Forstrom va publicar un cas d'estudi (2009) de gestió de registres electrònics personals a la institució que, juntament amb la informació aportada a la tesi de Gengenbach (2012, p. 27-31), formen la base del flux de treball de la Beinecke, el qual mostrem a la Figura 19.

Figura 19. Flux de treball a la Beinecke Rare Book and Manuscript Library, Yale University



Font: Gengenbach, 2012, p. 32. Traducció de l'autor

El cas d'estudi a la Beinecke es centra en l'arxiu de l'escriptor i crític George Whitmore. Aquest arxiu s'adquirí l'any 1996, i contenia correspondència professional, documentació de recerca, esborranys i diaris i fotografies personals. Com a material nascut digital, l'arxiu comptava amb 58 disquets de 5 ¼ polzades amb esborranys de les seves obres, notes de recerca, registres de diari, correspondència i altres. El programari que s'utilitzà per crear aquesta documentació digital fou Wordstar en el sistema operatiu CP/M, la qual cosa va fer necessari el recórrer a una empresa externa per tal de migrar els fitxers originals a formats ASCII. Els fitxers ja convertits foren ingestats a l'emmagatzematge digital de la Yale University, el Rescue Repository (un "arxiu fosc" que no permet l'accés públic als materials), i els suports físics van ser reallotjats a la col·lecció com a ítem fràgil i restringit. Una part molt interessant del procés és l'enregistrament de la informació administrativa, en què es van documentar el maquinari que va utilitzar Whitmore, un Kaypro (ordinador personal que es va començar a vendre l'any 1981); l'accés restringit dels disquets; la propietat física de la col·lecció, que pertany a la Beinecke; i les notes sobre el processat de la col·lecció, que especifiquen la presència dels 58 disquets i la migració que s'ha fet dels fitxers originals. Tota aquesta informació es pot consultar a la guia d'ajuda de la institució[298].

Hem de recordar, no obstant, que aquest cas d'estudi es va redactar abans que la Yale University finalitzés la seva participació al projecte AIMS i com a conseqüència d'això, s'han produït molts canvis dins el flux de treball, com l'ús de programari per a la creació d'imatges de disc com FTK Imager i maquinari especialitzat per recuperar el contingut de disquets com KryoFlux. Un cop feta la imatge, es genera un valor *hash* i es guarda per separat per tal d'assegurar que no s'hagin produït alteracions o corrupcions de dades i al mateix temps es realitzen comprovacions de presència de virus i identificació de formats de fitxer. L'anàlisi d'imatge de disc es fa amb fiwalk, que genera un fitxer XML segons l'esquema de metadades DFXML (com ja s'ha vist al capítol 3.3.2). Aquest fitxer XML i una còpia de la imatge de disc s'empaqueten segons l'especificació BagIt (vegeu el capítol 3.3.4), que permet fer un manifest dels continguts del *bag* que seran ingestats al Rescue Repository, i posteriorment validats i verificats mitjançant JHOVE.

[298] Yale University Library. *Guide to George Whitmore papers*. <http://goo.gl/X0rtzs>. [Consulta: 01/09/2016]

Una segona còpia es guarda a un emmagatzematge de xarxa per tal de ser processada amb el programari FTK, que permet fer una descripció i classificació dels fitxers, a més de definir els nivells d'accés dels usuaris. Un cop s'ha processat, el material accessible als investigadors es transfereix a un ordinador de la sala de lectura de la Beinecke.

4.2.6 BitCurator

BitCurator és un projecte conjunt de la School of Information and Library Science a la University of North Carolina i el Maryland Institute for Technology in the Humanities que s'inicià en octubre de 2011 gràcies al finançament de l'Andrew W. Mellon Foundation (Lee, 2012b; Lee et al., 2014) i es dissenyà per abordar dos problemes importants dins les institucions que reben col·leccions documentals (Lee et al., 2012; Lee; Woods, 2012):

- Integrar eines i mètodes forenses digitals als fluxos de treball i entorns de gestió de col·leccions. Això inclou l'exportació de dades forenses de manera que es puguin importar a sistemes descriptius, així com la modificació de tècniques forenses de selecció de dades que s'ajustin millor a les necessitats de les institucions
- Donar accés públic a les dades adquirides amb eines forenses. Un cas típic d'anàlisi forense digital és que les proves que s'han recollit d'una investigació criminal mai es donen a conèixer al públic. En canvi, les institucions que estan creant imatges de disc han d'afrontar el problema de com donar accés a les dades. Això no només inclou problemes tècnics sinó també ètics per la possible presència de dades confidencials

La primera fase del projecte (1 d'octubre de 2011 – 30 de setembre de 2013) va tenir com objectiu discutir motivacions, reptes i estratègies per a l'ús de tecnologies forenses i fluxos de treball en biblioteques, arxius i museus (Lee et al., 2013). Per aquesta raó es va comptar amb el suport de dos grups d'experts: el Professional Experts Panel, compost per arxivers i bibliotecaris d'institucions que reben materials nascuts digitals, i el Development Advisory Group, compost per tecnòlegs especialitzats en preservació digital (Gengenbach; Chassanoff; Olsen, 2012), que van col·laborar activament amb

l'equip principal del projecte (format per Christopher A. Lee, Kam Woods, Sunitha Misra, Alexandra Chassanoff, Matthew Kirschenbaum i Porter Olsen). Una de les reunions va coincidir amb l'acte CurateGear (Poole; Lee; Murillo, 2012) on es va fer la primera presentació de BitCurator com a projecte per construir eines i mètodes d'anàlisi forense digital adreçats a professionals que treballin amb col·leccions. Les decisions finals d'aquestes reunions foren l'aplicació dels següents criteris:

- Utilitzar tecnologies forenses de codi obert i amb possibilitat d'expansions
- Concentrar el desenvolupament en extensions, complements i paquets més que en desenvolupar programari des de zero
- Complir amb estàndards comuns de metadades forenses i facilitar conversions a esquemes de metadades rellevants de biblioteques i arxius

A partir d'aquests criteris, es va començar a desenvolupar i posar a prova una *suite* d'eines de codi obert en un entorn Linux sota llicència GPL que es podia executar com a una màquina virtual o bé com a un sistema operatiu instal·lable amb un fitxer ISO, amb una interfície gràfica basada en Ubuntu. Un avantatge d'aquestes eines és que es poden compilar fàcilment a entorns Windows i de fet, en la major part dels casos es distribueixen tant en forma de codi font com en forma d'arxius binaris per a Windows (Lee, 2014). L'entorn Ubuntu es trobaria optimitzat per gestionar suports, amb les funcionalitats de bloquejar l'alteració de dades per un *write blocker* integrat al programari, fer captura forense amb l'eina Guymager, analitzar dades amb les eines fiwalk i bulk_extractor i crear informes amb el generador intern de BitCurator (Woods; Lee; Misra, 2013). Aquests informes es generarien amb metadades DFXML que permeten conversions a altres estàndards (vegeu el capítol 3.3.2).

La segona fase del projecte (1 d'octubre de 2013 – 29 de setembre de 2014) es va centrar en augmentar les activitats dins les comunitats de professionals juntament amb la continuació del desenvolupament del programari (Lee et al., 2014). Aquestes activitats van consistir en difondre les eines de BitCurator, formar a professionals en els mètodes forenses i establir un grup d'usuaris. Dins aquesta fase es van fer diverses visites a responsables de diferents biblioteques i arxius dels EUA, de Regne Unit, dels Països Baixos, d'Alemanya i de Suècia, que consistien en una introducció a l'anàlisi forense digital, proves amb BitCurator i una conferència pública amb l'objectiu

d'introduir BitCurator a una audiència més àmplia. Les conclusions finals d'aquestes visites foren:

- La importància de compartir costos dins la comunitat, ja que això va permetre augmentar el nombre de visites institucionals i ajuda a tenir un compromís comú
- Concentrar-se en comunitats locals, ja que això va fer que més institucions properes geogràficament poguessin treballar conjuntament
- Invitar a molts agents a les activitats, atès que els arxivers i bibliotecaris necessiten de suport institucional

Quant al desenvolupament del programari, a la segona fase les eines van evolucionar significativament (en part gràcies a les contribucions de les visites institucionals). El generador d'informes de BitCurator va passar de ser un *script* Python que requeria introduir línies de comanament a una interfície gràfica que executava fiwalk, bulk_extractor i el mòdul de generador d'informes al mateix temps. Es va refinar la generació dels informes en format PDF i en Excel per tal de proporcionar una revisió ràpida de la imatge de disc i així facilitar el seu procés de revisió. També es va millorar el muntatge d'imatges de disc dins el sistema, que en un principi requeria l'ús de línies de comanament, mitjançant un menú senzill contextual tot seleccionant la imatge i clicant el botó dret. Per altra banda, es va incorporar la interfície gràfica Disk Image Access que permet la càrrega d'imatges forenses i *raw*, seleccionar i exportar fitxers (tant assignats com no assignats) d'una imatge de disc i visualitzar metadades d'aquesta imatge.

Un cop acabada aquesta fase del projecte, el projecte BitCurator es va donar per finalitzat. La versió més recent del programari en la data de redacció d'aquesta tesi (tant en la seva versió de màquina virtual com d'imatge ISO d'instal·lació) és la 1.8.0 amb data de publicació de 17 de març de 2017 del qual, en entorns de producció, es recomana la instal·lació de la imatge ISO en un maquinari amb processador Intel Core i5 o i7 amb 16 GB de RAM[299]. Actualment, el projecte es troba gestionat pel BitCurator

[299] BitCurator Consortium; UNC School of Information and Library Science (2017, Mar. 17*). BitCurator: quick start guide. Release(s): 1.8.0 and later* [presentació de Power Point]. <https://wiki.bitcurator.net/downloads/BitCurator-Quickstart.pdf>. [Consulta: 29/03/2017]

Consortium[300], una associació de diverses biblioteques i universitats que donen suport a la preservació de materials nascuts digitals mitjançant l'aplicació d'eines forenses en codi obert.

Hem de fer esment a un projecte que s'inicià immediatament després: el BitCurator Access[301] (1 d'octubre de 2014 – 30 de setembre de 2016). També va rebre finançament de l'Andrew W. Mellon Foundation, i el seu objectiu va ser desenvolupar programari de codi obert que permetés donar accés públic al contingut d'imatges de disc (Chassanoff; Woods; Lee, 2016) mitjançant tres mètodes:

- Elaborar eines per donar suport a serveis web
- Habilitar l'exportació de sistemes de fitxers i metadades associades
- Utilitzar entorns d'emulació

Addicionalment, BitCurator Access també va desenvolupar una eina per redactar fitxers, metadades del sistema de fitxers i seqüències de bytes dins imatges de disc. Aquest seria un concepte estrany a les investigacions criminals, ja que un investigador no tindria cap raó per alterar proves d'un cas, però des d'un punt de vista bibliotecari, és necessari eliminar les dades confidencials abans de la seva publicació (Wolverton, 2016).

En el cas d'accés a imatges de disc, un primer prototip fou DIMAC. Aquesta eina permetia explorar els continguts d'una imatge de disc mitjançant un navegador web i així facilitar als usuaris l'accés de forma remota (Misra; Lee; Woods, 2014). Una altra manera de donar accés fou l'ús de l'emulació amb el programari EaaS[302] i una aplicació web dedicada (Woods et al., 2015). El programari disponible actualment rep el nom de 'bitcurator-access-webtools'[303], i es troba dissenyat en Flask. Com que no forma part de la *suite* BitCurator, és necessari descarregar-lo i instal·lar-lo; a més, també cal descarregar i instal·lar els programaris VirtualBox i Vagrant.

[300] <https://www.bitcurator.net/bitcurator-consortium/>. [Consulta: 24/03/2017]

[301] <https://www.bitcurator.net/bitcurator-access/>. [Consulta: 24/03/2017

[302] *bwFLA — Emulation as a Service*. <http://bw-fla.uni-freiburg.de/>. [Consulta: 24/03/2017]

[303] <https://wiki.bitcurator.net/index.php?title=BitCurator_Access_Webtools>. [Consulta: 24/03/2017]

Quant a la redacció de continguts a imatges de disc, això es va aconseguir gràcies a l'elaboració del *script* Python 'iredact.py', que va permetre llegir informes generats per bulk_extractor i executar redaccions en imatges *raw*; la raó de no fer-ho en imatges forenses és que aquestes no poden ser alterades sense comprometre la validesa dels *checksums* incrustats a la imatge (Woods; Lee, 2015). Actualment, l'eina s'ha refinat amb el nom de 'bitcurator_access_redaction'[304] i es pot executar tant amb línies de comanament com per interfície gràfica.

Actualment, es troba en fase de desenvolupament el projecte BitCurator NLP[305] (iniciat l'1 d'octubre de 2016 i amb data prevista de finalització de 30 de setembre de 2018), el qual desenvoluparà programari d'extracció, anàlisi i producció d'informes sobre característiques d'interès en text extret de materials nascuts digitals. Aquest programari utilitzarà biblioteques de processament de llenguatge natural per a la identificació i generació d'informes d'aquells ítems rellevants per a les institucions que reben col·leccions. Algunes proves que ja s'han fet mostren que és possible generar informes sobre el nombre d'ocurrències de paraules clau, ordenades per categories, presents a més de mil fitxers de text[306].

Un dels primers casos d'ús de BitCurator amb materials nascuts digitals es realitzà l'any 2014 amb la col·lecció de l'escriptor John Updike, dipositats a la Houghton Library[307] de la Harvard University (Lee et al., 2014). El maquinari que es va emprar fou un Mac Mini amb 16 GB de RAM i una màquina virtual preinstal·lada amb BitCurator, amb una unitat externa de disquets de 3 ½ polzades. La col·lecció estava composta per aproximadament 50 disquets de 3 ½, sis disquets de 5 ¼ que eren discs d'instal·lació del programari Lotus Ami Pro, i una dotzena de CD-ROMs. Com que aquest cas d'ús era més posar a prova les funcionalitats de BitCurator que preservar tota la col·lecció, només es van capturar una dotzena de disquets, dels quals es va fer una anàlisi amb bulk_extractor i amb les eines de generació d'informes de BitCurator. La conclusió d'aquesta experiència fou molt profitosa per al personal de la Houghton, ja que es van

[304] <https://github.com/BitCurator/bitcurator-access-redaction>. [Consulta: 24/03/2017]

[305] <https://www.bitcurator.net/bitcurator-nlp/>. [Consulta: 24/03/2017]

[306] Woods, Kam (2017, Feb. 3). *BitCurator NLP: mining collections for NEs, relationships, and topics to enrich access* [presentació de Power Point]. <http://www.bitcurator.net/wp-content/uploads/2016/12/BCNLP-nlp4arc-v3.pdf>. [Consulta: 26/03/2017]

[307] <http://hcl.harvard.edu/libraries/houghton/>. [Consulta: 25/03/2017]

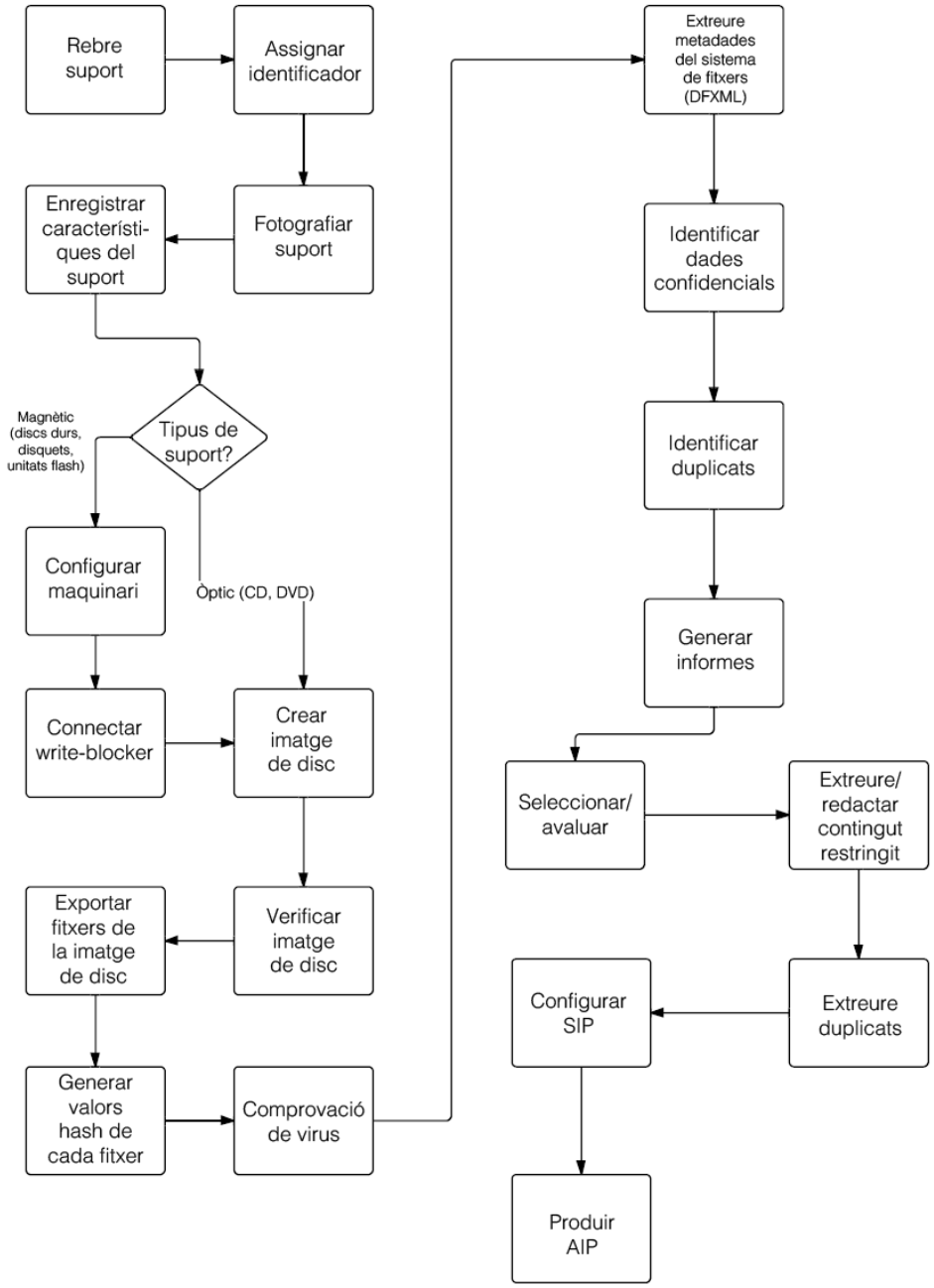
adonar que era possible rescatar els continguts dels disquets, però que això implicava futurs treballs de disseny de polítiques i de gestió dels fitxers, així com preparar els continguts per tal que els investigadors hi tinguin accés.

Un altre cas d'ús en què es va aplicar un flux de treball elaborat, que mostrem a la Figura 20, fou el de la col·lecció de la poeta Patricia Goedicke, un arxiu híbrid dipositat a la Maureen and Mike Mansfield Library[308] a la University of Montana, on es trobaven materials nascuts digitals en suports diversos com disquets (tant de 3 ½ com de 5 ¼ polzades), discs Zip, CDs, DVDs i una memòria USB (Meister; Chassanoff, 2014). Un maquinari addicional que es va emprar fou una unitat externa de disquets de 5 ¼, una targeta FC5025 per poder connectar aquesta unitat per USB, una unitat externa USB de disquets de 3 ½ i un *write blocker* USB Tableau T8-R2. Quant al programari, a més de l'entorn BitCurator es va utilitzar un sistema Windows per poder executar FTK Imager.

La primera operació del flux de treball consistí en documentar els suports a un nivell administratiu, on es van enregistrar les seves característiques físiques i es va assignar un identificador individual. A continuació es van crear les imatges forenses, on s'utilitzà Guymager per a la captura de la memòria USB, FTK Imager per a la captura de disquets de 3 ½ i el programari propi de FC5025 en el cas de disquets de 5 ¼. El format d'imatge forense escollit va ser l'AFF degut a què es tracta d'un format obert i extensible (per a més detalls, vegeu el capítol 4.3.1). Seguidament, es van exportar fitxers de la imatge de disc perquè es va decidir de forma interna que els AIPs estarien compostos tant per les imatges de disc com pels fitxers exportats. A la següent operació es van utilitzar les eines d'anàlisi de BitCurator, on es van extreure les metadades del sistema de fitxers amb fiwalk, la identificació de dades privades i sensibles amb bulk_extractor i la generació d'informes que presentin les dades analitzades (per a més detalls d'aquestes operacions, vegeu el capítol 4.3). Dins el cas d'estudi mostrat no es van executar les operacions concretes de redacció i extracció de dades confidencials, atès que en aquell moment encara estaven en desenvolupament les eines dedicades. No obstant, aquestes operacions es van incorporar al flux de treball degut a què formen part necessàriament de l'anàlisi de contingut dels suports.

[308] <http://www.lib.umt.edu/>. [Consulta: 25/03/2017]

Figura 20. Flux de treball de preservació de l'arxiu de Patricia Goedicke



Font: Meister; Chassanoff, 2014. Traducció de l'autor

4.3 Proves amb el programari forense BitCurator

Per tal de demostrar les possibilitats de les eines forenses i la seva viabilitat amb les necessitats de preservació del model s'han una realitzat una sèrie de proves. S'ha utilitzat per aquest fi una estació de treball amb el sistema operatiu BitCurator de 64 bits

en la versió 1.7.30 (publicada el 17 d'agost de 2016), el qual es troba sota la llicència GNU GPL versió 3[309], i amb un maquinari de 8 GB de RAM, processador Intel Core i7-4790 a 3,60 GHz i 1 TB de disc dur. Les raons per utilitzar BitCurator han estat, per una banda, a què es va crear per integrar eines i mètodes forenses dins els fluxos de treball de biblioteques i a què dóna suport a l'accés públic de dades adquirides de forma forense (Lee et al., 2012), i per altra, a què es tracta d'un sistema derivat de la distribució Linux Ubuntu, i per tant totes les eines són de codi obert i sense cost per a l'usuari. Val a dir que les proves haurien estat més esclaridores si s'hagués pogut utilitzar el maquinari FRED i el programari FTK i poder comparar resultats, però no ha estat possible pel seu cost elevat (vegeu les taules Taula 26 i Taula 27). Per instal·lar BitCurator a l'estació de treball s'ha utilitzat la imatge ISO; val a dir que aquesta imatge ISO està configurada per treballar amb una configuració regional dels EUA, per tant durant la instal·lació no s'haurà de modificar aquesta configuració (hora i data, idioma de teclat i idioma de sistema) perquè en cas contrari s'haurà de tornar a començar el procés. Un cop instal·lat BitCurator, es podrà modificar la configuració regional.

BitCurator dóna suport a quatre àrees dins el flux de treball arxivístic: la creació d'imatges forenses (que asseguren què no s'han produït canvis durant la cadena de custòdia abans de la ingesta), la identificació d'informació privada i sensible (amb reconeixement automatitzat i extracció per categories), l'avaluació i selecció de dades (que inclou identificació de formats de fitxer i de dades esborrades, així com la generació d'informes) i l'exportació de metadades com és el cas de DFXML. Tots aquests processos tenen com a fonaments els principis arxivístics de procedència (s'identifica i es guarda informació sobre el context i la creació de les dades), ordre original (es preserva l'estructura original de carpetes i fitxers) i cadena de custòdia (s'utilitza maquinari i programari per assegurar que les dades no s'han alterat).

4.3.1 Creació d'imatges forenses

Els beneficis que té la creació d'imatges forenses són diversos (Woods; Lee, 2012; Woods; Lee; Garfinkel, 2011): reducció de riscos de pèrdua de dades, extracció de dades sense realitzar canvis al suport original (sempre i quan s'utilitzi un *write blocker*),

[309] <https://www.gnu.org/licenses/gpl-3.0.html>. [Consulta: 02/09/2016]

retenció de metadades del sistema de fitxers original, documentació d'aplicacions que s'han utilitzat per crear els documents al suport i identificació d'informació esborrada i privada.

Abans de començar a parlar del procés, paga la pena comentar quines són les característiques dels tres tipus de formats d'imatges de disc que es poden crear amb BitCurator. En primer lloc, tenim la imatge RAW (amb extensió .dd), la qual bàsicament és una còpia bit a bit de les dades del suport original, sense metadades addicionals que es guardin dins el fitxer. En segon lloc, tenim el format EnCase Forensic, un format propietari de l'empresa Guidance Software, que guarda informació sobre dins la seva capçalera sobre el nom de l'investigador, el nom del cas, descripció del suport, informació de l'hora i dia en què s'ha creat la imatge, la versió del programari que s'ha utilitzat i el sistema operatiu que s'ha executat per crear la imatge[310]. A la seva versió 1 existia una limitació de 2 GB, la qual cosa obligava a crear múltiples fitxers per a la imatge forense (fitxer1.E01, fitxer2.E02, fitxer3.E03...), però BitCurator utilitza una versió més avançada, atès que amb les proves que s'han fet ha estat possible guardar una imatge de 160 GB en un sol fitxer.

Un altre format és l'Advanced Forensic Format (AFF), que té tots els avantatges del format EnCase però a més, es tracta d'un format no propietari, implementat sota codi obert, lliure de restriccions de propietat intel·lectual, amb un disseny senzill i que permet, a més de la generació de valors *hash* com MD5 i SHA-1, fer signatures digitals basats en certificats X.509 (Garfinkel et al., 2006; Garfinkel, 2009b). Els seus avantatges inclouen també guardar les metadades de la imatge forense de forma separada o bé dins del fitxer i la recuperació de dades si es produeixen errors durant la seva creació. Segons Garfinkel (2006), AFF permet emmagatzemar més d'un TeraByte d'imatges de disc en menys de 200 GB. A partir de la versió 3, és possible dividir una imatge forense AFF de gran mida en segments múltiples de 2 GB, en aquest cas amb l'extensió .afd[311]. La versió més recent, la 4, va incorporar la funció d'emmagatzemar diversos suports en un sol fitxer (Cohen; Garfinkel; Schatz, 2009) i una millora en la

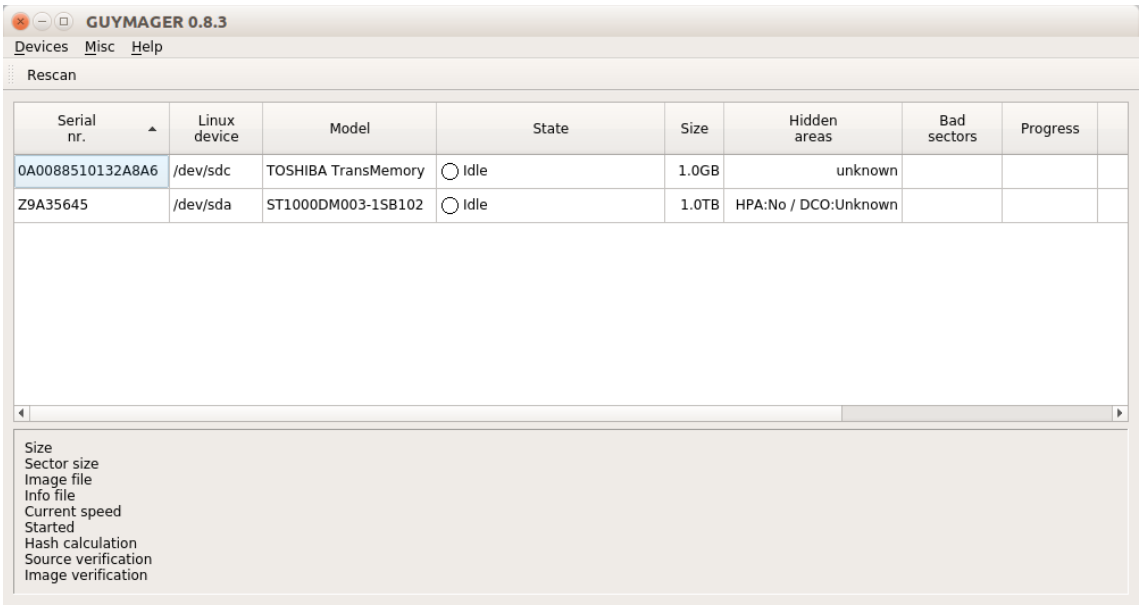
[310] Forensicsware (2016). *EnCase E01 file format explained.* <http://www.forensicsware.com/blog/e01-file-format.html>. [Consulta: 02/09/2016]

[311] *Advanced Forensic Format Disk Image, AFF Version 1.0*. <http://www.digitalpreservation.gov/formats/fdd/fdd000412.shtml>. [Consulta: 07/09/2016]

generació de valors *hash* (Cohen; Schatz, 2010). Un estudi de Kim i Ross (2012) mostra els avantatges d'AFF davant altres formats contenidors com tar i WARC, com la seva millor flexibilitat quant a la construcció de metadades.

Per fer les proves de creació d'imatge, s'ha utilitzat el programari Guymager sota BitCurator amb tres suports: un disc dur intern IDE Samsung de 160 GB connectat externament per USB, una memòria USB Toshiba d'1 GB i un CD-RW TDK de 650 MB. En tots tres suports s'ha emmagatzemat documentació que s'ha utilitzat amb aquesta tesi (documents de text, fulls de càlcul, etc.). No s'ha utilitzat cap maquinari de *write blocker* perquè BitCurator utilitza un sistema intern de bloquejat contra escriptura. Dins les tres opcions de format d'imatge forense que permet Guymager, s'han utilitzat EnCase i AFF, atès que les dues guarden metadades administratives d'adquisició i així poder comparar millor les seves característiques. Les imatges RAW, com s'ha esmenat més amunt, no retenen metadades addicionals sobre el procés de captura ni altres accions que s'hagin realitzat durant l'adquisició, la qual cosa dificulta la demostració del manteniment de la cadena de custòdia.

Figura 21. Interfície gràfica de Guymager



Font: Entorn BitCurator. Captura de l'autor

Quan obrim Guymager, ens dóna diversa informació sobre les unitats que reconeix, com la seva mida, el model i el número de sèrie del suport que s'ha introduït a l'estació de treball, una dada molt important ja que és únic per a cada dispositiu i permet identificar-

lo de forma segura. Un cop s'ha seleccionat el suport, es pot seleccionar el format d'imatge que es vol entre les tres disponibles, i opcionalment es poden afegir descripcions com el número de cas, el número d'evidència, el nom de l'examinador i una descripció del suport. Es poden generar tres tipus de valors *hash*, fer una segona lectura del suport per verificar-lo i verificar la creació de la imatge. Un cop hagi acabat el procés, obtindrem la imatge forense del suport i també un fitxer d'informació amb informació tècnica del procés de captura. Els resultats finals es poden consultar a la Taula 20.

Dins els resultats, es pot comprovar l'eficàcia del sistema BitCurator pel que fa al bloquejat contra escriptura, atès que els valors *hash* han coincidit perfectament amb els de la captura del disc dur i de la memòria USB. La raó per la qual el CD-RW ha donat resultats diferents és que durant la captura, Guymager va localitzar un sector del disc en mal estat i va substituir-lo per un valor 0.

Un cop acabades les proves, es podria concloure que el format més adient per al seu ús al nostre model seria l'AFF, ja que té una taxa de compressió una mica millor que EnCase, és un format de codi obert i permet un major nombre de metadades. Però també s'ha de dir que AFF ja no està desenvolupat pel seu creador, el Dr. Garfinkel[312], encara que continua experimentant millores i avenços gràcies als treballs de Schatz (2015). Per tant, més endavant justificarem quin format utilitzarem al nostre model, un cop hem exposat el model teòric de preservació al capítol 5.

[312] Guymager wiki. *AFF format deprecated.* <https://sourceforge.net/p/guymager/wiki/AFF%20format%20deprecated/>. [Consulta: 02/09/2016]

Taula 20. Comparativa de resultats de creació d'imatge forense amb Guymager

Suport	Format EnCase	Format AFF
Disc dur Samsung (160 GB)		
Valor *hash* MD5	3a572a4d679fe64e77394b208d9e88e8	3a572a4d679fe64e77394b208d9e88e8
Valor *hash* SHA-1	0d5c7465eb27dce7fd7a5b9ced5a9f3313873e8	0d5c7465eb27dce7fd7a5b9ced5a9f3313873e8
Valor *hash* SHA-256	4fbcb79375b6315da4bb50add3d7b106553208d1d4a11074e3386922e0fef990	4fbcb79375b6315da4bb50add3d7b106553208d1d4a11074e3386922e0fef990
Mida de la imatge	145,6 GB	145,3 GB
Temps de creació de la imatge	1 hora, 10 minuts, 58 segons	1 hora, 9 minuts, 14 segons
Temps de verificació de la imatge	1 hora, 9 minuts, 21 segons	1 hora, 20 minuts, 51 segons
S'han produït errors durant la creació de la imatge?	No	No
Memòria USB Toshiba (1 GB)		
Valor *hash* MD5	6fadab0c09fb79cb854f40b71d33fb71	6fadab0c09fb79cb854f40b71d33fb71
Valor *hash* SHA-1	3fe9986a1c3b1ce28f58d8d522508abda569b26e	3fe9986a1c3b1ce28f58d8d522508abda569b26e
Valor *hash* SHA-256	7e5c9c3b3703a2c28c431b42c0744a4bcb705a36f5ea99e7d4f1f78b2946bf29	7e5c9c3b3703a2c28c431b42c0744a4bcb705a36f5ea99e7d4f1f78b2946bf29
Mida de la imatge	883 MB	880 MB
Temps de creació de la imatge	1 minut, 37 segons	1 minut, 36 segons
Temps de verificació de la imatge	1 minut, 35 segons	1 minut, 35 segons
S'han produït errors durant la creació de la imatge?	No	No
CD-RW TDK (650 MB)		
Valor *hash* MD5	afedfd5eae1c894f3a29d5041226951f	e585da80a33319a71d2a32dfd24a6e2f
Valor *hash* SHA-1	fa3db7963a94e072a1e717a2b09b9d2c48597480	8a1a4a760e605e2d71f2dca3a00c3dc71adca4c5
Valor *hash* SHA-256	f9508523383d4acf7d42715cdf2b1d90dd1f9ee85bc2f0e165c18458a7c93a74	0541d17846e4aa4ae9e16841bef2a7db899c49b0a7042be1cbdc84090f85fa45
Mida de la imatge	435 MB	433 MB
Temps de creació de la imatge	7 minuts, 43 segons	7 minuts, 54 segons
Temps de verificació de la imatge	7 minuts, 42 segons	7 minuts, 54 segons
S'han produït errors durant la creació de la imatge?	Sí	Sí

Font: L'autor, a partir de sis fitxers .info de creació d'imatge forense generats per Guymager

4.3.2 Identificació d'informació privada i sensible

Un cop creada la imatge forense, podem extreure el suport, "muntar" la imatge i navegar dins els continguts en mode només lectura tal i com si estigués connectat el suport original. Els continguts seran exactament els mateixos, sempre i quan la verificació de la imatge hagi estat correcta. És en aquest moment que podrem fer un escanejat de virus mitjançant ClamAV, ja que si el suport original tenia virus, també estaran presents a la imatge forense. La interfície gràfica, ClamTK, no genera cap registre d'escanejat, així que seria necessari la utilització de ClamAV amb línies de comanaments i utilitzar les opcions necessàries per a aquest fi.

El següent pas serà utilitzar bulk_extractor, un programa dissenyat originalment per Garfinkel i Cox (2009) amb l'objectiu d'extreure dades privades i sensibles d'una imatge forense mitjançant un escanejat del contingut byte a byte i generar informes separats per categories. És capaç d'extreure correus electrònics, números de targetes de crèdit, URLs i altres tipus d'informació amb resultats molt satisfactoris en comparació amb altres programaris (Garfinkel, 2013). Hi ha diverses maneres d'executar aquest programari, però a les proves hem utilitzat la interfície gràfica Bulk Extractor Viewer, que ofereix diverses opcions d'escanejat les quals es mostren a la Figura 22. Aquests escàners generaran fitxers de text amb la informació recuperada, els quals es detallen a la Taula 21. Gràcies a aquesta eina, serà molt més senzill identificar informació privada i confidencial i bloquejar-la si és necessari.

Cadascun dels fitxers es troba en format tabulat, amb tres columnes on es presenten les següents característiques: posició, característica i context. La posició es refereix a la ubicació en la imatge forense tenint en compte el seu nombre de bytes, la característica és la informació trobada (p. ex., si és un correu electrònic, seria aquest correu electrònic) i el context dóna una referència a l'entorn en què s'ha trobat la informació (p. ex. la cadena de text on s'han trobat originalment les dades). A la Taula 22 mostrem un exemple d'extraccions de correu electrònic de l'autor.

Taula 21. Tipus d'escanejat que realitza Bulk Extractor (v. 1.6.0)

Nom de l'escanejat	Fitxer(s)	Descripció de la informació recuperada
accts	telephone.txt, ccn.txt, ccn_track2.txt, pii.txt	Comptes numèrics, com números de targetes de crèdit i números de telèfon
aes	aes_keys.txt	Informació encriptada en valors AES
base16	hex.txt	Informació codificada en base 16 (inclou valors MD5)
base64	Diversos fitxers descodificats	Codi en base 64
elf	elf.txt	Executables ELF
email	email.txt, rfc822.txt, domain.txt, url.txt	Adreces de correu electrònic
exif	exif.txt, jpeg.txt, jpeg_carved.txt	Metadades Exif d'imatges JPEG
facebook	url_facebook-address.txt, url_facebook-id.txt	Codi HTML de Facebook
find	find.txt	Expressions regulars
gps	gps.txt	Coordenades GPS de fitxers XML de dispositius GPS de Garmin
gzip	Diversos fitxers descodificats	Informació de fitxers GZIP
hiberfile	Diversos fitxers descodificats	Fragments de fitxers d'hibernació de Windows
httplogs	httplogs.txt	Fitxers de registre HTTP
json	json.txt	Fragments JSON extrets i validats
kml	kml.txt	Fitxers KML
lightgrep	ligthgrep.txt	Expressions regulars
msxml	Diversos fitxers descodificats	Microsoft XML Core Services
net	ip.txt, ether.txt, tcp.txt, domain.txt	Adreces IP, MAC i URLs
outlook	Diversos fitxers descodificats	Informació encriptada de fitxers del gestor de correu electrònic Outlook
pdf	Diversos fitxers descodificats	Text d'arxius PDF
rar	rar.txt, unrar_carved.txt	Informació de fitxers RAR
sceadan	No indicat	Utilitza el Systematic Classification Engine for Advanced Data ANalysis per analitzar dades en brut
sqlite	sqlite_carved.txt	Detecció de fitxers de bases de dades SQLite3
vcard	vcard.txt	Dades de vCard (format de targeta electrònica de negocis)
windirs	windirs.txt	Entrades de directoris FAT32 i NTFS
winlnk	Diversos fitxers descodificats	Fitxers LNK de Windows descodificats
winpe	winpe.txt	Executables previs a la instal·lació de Windows
winprefetch	winprefetch.txt	Fitxers i fragments de fitxer que s'executen en quan arrenca un ordinador amb Windows
wordlist	worldlist.txt, wordlist_histogram.txt	Desencriptat de contrasenyes

Nom de l'escanejat	Fitxer(s)	Descripció de la informació recuperada
xor	Diversos fitxers descodificats	Dades ocultes per la tècnica XOR
zip	zip.txt, unzip_carved.txt	Informació de fitxers ZIP

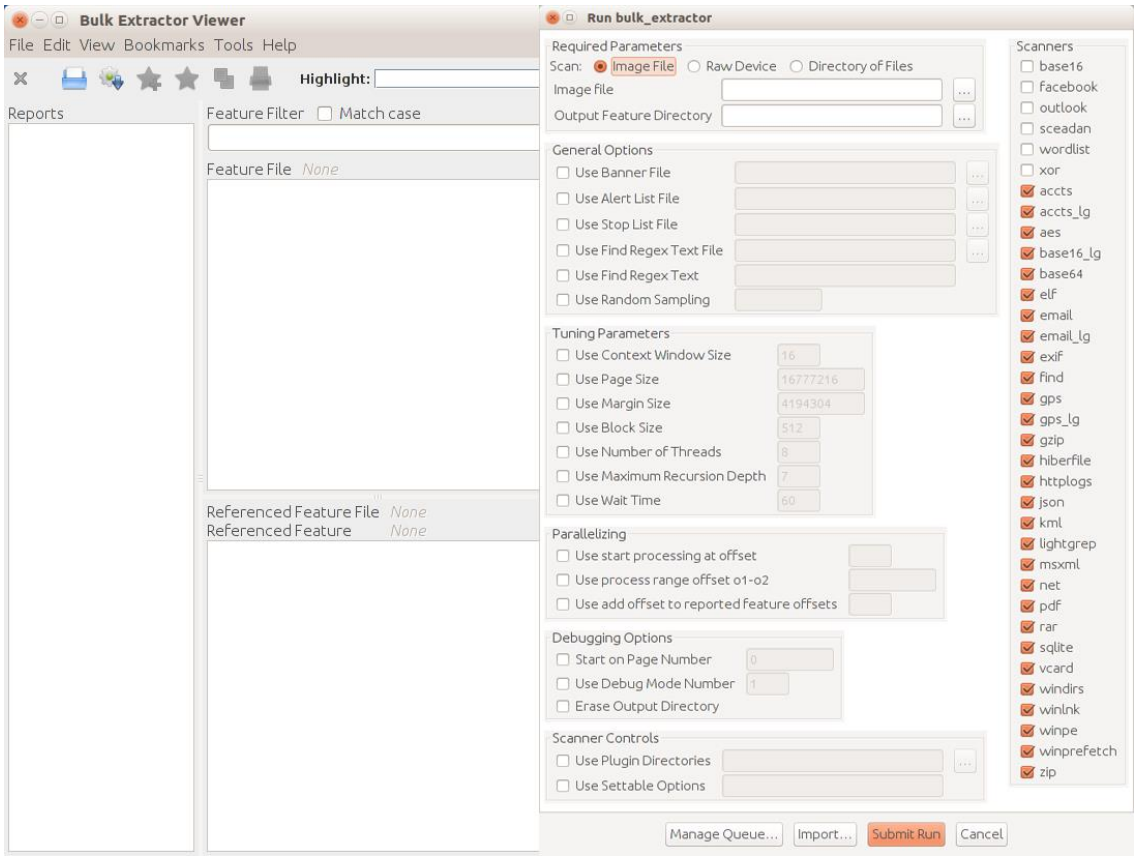
Font: BitCurator (2016). *Bulk Extractor scanners*. <http://wiki.bitcurator.net/index.php?title=Bulk_Extractor_Scanners>. [Consulta: 02/09/2016]; Bradley; Garfinkel, 2015; *GitHub - nbeebe/sceadan: Systematic Classification Engine for Advanced Data ANalysis.* <https://github.com/nbeebe/sceadan>. [Consulta: 02/09/2016]

Taula 22. Exemple d'extraccions d'informació privada i sensible amb bulk_extractor

Posició	Característica	Context
13950512	tlopezwi@gmail.com	ek\x0A\x0AWilderbeek\x0A\x0Atlopezwi@gmail.com\x0A\x0Atlopezwi@gmail
50547814	tlopezwi@gmail.com	\x0DCaracter\xEDstica\x0Dtlopezwi@gmail.com\x0DContext\x0Dreu ele
51718559	tlopezwi@gmail.com	reu electr\xF2nic tlopezwi@gmail.com\x07\x07\x07\x07\x0DDADES D\x92ACC

Font: L'autor, a partir d'una imatge forense en format EnCase

Figura 22. Interfície gràfica Bulk Extractor Viewer, amb detall de les opcions d'escanejat



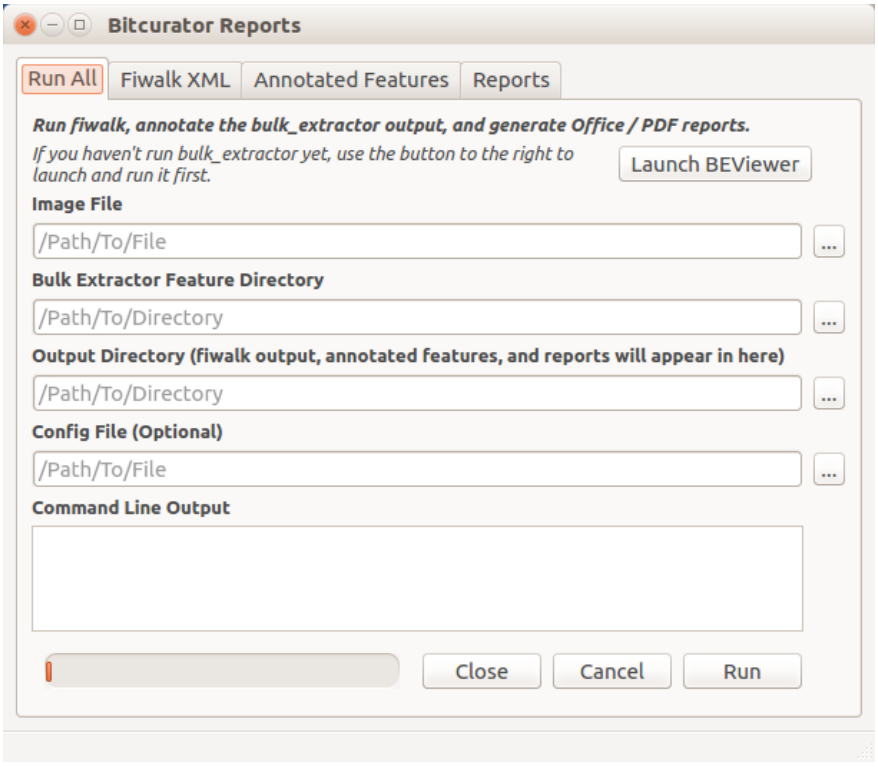
Font: Entorn BitCurator. Captura de l'autor

Per altra banda, Bulk Extractor Viewer genera un fitxer de metadades DFXML amb informació de les característiques que s'han extret de la imatge forense. Ens informa bàsicament sobre quins són els escanejats que s'han fet servir i els fitxers de text que s'han generat, així com els temps d'execució. Aquesta és una dada important quan s'ha de treballar amb imatges forenses de gran mida, ja que permet fer una previsió del temps necessari de processat que necessitaria una estació de treball.

4.3.3 Avaluació i selecció de dades

Un cop hem executat Bulk Extractor Viewer podem generar fitxers amb metadades DFXML i PREMIS i diferents tipus d'informes amb l'eina BitCurator Reports, que es troba escrit en llenguatge Python i permet llegir i processar metadades extretes de forma forense (Woods; Lee et al., 2013). Tal i com es mostra a la Figura 23, es tracta d'una interfície gràfica amb diverses opcions, però les que farem servir seran les de 'Run All' per tal de poder generar el fitxer de metadades PREMIS, la qual cosa no ha estat possible durant les proves que s'han realitzat amb les altres opcions.

Figura 23. Interfície gràfica de BitCurator Reports



Font: Entorn BitCurator. Captura de l'autor

Primer de tot, seleccionarem la imatge forense que volem analitzar. En segon lloc, seleccionem el directori amb el qual vam desar les dades extretes anteriorment mitjançant Bulk Extractor. Finalment, seleccionarem el directori en què volem desar els informes que es generaran amb BitCurator Reports. És important tenir en compte que aquest directori ha de ser nou, ja que per qüestions tècniques no es pot seleccionar un directori creat prèviament. Finalment executem el programari, que ens ha de generar el següent:

- Un fitxer amb metadades DFXML amb informació detallada sobre el sistema de fitxers, el nombre de fitxers i de directoris, dades esborrades, formats de fitxers i valors *hash* MD5 i SHA-1 de cadascun dels fitxers. Aquest fitxer rebrà per defecte el nom de 'fiwalk-output.xml'
- Directori 'annotated-features', on s'emmagatzemaran fitxers de text pla tabulat en funció dels escàners que hem seleccionat a Bulk Extractor Viewer. Aquests fitxers es generaran amb el prefix 'annotated' i ens donen informació addicional respecte als fitxers creats amb Bulk Extractor: el fitxer on es troben aquestes dades confidencials i el valor *hash* MD5 d'aquest fitxer. Per tant, els fitxers tabulats d'aquest directori són molt importants, perquè ens diuen exactament on localitzar les dades reconegudes com confidencials i d'aquesta manera, es podrà valorar si és necessari bloquejar o no els fitxers referenciats
- Directori 'reports', on s'emmagatzemaran els següents fitxers:
    - bc_format_bargraph.pdf. Diagrama de barres que representa el nombre de tipus de formats de fitxer reconeguts i el nombre d'ocurrències per a cada format
    - bulk_extractor_report.pdf. Informa dels fitxers que ha generat Bulk Extractor amb detalls del temps d'operació
    - fiwalk_deleted_files.pdf. Informa dels fitxers reconeguts com esborrats a la imatge forense
    - fiwalk_report.pdf. Informa de les metadades tècniques de la imatge forense, com número de partició, nombre de fitxers, nombre de directoris, nombre de fitxers esborrats o el sistema de fitxers
    - fiwalk-output.xml.xlsx. Dóna la mateixa informació que 'fiwalk.xml', però en format Microsoft Excel. Presenta un total d'11 columnes, que permeten veure de forma ràpida els continguts de la imatge forense.

Aquestes columnes informen del número de partició, el nom de fitxer, la seva extensió, la seva mida en bytes, el format de fitxer, la data d'accés, la data de creació, la data de modificació i els valors *hash* MD5 i SHA1

- format_table.pdf. Llistat dels formats reconeguts a la imatge forense i el nombre d'ocurrències per a cada format
- premis.xml. Aquest fitxer presenta metadades PREMIS que enregistrarà tres Esdeveniments: la captura de la imatge forense, l'anàlisi del sistema de fitxers i l'ús de Bulk Extractor per extreure dades confidencials
- Subdirectori 'features', on es desaran diversos fulls de càlcul en format Microsoft Excel que contindran tres columnes amb la informació del fitxer on s'han trobat les dades, la característica trobada i la posició

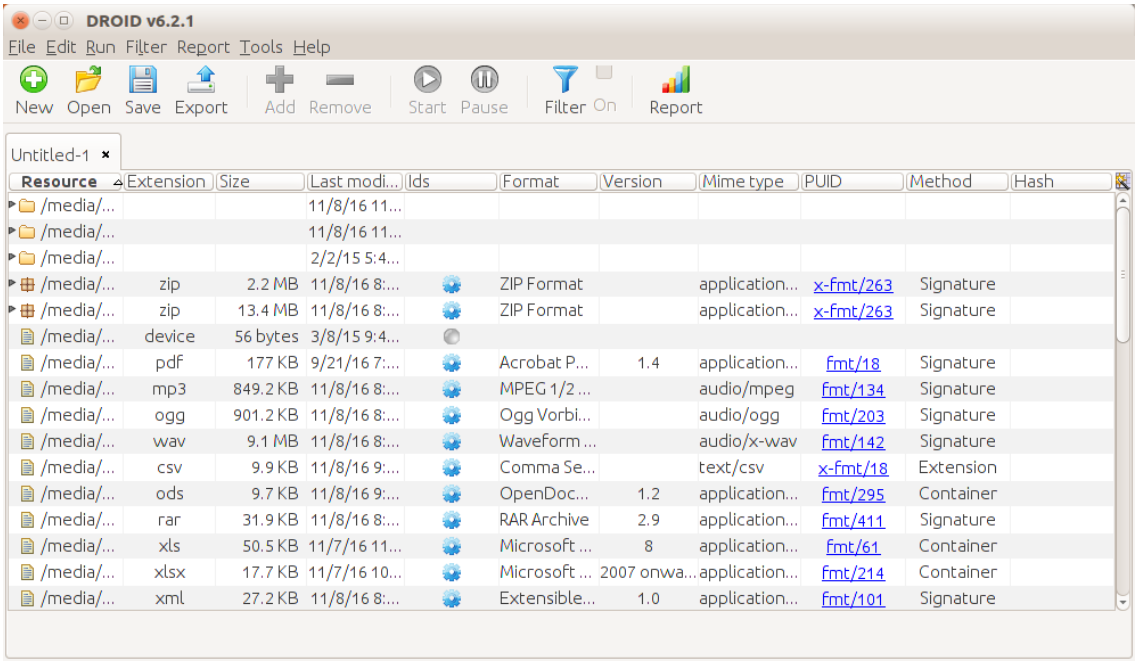
Es poden aprofitar els informes de formats de fitxer per fer normalització de formats per a la preservació si la institució que ha de preservar el contingut té alguna política al respecte. No obstant, hem fet algunes comparatives entre els resultats de BitCurator Reports, el qual utilitza fiwalk, i els que ens ha donat el programari DROID i els resultats amb aquest últim són molt més acurats, ja que es tracta d'una eina de codi obert sota llicència BSD programada en Java per The National Archives, arxiu nacional del Regne Unit, dissenyada per l'anàlisi i el reconeixement de formats de fitxer (Dearborn et al., 2014). Permet donar informació detallada sobre la mida del fitxer, les dates de darreres modificacions o informació sobre duplicats. Per reconèixer els formats fa servir el registre PRONOM[313] (que també es administrat per The National Archives), on s'emmagatzemen les dades tècniques de cada format. Per identificar els diferents formats i actualitzar les dades tècniques s'utilitzen els fitxers de signatures DROID[314], dels quals la darrera versió publicada a data de 6 de gener de 2017 és la 88.

S'ha de dir, però, que DROID no és un programari instal·lat de sèrie a BitCurator i per tant, és necessari descarregar-lo a la pàgina web de The National Archives. Ja que es tracta d'un programari en Java, es pot utilitzar a entorns Linux o Windows. En el cas de BitCurator, es pot executar mitjançant línia de comanament o mitjançant el *script* (amb extensió .sh) que obre la interfície gràfica que es mostra a la Figura 24.

[313] <http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx/>. [Consulta: 09/09/2016]

[314] DROID signature files. <https://www.nationalarchives.gov.uk/aboutapps/pronom/droid-signature-files.htm>. [Consulta: 06/01/2017]

Figura 24. Interfície gràfica de DROID



Font: Captura de l'autor

En el cas d'imatges forenses, podem realitzar l'anàlisi de formats seleccionant la ruta '/media' que es troba dins l'arrel del sistema, ja que és la ruta on es munten per defecte les imatges. Un cop realitzat l'anàlisi, es poden exportar els resultats en informes en els formats PDF, XML per a Planets (vegeu el capítol 4.2.1), text pla, XML i HTML. Com que no genera els informes a partir de la imatge forense, no es recuperarà informació com directoris o fitxers buits, cosa que sí pot fer BitCurator Reports. A la Taula 23 exposem una mostra d'etiquetes XML que pot generar DROID.

Tornant a les comparatives entre resultats de fiwalk i de DROID, es poden consultar els resultats a la Taula 24, on es pot comprovar que hi ha molts formats que fiwalk no reconeix directament, sinó que els considera directament com a "dades", ja que això passa en un total de 13 casos. En canvi, DROID no només reconeix més formats, sinó que ho fa en funció de tres criteris: per extensió de fitxer, per PUID i per MIME. Una possible raó és que BitCurator Reports reconeix als seus informes els fitxers esborrats i no assignats (que els cataloga com a "dades"), cosa que no passa amb DROID, ja que no fa una anàlisi de la imatge forense en si mateixa, sinó de la imatge muntada, per la qual cosa es passen per alt aquests fitxers esborrats.

Taula 23. Mostra d'etiquetes XML generades per DROID

Etiquetes i subetiquetes		Descripció
Title		Títol de l'informe
	Profile Id	Identificador del perfil
	CreatedDate	Data de creació del perfil
	Name	Nom del perfil
Resources		Recursos localitzats a l'anàlisi
Dir		Dades d'un directori concret
File		Dades d'un fitxer concret
	Size	Mida en bytes del directori o fitxer
	LastModifiedDate	Data de darrera modificació del directori o fitxer
	Extension	Extensió del fitxer
	Name	Nom del directori o fitxer
	Uri	URI del directori o fitxer
	Path	Ruta original del directori o fitxer
State		Estat de l'informe; si ha finalitzat, serà FINISHED
SignatureFileVersion		Versió del fitxer de signatures que s'ha utilitzat per fer l'informe
ContainerSignatureFileVersion		Versió del fitxer contenidor de signatures que s'ha utilitzat per fer l'informe
Filter		Informa si s'ha activat o no un filtre
SignatureFileName		Nom del fitxer de signatures
ContainerSignatureFileName		Nom del fitxer contenidor de signatures
EndDate		Data i hora de finalització de generació de l'informe
StartDate		Data i hora d'inici de generació de l'informe
GenerateHash		Indica si s'ha activat o no la generació de valors *hash*
HashAlgorithm		Tipus d'algoritme *hash* que s'ha utilitzat
ProcessArchiveFiles		Indica si s'han processat o no els fitxers contenidor com .rar
ProcessWebArchiveFiles		Indica si s'han processat o no els fitxers contenidor d'arxiu web
Group		Valors que identifiquen el format de fitxer
	Values	Valors identificats d'un format en concret; poden ser del tipus MIME o PUID
	ProfileSummaries	Estadístiques de mida en bytes d'un format concret

Font: Informe generat en format XML per l'autor

Taula 24. Formats i nombre de fitxers reconeguts a fiwalk i a DROID

Format de fitxer	fiwalk	DROID
Adobe Photoshop	1	9
CSS	-	1
CSV	-	7
Dades	780	-
EICAR virus test file	1	-
HTML	-	1
Interleaf	-	4
JPEG	1	11
MPEG-1	2	2
MS Excel	37	246
MS Office Encrypted Document	5	33
MS Office Owner File	-	1
MS Power Point	-	9
MS Word	7	86
OLE2 Compound Document Format	-	46
PDF	55	183
PDF/A	-	3
PDF/X	-	1
PNG	1	32
RTF	10	10
Stationery for Mac OS X	-	4
Text pla	16	24
Text pla ISO-8859	3	-
VML	-	3
WAV	1	1
WordPerfect	-	4
XML	29	5
ZIP	-	45

Font: L'autor, a partir d'una imatge forense en format EnCase

Per tant, es podria concloure que DROID dóna informació molt més acurada sobre els formats de fitxer. Però, un resultat molt interessant és que DROID no detectà un fitxer de prova EICAR, dissenyat per posar els programaris antivirus sense posar en risc el sistema amb virus autèntics. En canvi, BitCurator Reports sí que va reconèixer aquest fitxer (en forma de text pla amb extensió .txt). Atès que no és viable poder fer proves amb tots els tipus de formats de fitxer existents, la nostra conclusió és que el millor procediment és executar els dos programaris d'anàlisi i comparar resultats.

Per altra banda, BitCurator té en desenvolupament una eina per redactar imatges forenses que permet, si es desitja, bloquejar fitxers concrets o bé redactar contingut. Aquesta eina és el programari bitcurator_access_redaction, desenvolupat pel projecte

BitCurator Access (vegeu el capítol 4.2.6). Per fer-lo servir, és necessari descarregar-lo al repositori GitHub[315] i instal·lar-lo segons les instruccions indicades. Serà necessari tenir en compte que la redacció només serà possible amb una imatge *raw*, no amb la imatge forense. Per tant, haurem de recórrer novament a Guymager i crear una nova imatge, en aquest cas en format *raw*.

Un cop creada la imatge, s'ha d'executar el comanament 'redact-cli' al Terminal (la interfície per introduir comanaments a Ubuntu) amb les opcions corresponents, juntament amb un fitxer .txt de configuració, i podrem bloquejar fitxers concrets, redactar dades mitjançant expressions regulars, sobreescriure el contingut de fitxers amb zeros i ignorar fitxers concrets per evitar la seva redacció. Explicarem amb més detall el seu funcionament al capítol 5.2.4, però a tall orientatiu indiquem les possibilitats de redacció a la Taula 25.

Taula 25. Opcions de redacció de continguts amb bitcurator_access_redaction

Opció	Funció
FILE_NAME_MATCH	Busca un fitxer determinat dins l'estructura de fitxers i carpetes
FILE_MD5	Busca un fitxer amb un valor *hash* MD5 determinat dins l'estructura de fitxers i carpetes
FILE_SHA1	Busca un fitxer amb un valor *hash* SHA1 determinat dins l'estructura de fitxers i carpetes
SEQ_EQUAL	Busca una seqüència de text determinada
SEQ_MATCH	Busca una expressió regular
FILE_NAME_MATCH	Busca noms de fitxer amb un patró
FILE_DIRNAME_EQUAL	Busca tots els fitxers dins un directori concret
FUZZ	Evita l'execució (només per a fitxers .exe i .dll de l'entorn Windows)
FILL	Sobreescriu tot el contingut amb un codi ASCII
SCRUB	Sobreescriu tot el contingut amb zeros
IGNORE	Ignora fitxers
COMMIT	Executa la redacció

Font: BitCurator Access (2016, Dec. 13). *bitcurator-access-redaction: quick start guide*. <http://wiki.bitcurator.net/downloads/BCR-Quickstart.pdf>. [Consulta: 31/12/2016]

[315] <https://github.com/bitcurator/bitcurator-access-redaction>. [Consulta: 31/12/2016]

4.3.4 Exportació de metadades

BitCurator permet exportar metadades tècniques DFXML mitjançant fiwalk (més informació al capítol 3.3.2) que es poden incorporar altres estàndards de metadades com METS, PREMIS o EAD (Woods; Chassanoff; Lee, 2013) i també ajuda a configurar elements del model OAIS, com el Paquet d'Informació d'Enviament. A més, es generen metadades PREMIS que recuperen informació de preservació sobre el programari que s'ha utilitzat per crear la imatge i les anàlisis que s'han realitzat.

Figura 25. Mostra de metadades PREMIS generades per BitCurator

Font: L'autor, a partir d'una imatge forense EnCase

Si ens fixem en la Figura 25, podem veure que ens recupera una Entitat Objecte, amb el seu UUID, i l'Esdeveniment Captura, amb el detall de la ruta original de la ubicació de la imatge forense, l'hora de l'Esdeveniment, el programari que s'ha utilitzat i la mida de la imatge. Altres Esdeveniments que no es mostren a la figura són l'anàlisi de sistema de fitxers, que ens recupera el fitxer XML amb metadades DFXML que conté aquesta informació i el de l'extracció d'informació, que ens recupera la versió del programari que s'ha utilitzat (en aquest cas, Bulk Extractor v. 1.6.0). Val a dir que aquesta generació de metadades PREMIS encara pot enregistrar altres Esdeveniments, ja que al SAA Research Forum es va preveure un d'addicional, el de redacció d'informació privada i sensible[316]. Aquesta funció, però, encara no es troba integrada al flux general

[316] Chassanoff, Alexandra; Woods, Kam; Lee, Christopher A. (2013, Aug. 13) *Mapping digital forensics metadata to preservation events using BitCurator* [pòster]. <https://goo.gl/wq8dgb>. [Consulta: 11/12/2016]

de treball de BitCurator, tot i que sí que és possible realitzar-la amb les eines del projecte BitCurator Access (vegeu el capítol 4.2.6).

4.3.5 Síntesi

Tal com s'ha vist fins ara, el sistema BitCurator permet executar les següents operacions:

- Crear imatges forenses en tres formats (RAW, EnCase i AFF) amb fitxers d'informació associats que es poden exportar a altres estàndards de metadades
- Muntar imatges i navegar dins l'estructura de fitxers i carpetes
- Generar informes XML de fiwalk i de Bulk Extractor amb metadades DFXML
- Generar informes d'extracció de dades privades i sensibles amb Bulk Extractor
- Generar informes del suport original, que ens indiquen el sistema de fitxers, el nombre de fitxers i els formats de fitxer
- Fer identificació de formats de fitxer
- Generar metadades PREMIS relacionades amb la creació de la imatge

Es pot concloure que és un sistema prou eficaç de preservació, ja que permet comprovar fàcilment la integritat de les dades, detecció d'errors de lectura, comprovar virus, conèixer els formats de fitxer presents i extreure metadades de forma directa. Hem de recordar, no obstant, que DROID té una millor eficàcia que BitCurator Reports per reconèixer formats, així que recomanem utilitzar-lo per a aquesta finalitat. Una mostra de l'eficàcia de BitCurator és que ja s'ha utilitzat com a mínim en un cas judicial real, ja que va permetre aportar proves en un cas de frau de targetes de crèdit a la ciutat de San Luis Obispo, Califòrnia. Val a dir que l'examinador va tenir molt poc temps per poder fer la investigació forense, ja que només va poder disposar de la prova (un disc dur de 250 GB) fins el dia abans de l'audició preliminar (Bradley; Garfinkel, 2015). L'eina utilitzada fou bulk_extractor, que en només dues hores i mitja va localitzar evidències suficients (més de 10.000 números de targetes de crèdit) per empresonar els acusats.

5. Model de preservació de dades de recerca

Un cop s'han presentat les diferents funcionalitats que es poden aplicar al model de preservació, presentem la manera en què crearem aquest model amb explicacions detallades pas per pas. Primerament, indiquem el tipus de maquinari necessari per operar amb l'entorn BitCurator; en segon lloc, expliquem els processos en detall i els fluxos de treball corresponents a la creació i la ingesta de l'AIP; en tercer lloc, indiquem els processos en el cas de la preparació del DIP per al seu accés i en últim lloc, introduïm el programari de repositoris DSpace i justifiquem el seu ús dins la nostra proposta.

5.1 Adquisició de maquinari i programari

Abans de començar amb els processos, serà necessari l'adquisició del maquinari i del programari necessaris per preparar l'entorn de treball. Per tal de justificar millor l'aplicació de BitCurator al nostre model, hem realitzat un estudi de les solucions comercials existents i a continuació l'hem comparat amb BitCurator.

Dins l'anàlisi forense digital, el maquinari més potent al mercat el trobem a la sèrie d'estacions forenses FRED distribuïda per l'empresa Digital Intelligence, que ofereix unes opcions tècniques molt valuoses. La seva versió més bàsica[317] permet adquirir dades directament des de discs durs IDE, EIDE, ATA, SATA, ATAPI, SAS, Firewire i USB, a més de discs òptics (Blu-Ray, CD-ROM i DVD-ROM) i targetes de memòria (Compact Flash, Micro Drive, Smart Media, Memory Stick, Memory Stick Pro, xD Card, Secure Digital Media i MultiMedia Card). A més, incorpora dos sistemes operatius, Windows 10 i SUSE Professional Linux de 64 bits, amb sis *write blockers* de sèrie per a interfícies IDE, SATA, SAS, USB, Firewire i PCI. Compta amb un processador Intel Core i7-6800K a 3.4 GHz i una memòria RAM de 32 GB a 2133 MHz. Quant a l'emmagatzematge, la sèrie FRED té tres discs durs, dos de tipus sòlid i un de mecànic. Un dels discs durs sòlids s'utilitza per als sistemes operatius, l'adquisició forense i les eines de processat, l'altre disc sòlid per a dades temporals i de memòria cau i el disc dur mecànic s'empra per a l'espai de treball. Una altra avantatge de FRED és

[317] <http://www.digitalintelligence.com/products/fred>. [Consulta: 08/10/2016]

l'ús de safates per poder muntar i desmuntar de forma senzilla i ràpida els discs durs interns.

Quant el programari, hem estudiat FTK i EnCase perquè són *suites* populars dins l'anàlisi forense digital (Garfinkel, 2013; John, 2012; Vandeven, 2014). En el primer cas, es tracta d'una *suite* integrada de l'empresa AccessData[318] que permet la creació d'imatges forenses, anàlisi del registre del sistema, anàlisi gràfica de correus electrònics, desencriptació de fitxers PDF, recuperació de contrasenyes, biblioteca de valors *hash* i anàlisi avançada i automatitzada de la memòria volàtil. Una de les opcions més interessants és la visualització d'informació, ja que genera informes de diferents tipus, com gràfics circulars o amb clústers, en format HTML, PDF, XML o RTF. Una altra característica important és que no experimenta pèrdues d'informació degut al consum de recursos, ja que FTK està construït com a base de dades. El sistema de cerca utilitza índexs per trobar expressions regulars, amb reconeixement de més 700 formats de fitxers, múltiples sistemes de fitxers com Ext4, exFAT o Veritas File System. Incorpora també la tecnologia Explicit Image Detection per reconèixer de forma automàtica imatges pornogràfiques.

En el cas d'EnCase es tracta d'un producte de l'empresa Guidance Software, la qual ofereix diversos productes per a la investigació forense. En el cas que ens ocupa, el més adient seria EnCase Forensic, que també s'ha utilitzat per resoldre casos criminals[319]. Les seves funcionalitats són molt semblants als de FTK, amb la gran diferència que la seva interfície no és tan intuïtiva com la de FTK, i per tant requereix una major quantitat de temps de formació per dominar l'eina. Per aquesta raó, Guidance Software ofereix una sèrie de cursos amb tres nivells: introductori, intermedi i expert.

Hem elaborat les taules Taula 26 i Taula 27 per il·lustrar el pressupost que necessitaria un centre que optés per la via de maquinari i programari comercials. En el cas d'optar per l'opció més bàsica, una estació FRED-L system i una llicència EnCase per un any,

[318] <http://accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk>. [Consulta: 08/10/2016]

[319] Taub, Eric A (2006, Apr. 5). "Deleting may be easy, but your hard drive still tells all". *New York Times.* <http://www.nytimes.com/2006/04/05/technology/techspecial4/05forensic.html?_r=1&ref=techspecial4>. [Consulta: 08/10/2016]

es requeriria una inversió de 7.887 euros, una quantitat que no totes les institucions es poden permetre.

Taula 26. Mostra de maquinari especialitzat forense

Maquinari	Tipus	Preu (no inclou IVA ni despeses d'importació)
FRED-L system	Portàtil	4.679 euros
FRED system	Estació de treball bàsica	5.489 euros
FREDDIE system	Estació de treball portàtil	7.415 euros
FRED DX 2R system	Estació de treball avançada	10.662 euros

Font: Digital Intelligence. *Forensic workstations.* <https://goo.gl/7WCuGm>. [Consulta: 08/10/2016]

Taula 27. Mostra de programari especialitzat forense

Programari	Tipus	Preu (no inclou IVA ni despeses d'importació)
EnCase Forensic v8	*Suite* d'investigació forense	• 3.208 euros (amb servei de manteniment per un any) • 3.775 euros (amb servei de manteniment per dos anys) • 4.108 euros (amb servei de manteniment per tres anys)
FTK 6.0	*Suite* d'investigació forense	• 4.739 euros (llicència perpètua amb un any de servei de manteniment) • 5.569 euros (llicència perpètua amb dos anys de servei de manteniment) • 6.192 euros (llicència perpètua amb tres anys de servei de manteniment)

Fonts: Digital Intelligence. *EnCase Forensic v8* <https://goo.gl/ycBWF1>. [Consulta: 08/10/2016]; Forensic Store. *Forensic Toolkit 6.0* <http://forensicstore.com/product/forensic-software/forensic-toolkit-5>. [Consulta: 08/10/2016]

En canvi, en el cas d'utilitzar l'entorn BitCurator es requeriria l'adquisició d'una estació de treball que complís amb les característiques recomanades pel BitCurator Consortium; per tant, hauria d'incloure un processador Intel Core i7 amb 16 GB de RAM i un disc dur sòlid de 256 GB. Per tal de ser previsors amb l'arribada de grans volums de dades, hem seleccionat un ordinador amb un processador Intel i7 amb 32 GB de RAM, un disc dur SSD de 480 GB i dos discs durs SATA de 2 TB cadascun. Dins el maquinari a adquirir s'hauria d'afegir un *docking station* (Wilderbeek; Térmens, 2015) compatible per a l'entrada de discs durs interns amb interfícies d'entrada SATA o IDE per a discs durs de 2 ½ i 3 ½ polzades; un que compleix amb aquestes característiques és l'UNIDOCK2U. A més, serà necessari també adquirir una càmera rèflex per

documentar l'estat físic dels suports; en aquest cas, la Nikon D3300 té unes prestacions que inclouen una resolució de 24.2 megapíxels, possibilitat de connexió WiFi, connectivitat USB i una funció d'estabilització d'imatge ideal per a fotografia a curta distància. Opcionalment, també es pot adquirir un *write blocker* per a dispositius USB si la institució ho considera pertinent. Indiquem el maquinari recomanat a la Taula 28, on el total de pressupost necessari seria de 2.620,59 euros.

Taula 28. Llistat de maquinari recomanat per a l'entorn de treball BitCurator

Maquinari	Tipus	Preu
PcCom WorkStation III Intel i7-6800K/32GB/4TB+480GB SSD/GTX1060	Estació de treball d'ús general	1.929 euros
UNIDOCK2U	*Docking station* amb connexions SATA i IDE	81,59 euros
Nikon D3300	Càmera rèflex	356 euros
Wiebetech USB WriteBlocker	*Write blocker*	254 euros

Fonts: PC Componentes. *PcCom WorkStation III Intel i7-6800K/32GB/4TB+480GB SSD/GTX1060*. <https://goo.gl/iokPjk>. [Consulta: 08/10/2016]; StarTech. *USB to SATA IDE hard drive docking Station for 2.5in or 3.5in HDD dock*. <https://goo.gl/TV5ZbH>. [Consulta: 08/10/2016]; PC Componentes. *Nikon D3300 24.2 MP Solo Cuerpo*. <https://goo.gl/xmbXjX>. [Consulta: 08/10/2016]; Amazon UK. *Wiebetech USB writeblocker*. <https://goo.gl/0rHsZ1>. [Consulta: 08/10/2016]

Si comparem una estació FRED amb l'estació de treball que indiquem a la Taula 28, és evident que aquesta última és una unitat més senzilla quant a prestacions i en conseqüència el seu preu és molt més baix. Encara que no compta amb tantes funcionalitats com les estacions FRED, el seu ús és coherent dins les limitacions de la nostra proposta, atès que compta amb una unitat DVD-ROM de sèrie i quatre entrades per a dispositius USB.

Tot i que l'ús de programari comercial pot ser profitós, no totes les institucions es poden permetre l'adquisició de llicències EnCase i FTK (Lee; Woods, 2011). Per altra banda, hem de recordar que aquests programaris es van dissenyar per a la investigació criminal i aquest no és l'ús primordial que es requereix, sinó més aviat la possibilitat de capturar dades de forma fiable, fer el seu anàlisi i permetre l'accés ulterior i tal i com s'ha demostrat al capítol 4.3, aquestes funcions les compleix BitCurator amb solvència. Tal i com recomana el BitCurator Consortium, s'haurà de descarregar i instal·lar la versió en imatge ISO, ja que treballarem en un entorn de producció. Per instal·lar la imatge en

l'estació de treball, es recomana crear una memòria USB autoarrancable amb la imatge ISO[299].

5.2 Preparació de l'AIP

Per tal de deixar clar els processos des del principi, s'han establert uns límits dins la nostra proposta, ja que no s'ha dissenyat per satisfer tots els casos possibles de preservació de dades de recerca, sinó alguns concrets. Per tant, els procediments de treball s'han construït tenint en compte aquests límits, que es detallen a la Taula 29.

Taula 29. Límits de la proposta de preservació

Característica	Límit
Àmbit temàtic	Ciències socials i humanitats
Tipus de fitxers admesos	Text, imatges rasteritzades, gràfics vectorials, imatges 3D, àudio, vídeo, dades GIS, dades tabulars i text amb llenguatge de marques
Enviament del SIP	*Offline*; el dipositant enviarà les seves dades en un suport físic
Enviament del Submission Agreement	*Online* i *offline*; la institució facilitarà el formulari de lliurament de dades a l'investigador i el lliurarà omplert i imprès a la institució per signar-lo
Programari del repositori	DSpace
Tipus de repositori	Institucional
Consulta directa o indirecta després de petició	Indirecta
DSpace dedicat o integrat en el repositori Open Access	Dedicat
Suports admesos	Memòries USB, discos òptics, discs durs interns (SATA i IDE) i discs durs externs USB

Font: L'autor

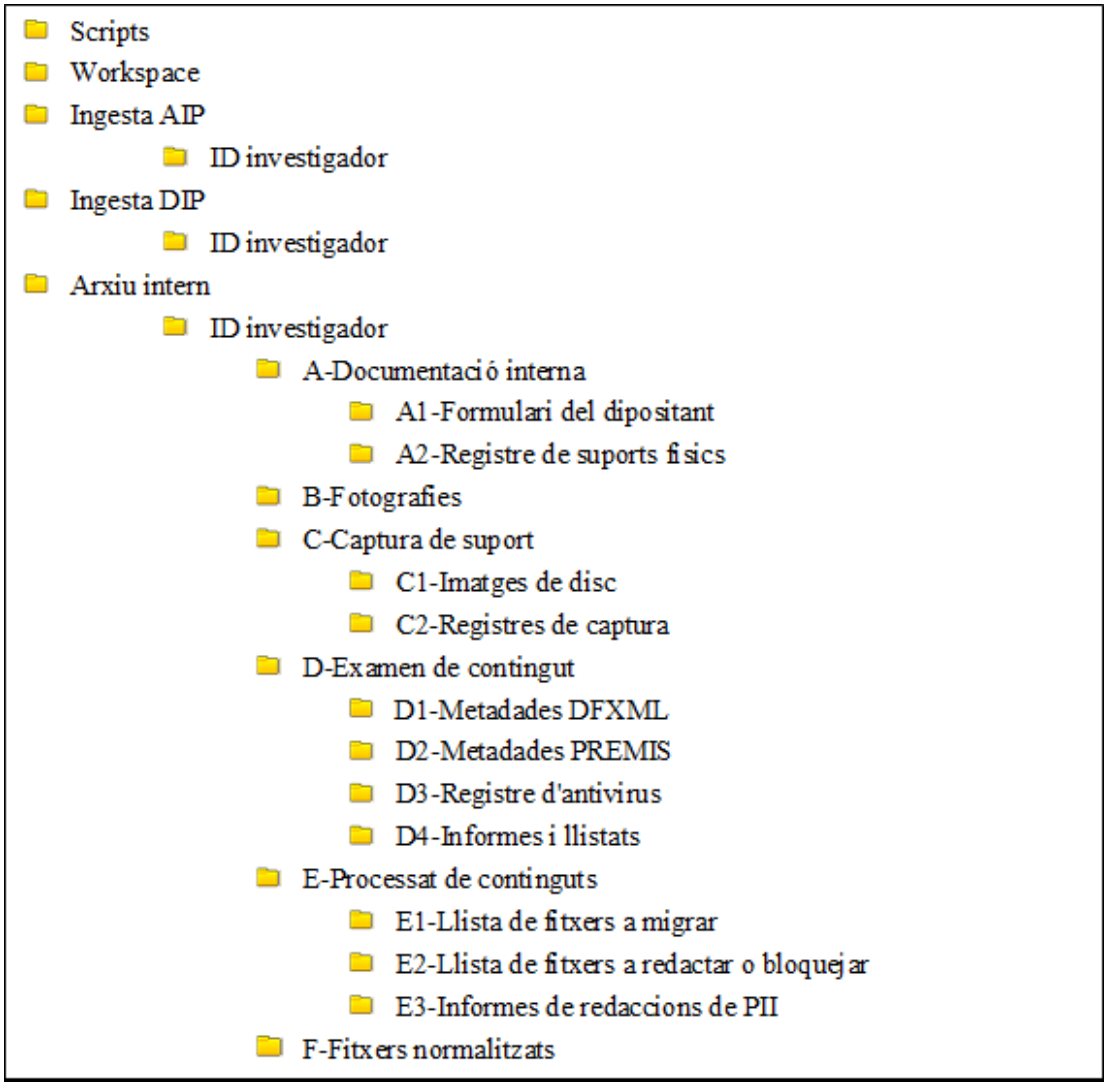
Com ja s'ha explicat al capítol 1.2.4, l'àmbit temàtic de la proposta s'ha limitat a les ciències socials i les humanitats degut als nostres millors coneixements d'aquestes disciplines, raó per la qual hem especificat formats de fitxer admesos al model (això no vol dir que altres formats es trobin exclosos), en funció de les tipologies de formats de fitxers que mostrem a la Taula 10. L'enviament del SIP no es podrà fer en línia degut a la gran mida que estimem que tindrien les dades i per tant haurà de ser l'investigador qui dipositi el seu(s) suport(s) a la institució, ja sigui presencialment o bé per un servei

de paqueteria i missatgeria. Quant al Submission Agreement, en un primer moment la institució facilitarà a l'investigador el formulari en línia (que pot ser en un document Word o bé un formulari web per facilitar la gestió) per tal que l'ompli amb antelació, l'imprimeixi, el signi i el lliuri a la institució; igual que amb el(s) suport(s), podrà fer-ho presencialment o bé per un servei de missatgeria. El programari del repositori serà DSpace per les raons explicades al capítol 1.2.4 i estaria dissenyat com a repositori institucional per recollir les dades de recerca dels seus investigadors. La consulta de les dades de recerca al repositori seria indirecta; això vol dir que l'usuari (Consumidor en terminologia OAIS) no podrà accedir als AIPs originals, sinó que haurà de demanar al repositori (ja sigui per correu electrònic, telèfon o altres vies disponibles) que faciliti el seu accés als continguts, la qual cosa requerirà que es prepari un DIP, del qual donem més detalls al capítol 5.3. Per altra banda, el repositori ha de ser dedicat i no integrat, perquè el gran volum de dades que gestionaria el nostre model complicaria en gran mesura una integració a un repositori institucional preexistent, a més que els requeriments de maquinari i programari que necessitem són de un nivell molt més elevat que el d'un repositori de publicacions en obert. Per tant, el repositori del nostre model hauria de ser dedicat per a dades de recerca. Per últim, els suports admesos que s'indiquen s'han seleccionat en funció de dos criteris: la seva compatibilitat amb maquinaris actuals (ja sigui de forma directa o mitjançant un *docking station*) i la seva gran capacitat d'emmagatzematge. S'han exclòs suports com disquets perquè es consideren obsolets (Breeding, 2012; Layne et al., 2012; Schumacher; VandeCreek, 2015) i perquè la seva capacitat és molt inferior als suports acceptats.

Així doncs, descrivim a continuació els passos que seguirà el nostre model amb la utilització de l'entorn BitCurator, on hem agafat de base els casos que hem exposat al capítol 4.3 i tenint en compte el model de referència OAIS. El maquinari que hem utilitzat ha estat el mateix que el del capítol 4.3 i en aquest cas, amb la versió de BitCurator 1.7.98 publicada el 22 de desembre de 2016.

Hem dividit els processos en sis parts: preparatius inicials, captura de les dades, examen i anàlisi de les dades, processat de continguts, preparació de paquets i la ingesta final al repositori, seguit d'un capítol dedicat a com s'ha modelat l'accés als continguts.

Figura 26. Estructura de carpetes a l'espai de treball



Font: L'autor

S'ha tingut en compte la possibilitat de dos escenaris quan es rebi el SIP: o bé una part dels continguts tindran accés restringit o bé el contingut serà totalment obert, ja que pot ser que l'investigador (bé per requeriments de l'agència de finançament o bé per iniciativa pròpia) hagi realitzat els passos necessaris per protegir el dret a la privacitat dins les dades de recerca. Això canvia els procediments de treball dins el model pel que respecta a l'examen de contingut, el processat i la preparació de paquets per a la ingesta. Per aquest motiu, s'han indicat els passos a realitzar en cada cas quan ha estat necessari.

Durant els diferents processos, es crearan diferents ítems digitals. Alguns seran de caire administratiu, però altres seran de contingut tècnic, com metadades o registres de processos. Tots aquests processos es realitzaran en una estació de treball local mitjançant l'ús de tres directoris que es crearan dins el sistema BitCurator a la ruta

'/home/bcadmin', com a espai intern de treball per a aquests documents. Val a dir que no tots els ítems s'ingestaran finalment al repositori, atès que alguns es dissenyaran exclusivament per a la Gestió i no per a la consulta dels Consumidors. Per tal de tenir una bona organització dins els fitxers d'aquest espai de treball, s'ha creat una estructura de carpetes en funció dels tipus de documents implicats en els processos i que mostrem a la Figura 26.

Per altra banda, treballarem d'inici amb la carpeta 'Scripts', on es guarden tots els *scripts* (programats en el llenguatge Bash) necessaris per executar el flux de treball. Aquests *scripts* no han vingut preinstal·lats a BitCurator, sinó que s'han creat al llarg de l'elaboració del model. Expliquem a la Taula 30 el nom de cada fitxer i la seva funció.

Taula 30. Fitxers inclosos a la carpeta 'Scripts'

Nom de fitxer	Funció
0-Creació d'estructura de carpetes.sh	Crea l'estructura de carpetes necessària per al flux de treball a la ruta '/home/bcadmin/'
1-Guymager.sh	Executa la interfície gràfica del programari Guymager
2-Muntar imatge forense.sh	Munta la imatge forense
3-ClamAV.sh	Executa l'anàlisi de ClamAV dins els fitxers muntats a la ruta '/media'
4-bulk_extractor.sh	Executa l'anàlisi de bulk_extractor a la imatge forense 'imatge.E01'
5-BitCurator_Reports.sh	Executa la interfície gràfica de BitCurator Reports
6-DROID.sh	Executa la interfície gràfica de DROID
7-Desmuntar imatge forense.sh	Desmunta la imatge forense
8-Watermark.sh	Afegeix una marca d'aigua a informes creats per DROID i BitCurator Reports en funció de la configuració indicada a 'stamp.ps'
9-Redacció de fitxers.sh	Executa bitcurator_access_redaction en funció de la configuració indicada a 'imatge_raw_config.txt'
10-Muntar imatge redactada.sh	Munta la imatge *raw* redactada 'imatge_raw_redacted.dd'
11-Desmuntar imatge redactada.sh	Desmunta la imatge *raw* redactada 'imatge_raw_redacted.dd'
12-Disk Image Access Interface.sh	Executa la interfície gràfica Disk Image Access Interface
13-Migration.sh	Migra formats PNG i PSD a TIFF
14-Reanomenar i moure fitxers.sh	Canvia de nom tots els fitxers generats al nom normalitzat i els mou o copia a les rutes finals
15-BagIt.sh	Genera els paquets BagIt
imatge_raw_config.txt	Fitxer de suport per redactar contingut amb dades confidencials
stamp.ps	Fitxer de suport per crear marques d'aigua

Font: L'autor

Dins l'estructura de carpetes, hi ha cinc de base: 'Scripts', on es trobaran els *scripts* ja creats per poder operar ràpidament en el flux de treball; 'Workspace', que utilitzarem per guardar els ítems digitals que hem generat; 'Ingesta AIP', que utilitzarem per dipositar els ítems ja llestos per a la ingesta al repositori; 'DIP', per als paquets d'accés que es crearan en resposta a les peticions dels usuaris i 'Arxiu intern', on guardarem tots els ítems generats durant el flux de treball amb normalització de noms de fitxer, dels quals alguns seran preservats al repositori i altres seran utilitzats per a la gestió interna i per tant no seran preservats.

Per normalitzar els noms dels fitxers, tal com es pot consultar a la Taula 31, s'ha seguit un patró consistent en els següents criteris, basats en bones pràctiques de les Stanford University Libraries[320]:

- Un prefix amb la data en què s'ha creat el document o fitxer amb el format YYYYMMDD
- Un cos central amb l'identificador de l'investigador o del suport
- Un sufix que pot ser dels següents tipus:
    - Tipus de document o fitxer. S'han utilitzat termes en anglès perquè l'ús de termes en català requeriria de un nombre excessiu de caràcters en el nom de fitxer
    - Nom del programari utilitzat per crear el document o fitxer
    - Tipus del programari utilitzat per crear el document o fitxer
    - Format de metadades en què es troba al document o fitxer

No utilitzarem aquesta normalització de fitxers directament, ja que en darrera instància utilitzarem un *script* que canviarà de nom els fitxers i els mourà a les rutes corresponents. Per tant, tots els fitxers els guardarem primer a 'Workspace' amb un nom concret, que indiquem a la Taula 31 i els deixarem en aquesta ruta durant tots els passos del flux de treball fins executar el *script*. Hem optat per aquesta forma de treballar perquè BitCurator treballa millor amb directoris sense espais ni guions, i per altra banda és més senzill treballar amb un sol directori.

[320] *Best practices for file naming.* <https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming>. [Consulta: 04/12//2016]

Taula 31. Explicació dels continguts a l'estructura de fitxers i carpetes

Directori	Descripció del contingut	Estructura de fitxer/directori	Comentaris
Scripts	*Scripts* per executar línies de comanament	Programari.sh o acció (P. ex., 'clamav.sh'; 'muntar imatge forense.sh')	
Workspace	Fitxers generats durant el flux de treball sense normalització de nom		
Ingesta AIP	Paquets AIP que es preservaran al repositori	Conté el subdirectori 'ID investigador'	Els paquets AIP seran fitxers .zip segons l'especificació BagIt
DIP	Continguts que formaran part del DIP	Conté el subdirectori 'ID investigador'	Els DIP estaran formats per fitxers i carpetes individuals d'accés obert
Arxiu intern	Fitxers generats durant el flux de treball amb normalització de nom	Conté el subdirectori 'ID investigador'	
ID investigador	Codi identificador de l'investigador que diposita les dades	Seqüència alfanumèrica (P. ex., 12345678). Conté subdirectoris A, B, C, D, E i F	
A-Documentació interna	Documentació d'ús intern de la institució	Conté subdirectoris A1 i A2	
A1-Formulari del dipositant	Conté formularis de lliurament de dades (model a l'annex A)	YYYYMMDD-ID investigador-Form (P. ex., 20161115-12345678-Form)	El guardarem inicialment amb el nom de fitxer 'Form.pdf' dins 'Workspace'
A2-Registre de suports físics	Conté registres de suports físics	YYYYMMDD-ID investigador-Medialog (P. ex., 20161115-12345678-Medialog)	El guardarem inicialment amb el nom de fitxer 'Medialog.ods' dins 'Workspace'
B-Fotografies	Conté fotografies dels suports físics	YYYYMMDD-ID suport_XX (P. ex., 20161115-12345678USB01_01, 20161115-12345678USB01_02...)	Els guardarem inicialment amb els noms de fitxer els noms de fitxer 'ID suport_01.jpg', 'ID suport_02.jpg'... dins 'Workspace'
C-Captura de suport	Fitxers generats per Guymager	Conté subdirectoris C1 i C2	
C1-Imatges de disc	Conté els fitxers d'imatge de disc	YYYYMMDD-ID suport (P. ex., 20161115-12345678USB01)	El guardarem inicialment amb el nom de fitxer 'imatge.E01' i (si escau) 'imatge_raw.dd' dins 'Workspace'

Directori	Descripció del contingut	Estructura de fitxer/directori	Comentaris
C2-Registres de captura	Conté els fitxers de registre de captura forense	YYYYMMDD-ID suport (P. ex., 20161115-12345678USB01)	El guardarem inicialment amb el nom de fitxer 'imatge.info'
D-Examen de contingut	Fitxers generats per l'examen de contingut	Conté subdirectoris D1 a D4	
D1-Metadades DFXML	Conté fitxers de metadades DFXML	YYYYMMDD-ID suport-programari-DFXML (P. ex., 2016115-12345678USB01-fiwalk-DFXML)	Generats per bulk_extractor i per BitCurator Reports; es guardaran inicialment a les rutes ~/Workspace/bulk_extractor i ~/Workspace/bitcurator_reports
D2-Metadades PREMIS	Conté fitxers de metadades PREMIS	YYYYMMDD-ID suport-PREMIS (P. ex., 20161115-12345678USB01-PREMIS)	Generat per BitCurator Reports; es guardarà inicialment a la ruta ~/Workspace/bitcurator_reports
D3-Registre d'antivirus	Conté fitxers de registre d'antivirus	YYYYMMDD-ID suport-clamav (P. ex., 20161115-12345678USB01-clamav)	Aquest fitxer és generat directament per un *script*
D4-Informes i llistats	Conté informes de fiwalk i Bulk Extractor	YYYYMMDD-ID suport-programari-TYPE (P. ex., 20161115-12345678USB01-fiwalk-Deleted). TYPE té les següents categories: -'Report' per a informes tècnics -'Formats' per a llistats de formats -'Filelist' per a llistats de fitxers	Generats per BitCurator Reports; es guardaran a la ruta ~/Workspace/bitcurator_reports Els fitxers generats per DROID es guardaran com a 'droid.xml' i 'droid.pdf'
E-Processat de continguts	Fitxers generats pel processat de continguts	Conté subdirectoris E1 a E3	
E1-Llista de fitxers a migrar	Conté la llista de fitxers que s'han de migrar a altres formats	YYYYMMDD-ID suport-Migration	El guardarem inicialment amb el nom de fitxer 'Migration.ods' dins 'Workspace'
E2-Llista de fitxers a redactar o bloquejar	Conté la llista de fitxers que s'han de redactar o bloquejar	YYYYMMDD-ID suport-Redaction	El guardarem inicialment amb el nom de fitxer 'Redaction.ods' dins 'Workspace'

Directori	Descripció del contingut	Estructura de fitxer/directori	Comentaris
E3-Informes de redaccions de dades confidencials	Conté informes de redaccions de fitxers degut a presència de dades confidencials	YYYYMMDD-ID suport-programari (P. ex., 20161115-12345678USB01-bitcurator-access-redaction)	Aquest fitxer és generat directament per un *script*
F-Fitxers normalitzats	Fitxers normalitzats o migrats de la imatge forense	Mateixos noms que els originals, amb el sufix '_migrated'	Aquest fitxers és generen per un *script*

Font: L'autor

Un cop haguem acabat tots els processos, els fitxers que s'hauran de preservar al repositori es guardaran a una carpeta diferent, també a la ruta '/home/bcadmin/, que rebrà el nom d''Ingesta AIP'. La raó de no tenir-los dins 'Workspace' és que els fitxers a preservar seran empaquetats segons l'especificació BagIt, i guardar-los a una altra ruta facilitarà els processos. Mostrem l'estructura de carpetes a la Figura 27 i comentem amb més detall el contingut de les carpetes a la Taula 32.

Figura 27. Estructura de carpetes per a la ingesta d'AIPs

Font: L'autor

Taula 32. Explicació dels continguts a l'estructura de fitxers i carpetes a l'espai intern d'ingesta dels AIPs

Directori	Descripció	Estructura de fitxer/directori	Comentaris
ID investigador	Codi identificador de l'investigador que diposita les dades	Seqüència alfanumèrica (P. ex., 12345678). Contindrà fitxers de paquets BagIt amb l'estructura 'ID suport_AIP.zip'	Els fitxers .zip es generen amb un *script*

Font: L'autor

De manera semblant, utilitzarem la carpeta 'DIP' per als continguts oberts que es difondran en resposta a les peticions d'accés dels usuaris. Mostrem la seva estructura a la Figura 28 i expliquem el seu contingut a la Taula 33.

Figura 28. Estructura de carpetes per als DIPs

Font: L'autor

Taula 33. Explicació dels continguts a l'estructura de fitxers i carpetes a l'espai intern d'ingesta dels DIPs

Directori	Descripció	Estructura de fitxer/directori	Comentaris
ID investigador	Codi identificador de l'investigador que diposita les dades	Seqüència alfanumèrica (P. ex., 12345678). Contindrà fitxers considerats com d'accés obert	Els fitxers de contingut obert s'exportaran de les imatges forenses o imatges *raw* redactades, si escau

Font: L'autor

5.2.1 Preparatius inicials

Els preparatius inicials, dintre del model OAIS, correspondrien amb les negociacions amb el Productor per preparar el Paquet d'Informació d'Enviament (o SIP en terminologia OAIS en anglès), el qual és definit com la representació d'objectes empaquetats per ser dipositats al repositori.

Tal com hem estipulat a la Taula 29, serà el dipositant (o Productor en terminologia OAIS), qui haurà d'emplenar el formulari (veure Annex B) prèviament i haurà de signar-lo en suport paper. També haurà de dipositar els suports físics presencialment o per missatgeria.

Recepció del formulari i del(s) suport(s)

Gràcies a la informació aportada pel dipositant al formulari de lliurament de dades, tindrem informació suficient per començar a planejar els processos (no és el mateix processar un fons de CD-ROMs que un fons de discs durs). Serà molt important tenir present la mida total en GB dels suports, ja que si aquest és superior en un 75% a l'espai total del nostre espai intern de treball, haurem d'utilitzar vies alternatives com l'ús d'un segon disc dur de més capacitat.

A continuació indiquem cadascun dels camps i justifiquem la seva inclusió al formulari:

- Tipus de suport. El model només contempla el dipòsit de discs durs (interns o externs), memòries USB i discs òptics. Com ja hem comentat més amunt, no s'han considerat suports obsolets com disquets
- Sistemes operatius. Cada sistema operatiu utilitza de forma obligatòria un o diversos sistemes de fitxers. Windows fa servir FAT i NTFS, OS X d'Apple fa servir HFS Plus i Linux, Ext4. Conèixer aquesta informació permetrà resoldre problemes d'accessibilitat, ja que per exemple, el sistema Ext4 no es pot llegir a un sistema Windows
- Tipologies de dades. Al formulari s'ha indicat una tipologia de dades, considerades les més recurrents dins les ciències socials i les humanitats que seran susceptibles de preservació a llarg termini, sense excloure altres tipus de dades
- Formats de fitxer. Com ja s'ha vist al capítol 2.3, hi ha una sèrie de formats de fitxer especialment aptes per a la preservació. Saber si l'investigador ha fet ús d'aquests formats permetrà conèixer si és necessari fer alguna migració en el futur per assegurar la preservació a llarg termini
- Estàndards de metadades. Si dins la col·lecció de dades ja s'ha fet un treball de descripció amb algun estàndard, es podrà aprofitar per migrar-lo o adaptar-lo als estàndards del nostre model
- Estructuració i/o normalització de carpetes i fitxers. Si l'investigador ha utilitzat algun patró per a la nomenclatura de les seves dades, es pot aprofitar aquest patró per adaptar-lo a les directrius de la institució dipositant
- Mida total de les dades. Conèixer la mida total de les dades permetrà saber el temps aproximat d'ingesta i altres qüestions tècniques, com si és necessari dividir la imatge forense en múltiples parts
- Nombre total de fitxers. Conèixer la mida total de les dades permetrà saber el temps aproximat de processat

Els suports de dades acceptats són tres: discs òptics (CDs i DVDs), discs durs (interns de connexió IDE o SATA o bé externs amb connexió USB) i memòries USB. En els casos de dispositius amb connexió USB s'acceptarien aquells conformes a l'especificació 3.0 o inferior.

Creació de l'estructura de carpetes

El primer pas que es farà és la creació de l'estructura de carpetes dins l'espai de treball, d'ingesta i d'arxiu intern. Per fer-ho, hem creat el *script* '0-Creació d'estructura de carpetes.sh' que generarà aquesta estructura de manera automàtica, que mostrem en part a la Figura 29.

Dins el *script*, la primera línia indica al sistema operatiu com s'ha d'executar aquest *script*, que en aquest cas és el Terminal. A la segona línia s'ha indicat en un comentari la funcionalitat del *script* mitjançant el caràcter de coixinet (#), que és la d'executar Guymager. A la tercera línia s'utilitza el comanament 'echo', que s'utilitza per mostrar el text indicat dins el Terminal mentre s'executa el *script*; el text que s'ha utilitzat pretén servir com a explicació tècnica bàsica de com s'executa del *script*. Tots aquests paràmetres (línia que indica la configuració d'execució en Terminal, línia amb caràcter de coixinet, línia amb el comanament 'echo' i línies successives amb comanaments d'execució) es repetiran a la resta de *scripts*.

Figura 29. *Script* utilitzat per crear l'estructura de carpetes de l'espai de treball i d'ingesta (detall)

```
#!/bin/sh

# script per crear estructura de carpetes de treball

echo Es crea una estructura de carpetes de treball a la ruta /home/
bcadmin
mkdir -v -pv "/home/bcadmin/Workspace/Pre-migration"
mkdir -v -pv "/home/bcadmin/Workspace/Post-migration"
mkdir -v -pv "/home/bcadmin/Arxiu intern/ID investigador/A-Documentació
interna/A1-Formulari del dipositant"
mkdir -v -pv "/home/bcadmin/Arxiu intern/ID investigador/A-Documentació
interna/A2-Registre de suports físics"
mkdir -v -pv "/home/bcadmin/Arxiu intern/ID investigador/B-Fotografies"
```

Font: L'autor

Per executar el *script*, s'ha de verificar que tinguem configurat correctament a BitCurator l'opció d'executar els fitxers de text executables, que es troba dins les preferències de fitxer. La ruta completa, dins un navegador de fitxers (que rep el nom de Nautilus dins Ubuntu), seria 'Edit' > 'Preferences' > 'Behavior' > 'Run executable text files when they are opened'. Si es prefereix, es pot marcar l'opció 'Ask each time' per visualitzar la finestra del Terminal i així comprovar el bon funcionament del *script*.

Per tal de crear l'estructura de carpetes, s'ha utilitzat el comanament 'mkdir' amb l'opció '-v' que mostra cadascun dels directoris que es creen dins el Terminal i l'opció '-pv' que serveix per crear subdirectoris si aquests no existeixen amb anterioritat. Per tal que 'mkdir' generi directoris amb espais dins el nom, s'han utilitzar les cometes dins la ruta indicada.

És important tenir en compte que s'haurà d'editar la cadena de text 'ID investigador' segons les necessitats de cada cas, ja que s'haurà de reemplaçar per l'identificador de l'investigador corresponent que dipositi les dades.

Escanejar i arxivar el formulari

Un cop el personal ha comprovat que el formulari es troba emplenament correctament, s'encarregarà d'escanejar-lo per tal de tenir una còpia dins l'espai intern de treball. El fitxer escanejat serà en format PDF amb el reconeixement de caràcters OCR per facilitar la recerca interna del text i es guardarà dins 'Workspace' amb el nom 'Form'.

Un cop estigui guardat el formulari escanejat, l'original en paper s'haurà d'arxivar segons els procediments d'arxiu estàndards de la institució.

Assignar identificador(s) al(s) suport(s)

Cadascun d'aquests suports rebrà un identificador unívoc sense utilitzar guions, que constarà d'un prefix amb l'identificador de l'investigador, unes segles al cos central basades amb el tipus de suport i un sufix amb una xifra que representi el nombre de suports d'aquest mateix tipus que ha rebut la institució. Les segles que s'han definit per als tipus de suport han estat els següents:

- USB: per a memòries USB
- HDINTIDE: per a discs durs interns de connexió IDE
- HDINTSATA: per a discs durs interns de connexió SATA
- HDEXTUSB: per a discs durs externs amb connexió USB
- CD: per a CD-ROMs

- DVD: per a DVD-ROMs

Per exemple, en el cas d'un Productor que hagi dipositat dos CD-ROMs i un disc dur extern USB s'utilitzarien els identificadors: 12345678CD01, 12345678CD02 i 12345678HDEXTUSB01. El següent pas seria crear un teixell amb l'identificador corresponent i pegar-lo al suport.

Fotografiar suport(s)

Es faran com a mínim dues fotografies, una per a la part anterior i una altra per a la part posterior de cada dispositiu, amb fotografies addicionals dels cantons si aporten nova informació, com l'etiqueta del número de sèrie. La càmera més adient per a aquests casos seria una rèflex perquè està dissenyada per recollir imatges de forma que la fotografia seria un resultat pràcticament idèntic a la imatge observada pel fotògraf i per tant, serà molt més senzill tenir imatges que permetin apreciar els detalls físics del suport i així poder documentar-los adequadament. Un cop fetes les dues fotografies mínimes, es guardaran dins la carpeta 'Workspace' amb els noms de fitxer 'ID suport_01.jpg', 'ID suport_02.jpg' i així successivament. Dins el flux de treball, hem considerat que aquestes fotografies haurien d'estar en format JPEG, ja que el seu ús principal serà el de consulta. Per altra banda, hem contemplat al nostre *script* de reanomenament de fitxers un màxim de quatre fotografies; si s'utilitzen més, s'haurà de modificar aquest *script* o bé guardar-les directament a la carpeta 'B-Fotografies' amb la normalització de nom corresponent.

Examen físic del(s) suport(s)

A continuació es farà un examen físic dels suports, on s'enregistrarà la següent informació:

- Tipus de suport. Serà un dels sis que hem definit anteriorment
- Fabricant. Si és visible de forma externa, s'indicarà el fabricant del suport. En cas contrari, s'indicarà "Desconegut"
- ID suport. S'indicarà l'identificador unívoc del suport, tal com s'ha indicat més amunt

- Observacions. Si el suport presenta algun estat físic que pugui presentar dificultats per recuperar contingut, com ratlladures a la superfície d'un disc, s'anotaran en aquesta columna
- Topogràfic. Indicarem el codi que farà servir la institució per la ubicació física on s'arxivarà el suport. Com que aquest topogràfic serà d'ús exclusiu del personal del repositori, no necessita estar basat en cap sistema de classificació existent, sinó que es pot utilitzar un fet a mida

La forma d'enregistrar aquesta informació de manera òptima seria en una base de dades automatitzada. Una altra manera és utilitzar un full de càlcul que ens servirà per il·lustrar els procediments, ja que es pot emplenar de forma provisional i després guardar les dades en format CSV, el qual permet importar fàcilment dades a bases de dades. Amb aquest fi, es pot fer servir la *suite* ofimàtica LibreOffice, el qual inclou Calc per crear i editar fulls de càlcul. No cal instal·lar res, ja que BitCurator incorpora de sèrie aquest programari. En el cas que s'utilitzi un full de càlcul, el guardarem a la carpeta 'Workspace' amb el nom 'Medialog'. A la Taula 34 mostrem un exemple de com es faria un registre de suports rebuts en el full de càlcul.

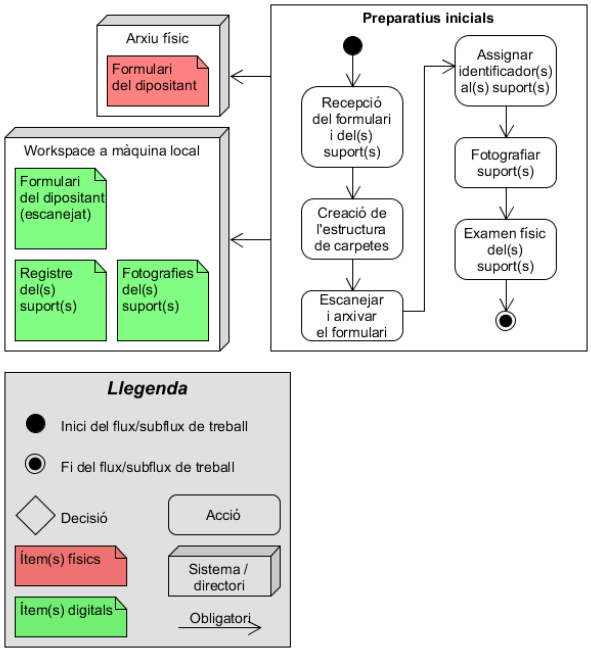
Taula 34. Exemple de registre de suports

Suport	Fabricant	ID suport	Observacions	Topogràfic
USB	Sandisk	12345678USB01	Etiquetat amb la inscripció "Dades"	USB01
HDINTIDE	Samsung	12345678HDINTIDE01	Anotació "Vídeos" a l'etiqueta	HD01
HDEXTUSB	Western Digital	12345678HDEXTUSB01	Anotació "Dades" a l'etiqueta	HD02
CD	Sony	12345678CD01	Anotació "Fotografies" a l'etiqueta	CD01
CD	Sony	12345678CD02	Anotació "Àudios" a l'etiqueta	CD02

Font: L'autor

Com a síntesi d'aquest conjunt de tasques, s'ha rebut i escanejat el formulari de dipòsit de dades, s'han creat les fotografies dels suports, s'ha creat el registre de suports físics que s'han rebut i s'ha arxivat el formulari original en paper dins l'arxiu físic de la institució. A la Figura 30 mostrem com els passos per realitzar aquest subflux de treball amb els ítems que es generarien.

Figura 30. Subflux de treball corresponent als preparatius inicials



Font: L'autor

5.2.2 Captura de suport(s)

Aquest conjunt de processos es concentren en adquirir de forma íntegra els continguts dels suports aportats pel Productor en forma d'un fitxer d'imatge forense i un fitxer de registre que ens recuperarà dades tècniques sobre el procés de captura.

Configurar maquinari i programari

Els passos a executar dependran del tipus de suport. Si es tracta de discs durs interns haurem de connectar-los per USB mitjançant el *docking station*, mentre que en el cas de discs durs externs USB podrem connectar-los directament de la mateixa manera que amb les memòries USB. Amb els discs òptics, es podrà utilitzar directament la unitat de lectura de l'estació de treball. És important, abans de connectar res, comprovar que l'estació de treball està en mode només lectura, per tal d'assegurar que no s'altera cap contingut digital posteriorment.

Connectar *write blocker*

Opcionalment, per a millor seguretat, es pot utilitzar un dispositiu *write blocker* en el cas de dispositius USB. Tal com hem mostrat a la Taula 28, una opció és el Wiebetech USB Write Blocker, que és compatible amb qualsevol dispositiu reconegut com a emmagatzematge massiu USB.

Crear i verificar imatge forense (Guymager)

Seguidament, obrirem el programari Guymager mitjançant el *script* '1-Guymager.sh' per començar a realitzar la captura d'imatge forense, el qual només executa el programari amb drets d'administrador perquè Guymager només es pot utilitzar amb interfície gràfica. En aquest cas s'utilitza el comanament 'gksudo', que serveix per executar interfícies gràfiques amb drets d'administrador.

Figura 31. *Script* utilitzat per executar Guymager

```
#!/bin/sh
# script per executar Guymager
echo La instrucció gksudo executa la interfície gràfica Guymager amb drets de root
gksudo guymager
```

Font: L'autor

Com que crearem una imatge forense, haurem d'utilitzar el format EnCase. Encara que és possible fer-ho també amb AFF, Guymager ja no permès l'opció de crear imatges amb aquest format per defecte. Si l'usuari realment necessita utilitza el format AFF, necessita editar les opcions de configuració de Guymager, la qual cosa requereix permisos d'administració de BitCurator per poder editar aquest fitxer (ubicat a la ruta '/etc/guymager/guymager.cfg'). Per tant, el format d'imatge que s'utilitzarà serà EnCase, que té l'extensió .Exx degut a que genera fitxers de forma seqüencial si configurem Guymager amb aquesta finalitat, com imatge.E01, imatge.E02, imatge.E03, etc. En el nostre cas, no crearem una imatge forense en múltiples fitxers sinó que utilitzarem un de sol. Per tant, hem de configurar les opcions de Guymager; si per exemple generem una imatge forense de 100 GB, haurem de canviar l'opció de múltiples fitxers per tal que només ho faci a partir de la següent escala de bytes, que en aquest cas serien TB.

Per crear la imatge forense i el fitxer .info amb informació tècnica del procés, utilitzarem el directori de l'espai de treball a la carpeta 'Workspace' i utilitzarem el nom de fitxer 'imatge', que crearà el fitxer 'imatge.E01'. Un cop creada la imatge forense es generarà un fitxer amb l'extensió .info amb el mateix nom de la imatge forense que contindrà la següent informació tècnica del procés:

- Mida del dispositiu
- Format de la imatge forense
- Metadades pròpies de la imatge forense (opcionals)
    - Número de cas
    - Número de prova
    - Nom de l'examinador
    - Descripció
    - Notes. Per defecte, Guymager reconeix i afegeix de forma automàtica el número de sèrie del suport
- Ruta i nom del fitxer de la imatge forense
- Verificació de la font (activada o desactivada)
- Verificació de la imatge (activada o desactivada)
- Informació sobre descobriment de sectors corruptes durant l'adquisició
- Informació sobre descobriment de sectors corruptes durant la verificació
- Informació de la finalització del procés (realitzat o no realitzat)
- Valors hash MD5, SHA-1 i SHA-256 de:
    - Suport original
    - Verificació de la font
    - Verificació de la imatge forense
- Informació de la verificació correcta del suport i de la imatge
- Data i hora de l'inici del procés d'adquisició
- Data i hora de l'inici del procés de verificació
- Data i hora de la finalització dels processos d'adquisició i de verificació
- Velocitat de l'adquisició
- Velocitat de la verificació

Errors en el procés?

El registre de captura de Guymager haurà de mostrar si s'han produït errors. Si no és el cas, ens mostrarà "Source verification OK", que ens verificarà que el suport tenia les mateixes dades durant l'adquisició i la verificació, i "Image verification OK", que ens verificarà que la imatge conté exactament les mateixes dades.

Provar opcions alternatives

Serà necessari intentar crear una imatge correcta en el cas d'errors. Per tant haurem de provar opcions alternatives com utilitzar un maquinari diferent. Per exemple, podem utilitzar una ranura diferent USB en el cas de memòries USB o un altre *docking station* en el cas de discs durs interns. Una altra opció és configurar diferents programaris de creació d'imatge com dcfldd[321], una versió millorada del programari dd i desenvolupada pel Defense Computer Forensics Laboratory del Departament de Defensa dels EUA. Un experiment de Woods i Brown (2009) va demostrar la seva eficàcia, ja que va poder generar 81 imatges de CD-ROMs, les quals no es van aconseguir generar per altres eines.

Requerir al dipositant un nou SIP

El fitxer de captura forense i les fotografies del suport seran enviats al Productor per tal d'informar-li de la impossibilitat de completar una captura exitosa i demanar-li l'enviament d'un nou suport amb les mateixes dades. En aquest cas, s'haurà de concertar com s'enviarà el suport, que pot requerir d'un servei de missatgeria, els costos del qual quedaran a càrrec del Productor, sempre i quan la institució del repositori no tingui algun altre acord diferent al respecte.

Guardar fitxers generats

Fins que rebem una resposta del Productor quant a la nostra sol·licitud per a un nova versió del SIP, harem de guardar els fitxers que s'hagin generat fins ara a la nostra classificació de carpetes. Per fer-ho, farem ús del *script* '14-Reanomenar i moure fitxers' per a aquesta finalitat, que mostrem parcialment a la Figura 32.

[321] <http://dcfldd.sourceforge.net/>. [Consulta: 11/12/2016]

Per canviar de nom i moure els fitxers, hem fet ús de la instrucció 'mv' juntament amb 'sudo' per tal de tenir drets d'administrador, on indiquem la ruta original on es troben ubicats els fitxers i la ruta on els volem moure, amb el nom normalitzat. Gràcies a les aportacions de la comunitat Ask Ubuntu[322], hem aconseguit incloure com a prefix la data de darrera modificació del fitxer, amb l'opció 'date +Y%m%d -r /home/bcadmin/Workspace/nomfitxer.xxx', que permet afegir-la com a prefix amb el format YYYYMMDD, tot especificant en cada cas el fitxer al qual es fa referència per tal que detecti aquell del qual es vol capturar la data de darrera modificació. Evidentment, s'hauran de canviar les cadenes de text 'ID investigador' i 'ID suport' a aquelles utilitzades dins la institució. Com que el *script* és un fitxer de text pla, és senzill de fer amb qualsevol editor de text. En el cas de Windows, recomanem utilitzar Notepad++ i en el cas de Linux, gedit.

Figura 32. *Script* utilitzat per canviar de nom i moure els fitxers (detall)

```
#!/bin/sh

# script per moure els diferents fitxers a la ruta dessignada i renombrar-los

echo Els fitxers es mouen a la ruta assignada i es canvia el nom de fitxer
en funció de la data de modificació i la forma normalitzada
sudo mv /home/bcadmin/Workspace/Form.pdf "/home/bcadmin/Arxiu intern/ID
investigador/A-Documentació interna/A1-Formulari del dipositant/$(date +%Y%m%
d -r /home/bcadmin/Workspace/Form.pdf)-ID investigador-Form.pdf"
sudo mv /home/bcadmin/Workspace/Medialog.ods "/home/bcadmin/Arxiu intern/ID
investigador/A-Documentació interna/A2-Registre de suports físics/$(date +%Y%
m%d -r /home/bcadmin/Workspace/Medialog.ods)-ID investigador-Medialog.ods"
sudo mv /home/bcadmin/Workspace/Medialog.csv "/home/bcadmin/Arxiu intern/ID
investigador/A-Documentació interna/A2-Registre de suports físics/$(date +%Y%
m%d -r /home/bcadmin/Workspace/Medialog.csv)-ID investigador-Medialog.csv"
sudo mv /home/bcadmin/Workspace/01.jpg "/home/bcadmin/Arxiu intern/ID
investigador/B-Fotografies/$(date +%Y%m%d -r /home/bcadmin/Workspace/01.jpg)-
ID suport_01.jpg"
sudo mv /home/bcadmin/Workspace/02.jpg "/home/bcadmin/Arxiu intern/ID
investigador/B-Fotografies/$(date +%Y%m%d -r /home/bcadmin/Workspace/02.jpg)-
ID suport_02.jpg"
sudo mv /home/bcadmin/Workspace/03.jpg "/home/bcadmin/Arxiu intern/ID
investigador/B-Fotografies/$(date +%Y%m%d -r /home/bcadmin/Workspace/03.jpg)-
ID suport_03.jpg"
sudo mv /home/bcadmin/Workspace/04.jpg "/home/bcadmin/Arxiu intern/ID
investigador/B-Fotografies/$(date +%Y%m%d -r /home/bcadmin/Workspace/04.jpg)-
ID suport_04.jpg"
```

Font: L'autor

[322] *Renaming files with last modified time on file name*. <http://askubuntu.com/questions/866738/renaming-files-with-last-modified-time-on-file-name/>. [Consulta: 03/01/2017]

Recordem que només guardarem aquests fitxers en el cas de presència d'errors en el procés, ja que en cas contrari hauran de romandre a la ruta '/home/bcadmin/Workspace' per poder continuar amb el flux de treball.

Presència de dades confidencials?

Per saber si existeixen dades confidencials dins el suport amb dades, consultarem el formulari del dipositant. En cas afirmatiu, crearem una imatge *raw* amb Guymager que ens servirà per fer les redaccions i/o bloquejos que comentarem més endavant, al capítol 5.2.4. En cas negatiu, retirarem els suports.

Crear i verificar imatge *raw* (Guymager)

Si hem tancat la interfície gràfica Guymager, la tornarem a obrir executant el *script* corresponent. De la mateixa manera que amb la imatge forense hem de crear una imatge d'un únic fitxer. Per tant, si a Guymager està marcada l'opció de dividir la imatge en múltiples fitxers, l'haurem de desmarcar. Un cop fet això, serà necessari utilitzar un nom de fitxer diferent per evitar conflictes amb la imatge forense, així que utilitzarem 'imatge_raw', que Guymager crearà per defecte amb l'extensió .dd. En el cas d'errors en la verificació, haurem de repetir els passos que hem vist més amunt per provar opcions alternatives, només que en aquest cas volem aconseguir una imatge *raw*.

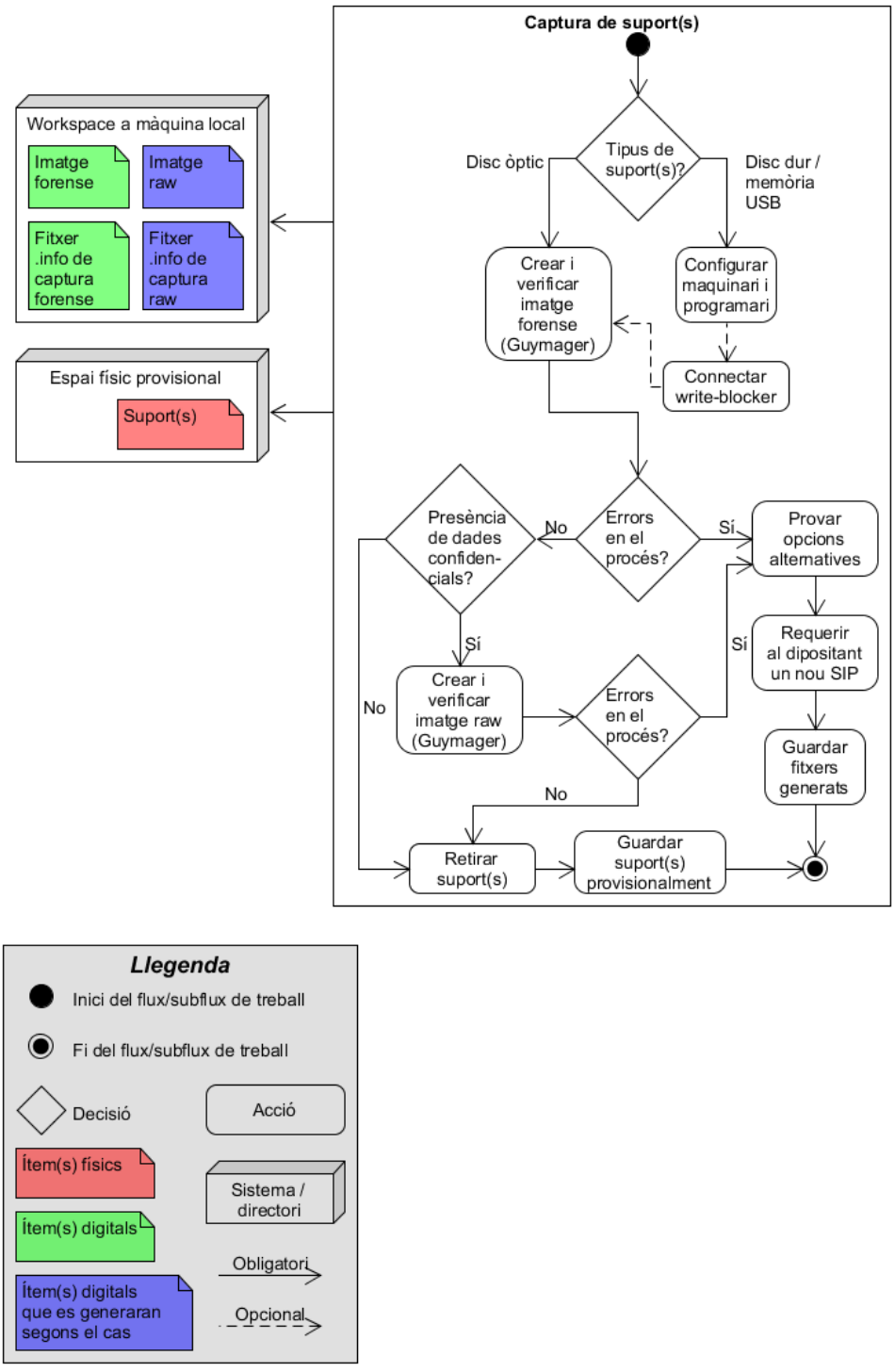
Retirar suport(s)

Un cop completat el procés de captura forense amb tots els suports, podrem desconnectar-los de l'estació de treball (com seria el cas de memòries USB), o bé retirar-los (com seria el cas de discs òptics).

Guardar suport(s) provisionalment

No depositarem encara els suports en la ubicació definida pel seu topogràfic, sinó que els desarem en un lloc provisional per si els necessitem en algun moment del flux de treball. És recomanable que la institució disposi d'ubicacions adequades.

Figura 33. Subflux de treball corresponent a la captura de suport(s)



Font: L'autor

Com a síntesi d'aquest conjunt de tasques, hem configurat el maquinari per tal de bloquejar l'escriptura de les dades originals i en els casos que ha estat necessari, hem utilitzat un *docking station* per accedir als continguts de discs durs interns. No és obligatori, encara que sí és aconsellable, l'ús de dispositius *write blockers* en els casos

de memòries USB i discs durs. Quant a ítems digitals, hem obtingut la imatge forense que conté tots els continguts íntegres dels suports. En els casos amb errors d'adquisició, s'ha recomanat provar diferents configuracions de maquinari. A la Figura 33 mostrem com els passos per realitzar aquest subflux de treball amb els ítems que es generarien.

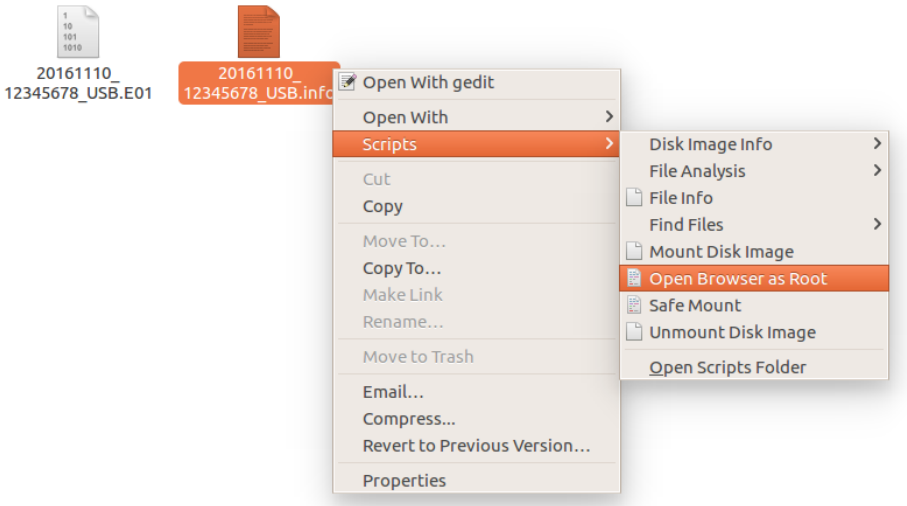
5.2.3 Examen i anàlisi de contingut

Arribats a aquest punt, passarem a l'examen de la imatge forense, que constarà de tres parts: l'escanejat antivirus, l'extracció de dades confidencials i la identificació dels formats de fitxers presents a la imatge.

Muntar imatge forense

Per muntar la imatge, només cal marcar el fitxer amb el botó dret del ratolí i dins l'opció de 'Scripts' es troba l'opció 'Mount Disk Image', que per defecte munta les imatges forenses a la ruta '/media'. Tal i com mostra la Figura 34, hi ha altres opcions, com la informació de fitxer que mostra els valors *hash* i una de les més importants, 'Open Browser as Root' que permet obrir el navegador de fitxers com a administrador, la qual cosa ens proporcionarà permisos per moure o esborrar fitxers que necessitin d'aquests permisos.

Figura 34. Opcions de 'Scripts' a l'entorn BitCurator



Font: Entorn BitCurator. Captura de l'autor

Si es prefereix, també es pot utilitzar l'opció d'executar el *script* '2-Muntar imatge forense.sh' que hem adaptat per al flux de treball. Es mostra en detall a la Figura 35, on s'executa el comanament 'fmount' amb drets d'administrador que serveix per muntar una imatge forense dins la ruta del nostre espai de treball intern.

Figura 35. *Script* utilitzat per muntar una imatge forense

```
#!/bin/sh

# script que munta la imatge forense imatge.Exx que s'ha de trobar a la
ruta /home/bcadmin/Workspace
#pkexec --user root fmount -t /home/bcadmin/Workspace/imatge.E01

echo La instrucció fmount munta la imatge forense present a /home/
bcadmin/imatge.E01

sudo fmount -t /home/bcadmin/Workspace/imatge.E01
```

Font: L'autor, a partir del *script* 'Mount Disk Image' de l'entorn BitCurator

Escanejar presència de virus (ClamAV)

Per fer l'escanejat antivirus, utilitzarem el programa ClamAV mitjançant el *script* '3-ClamAV.sh'. Tal com mostra la Figura 36, el comanament 'sudo' (marcat en negreta) fa que s'executi ClamAV amb permisos d'administrador, cerca tots els fitxers i carpetes presents a la ruta '/media' i reconeix i enregistra els fitxers que estiguin lliures de virus (aquells que ClamAV reconeix amb el patró OK) dins el fitxer 'clamav.log' que es genera a la ruta '/home/bcadmin/Workspace'. El següent comanament és un altre 'echo', que mostrarà el text indicat a la figura per informar-nos del procés de reconèixer i enregistrar els fitxers infectats, que en essència és igual que el que hem vist amb anterioritat, amb la diferència que no caldrà indicar cap patró i només indicarem amb l'opció -l que s'enregistri aquesta acció al fitxer 'clamav.log'. El patró 'date' servirà per què quedi enregistrada la data i hora en què s'ha executat el *script* dins el registre.

Es generarà el fitxer 'clamav.log' a la ruta especificada, el qual ens donarà dues llistes separades per guions: una contindrà els fitxers lliures de virus (ClamAV els marcarà com OK) i una altra, aquells fitxers que s'hagin reconegut com infectats (ClamAV els marcarà com FOUND). Al final del registre tindrem un sumari que ens informarà del nombre de virus que reconeix ClamAV a la seva base de dades, la versió de motor que ha utilitzat, els nombre de directoris i de fitxers que s'han escanejat, el nombre de fitxers infectats, la mida en MB que s'ha escanejat, el temps total que ha requerit

l'escanejat i la data i hora de l'escanejat. La Figura 37 mostra un exemple de com quedaria el registre.

Figura 36. *Script* utilitzat per a l'escanejat antivirus amb ClamAV

```
#!/bin/sh

# script per escanejar virus amb ClamAV dins fitxers d'imatges forenses
muntades a /media

echo ClamAV escaneja els fitxers de la imatge muntada a /media,
reconeix i enregistra els fitxers lliures de virus
sudo clamscan -r /media/imatge_vol02/ | grep OK >> ~/Workspace/
clamav.log
echo ClamAV escaneja els fitxers de la imatge muntada a /media,
reconeix i enregistra els fitxers infectats, fa un sumari i enregistra
la data i hora actuals
sudo clamscan -r /media/imatge_vol02/ -l ~/Workspace/clamav.log; date
>> /home/bcadmin/Workspace/clamav.log
```

Font: Fedoraforum.org. *clamav log file*. <http://forums.fedoraforum.org/showthread.php?t=296952>. [Consulta: 10/04/2017]

Figura 37. Mostra de registre de ClamAV

```
/media/imatge_vol02/FTK Imager Lite 3.1.1/langs/sve_ftki.dll: OK
/media/imatge_vol02/FTK Imager Lite 3.1.1/langs/trk_adencrypt.dll: OK
/media/imatge_vol02/FTK Imager Lite 3.1.1/langs/trk_adshattrdefs.dll: OK
/media/imatge_vol02/FTK Imager Lite 3.1.1/langs/trk_ftki.dll: OK
/media/imatge_vol02/pruebas redact.txt: OK

----------------------------------------------------------------------

/media/imatge_vol02/antivirus-testfile.txt: Eicar-Test-Signature FOUND

----------- SCAN SUMMARY -----------
Known viruses: 5402053
Engine version: 0.99.2
Scanned directories: 15
Scanned files: 162
Infected files: 1
Data scanned: 315.04 MB
Data read: 228.53 MB (ratio 1.38:1)
Time: 22.924 sec (0 m 22 s)
Tue Dec 27 16:23:32 CET 2016
```

Font: L'autor

Presència de virus?

En el cas que ClamAV detecti algun fitxer infectat, es poden emprar diverses estratègies. Una seria prendre nota de la ruta on es troben aquests fitxers infectats i requerir al dipositant una nova versió del seu SIP que estigui lliure de virus (Barrera-Gomez; Erway, 2013). Una altra seria extreure només els fitxers sense virus i ingestar-los de

forma individual, però si fem això haurem de descartar la ingesta de la imatge forense al repositori degut a l'existència de virus en la mateixa.

Requerir al dipositant un nou SIP

S'ha triat, doncs, sol·licitar una nova versió del SIP al Productor, tot indicant els fitxers infectats i la seva ruta original dins el suport. Tal com s'ha indicat al capítol 5.2.2, s'haurà de concertar un sistema de devolució dels suports que pot o no requerir el pagament dels ports per part del Productor.

Guardar fitxers generats

Com que s'ha detectat la presència de virus, no conservarem la imatge forense ni la imatge *raw* si s'ha generat, així com els registres de captura. Un cop esborrats aquests fitxers, mourem la resta (formulari del dipositant, registre de captura forense, fotografies dels suports i registre de virus) a la carpeta 'Arxiu intern', fent servir el *script* indicat al capítol 5.2.2. El flux de treball finalitzarà pel que respecta a la preservació del suport fins que arribi un nou SIP a la institució.

Extreure dades confidencials (bulk_extractor)

Tant si tenim presència de dades confidencials com si no, serà necessari extreure les dades confidencials amb bulk_extractor degut a què BitCurator necessita els diversos fitxers generats per aquesta eina per poder fer l'anàlisi de tota la imatge i així poder generar els informes tècnics corresponents i les metadades DFXML i PREMIS.

Tal com s'ha indicat al capítol 4.3.2, és possible executar Bulk Extractor Viewer per tal que analitzi la imatge forense i extregui diferents fitxers de text, en funció de les opcions d'escanejat que s'hagin seleccionat, tal i com es mostra a la Figura 22. No obstant, hi ha una opció més ràpida i senzilla: l'ús del *script* '4-bulk_extractor.sh' que hem creat expressament i que es mostra a la Figura 38.

La seva funció és executar bulk_extractor dins la ruta ~/Workspace (el símbol ~, a l'entorn BitCurator, serveix per substituir la ruta '/home/usuari', que en el nostre cas és '/home/bcadmin') amb l'opció -o que serveix per indicar la ruta on es generaran els

fitxers tabulats, que ha de ser un directori nou. Per tal d'identificar sense ambigüitats el programari utilitzat, hem indicat que aquest directori rebrà el nom de 'bulk_extractor'.

Figura 38. *Script* utilitzat per a l'extracció de dades confidencials amb Bulk Extractor

```
#!/bin/sh

# script per extreure dades confidencials d'una imatge forense

echo bulk_extractor analitza una imatge forense,
extreu dades confidencials en fitxers de text pla tabulat
i genera un informe en metadades DFXML

sudo bulk_extractor ~/Workspace/imatge.E01 -o ~/Workspace/bulk_extractor
```

Font: L'autor

Recordem que cadascun dels fitxers generats contindrà tres columnes: posició dins la imatge forense, informació trobada i context en què es va trobar la informació.

Val a dir que aquesta configuració del *script* només activa els escàners configurats per defecte, i per tant inhabilita els escàners base16, facebook, outlook, sceadan, wordlist i xor. És possible activar-los, però, amb l'ús de l'opció -e i si cal és possible refinar més l'extracció, com inhabilitar la generació d'histogrames o ajustar el mínim de dígits que hauria de tenir un número de telèfon per tal de ser detectat (Bradley; Garfinkel, 2015, p. 58-60). De la mateixa manera, també és possible desactivar els escàners que no volem utilitzar amb l'opció -x.

Juntament amb els fitxers de text tabulat, es generarà un fitxer de metadades DFXML que rebrà per defecte el nom de 'report.xml'. Aquest és el primer exemple que tenim de metadades DFXML i presentem les seves etiquetes més rellevants a la Taula 35. Per tal de representar la jerarquia d'etiquetes, hem introduït una segona columna més a la dreta dins la columna 'Etiqueta' quan una etiqueta penja d'una altra.

Taula 35. Etiquetes presents al fitxer de metadades DFXML generat per bulk_extractor

Etiquetes i subetiquetes	Descripció
<metadata>	Informació de la capçalera que defineix les metadades DFXML
<dc:type>	Format del fitxer
<creator>	Documentació sobre el programa i l'entorn en què s'ha fet la captura i/o l'anàlisi del disc
<program>	Nom del programa que genera el fitxer XML
<version>	Número de versió del programa d'anàlisi forense
<build_environment>	Informació sobre l'entorn que s'ha utilitzat per generar el fitxer XML
<execution_environment>	Informació sobre el sistema en què es van capturar les dades forenses
<configuration>	Configuració de Bulk Extractor
<threads>	Nombre de de fils d'execució o subprocessos que s'han utilitzat
<pagesize>	Nombre de bytes que busca Bulk Extractor en cada cerca
<marginsize>	Nombre de bytes que es poden superposar, per tal d'evitar pèrdues de dades
<scanners>	Escàners que s'han habilitat per fer l'extracció
<provided_filename>	Ruta del fitxer d'imatge forense que s'ha utilitzat per l'extracció
<runtime>	Estadístiques de temps d'execució dels subprocessos
<source>	Font de les dades forenses
<image_filename>	Nom de fitxer complet, amb la seva ruta original, de la imatge forense
<image_size>	Mida en bytes de la imatge forense
<hashdigest>	Valor *hash* MD5 de la imatge forense
<feature_files>	Fitxers que s'han extret amb dades confidencials
<name>	Nom del fitxer
<count>	Nombre d'ocurrències que s'han identificat i extret en cada fitxer
<report>	Informe sobre l'operació d'extracció
<total_bytes>	Mida en bytes de la imatge forense
<elapsed_seconds>	Temps en segons que s'ha requerit per fer l'extracció
<max_depth_seen>	Nivell màxim de recursivitat
<dup_data_encountered>	Dades duplicades que s'han trobat
<scanner_times>	Estadístiques de temps quant a l'ús dels escàners
<name>	Nom de l'escàner
<calls>	Nombre de crides de sistema que s'han utilitzat per a cada escàner
<seconds>	Nombre de segons que s'han utilitzat per a cada escàner
<rusage>	Estadístiques d'ús de memòria i d'eficiència d'execució del processador
<utime>	Temps acumulatiu en *jiffies* (aprox. 10 milisegons) que s'ha utilitzat per executar el codi d'usuari
<stime>	Temps acumulatiu en *jiffies* (aprox. 10 milisegons) que s'ha utilitzat per executar el codi de sistema
<maxrss>	Mesura d'ús de memòria per part del sistema durant el procés d'extracció

Etiquetes i subetiquetes	Descripció
<minflt>	Nombre en KB d'errors mínims de pàgina que es mapen a un espai d'adreça virtual però que no es mapen a la memòria física
<majflt>	Nombre en KB d'errors majors de pàgina que es mapen a un espai d'adreça virtual però que no es mapen a la memòria física
<nswap>	Transferències de contingut de memòria al disc degut a què la memòria està plena. Amb sistemes moderns amb molta memòria, el resultat habitualment serà 0
<inblock>	Mesura en blocs que s'han afegit a la memòria cau del sistema de fitxers durant el procés d'extracció
<oublock>	Mesura en blocs que s'han escrit i assignat a la llista lliure de la memòria disponible durant el procés d'extracció
<clocktime>	Temps necessari per executar el programa en funció de la velocitat de rellotge del processador

Fonts: L'autor; DFXML tag library: v4. <https://goo.gl/BTSTvs>. [Consulta: 31/12/2016]; *Find sensitive data with Bulk Extractor*. <https://warroom.securestate.com/find-sensitive-data-with-bulk-extractor>. [Consulta: 31/12/2016]

Podem diferenciar les següents categories d'etiquetes:

- <metadata> conté informació de capçalera que defineix les metadades en el document DFXM
- <creator> ens recupera informació sobre el programari i l'entorn utilitzats per crear el fitxer xml
- <configuration> conté la configuració que s'ha emprat per executar el programari
- <provided_filename> és una referència a la ruta de la imatge forense de la qual s'han extret les dades
- <runtime> conté diferents dades quant al temps d'execució dels subprocessos
- <source> conté la font de les dades forenses, que pot contenir informació sobre el suport original i de la imatge forense que es crea
- <feature_files> conté informació detallada dels fitxers que s'han generat amb l'extracció. Una subetiqueta, <count>, ens recupera una dada molt interessant: el nombre d'ocurrències que Bulk Extractor ha detectat. P. ex., podem localitzar ràpidament el nombre de correus electrònics recuperats
- <report> dona dades tècniques sobre l'operació d'extracció en general, com la mida en bytes de la imatge forense original o el nivell màxim de recursivitat;

això vol dir el nombre màxim de vegades que una funció es crida a si mateixa de forma recursiva

- <scanner_times> recupera dades tècniques quant els temps que han necessitat els escàners per fer l'extracció
- <rusage> conté informació de la durada en l'execució d'ús del processador

Executar BitCurator Reports

Seguidament executem BitCurator Reports mitjançant el *script* '5-BitCurator_Reports.sh', el qual executa la interfície gràfica de BitCurator Reports. S'ha de tenir en compte que es tracta d'un programari realitzat sota Python 3, el qual executa altres *scripts* Python per generar tots els fitxers esmenats al capítol 4.3.3. A tall il·lustratiu de com s'hauria d'executar un fitxer .py, s'ha utilitzat la instrucció 'python3 [ruta del fitxer]', però també es possible executar-lo amb la instrucció directa 'bc_reports_tab_py'.

Figura 39. *Script* utilitzat per executar BitCurator Reports

```
#!/bin/sh

# script per obrir BitCurator Reports

echo python 3 obre la interfície gràfica BitCurator Reports

python3 /usr/local/bin/bc_reports_tab.py
```

Font: L'autor

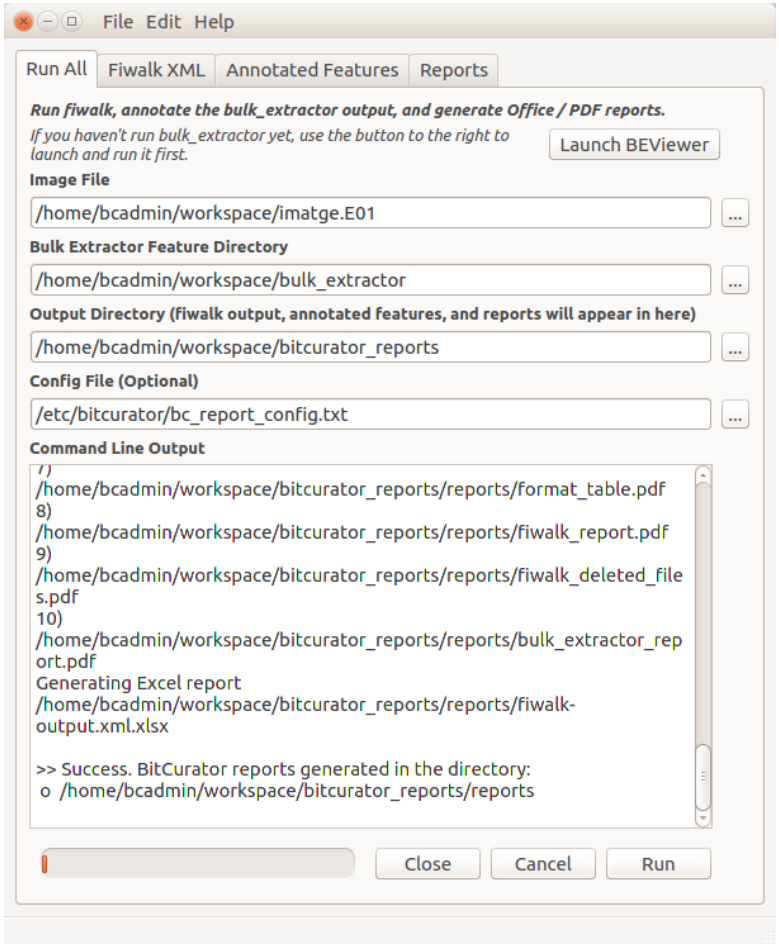
La raó de no utilitzar un *script* que generi directament els fitxers amb línies de comanament és que BitCurator Reports no està dissenyat per ser executat íntegrament amb comanaments. Sí que és possible generar fitxers DFXML executant fiwalk, generar fitxers anotats i informes amb *scripts* Python, però no és possible crear les metadades PREMIS amb l'ús d'un *script*, tot i que sí que existeix un *script* Python amb aquesta finalitat, però després de diferents proves no s'han aconseguit resultats satisfactoris degut a què aquest *script* porta molt de temps sense actualitzar, segons m'indicà Kam Woods (membre de l'equip BitCurator) a una conversa dins la llista de correu BitCurator Users[323]. A més, en les proves que s'han fet hem comprovat que executar

[323] *Generating PREMIS xml file via command line*. <https://groups.google.com/forum/#!topic/bitcurator-users/Ku7e4WsdCys/>. [Consulta: 31/12/2016]

BitCurator Reports en interfície gràfica donava millors resultats, ja que les operacions es completaven més ràpidament.

Un cop obert, utilitzarem l'opció 'Run All' i seleccionarem la imatge forense, la ruta on hem generat els fitxers extrets per bulk_extractor i seleccionarem una carpeta nova dins 'Workspace' on es guardaran tots els nous fitxers. La Figura 40 mostra com haurien de quedar totes les opcions de rutes dins la pestanya 'Run All'.

Figura 40. Interfície gràfica de BitCurator Reports amb les opcions ja emplenades per exportar informes



Font: Entorn BitCurator. Captura de l'autor

A la primera opció, hem de seleccionar la imatge forense creada amb anterioritat, 'imatge.E01'. A la segona opció, seleccionem el directori on hem generat els fitxers d'extracció de dades confidencials que hem creat amb el *script* de bulk_extractor. A la tercera opció, hem de navegar dins 'Workspace' i indicar el nom de la nova carpeta on es guardaran els informes i llistats de BitCurator Reports; aquest nom serà

'bitcurator_reports'. Finalment, tenim l'opció d'indicar la ruta d'un fitxer de configuració, que per defecte es troba ubicat a la ruta '/etc/bitcurator/bc_report_config.txt', el qual permet seleccionar o desseleccionar els informes en PDF i en Excel (Lee et al., 2014).

Tal com hem indicat al capítol 4.3.3, l'execució de BitCurator Reports generarà un fitxer de metadades DFXML generat per fiwalk, un directori 'annotated-features' i un directori 'reports' amb diferents subdirectoris i fitxers. Analitzem a continuació amb més detall tots els fitxers que s'han generat.

En primer lloc, tenim el fitxer XML de metadades DFXML que per defecte s'anomena fiwalk-output.xml. Dins la Taula 36 es mostren les etiquetes principals i la seva descripció, on si el comparem amb el fitxer DFXML generat per Bulk Extractor, tenim un grup d'etiquetes sota <volume>, on trobem dades tècniques de cada fitxer individual de la font original.

El directori 'annotated-features' contindrà diversos fitxers de text tabulat, on cadascun tindrà una estructura de nom amb el prefix 'annotated', seguit d'un guió baix i el nom del fitxer de text que Bulk Extractor generà amb anterioritat. Cada fitxer té cinc columnes (a diferència dels fitxers de bulk_extractor que contenien tres columnes), amb el següent contingut:

- Posició. Igual que amb bulk_extractor, fa referència a la ubicació en nombre de bytes de la dada dins la imatge forense
- Característica. Igual que amb bulk_extractor, fa referència a les dades que s'han localitzat dins la imatge forense
- Context. Igual que amb bulk_extractor, fa referència al context on s'han localitzat les dades
- Nom de fitxer. Ens recupera el nom de fitxer concret on es troben les dades confidencials
- MD5. Valor *hash* MD5 del fitxer on s'han localitzat les dades

Taula 36. Etiquetes presents al fitxer de metadades DFXML de fiwalk

Etiquetes i subetiquetes	Descripció
<metadata>	Informació de la capçalera que defineix les metadades DFXML
<dc:type>	Format del fitxer
<creator>	Documentació sobre el programa i l'entorn en què s'ha fet la captura i/o l'anàlisi del disc
<program>	Nom del programa que genera el fitxer XML
<version>	Número de versió del programa d'anàlisi forense
<build_environment>	Informació sobre l'entorn que s'ha utilitzat per generar el fitxer XML
<execution_environment>	Informació sobre el sistema en què es van capturar les dades forenses
<source>	Font de les dades forenses
<image_filename>	Nom de fitxer complet, amb la seva ruta original, de la imatge forense
<volume>	Informació sobre fitxers individuals a un fitxer DFXML
<partition_offset>	Localització d'inici de la partició dins la imatge de disc, mesurada en bytes
<sector_size>	Mida en bytes d'un sector de disc del volum
<block_size>	Mida en bytes d'un bloc individual de dades en un volum, tal com es defineix al sistema de fitxers
<ftype>	Identificador numèric que representa el sistema de fitxers present a la partició del volum
<ftype_str>	Cadena de text corresponent al sistema de fitxers que es representa a <ftype>
<block_count>	Llista del nombre total de blocs en el volum de destinació
<first_block>	Adreça del primer bloc del sistema de fitxers, en bytes
<last_block>	Número del darrer bloc dins el volum
<fileobject>	Informació del fitxer i les seves metadades, amb informació de la mida i tipus de fitxer, valors hash i informació de procedència
<rusage>	Estadístiques d'ús de memòria i d'eficiència d'execució del processador
<utime>	Temps acumulatiu en *jiffies* (aprox. 10 milisegons) que s'ha utilitzat per executar el codi d'usuari
<stime>	Temps acumulatiu en *jiffies* (aprox. 10 milisegons) que s'ha utilitzat per executar el codi de sistema
<maxrss>	Mesura d'ús de memòria per part del sistema durant el procés de captura
<minflt>	Nombre en KB d'errors mínims de pàgina que es mapen a un espai d'adreça virtual però que no es mapen a la memòria física
<majflt>	Nombre en KB d'errors majors de pàgina que es mapen a un espai d'adreça virtual però que no es mapen a la memòria física
<nswap>	Transferències de contingut de memòria al disc degut a què la memòria està plena. Amb sistemes moderns amb molta memòria, el resultat habitualment serà 0

Etiquetes i subetiquetes	Descripció
<inblock>	Mesura en blocs que s'han afegit a la memòria cau del sistema de fitxers durant el procés de captura
<oublock>	Mesura en blocs que s'han escrit i assignat a la llista lliure de la memòria disponible durant el procés de captura
<clocktime>	Temps necessari per executar el programa en funció de la velocitat de rellotge del processador

Fonts: L'autor; DFXML tag library: v4. <https://goo.gl/BTSTvs>. [Consulta: 31/12/2016]

El directori 'reports' contindrà el següents fitxers, el contingut dels quals ja s'ha comentat al capítol 4.3.3:

- bc_format_bargraph.pdf
- bulk_extractor_report.pdf
- fiwalk_deleted_files.pdf
- fiwalk_report.pdf
- fiwalk-output.xml.xlsx
- format_table.pdf
- premis.xml
- Subdirectori 'features' amb fulls de càlcul Excel generats a partir dels fitxers generats per Bulk Extractor, però només amb tres columnes: nom de fitxer, característica i posició

Quant al fitxer de metadades PREMIS, BitCurator Reports identifica dos tipus d'entitats: Objecte i Esdeveniment. En el primer cas, l'Objecte es refereix al fitxer d'imatge forense, i en el segon cas, hi ha tres Esdeveniments: el primer és el procés de captura, que recupera informació sobre el format de la imatge forense, el programari utilitzat, la mida de la imatge i la data i hora de captura; el segon tracta de l'anàlisi del sistema de fitxers, que recupera informació sobre el fitxer DFXML amb la ruta completa de la seva ubicació i la data i hora de l'anàlisi; el tercer Esdeveniment és l'extracció en massa d'informació personal on s'ha utilitzat el programari bulk_extractor. Mostrem a la Taula 37 la relació d'etiquetes PREMIS que s'utilitzen en aquest cas, amb la seva descripció.

Taula 37. Etiquetes presents al fitxer de metadades PREMIS generat per BitCurator Reports

Etiquetes i subetiquetes	Descripció
<object>	Informació sobre l'Objecte digital que gestiona un repositori de preservació i descripció de les característiques rellevants per a la seva gestió
<originalName>	Nom de l'Objecte tal i com s'envia o es recull per part del repositori, abans que el repositori faci un canvi de nom
<objectIdentifier>	Designació que s'utilitza per identificar de forma unívoca l'Objecte dins el sistema de preservació del repositori en què s'emmagatzema
<event>	Informació vers una acció que afecta un o més Objectes
<eventIdentifier>	Designació que s'utilitza per identificar de forma unívoca l'Esdeveniment dins el sistema de preservació del repositori
<eventType>	Categorització de la naturalesa de l'Esdeveniment
<eventDetail>	Informació addicional sobre l'Esdeveniment
<eventDateTime>	Data i hora en què va succeir l'Esdeveniment
<eventOutcomeInformation>	Informació sobre el resultat d'un Esdeveniment
<eventOutcomeDetail>	Descripció detallada del resultat o producte de l'Esdeveniment

Fonts: L'autor; PREMIS Editorial Committee, 2015

Generar informes de formats de fitxer (DROID)

Executarem el programari DROID per generar informes de formats de fitxer per tal que el personal pugui comparar els resultats amb els informes de BitCurator, tal i com es va exposar al capítol 4.3.3. Hem de recordar que no és un programari instal·lat per defecte a BitCurator, i per tant s'haurà de descarregar i descomprimir. En el nostre cas, l'hem desat a la ruta '/home/bcadmin/utils/DROID' i hem utilitzat el *script* '6-DROID.sh' per executar-lo directament, el qual mostrem a la Figura 41. Evidentment, la ruta indicada al *script* s'haurà de modificar si DROID es desa en una ruta diferent.

Encara que es pot executar per interfície gràfica tal i com vam exposar al capítol 4.3.3, és més eficient i ràpid executar-lo mitjançant línia de comanaments i és així com hem configurat el *script*. En primer lloc, s'ha d'emprar la instrucció 'java -jar' amb el fitxer 'droid-command-line-6.2.1.jar', on amb l'opció -a seleccionem la ruta on muntem la imatge forense, que per defecte a BitCurator es fa a '/media/[nom d'imatge]_vol02'. Com que al flux de treball sempre utilitzem el nom 'imatge.E01' per a la imatge forense, la ruta ha de ser '/media/imatge_vol02'. L'opció -p serveix per crear un perfil DROID, la

qual cosa és obligatòria per crear informes. En el nostre cas, hem creat el perfil 'bitcurator.droid' a la ruta '/home/bcadmin/utils/DROID'.

Les dues línies següents fan l'exportació d'informes, que amb l'opció -p seleccionen el perfil que hem creat anteriorment, l'opció -n indica el tipus d'informe, l'opció -t selecciona el format de l'informe i l'opció -r indica la ruta i el nom de fitxer que s'utilitzarà. Per al nostre flux de treball, generarem els fitxers 'droid.pdf' i 'droid.xml' del tipus 'Comprehensive breakdown', la més completa pel que fa a la informació exportada, que es guardaran a '/home/bcadmin/Workspace'.

Figura 41. Script utilitzat per executar DROID

```
#!/bin/sh

# script per crear un perfil DROID de la imatge muntada a '/media/
imatge_vol02' i exportar informes

echo DROID crea amb línia de comanaments un perfil en funció dels
fitxers localitzats a la imatge muntada i exporta informes en PDF i en
XML

cd /home/bcadmin/utils/DROID
java -jar droid-command-line-6.2.1.jar -R -a "/media/imatge_vol02/" -p
"/home/bcadmin/utils/DROID/bitcurator.droid"
java -jar droid-command-line-6.2.1.jar -p "/home/bcadmin/utils/DROID/
bitcurator.droid" -n "Comprehensive breakdown" -t "Pdf" -r "/home/
bcadmin/Workspace/droid.pdf"
java -jar droid-command-line-6.2.1.jar -p "/home/bcadmin/utils/DROID/
bitcurator.droid" -n "Comprehensive breakdown" -t "DROID Report XML" -r
"/home/bcadmin/Workspace/droid.xml"
```

Font: L'autor

Afegir marca d'aigua als informes

Com que els informes en PDF generats per BitCurator Reports i DROID no aporten suficient informació per identificar amb seguretat la seva procedència, s'ha creat el *script* '8-Watermark.sh' per tal d'afegir una marca d'aigua amb la identificació del suport, el qual executa el programari Ghostscript (ja instal·lat per defecte a BitCurator), que afegeix el text que ens interessa mitjançant un fitxer PostScript, al qual hem nomenat 'stamp.ps'. Mostrem el *script* a la Figura 42 i el fitxer PostScript a la Figura 43.

Figura 42. *Script* utilitzat per executar Ghostscript

```
#!/bin/sh

# script per afegir marca d'aigua a fitxers pdf mitjançant Ghostscript

echo Ghostscript localitza el fitxer pdf i afegeix el text indicat al fitxer
PostScript stamp.ps

gs -dBATCH -dNOPAUSE -dAutoRotatePages=/None -sDEVICE=pdfwrite -sOutputFile=/
home/bcadmin/Workspace/bitcurator_reports/reports/
bc_format_bargraph_mark.pdf stamp.ps -f /home/bcadmin/Workspace/
bitcurator_reports/reports/bc_format_bargraph.pdf
gs -dBATCH -dNOPAUSE -dAutoRotatePages=/None -sDEVICE=pdfwrite -sOutputFile=/
home/bcadmin/Workspace/bitcurator_reports/reports/
bulk_extractor_report_mark.pdf stamp.ps -f /home/bcadmin/Workspace/
bitcurator_reports/reports/bulk_extractor_report.pdf
gs -dBATCH -dNOPAUSE -dAutoRotatePages=/None -sDEVICE=pdfwrite -sOutputFile=/
home/bcadmin/Workspace/bitcurator_reports/reports/fiwalk_report_mark.pdf
stamp.ps -f /home/bcadmin/Workspace/bitcurator_reports/reports/
fiwalk_report.pdf
gs -dBATCH -dNOPAUSE -dAutoRotatePages=/None -sDEVICE=pdfwrite -sOutputFile=/
home/bcadmin/Workspace/bitcurator_reports/reports/format_table_mark.pdf
stamp.ps -f /home/bcadmin/Workspace/bitcurator_reports/reports/
format_table.pdf
gs -dBATCH -dNOPAUSE -dAutoRotatePages=/None -sDEVICE=pdfwrite -sOutputFile=/
home/bcadmin/Workspace/droid_mark.pdf stamp.ps -f /home/bcadmin/Workspace/
droid.pdf
```

Font: L'autor

El fitxer PostScript està creat per tal que afegeixi l'identificador del suport (a tall d'exemple, hem utilitzat l'estructura 12345678USB01) a la part inferior esquerra de cadascuna de les pàgines dels informes PDF indicats al *script* de la Figura 42. Com que Ghostscript no sobreescriu els informes, és necessari crear uns de nous amb el sufix '_mark'. Aquests nous fitxers més endavant es canviaran de nom a la forma normalitzada i s'esborraran els fitxers anteriors mitjançant el *script* indicat al capítol 5.2.2.

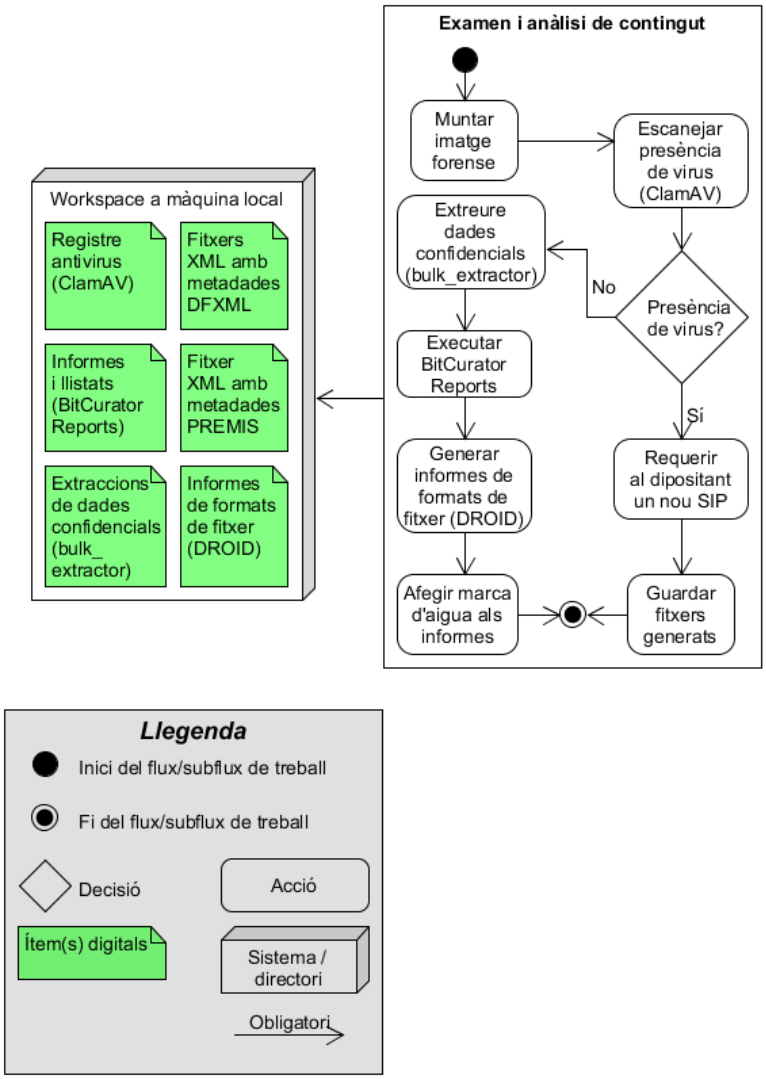
Figura 43. Fitxer PostScript 'stamp.ps'

```
<<
   /EndPage
   {
     2 eq { pop false }
     {
         gsave
         /NimbusSanL-Reg 12 selectfont
         55 20 moveto (12345678USB01) show
         grestore
         true
     } ifelse
   } bind
>> setpagedevice
```

Font: L'autor

Com a síntesi d'aquest conjunt de tasques, hem muntat la imatge forense per poder fer l'escanejat de virus amb ClamAV, que ha generat un registre a Workspace. S'ha executat bulk_extractor, que ha generat extraccions de dades confidencials de la imatge forense i ha generat un informe en metadades DFXML. Posteriorment hem executat BitCurator Reports, que ha generat diversos informes i llistats, que inclouen un informe en metadades DFXML de fiwalk i un informe d'Esdeveniments en format PREMIS. Finalment, s'han generat informes de formats de fitxer amb DROID. En el cas de presència de virus, s'ha contemplat requerir al dipositant un nou SIP, i per tant es guarden els fitxers amb noms normalitzats a la carpeta 'Arxiu intern'. Mostrem tot el subflux de treball a la Figura 44.

Figura 44. Subflux de treball corresponent a l'examen i anàlisi de contingut



Font: L'autor

5.2.4 Processat de continguts

El processat implica fer una funció de Preservació, en terminologia OAIS. Això implica la creació d'un nou Paquet d'Informació d'Arxiu (Lee, 2005, p. 132), ja que es produiran canvis en la Informació de Contingut, atès que realitzarem el bloqueig d'informació privada i sensible i farem una normalització de formats de fitxer, en funció de la política de preservació de la institució. Per una banda, s'hauran d'identificar els fitxers a esborrar i/o bloquejar i per altra, s'identificaran aquells fitxers amb formats que s'hagin de migrar a altres aptes per a la preservació. Proposem utilitzar en un full de càlcul per a cada operació de forma separada. A continuació descriurem les operacions amb més detall.

Presència de dades confidencials?

Per saber si existeixen dades confidencials dins el suport amb dades, consultarem el formulari del dipositant. En cas afirmatiu, haurem d'identificar i llistar els fitxers a redactar i bloquejar. En cas negatiu, comprovarem si s'han de normalitzar formats.

Identificar fitxers a redactar i bloquejar

Primer començarem amb identificar els fitxers amb informació privada. Amb aquest objectiu utilitzarem el formulari de donació per localitzar els fitxers i els contrastarem amb els informes que ens ha donat bulk_extractor i BitCurator Reports per tal de valorar quins fitxers seran susceptibles de ser restringits totalment i aquells que necessitaran només d'una redacció. Per comprovar si és possible fer-ne una redacció de cadenes de text, consultarem els fitxers identificats a la imatge muntada i els analitzarem amb l'editor hexadecimal. L'entorn BitCurator permet fer-ho fàcilment, marcant el fitxer i amb el botó dret utilitzar les opcions Scripts > File Analysis > View in Hexeditor. Farem cerques de les cadenes de text que volem bloquejar, com el número de NIF o el correu electrònic. Si trobem aquestes cadenes, serà possible redactar-los. En cas contrari, els haurem de bloquejar completament.

A continuació, en un nou full de càlcul LibreOffice Calc, que guardarem a 'Workspace' amb el nom 'Redaction.ods' i també en format CSV, es crearan sis columnes que tindran el següent contingut:

- Nom del fitxer original
- Valoració
- Acció
- Valor *hash* MD5
- Valor *hash* SHA1
- Comentaris

Taula 38. Exemple de llista de fitxers a redactar i bloquejar

Nom del fitxer original	Valoració	Acció	MD5 hash	SHA1 hash	Comentaris
Umlet.exe	Restringit	FUZZ	c700e8937d286fa9e9353540768851fa	b038a893a2b641a3728c5380d02499a06e348cc5	Programari bloquejat a petició de l'investigador
2016-2017 – Full Matrícula UB.pdf	Restringit	SCRUB	4a0a0ccde6b33ded84d9ea9a6a44f2ec	99ff4ad948b9d090812ab8813c3c493a0f106f3c	No permet redacció, s'ha de bloquejar
inscripció doctorat.doc	Per redactar	FILL	99d9ef0e23a3b6633d686a3d6f1acb72	e00c67e55c78d3f974dc4eedc8330c404ee6ef13	S'han de redactar noms, cognoms i NIF

Font: L'autor

Podrem extreure aquesta informació del full de càlcul generat per BitCurator Reports, que es trobarà a la ruta '~/Workspace/bitcurator_reports/reports/fiwalk-output.xml.xlsx'. A la primera columna indicarem el nom del fitxer amb dades confidencials, que inclourà la ruta original si no es troba a l'arrel. Dins la segona columna s'escriurà, un cop s'ha valorat el contingut del fitxer, una d'aquestes dues opcions: 'Per redactar', si només es necessari fer-ne un reescrit parcial, o 'Restringit', si el fitxer conté essencialment informació privada (com seria el cas d'un formulari de cessió de dades). Dins la columna 'Acció', s'escriuran les operacions a realitzar en el cas dels fitxers que s'han de redactar, segons les opcions del fitxer de configuració de bitcurator_access_redaction que vam exposar a la Taula 25. La quarta i cinquena columnes inclouran els valors *hash* MD5 i SHA1, que ajudaran a configurar les opcions de redacció per identificar els fitxers a redactar o bloquejar. Finalment, la columna 'Comentaris' servirà per anotar qualsevol incidència que s'hagi produït com la impossibilitat de fer una redacció i per tant s'ha hagut de bloquejar completament el fitxer, o instruccions concretes per

redactar el fitxer. A la Taula 38 mostrem un exemple de com es podria crear aquesta llista.

Desmuntar imatge forense

Com que ja no és necessari tenir la imatge forense muntada, podem desmuntar-la. De la mateixa manera que amb el muntatge de la imatge forense podem fer-ho amb l'opció 'Unmount' integrada a BitCurator (vegeu la Figura 34) o amb el *script* '7-Desmuntar imatge forense' que mostrem a la Figura 45.

Figura 45. *Script* utilitzat per desmuntar la imatge forense

```
#!/bin/sh

# script que desmunta la imatge forense imatge.Exx que s'ha de trobar a
la ruta /home/bcadmin/Workspace
#pkexec --user root fmount -u /home/bcadmin/Workspace/imatge.Exx

echo La instrucció fmount desmunta la imatge forense present a /home/
media/imatge_vol02

sudo fmount -u /home/bcadmin/Workspace/imatge.E01
```

Font: L'autor, a partir del *script* 'Unmount Disk Image' de l'entorn BitCurator

Configurar i executar bitcurator_access_redaction

Tal i com es va exposar al capítol 4.3.3, l'eina bitcurator_access_redaction ja permet fer un redacció automàtica dels continguts d'una imatge *raw*, i per tant serà necessari instal·lar aquest programari. Un cop fet això, utilitzarem dos fitxers per poder executar-lo segons les nostres necessitats: per una banda, farem servir el *script* '9-Redacció de fitxers.sh' que mostrem a la Figura 46, i el fitxer de configuració 'imatge_raw_config.txt'. Per poder executar-lo adequadament és important verificar que la imatge *raw* que hem creat al capítol 5.2.2 no es trobi muntada dins el sistema.

El comanament que s'executa és 'redact-cli' que, amb l'opció -c, serveix per executar el fitxer de configuració 'imatge_raw_config.txt' que mostrem a la Figura 47. Dins la capçalera d'aquest fitxer, la primera línia indica la ruta completa de la imatge *raw* de base amb la sintaxi INPUT_FILE. A la segona línia indiquem la ruta i el nom que rebrà el nom de la imatge *raw* redactada, que serà 'imatge_raw_redacted.dd' amb la sintaxi OUTPUT_FILE. Finalment, a la tercera línia indiquem la ruta i el nom que rebrà

l'informe en format JSON que es generarà de forma automàtica i que ens informa dels fitxers que es redactaran.

Figura 46. *Script* utilitzat per executar bitcurator_access_redaction

```
#!/bin/sh

# script per bloquejar o redactar fitxers amb PII

echo El comanament redact-cli fa les redaccions en funció de la
configuració indicada al fitxer imatge_raw_config.txt, el qual genera
una nova imatge raw redactada i un informe de les redaccions en format
JSON

redact-cli -c ~/Scripts/imatge_raw_config.txt
```

Font: L'autor

Figura 47. Fitxer de configuració per redactar continguts

```
INPUT_FILE /home/bcadmin/Workspace/imatge_raw.dd
OUTPUT_FILE /home/bcadmin/Workspace/imatge_raw_redacted.dd
REPORT_FILE /home/bcadmin/Workspace/imatge_raw_redacted.json

# Busca un fitxer determinat i evita la seva execució (només per a fitxers .exe i .dll de
l'entorn Windows)
FILE_NAME_MATCH Umlet/Umlet.exe FUZZ

# Busca el fitxer amb el valor hash indicat i sobreescriu el contingut amb codis ASCII
0x2a (caràcter "*")
FILE_MD5 4a0a0ccde6b33ded84d9ea9a6a44f2ec FILL 0x2a

# Busca el fitxer amb el valor hash indicat i sobreescriu el contingut amb codis ASCII
0x2a (caràcter "*")
FILE_SHA1 a68c5a9cec927d8da3a873c5147dda73e6f22532 FILL 0x2a

# Omple seqüències que continguin Wilderbeeek amb el codi ASCII 0x2a (caràcter "*")
SEQ_EQUAL Wilderbeek FILL 0x2a

# Omple seqüències que continguin tlopezwi@gmail.com amb el codi ASCII 0x2a (caràcter
"*")
SEQ_EQUAL tlopezwi@gmail.com FILL 0x2a

# Omple seqüències amb l'expressió regular corresponent al DNI i al NIE amb el codi ASCII
0x2a (caràcter "*")
SEQ_MATCH (([X-Z]{1})([-]?)(\d{7})([-]?)([A-Z]{1}))|((\d{8})([-]?)([A-Z]{1})) FILL 0x2a

# Omple seqüències amb l'expressió regular corresponent a números de telèfon d'Espanya
amb el codi ASCII 0x2a (caràcter "*")
SEQ_MATCH [9|6|7][0-9]{8} FILL 0x2a

# Busca tots els fitxers amb el patró indicat i els sobreescriu amb zeros
FILE_NAME_MATCH acronims.* SCRUB

# Busca tots els fitxers dins del directori indicat i els sobreescriu amb zeros
FILE_DIRNAME_EQUAL geotiff SCRUB

# Ignora els fitxers els noms dels quals tinguin l'expressió regular indicada (repetible)
IGNORE *.tif

# Executa la redacció (genera una imatge de disc raw redactada)
COMMIT
```

Font: L'autor

Les opcions de redacció ja es van esmenar a la Taula 25, però indicarem que amb les opcions de cerca de seqüències de text, com és el cas de SEQ_EQUAL o SEQ_MATCH,

és necessari ser molt curosos amb les expressions regulars, ja que amb el cas de seqüències com els documents d'identitat del tipus 12345678-X, el programari redactarà totes aquestes seqüències. Si fem servir aquestes opcions a tots els fitxers de la imatge *raw*, es pot córrer el risc de esborrar informació pertinent. De fet, la redacció d'una imatge *raw* té el potencial de deixar el sistema de fitxers (o fitxers que es trobin dins el sistema de fitxers) en un estat inconsistent (Woods; Lee, 2015). Per tant, és necessari que aquest fitxer de configuració s'elabori amb cura i sempre comprovant amb posterioritat les modificacions fent servir l'informe en format JSON. Si es necessita, aquest fitxer es pot migrar a un nou informe en format XML mitjançant eines en línia de conversió com *XML to JSON and JSON to XML converter online*[324].

Muntar imatge *raw* redactada

Per poder comprovar si les redaccions s'han realitzat correctament, hem de muntar la imatge *raw* redactada, que s'haurà generat dins 'Workspace' amb el nom 'imatge_raw_redacted.dd'. Podem fer-ho de dues maneres: amb les opcions de *script* de l'entorn BitCurator que es mostren a la Figura 34 o bé crear un nou *script* per muntar la imatge redactada. Nosaltres hem creat el *script* '10-Muntar imatge redactada.sh' que és essencialment igual que el mostrat a la Figura 35, però canviant la ruta de la imatge forense per la ruta de la imatge *raw* redactada.

Comprovar redaccions

Per tal de comprovar redaccions, obrirem el fitxer JSON que es va generar amb el nom 'imatge_raw_redacted.json', amb especial atenció a l'etiqueta 'filename', que és la que indica el fitxer que ha estat redactat, i consultarem que les redaccions s'hagin realitzat correctament en funció de la configuració que hem indicat al nostre fitxer 'imatge_raw_config.txt' navegant dins els continguts de la imatge muntada i comprovant els fitxers un per un. És molt important ser curosos durant aquest procés, ja que si s'han utilitzat expressions regulars al fitxer de configuració és molt possible que s'hagin redactat fitxers que no havien de ser modificats.

[324] <http://www.utilities-online.info/xmltojson>. [Consulta: 07/01/2017]

Errors a la redacció?

En el cas en què s'hagin produït errors a la redacció, com que s'hagin bloquejat fitxers sense dades confidencials o que no s'hagin redactat adequadament, haurem de desmuntar la imatge redactada, revisar el fitxer de configuració, canviar les opcions corresponents i tornar a executar bitcurator_access_redaction. En cas negatiu, passarem al procés de normalització de formats.

Desmuntar imatge *raw* redactada

Desmuntarem la imatge redactada de la mateixa manera que amb la imatge forense, ja sigui amb el nostre *script* '11-Desmuntar imatge redactada.sh' o bé amb l'opció inclosa a l'entorn BitCurator. Aquest *script* també és essencialment igual que el mostrat a la Figura 45, però indicant la imatge *raw* redactada.

Revisar configuració

Els errors dins el processat es poden donar per diversos factors com no haver indicat correctament els valors *hash*, atès que la diferència d'un sol caràcter dins el valor pot invalidar el resultat. Juntament amb la llista de fitxers amb dades confidencials que hem elaborat i l'examen del fitxer JSON, revisarem el fitxer de text de configuració i repetirem el procés anterior.

Identificar fitxers amb formats no preferits

L'operació final serà la d'identificar quins fitxers s'han de migrar a formats sostenibles per a la preservació. Evidentment, cada institució tindrà polítiques diferents, però a tall orientatiu, es realitzarà aquesta operació tot basant-nos en els resultats observats a la Taula 10, que van permetre fer una síntesi de formats preferits per a la preservació. Detallem a continuació les tipologies de fitxers i els formats concrets:

- Text (amb i sense llenguatge de marques): utilitzarem PDF, text pla, XML, SGML i Open Document Text
- Fulls de càlcul i dades tabulars: utilitzarem OpenDocument Spreadsheet, CSV i SPSS
- Imatges: utilitzarem TIFF i JPEG

- Gràfics vectorials: utilitzarem SVG i AutoCAD
- Àudio: utilitzarem WAV, BWF i FLAC
- Vídeo: utilitzarem formats amb còdecs MPEG-2 i MPEG-4
- Dades geoespacials: utilitzarem ESRI Shapefile i GeoTIFF
- Gràfics 3D: utilitzarem X3D

D'aquesta operació de migració de fitxers s'exclouran aquells que no es trobin dins aquestes categories i quant als fitxers amb informació privada, es migraran també els fitxers restringits i els que ja hagin estat redactats. Un cop identificats els fitxers amb formats que s'han de migrar per ajudar a la seva preservació, es llistaran aquests fitxers en un nou full de càlcul que rebrà el nom de 'Migration.ods' (que també guardarem amb format CSV) dins 'Workspace', i contindrà la següent informació:

- Nom del fitxer original
- MIME type (del fitxer original)
- Programari (que s'utilitzarà per a la migració)
- Nom del fitxer migrat
- MIME type (del fitxer migrat)

Un exemple de com podria quedar el full de càlcul el mostrem a la Taula 39. Per a aquest cas s'han seleccionat dos fitxers d'imatge, un en format Photoshop i un altre en format PNG i per migrar de format al més apte per a la preservació a llarg termini (TIFF) s'utilitzarà ImageMagick (ja instal·lat per defecte a BitCurator).

Taula 39. Exemple de llista de fitxers a migrar

Nom del fitxer original	MIME type	Programari	Nom del fitxer migrat	MIME type
digital lives esquema.psd	image/vnd.adobe.photoshop	ImageMagick	digital lives esquema_migrated.psd	image/tiff
workflow.png	image/png	ImageMagick	workflow_migrated.tif	image/tiff

Font: L'autor

Executar Disk Image Access Interface

Un cop ja sabem els fitxers que volem normalitzar, utilitzarem la interfície Disk Image Access Interface per navegar dins l'estructura de fitxers d'una imatge forense i extreure els fitxers desitjats. Amb aquest objectiu, executarem un *script* que mostrem a la Figura 48.

Figura 48. *Script* utilitzat per executar Disk Image Access Interface

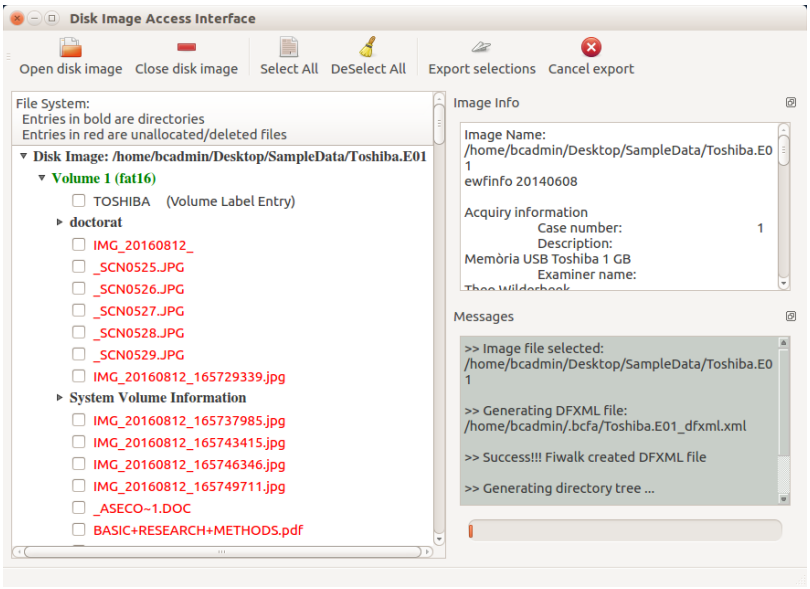
```
#!/bin/sh

# script per executar Disk Image Access Interface

echo Python 3 executa la interfície gràfica Disk Image Access Interface
python3 /usr/local/bin/bc_disk_access_v2.py
```

Font: L'autor

Extreure fitxers a normalitzar

Haurem de seleccionar la imatge forense de la qual volem extreure els fitxers, i un cop marcats aquells que ens interessen, els guardarem a 'Workspace' dins el subdirectori 'Pre-migration'. La raó d'això és que això permetrà automatitzar el procés de normalització de formats si tenim tots els fitxers que volem migrar a una mateixa carpeta. La Figura 49 mostra com es visualitzarien els fitxers un cop oberta una imatge de disc.

Figura 49. Interfície gràfica de Disk Image Access Interface



Font: Entorn BitCurator. Captura de l'autor

Normalitzar formats

Un cop extrets els fitxers a 'Pre-migration', utilitzarem el *script* '13-Migration.sh' per migrar els formats Photoshop i PNG a TIFF, que utilitza ImageMagick amb la instrucció 'mogrify' per crear fitxers nous amb el format normalitzat; aquests nous fitxers tenen el sufix '_migrated' per tal de tenir constància evident que aquests fitxers han passat per aquest procés. Tal com es mostra a la Figura 50, es treballa amb els fitxers dins la ruta '/home/bcadmin/Workspace/Pre-migration', on es reconeixen els fitxers amb extensió .psd (el 0 indica que s'han de agrupar totes les capes de la imatge en un sol fitxer, i no processar cada capa en imatges separades) i .png. Els fitxers nous es mouen a la ruta '/home/bcadmin/Workspace/Post-migration' on s'afegeix el sufix '_migrated'.

Figura 50. *Script* utilitzat per migrar fitxers d'imatges rasteritzades

```
#!/bin/sh

# script per migrar fitxers PNG i PSD a TIF

echo Els fitxers es converteixen a TIF, es mouen a la ruta assignada a
Workspace i es renombren amb el sufix _migrated
cd /home/bcadmin/Workspace/Pre-migration
mogrify -format tif *.psd[0]
mogrify -format tif *.png
mv *.tif /home/bcadmin/Workspace/Post-migration
cd /home/bcadmin/Workspace/Post-migration
rename 's/\.tif$/_migrated.tif/' *.tif
```

Font: L'autor

Hem preferit mostrar només un exemple per migrar imatges en lloc de mostrar cas per cas totes les opcions de migració per a àudio, vídeo, text, dades geoespacials o gràfics 3D perquè seria un treball massa llarg i complex i tampoc és l'objectiu d'aquesta tesi. Recomanem, sempre que sigui possible, fer aquests processos de forma automatitzada, tal com ja es fa a la Koninklijke Bibliotheek dels Països Baixos (Strodl et al., 2007), en què es fa migració a formats PDF/A per a text i TIFF per a imatges. A la Taula 9 també es poden consultar algunes eines de codi obert per a normalització de formats que fa servir el programari Archivematica.

Guardar fitxers generats

Ja s'han generat tots els fitxers digitals possibles dins el flux de treball i per tant és el moment de guardar-los en les rutes corresponents mitjançant el *script* '14-Reanomenar i moure fitxers.sh' que vam mostrar a la Figura 32. Per tal de tenir una panoràmica completa d'aquests fitxers, mostrem les rutes i els fitxers on es generen per defecte i on s'arxivaran a la Taula 40.

Entre tots els fitxers, alguns no es preservaran degut a diversos factors i per tant s'esborraran. Esmenarem a continuació aquests fitxers individuals o grups de fitxers i expliquem les causes del seu esborrat:

- imatge_raw.dd. Aquest fitxer és una imatge *raw* que es crea amb Guymager només per poder crear més endavant la imatge redactada imatge_raw_redacted.dd. Un cop tenim la imatge redactada ja no necessitem aquesta imatge *raw*
- imatge_raw.info. Aquest fitxer és un registre de captura, que es genera automàticament amb Guymager quan generem imatge_raw.dd. L'esborrem perquè és informació redundant, atès que ja tenim un registre de captura amb la imatge forense imatge.E01 que tindrà la mateixa informació
- fiwalk_deleted_files.pdf. Aquest fitxer el crea bitcurator_reports com a part del conjunt d'informes de l'anàlisi de contingut de la imatge forense. La raó d'esborrar aquest informe és degut a què consisteix en una llista de fitxers esborrats o localitzats en parts no assignades, la qual aporta una informació que el centre de preservació no necessita, ja que l'objecte de preservació són les dades de recerca que es troben efectivament al suport, i no altres fitxers que no tenen res a veure amb aquest objecte de preservació
- fiwalk_report.pdf. La raó d'esborrar aquest informe és que ja tenim una nova versió amb la marca d'aigua de l'ID del suport, fiwalk_report_mark.pdf, i per tant no es necessita preservar aquest fitxer amb informació redundant
- bulk_extractor_report.pdf. Mateix cas que amb fiwalk_report.pdf
- bc_format_bargraph.pdf. Mateix cas que amb fiwalk_report.pdf
- format_table.pdf. Mateix cas que amb fiwalk_report.pdf
- droid.pdf. Mateix cas que amb fiwalk_report.pdf

- Extraccions de dades confidencials. Un cop ja hem verificat les dades confidencials i les hem bloquejat o redactat, no caldrà retenir les extraccions individuals que ha fet BitCurator
- Fitxers seleccionats per ser migrats. Aquests fitxers s'han extret de la imatge forense mitjançant Disk Image Access Interface i els esborrarem perquè seria informació redundant; aquests fitxers continuaran existint dins la imatge forense

Taula 40. Rutes i fitxers d'origen i d'arxiu al flux de treball

Ruta original	Nom original del fitxer	Ruta d'arxiu	Nom normalitzat del fitxer
/home/bcadmin/Workspace	Form.pdf	/home/bcadmin/Arxiu intern/ID investigador/A-Documentació interna/A1-Formulari del dipositant	YYYYMMDD-ID investigador-Form.pdf
/home/bcadmin/Workspace	Medialog.ods	/home/bcadmin/Arxiu intern/ID investigador/A-Documentació interna/A2-Registre de suports físics	YYYYMMDD-ID investigador-Medialog.ods
/home/bcadmin/Workspace	ID suport_01.jpg	/home/bcadmin/Arxiu intern/ID investigador/B-Fotografies	YYYYMMDD-ID suport_01.jpg
/home/bcadmin/Workspace	ID suport_02.jpg	/home/bcadmin/Arxiu intern/ID investigador/B-Fotografies	YYYYMMDD-ID suport_02.jpg
/home/bcadmin/Workspace	ID suport_03.jpg	/home/bcadmin/Arxiu intern/ID investigador/B-Fotografies	YYYYMMDD-ID suport_03.jpg
/home/bcadmin/Workspace	ID suport_04.jpg	/home/bcadmin/Arxiu intern/ID investigador/B-Fotografies	YYYYMMDD-ID suport_04.jpg
/home/bcadmin/Workspace	imatge.E01	/home/bcadmin/Arxiu intern/ID investigador/C-Captura de suport/C1-Imatges de disc	YYYYMMDD-ID suport.E01
/home/bcadmin/Workspace	imatge_raw.dd	No s'arxiva	No aplicable
/home/bcadmin/Workspace	imatge_raw.info	No s'arxiva	No aplicable
/home/bcadmin/Workspace	imatge_raw_redacted.dd	/home/bcadmin/Arxiu intern/ID investigador/C-Captura de suport/C1-Imatges de disc	YYYYMMDD-ID suport-redacted.dd
/home/bcadmin/Workspace	imatge.info	/home/bcadmin/Arxiu intern/ID investigador/C-Captura de suport/C2-Registres de captura	YYYYMMDD-ID suport.info
/home/bcadmin/Workspace/bitcurator_reports	fiwalk-output.xml	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D1-Metadades DFXML	YYYYMMDD-ID suport-fiwalk-DFXML.xml
/home/bcadmin/Workspace/bulk_extractor	report.xml	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D1-Metadades DFXML	YYYYMMDD-ID suport-bulk_extractor-DFXML.xml

Ruta original	Nom original del fitxer	Ruta d'arxiu	Nom normalitzat del fitxer
/home/bcadmin/Workspace/bitcurator_reports/reports	premis.xml	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D2-Metadades PREMIS	YYYYMMDD-ID suport-PREMIS.xml
/home/bcadmin/Workspace	clamav.log	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D3-Registre d'antivirus	YYYYMMDD-ID suport-clamav.log
/home/bcadmin/Workspace/bulk_extractor	Fitxers amb dades confidencials en formats diversos	No s'arxiva	No aplicable
/home/bcadmin/Workspace/bitcurator_reports/annotated-features	Fitxers amb dades confidencials en text pla	No s'arxiva	No aplicable
/home/bcadmin/Workspace/bitcurator_reports/reports/features	Fitxers amb dades confidencials en Excel	No s'arxiva	No aplicable
/home/bcadmin/Workspace/bitcurator_reports	fiwalk_deleted_files.pdf	No s'arxiva	No aplicable
/home/bcadmin/Workspace/bitcurator_reports/reports	fiwalk_report.pdf	No s'arxiva	No aplicable
/home/bcadmin/Workspace/bitcurator_reports/reports	fiwalk_report_mark.pdf	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D4-Informes i llistats	YYYYMMDD-ID suport-fiwalk-Report.pdf
/home/bcadmin/Workspace/bitcurator_reports/reports	fiwalk-output.xml.xlsx	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D4-Informes i llistats	YYYYMMDD-ID suport-fiwalk-Filelist.xlsx
/home/bcadmin/Workspace/bitcurator_reports/reports	bulk_extractor_report.pdf	No s'arxiva	No aplicable
/home/bcadmin/Workspace/bitcurator_reports/reports	bulk_extractor_report_mark.pdf	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D4-Informes i llistats	YYYYMMDD-ID suport-bulk_extractor-Report.pdf
/home/bcadmin/Workspace/bitcurator_reports/reports	bc_format_bargraph.pdf	No s'arxiva	No aplicable
/home/bcadmin/Workspace/bitcurator_reports/reports	bc_format_bargraph_mark.pdf	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D4-Informes i llistats	YYYYMMDD-ID suport-bulk_extractor-Formats.pdf
/home/bcadmin/Workspace/bitcurator_reports/reports	format_table.pdf	No s'arxiva	No aplicable

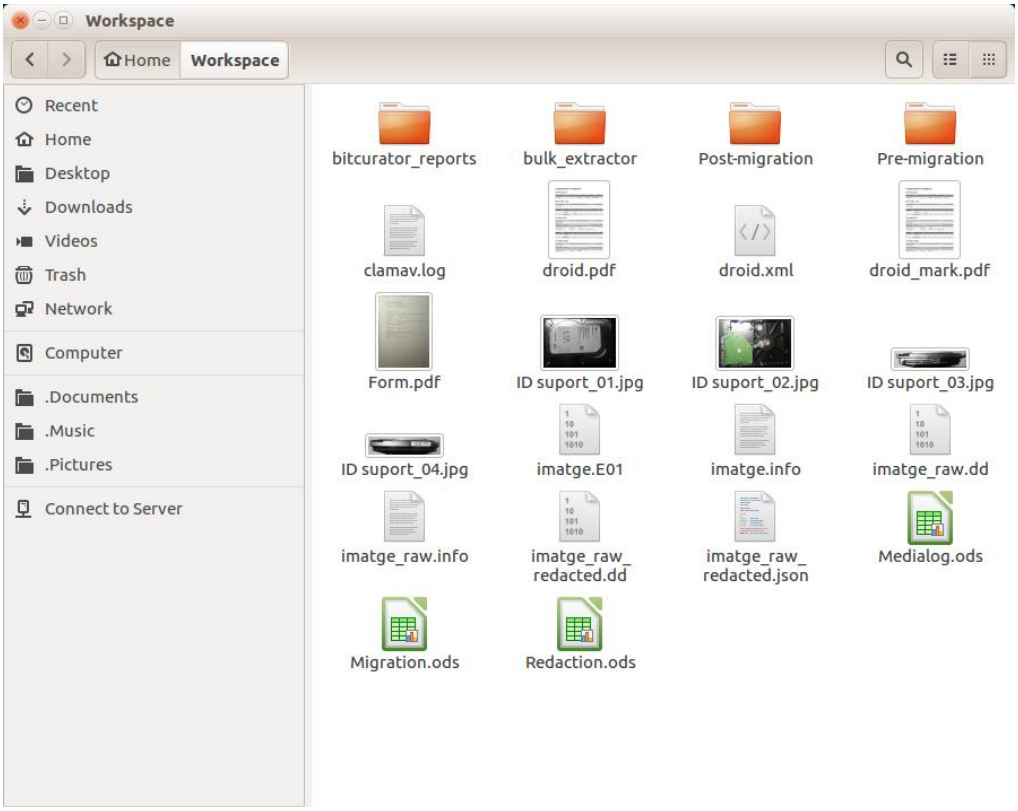
Ruta original	Nom original del fitxer	Ruta d'arxiu	Nom normalitzat del fitxer
/home/bcadmin/Workspace/bitcurator_reports/reports	format_table_mark.pdf	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D4-Informes i llistats	YYYYMMDD-ID suport-bitcurator_reports-Formats
/home/bcadmin/Workspace	droid.xml	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D4-Informes i llistats	YYYYMMDD-ID suport-DROID-Formats.xml
/home/bcadmin/Workspace	droid.pdf	No s'arxiva	No aplicable
/home/bcadmin/Workspace	droid_mark.pdf	/home/bcadmin/Arxiu intern/ID investigador/D-Examen de contingut/D4-Informes i llistats	YYYYMMDD-ID suport-DROID-Formats.pdf
/home/bcadmin/Workspace	Migration.ods	/home/bcadmin/Arxiu intern/ID investigador/E-Processat de continguts/E1-Llista de fitxers a migrar	YYYYMMDD-ID suport-Migration.ods
/home/bcadmin/Workspace	Migration.csv	/home/bcadmin/Arxiu intern/ID investigador/E-Processat de continguts/E1-Llista de fitxers a migrar	YYYYMMDD-ID suport-Migration.csv
/home/bcadmin/Workspace	Redaction.ods	/home/bcadmin/Arxiu intern/ID investigador/E-Processat de continguts/E2-Llista de fitxers a redactar o bloquejar	YYYYMMDD-ID suport-Redaction.ods
/home/bcadmin/Workspace	Redaction.csv	/home/bcadmin/Arxiu intern/ID investigador/E-Processat de continguts/E2-Llista de fitxers a redactar o bloquejar	YYYYMMDD-ID suport-Redaction.csv
/home/bcadmin/Workspace	imatge_raw_redacted.json	/home/bcadmin/Arxiu intern/ID investigador/E-Processat de continguts/E3-Informes de redaccions de dades confidencials	YYYYMMDD-ID suport-bitcurator_access_redaction.json
/home/bcadmin/Workspace/Pre-migration	Fitxers seleccionats per ser migrats	No s'arxiven	No aplicable
/home/bcadmin/Workspace/Post-migration	Fitxers migrats	/home/bcadmin/Arxiu intern/ID investigador/F-Fitxers normalitzats	Addició del sufix '_migrated' al nom original del fitxer

Font: L'autor

Per altra banda, el *script* també selecciona aquells fitxers que s'hagin seleccionat per ser preservats al repositori i els comprimeix a un fitxer ZIP, amb el nom de fitxer normalitzat 'ID suport.zip'. Aquest fitxer contindrà la imatge forense juntament amb una imatge *raw* redactada si el suport original contenia dades confidencials. A més, hi ha el cas dels fitxers migrats que no caldrà generar en tots els casos; tot dependrà de la

política de la institució. A la Figura 51 mostrem com quedaria l'estructura de fitxers i carpetes al final de tot el flux de treball.

Figura 51. Estructura de carpetes i fitxers a 'Workspace' al final del flux de treball



Font: L'autor

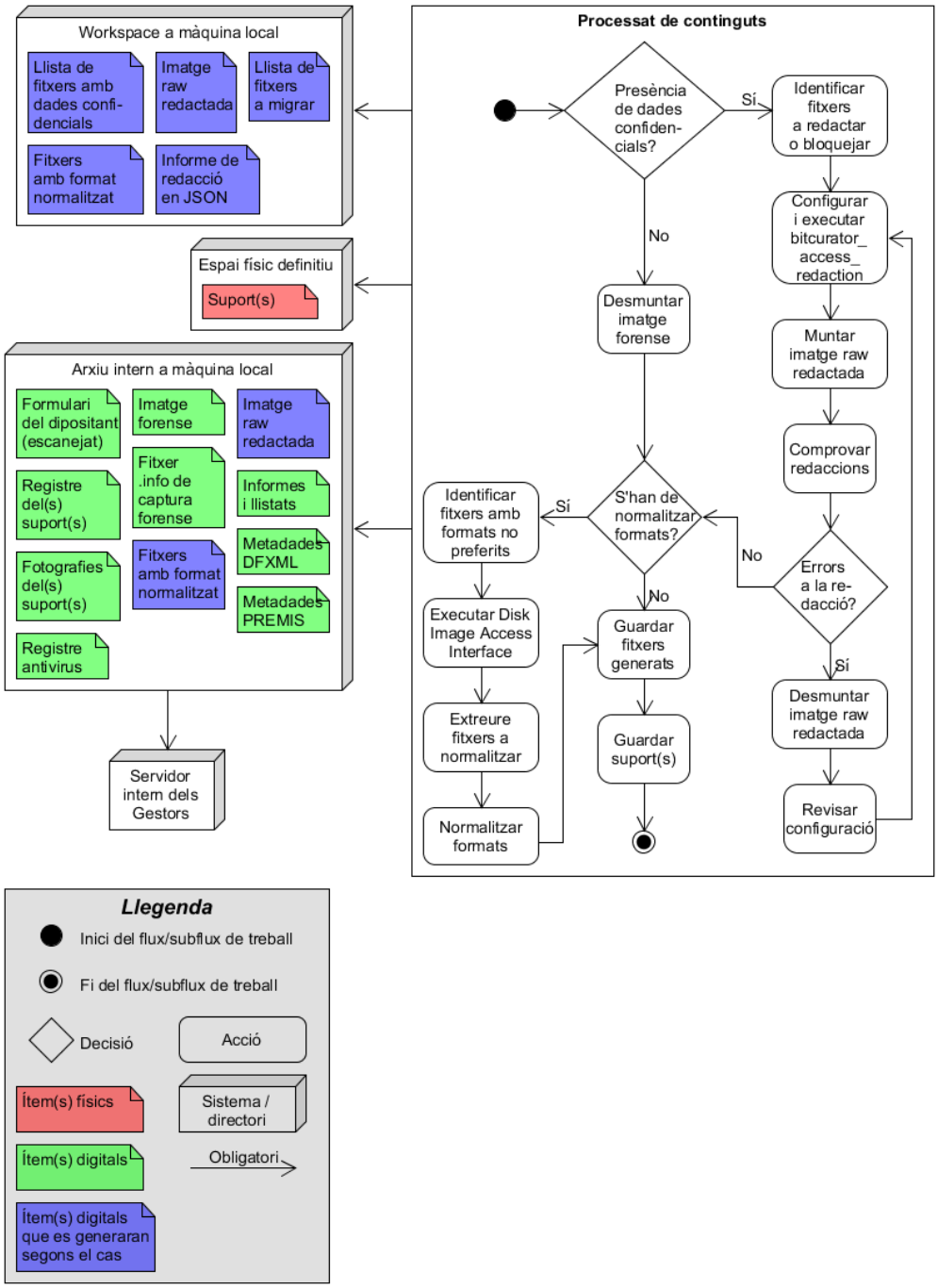
Guardar suport(s)

Un cop finalitzades les operacions de guardar els fitxers generats a les rutes definitives, serà el moment de guardar el(s) suport(s) de forma física a un lloc adequat. Tal i com vam indicar al fitxer 'Medialog.ods', guardarem els suports en funció del seu topogràfic.

Aquest conjunt de tasques ha constat inicialment de dues parts; en la primera s'ha treballat amb dades confidencials en què s'ha creat una llista de fitxers a redactar o bloquejar degut a la presència de dades confidencials, s'ha configurat i executat l'eina bitcurator_access_redaction que ha generat una nova imatge *raw* redactada i s'han comprovat les redaccions dins la nova imatge. En la segona part, s'han identificat fitxers imatge que s'han de migrar a formats adients per a la preservació, s'han extret aquests fitxers a una carpeta concreta de 'Workspace' i seguidament s'ha executat un *script* per migrar de forma ràpida i senzilla les imatges a TIFF i moure-les a una altra carpeta dins

'Workspace'. Finalment, s'han guardat els fitxers generats al flux de treball a les seves rutes definitives mitjançant un *script* en què s'han esborrat alguns fitxers amb informació redundant o no adequada per a la preservació i s'han guardat definitivament els suports físics a la seva ubicació segons el topogràfic indicat al fitxer 'Medialog.ods'. A la Figura 52 mostrem el subflux de treball corresponent.

Figura 52. Subflux de treball corresponent al processat de continguts



Font: L'autor

5.2.5 Preparació dels paquets AIP per a la seva ingesta

Arribats a aquest punt, ja tenim tots els fitxers necessaris per a la ingesta final dels paquets AIP. No tots els fitxers que s'han adquirit dins els procés estaran destinats a la preservació, ja que tenim fitxers d'ús intern per al personal del repositori (com és el cas dels fulls de càlcul) i altres fitxers que s'ingestaran al repositori com a AIP i així poder preparar el DIP posteriorment. A la Taula 41 es mostren tots els fitxers que s'han creat o adquirit durant els procediments i s'expliquen les raons per les quals alguns s'ingestaran o no al repositori.

Taula 41. Diferents nivells de destinació dels fitxers creats o adquirits durant el flux de treball

Fitxer	Tipologia	Destinació	Observacions
Form.pdf	Formulari del dipositant	AIP	Accés restringit a l'administrador
Medialog.ods	Registre del suport (LibreOffice)	AIP	Accés restringit a l'administrador
ID suport_01.jpg i successius	Fotografies del suport físic	AIP	Accés restringit a l'administrador
imatge.E01	Imatge forense	AIP DIP (continguts seleccionats)	Accés restringit a l'administrador. Els continguts formaran part del DIP, prèvia exportació, si estan lliures de dades confidencials
imatge_raw.dd	Imatge *raw*	S'esborra	Imatge *raw* que s'esborra un cop s'ha creat imatge_raw_redacted.dd
imatge_raw_redacted.dd	Imatge *raw* amb redacció de dades confidencials	AIP DIP (continguts seleccionats)	Els continguts formaran part del DIP prèvia exportació
imatge.info	Fitxer .info de captura forense	AIP	Accés restringit a l'administrador
imatge_raw.info	Fitxer .info de captura d'imatge *raw*	S'esborra	Informació redundant
fiwalk-output.xml	Metadades tècniques DFXML de captura forense	AIP	Accés restringit a l'administrador. No formarà part del DIP
fiwalk_deleted_files.pdf	Informe de fitxers esborrats al suport original	S'esborra	Llistat de fitxers esborrats al suport; dades confidencials que no es preserven
report.xml	Metadades tècniques DFXML d'extracció de dades confidencials	AIP	Accés restringit a l'administrador. No formarà part del DIP

Fitxer	Tipologia	Destinació	Observacions
premis.xml	Metadades de preservació PREMIS	AIP	No formarà part del DIP
clamav.log	Registre antivirus	AIP	Accés restringit a l'administrador
fiwalk_report.pdf	Informe tècnic de particions	S'esborra	Es genera una nova versió amb marca d'aigua
fiwalk_report_mark.pdf	Informe tècnic de particions	Arxiu intern de la institució	Ús intern del personal del repositori
fiwalk-output.xlsx	Llistat de fitxers presents a la imatge forense	AIP	No formarà part del DIP
bulk_extractor_report.pdf	Informe d'extraccions de dades confidencials	S'esborra	Es genera una nova versió amb marca d'aigua
bulk_extractor_report_mark.pdf	Informe d'extraccions de dades confidencials	Arxiu intern de la institució	Ús intern del personal del repositori
bc_format_bargraph.pdf	Diagrama de barres de formats de fitxer presents a la imatge forense	S'esborra	Es genera una nova versió amb marca d'aigua
bc_format_bargraph_mark.pdf	Diagrama de barres de formats de fitxer presents a la imatge forense	AIP DIP	Formarà part del DIP; aporta informació sobre els continguts del *dataset*
format_table.pdf	Llistat de formats de fitxer presents a la imatge forense	S'esborra	Es genera una nova versió amb marca d'aigua
format_table_mark.pdf	Llistat de formats de fitxer presents a la imatge forense	AIP DIP	Formarà part del DIP; aporta informació sobre els continguts del *dataset*
droid.xml	Llistat de formats de fitxer presents a la imatge forense	AIP	No formarà part del DIP
droid.pdf	Llistat de formats de fitxer presents a la imatge forense	S'esborra	Es genera una nova versió amb marca d'aigua
droid_mark.pdf	Llistat de formats de fitxer presents a la imatge forense	AIP DIP	Formarà part del DIP; aporta informació sobre els continguts del *dataset*
Migration.ods	Llista de fitxers seleccionats per ser migrats (LibreOffice)	Servidor intern de la institució	Ús intern del personal del repositori
Redaction.ods	Llista de fitxers seleccionats per ser redactats o bloquejats (LibreOffice)	Arxiu intern de la institució	Ús intern del personal del repositori
imatge_raw_redacted.json	Informe de redaccions a la imatge *raw*	Arxiu intern de la institució	Ús intern del personal del repositori

Fitxer	Tipologia	Destinació	Observacions
No definit (addició del sufix _migrated)	Fitxers migrats	AIP	No formaran part del DIP; són formats dissenyats per a la preservació
Fitxers diversos	Extraccions de dades confidencials	S'esborren	Les dades confidencials no són objecte de preservació

Font: L'autor

Així doncs, podem diferenciar tres grups de fitxers: en primer lloc, tenim fitxers que s'esborren degut a què contenen dades confidencials o bé que es tracten de fitxers intermedis; en segon lloc, hi ha fitxers que no formaran part de l'AIP degut al seu ús intern i per tant la institució els arxivarà al seu propi servidor i en tercer lloc tenim aquells fitxers que sí s'ingestaran al repositori com a part d'un AIP i que formaran part o no del DIP accessible als Consumidors. El criteri que s'ha seguit per alliberar continguts ha estat principalment facilitar la visualització dels continguts del suport original a l'usuari i aportar-li informació rellevant; és per aquesta raó que hem deixat oberts els informes de formats de fitxer i llistats de fitxers.

Presència de dades confidencials/fitxers migrats?

Hem de tenir en compte que si s'ha creat una imatge *raw* redactada per la presència de dades confidencials i/o s'ha executat el procés de migració de formats, l'execució dels processos del capítol 5.2.5 pot variar. En el primer cas, es pot produir el cas de crear un paquet AIP que inclogui dues imatges de disc (la forense i la *raw* redactada) i en el segon cas, l'addició dels fitxers migrats.

Crear paquet AIP (BagIt)

Per tal de facilitar la ingesta, utilitzarem l'especificació BagIt que permet la creació de paquets que conserven l'estructura original de carpetes i que crea metadades sobre el contingut del paquet (vegeu el capítol 3.3.4), així com manifestos amb la relació de fitxers i de carpetes, dels valors *hash* i de la mida total del paquet. Amb aquest objectiu, s'ha creat un *script* que genera un paquet BagIt, que mostrem a la Figura 53 (a tall il·lustratiu, l'hem omplert amb dades fictícies) i que es pot editar per a cada cas diferent

d'ingesta. Per tant, aquest seria un procés manual pel que respecta a l'edició de metadades.

Figura 53. *Script* utilitzat per generar paquets BagIt

```
#!/bin/sh

# script que genera els paquets BagIt
echo Python executa el script bagit.py a la ruta especificada
cd /usr/local/bin
sudo python3 bagit.py --md5 --sha1 --source-organization 'Universitat de
Catalunya' --organization-address 'Carrer de Balmes, 500, 08022
Barcelona' --contact-name 'Dídac Reguant' --contact-phone '+34 934345578'
--contact-email 'dreguant@uc.cat' --external-description 'Imatge forense
Encase, imatge raw redactada i fitxers migrats per fer les proves de
workflow de preservació amb anàlisi forense digital' --external-
identifier '12345678USB01_1' --bag-group-identifier '12345678USB01' --
bag-count '1 de 1' --bag-size '1.7 GB' "/home/bcadmin/Ingesta AIP/ID
investigador"
```

Font: L'autor

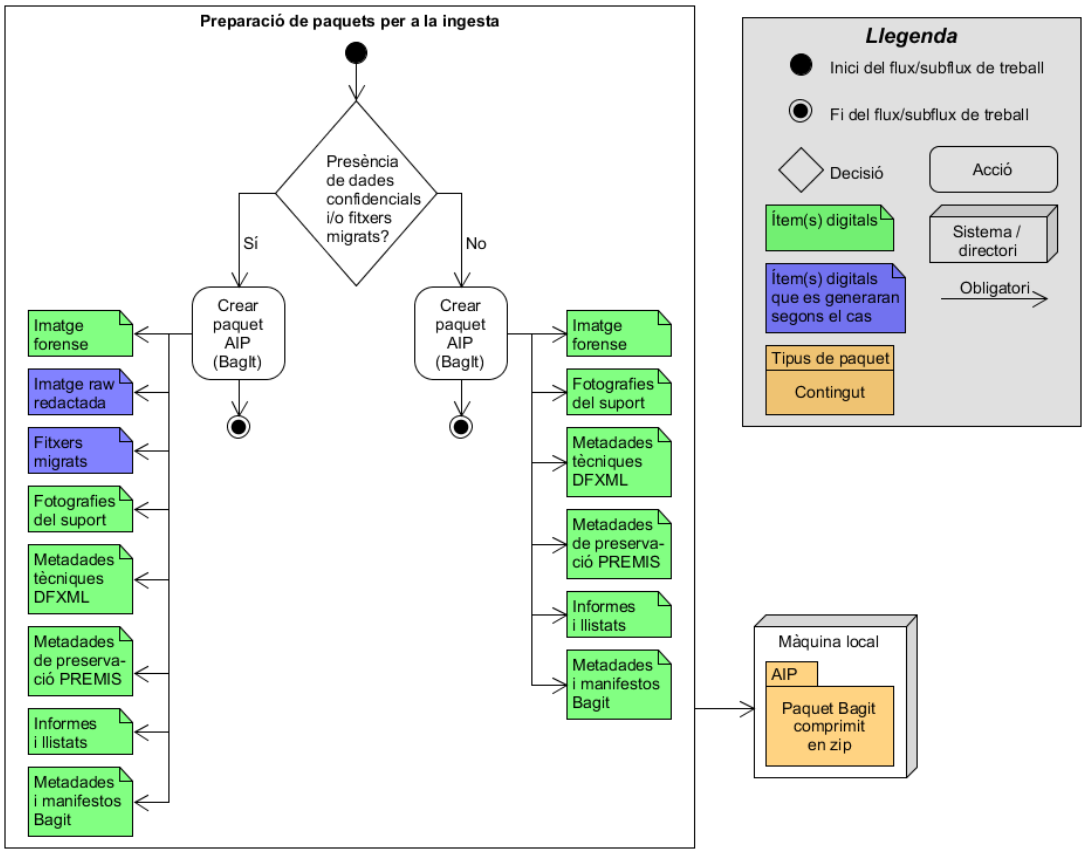
Aquest *script* localitza 'bagit.py', que es troba per defecte a BitCurator dins la ruta /usr/local/bin i l'executa a la ruta indicada dins '/home/bcadmin/Ingesta AIP'. Tal com indiquem al capítol 3.3.4, hi ha diferents opcions per crear metadades; per al nostre cas hem seleccionat que es generin valors *hash* MD5 i SHA1 dels continguts de cada paquet, hem indicat l'organització que rep les dades de recerca, l'adreça de l'organització, el nom i el telèfon del contacte de l'organització, una descripció externa del contingut de cada paquet, l'identificador del suport com identificador del grup de *bags*, que serà el mateix per a cada *bag* amb l'addició del número de *bag* en el cas de l'identificador de cada *bag* individual, el número de paquet dins el total de *bags* (1 de 2, 2 de 2, etc.) i la mida del *bag*. Sobre l'edició d'aquestes metadades, hem d'indicar que el corresponent a 'Bag Group Identifier' només l'utilitzarem quan haguem d'ingestar més d'un suport al repositori d'un mateix investigador i per tant, si només tenim un suport no caldrà utilitzar-lo. Per tant, aquest *script* s'haurà de tornar a editar amb el segon i successius suports.

El procés de preparar els paquets per a la ingesta el mostrem a la Figura 54. Com a resultat final, tindrem els següents fitxers, generats a '/home/bcadmin/Ingesta AIP/ID investigador':

- Directori 'data' amb els fitxers seleccionats per a la preservació. En el nostre cas, serien fitxers .zip

- Fitxer bag-info.txt amb metadades BagIt
- Fitxer bagit.txt amb informació de la versió de BagIt i la codificació de caràcters del fitxer d'etiquetes
- Fitxer manifest-md5.txt amb el llistat dels fitxers seleccionats per a la preservació i els seus valors *hash* MD5
- Fitxer tagmanifest-md5.txt amb el llistat de fitxers amb etiquetes creats per BagIt i els seus valors *hash* MD5
- Fitxer manifest-sha1.txt amb el llistat dels fitxers seleccionats per a la preservació i els seus valors *hash* SHA1
- Fitxer tagmanifest-sha1.txt amb el llistat de fitxers amb etiquetes creats per BagIt i els seus valors *hash* SHA1

Figura 54. Subflux de treball corresponent a la preparació de paquets per a la ingesta



Font: L'autor

5.2.6 Ingesta al repositori

En aquest capítol tractem la ingesta final dels paquets AIP al repositori i les dificultats tècniques que s'hauran de solucionar. Mostrem aquest subflux de treball a la Figura 55.

S'ha dipositat més d'un suport?

Si l'investigador ha dipositat més d'un suport amb dades de recerca, reiniciarem el flux de treball des del subflux captura de suport(s) del capítol 5.2.2. En cas contrari, podrem passar a la ingesta del paquet AIP al repositori, tot tenint en compte els nivells de privilegi necessari per a l'accés. En aquest cas, l'AIP no serà el mateix que el DIP, ja que els continguts seran accessibles o no al Consumidor segons certes condicions que veurem al capítol 5.3.

AIP superior a 4 GB?

Per ingestar l'AIP en format BagIt caldria que el nostre DSpace tingui configurat i habilitat el complement Replication Task Suite[325], i s'hauria de fer la ingesta per una via diferent de SWORD, el servei de dipòsit web que està configurat per defecte a un repositori DSpace, ja que estem tractant amb paquets de gran mida que el protocol HTTP no accepta [326]. Específicament, DSpace no accepta l'enviament de fitxers superiors a 4 GB pel sistema d'enviament per web (vegeu el capítol 5.4.2).

Per tant, seria necessari que el repositori permeti l'enviament de fitxers amb altres tipus de protocols, com FTP, SFTP o SCP[327]. Per manca de recursos, no hem pogut fer proves d'ingesta per FTP però A. J. Prieto[328] ha indicat que tècnicament es pot fer aquest tipus d'ingesta a un repositori DSpace i segons Nash i Wheeler (2016), ja s'han realitzat experiències d'ingesta mitjançant el protocol FTP amb èxit. Un exemple és la University of Texas (que utilitzen un repositori DSpace), on s'ha ingestat una imatge de

[325] DuraSpace wiki. *Replication Task Suite*. <https://wiki.duraspace.org/display/DSPACE/ReplicationTaskSuite>. [Consulta: 24/10/2016]

[326] R. de la Vega, entrevista, 9 de març de 2017

[327] Pottinger, Hardy J. (2012, Aug. 30). *Ingesting large data set* [missatge a grup de discussió]. <http://dspace.2283337.n4.nabble.com/Ingesting-large-data-set-td4657204.html>. [Consulta: 02/04/2017]

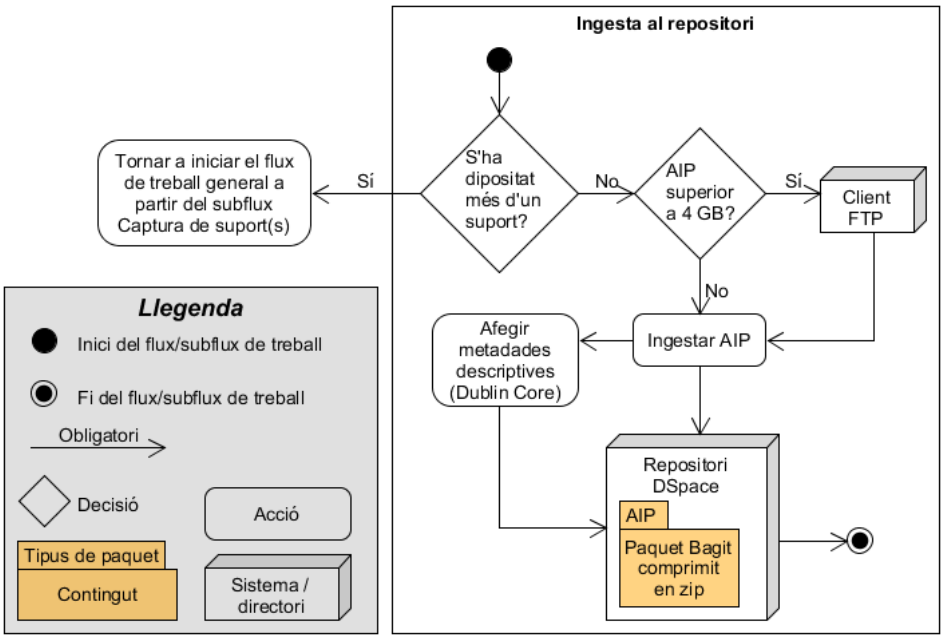
[328] Entrevista, 14 de febrer de 2017

disc dur de 40 GB corresponent als arxius personals del músic George Sanger[329], encara que el seu accés estigui restringit als administradors. Així doncs, indiquem al flux de treball l'ús d'un client FTP per pujar els continguts al servidor on es trobi allotjat el repositori.

Ingestar AIP

Un cop s'ha verificat que la ingesta és correcta, ja podrem esborrar el paquet BagIt de la màquina local o conservar-lo com a còpia de seguretat.

Figura 55. Subflux de treball corresponent a la ingesta al repositori



Font: L'autor

Afegir metadades descriptives (Dublin Core)

Si el repositori DSpace no té camps adequats per a totes les dades pròpies del pla de recerca (nom del pla de recerca, codi identificador, disciplina/àrea de coneixement, etc.), crearem aquests camps de forma manual en la nostra màquina local i en funció de certs criteris que exposarem al capítol 5.4.4. Recordem que al formulari també s'han inclòs camps relacionats amb informació de drets d'autor; un respecte a les llicències d'ús (si no s'indica, s'entendrà que les dades són obertes sense cap tipus de restricció d'ús) i un

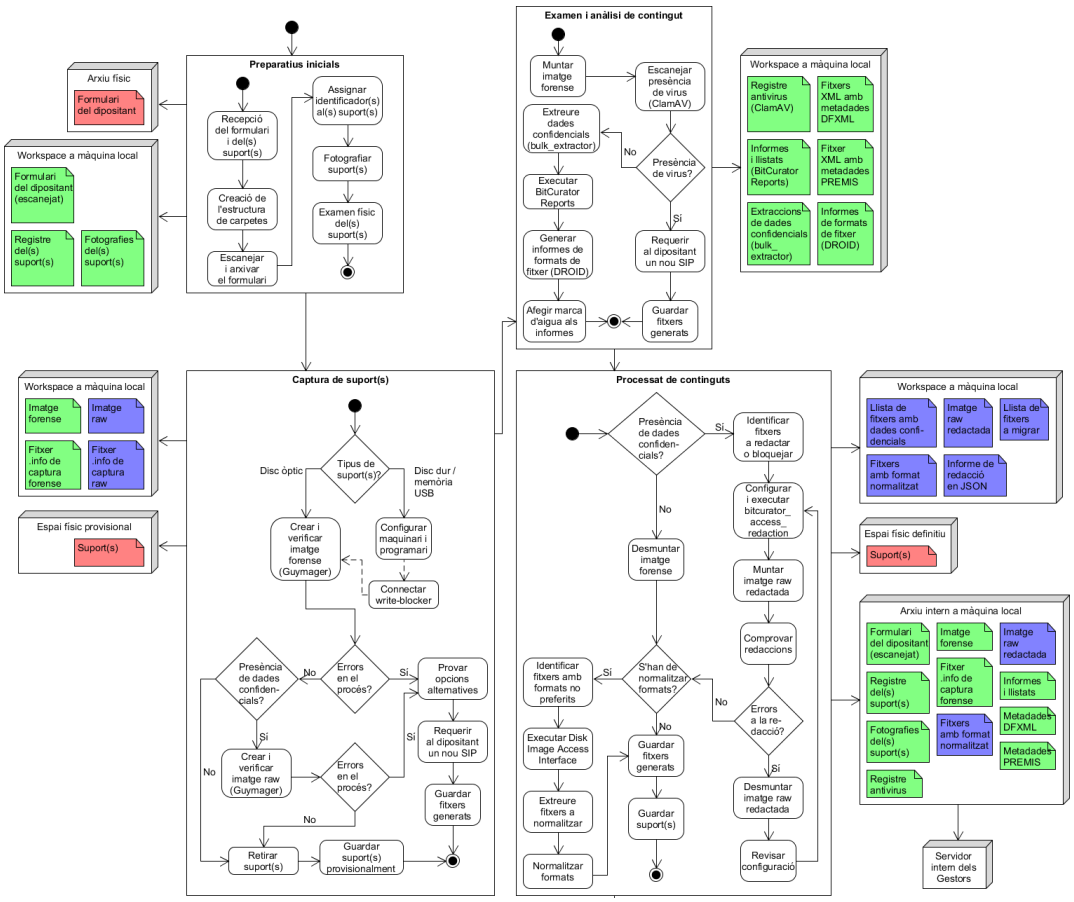
[329] Austin, Graham *et al. When the "fat-man" sings: archiving the George Sanger 8600/300 workstation*. <https://ford.ischool.utexas.edu/handle/2081/32088>. [Consulta: 05/04/2017]

altre respecte a fitxers amb informació personal i/o sensible que haurà de ser bloquejada o restringida. Aquesta informació aportada pel dipositant s'haurà d'incloure, en el primer cas, en el camp adequat de DSpace sobre llicències i en el segon cas, en un camp de Notes.

5.2.7 Flux de treball final

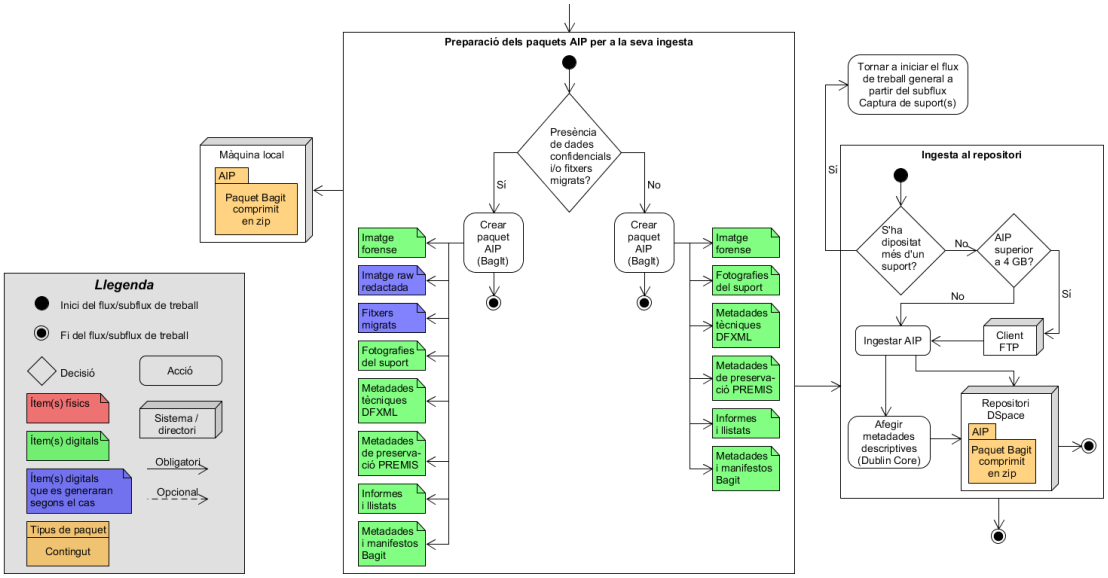
Un cop hem exposat com executarem els diferents processos de preparació de l'AIP on hem il·lustrat els seus subfluxos, mostrem el flux de treball final de tots els processos a la Figura 56, la qual, degut a la seva mida, hem dividit en dues parts.

Figura 56. Flux de treball del model de preservació



Font: L'autor, a partir dels subfluxos presentats al capítol 5.2

Figura 56. Flux de treball del model de preservació (continuació)



Font: L'autor, a partir dels subfluxos presentats al capítol 5.2

5.2.8 Síntesi d'operacions en terminologia OAIS

Els processos es poden categoritzar en set categories dins el model de referència OAIS, on cada tasca del flux de treball té un paper dins la gestió i/o creació dels Paquets d'Arxiu. La Taula 42 descriu les diferents categories que s'han contemplat al model teòric.

Taula 42. Categories OAIS dins el model teòric de preservació

Categoria	Tasca
1. Recepció del SIP	Recepció del formulari del dipositant Recepció dels suports
2. Revisió del SIP	Examen físic del suport Fotografiar suport
3. Captura del SIP	Configurar maquinari i programari Crear imatge forense Verificar imatge forense
4. Quarantena del SIP	Escanejar presència de virus
5. Anàlisi del SIP	Generar metadades Extreure dades privades i sensibles Generar informes de formats de fitxer
6. Preparació de l'AIP	Extreure fitxers de la imatge forense Identificar dades privades Bloquejar o redactar dades privades amb creació d'imatge *raw* Identificar fitxers amb formats no preferits per a la preservació Normalitzar fitxers
7. Emmagatzematge de l'AIP	Crear paquets i metadades BagIt Ingesta al repositori Afegir metadades descriptives

Font: L'autor

5.3 Preparació del DIP per al seu accés

Després de definir com es farà la ingesta del paquet AIP, expliquem a continuació els passos que es faran quan un usuari vulgui accedir als seus continguts, tant des del punt de vista de l'usuari com del personal del repositori. Aquesta versió de l'AIP, com ja hem explicat al capítol 3.2.1, és el DIP o Paquet d'Informació de Difusió, definit com "Versió del paquet d'informació que s'entrega al Consumidor en resposta a una petició d'accés".

5.3.1 Accions del Consumidor

Recordem que el Consumidor, en terminologia OAIS, es defineix com "Individus, organitzacions o sistemes que consumeixen, o utilitzen, la informació preservada a l'OAIS". Com ja vam exposar al capítol 3.2.1, són els usuaris d'un repositori o arxiu OAIS.

Consulta de continguts

El Consumidor navegarà dins el repositori i accedirà al registre de metadades Dublin Core del *dataset* on podrà consultar la seva descripció. L'AIP està enllaçat a aquest registre, però l'accés està restringit a l'administrador.

Sol·licitud d'accés

L'usuari haurà de demanar accés als continguts de l'AIP per qualsevol via permesa a les opcions del repositori (com per correu electrònic).

5.3.2 Accions de l'Administració

Recordem que l'Administració, en terminologia OAIS, s'encarrega de gestionar les operacions diàries de l'arxiu. Per tant, és el personal del repositori encarregat, en aquest cas, de gestionar les sol·licituds de DIPs i preparar-los per al seu accés.

Recepció de la sol·licitud

En aquesta fase s'inicia el flux de treball del personal del repositori, amb la recepció de sol·licitud d'accés.

Preparació del DIP per al Consumidor

El primer pas consistirà en executar el Disk Image Access Interface i s'haurà de consultar el contingut original de l'AIP; si conté una imatge *raw* redactada, s'haurà d'obrir aquesta última. En cas contrari, s'obrirà la imatge forense ja que aquesta s'entendrà com lliure de dades confidencials. Per tal d'optimitzar els recursos, s'haurà d'obrir la imatge corresponent als fitxers emmagatzemats a l'arxiu intern de la institució. Aquests fitxers i carpetes s'exportaran a la ruta '/home/bcadmin/DIP/ID investigador', a més d'aquells fitxers que s'han considerat d'accés obert dins la Taula 41.

Crear enllaç (emmagatzematge al núvol)

Un cop tinguem tots els continguts preparats, haurem de generar l'enllaç que servirà per compartir els continguts amb el Consumidor. Per tal de generar l'enllaç, la institució haurà de tenir contractar algun servei d'emmagatzematge al núvol, com Dropbox[330], Google Drive[331] o OneDrive[332].

Avís al Consumidor

Un cop s'hagi generat l'enllaç, es donarà avís al Consumidor de la disponibilitat dels continguts per la via que la institució consideri pertinent (com per correu electrònic). En funció del pla que tingui contractat la institució, pot donar un temps límit al Consumidor per accedir als continguts.

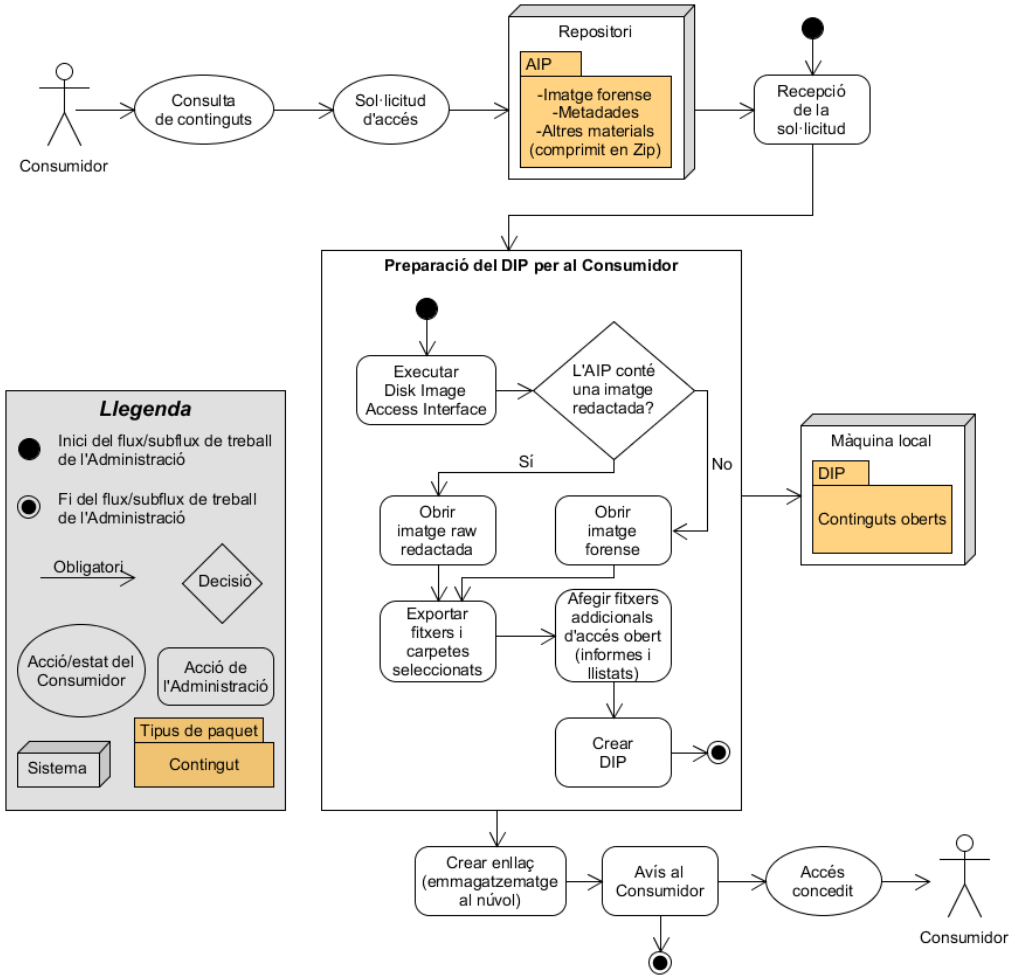
Amb això acabaria el flux de treball per donar accés als continguts., que mostrem en la seva totalitat en la Figura 57.

[330] <http://dropbox.com/>. [Consulta: 23/04/2017]

[331] <https://www.google.com/drive/>. [Consulta: 23/04/2017]

[332] <https://onedrive.live.com/>. [Consulta: 23/04/2017]

Figura 57. Flux de treball de l'accés als continguts al repositori



Font: L'autor

5.4 Com encaixa DSpace a la nostra proposta de preservació?

DSpace fou publicat l'any 2002 sota la llicència de codi obert BSD (Smith et al., 2003) com a resultat d'un projecte conjunt de Hewlett Packard Labs i el Massachusetts Institute of Technology (Kurtz, 2010), el qual va ser possible gràcies a una subvenció de l'Andrew W. Mellon Foundation (Chen; Zhang, 2014). Consisteix en un programari de codi obert, el qual tenia com a funcions inicials un model de dades per a la seva organització, metadades, *e-people* (usuaris i sistemes informatitzats), autorització, ingesta d'arxius, *workflow* i sistema *handle* de citacions (Tansley et al., 2003); aquest últim permet identificar els continguts de forma única i persistent (Kowalczyk; Shankar, 2011). Actualment existeixen les funcions addicionals de recollir estadístiques, Dublin

Core Meta Toolkit (mòdul que permet convertir informació d'altres bases de dades en metadades Dublin Core), temes Manakin (mòdul per personalitzar la interfície de repositori creat per la Texas Digital Library), funció d'embargament (que permet restriccions temporals d'accés), paquets d'idiomes i SWORD (protocol estàndard per fer un dipòsit de diferents fonts i tipus de fitxers). Avui en dia compta amb la comunitat d'usuaris de repositori més àmplia en tot el món i és el més utilitzat per part de biblioteques de recerca com a repositori institucional (Chen; Zhang, 2014). Compta amb el suport de DuraSpace [333], organització sense ànim de lucre que col·labora amb comunitats acadèmiques i científiques per facilitar l'accés al patrimoni digital.

Dins DSpace l'entitat bàsica és un ítem, que inclou tant metadades del tipus Dublin Core qualificat i contingut digital, que s'identifiquen amb el sistema de *handles* CNRI. L'organització dels ítems es fa per comunitats i col·leccions i l'emmagatzematge de dades es fa en una base de dades relacional PostgreSQL o bé Oracle. La cerca d'ítems és possible gràcies a la indexació de metadades bàsiques, que per defecte són del tipus títol, autor i data. Per tal de compartir fàcilment les metadades amb altre repositoris s'utilitza el protocol OAI-PMH amb suport d'OpenURL que facilita enllaços per a cada pàgina on es troba un ítem (Pirounakis; Nikolaidou, 2009). La seva instal·lació és una tasca prou complexa i requereix del programari addicional Java, Apache, PostgreSQL i TomCat (Rodríguez Gairín; Sulé Duesa, 2008).

Un 21% dels repositoris DSpace contenen *datasets*, i la tendència és que la gestió de dades poc a poc està augmentant progressivament (Chen; Zhang, 2014) i és previsible que en el futur pròxim més repositoris inclouran conjunts de dades de recerca. Un exemple d'integració de Data Management Plan als repositoris DSpace el trobem a la Texas Digital Library [334] i el seu repositori institucional Digital Collections Repository[335], que recull les funcions de captura de metadades bàsiques, informació de citació, informació de llicències de dades i paraules clau. Les dades queden emmagatzemades en dues còpies: una en línia dins el centre de dades i una altra en un arxiu no connectat a la xarxa.

[333] <http://www.duraspace.org/>. [Consulta: 05/06/2015]

[334] *Data Management and the TDL*. <http://www.tdl.org/members/resources/data-management/tdl-support/>. [Consulta: 10/06/2015]

[335] <https://digital.library.txstate.edu/>. [Consulta: 10/06/2015]

Per tal de demostrar fins a quin nivell DSpace és una bona solució quant a la funció de dipòsit de dades del nostre model, exposarem a continuació els resultats de les entrevistes que vam realitzar a responsables de repositoris i analitzarem els punts crítics que s'haurien de modificar dins el codi del programari. Per estudiar millor els límits de DSpace, vam instal·lar un repositori DSpace versió 5.1, fent servir una base de dades PosgreSQL, un servidor web Tomcat i un contenidor de *servlets* Tomcat a una màquina local sota el sistema operatiu Windows 7.

5.4.1 Entrevistes amb responsables de repositoris

Per tal de poder implementar DSpace com a repositori del nostre model, és necessari que el programari compleixi una sèrie de requisits. Amb l'objectiu de contrastar dades i verificar si la nostra proposta era realitzable, es van realitzar un total de sis entrevistes (per consultar el contingut del qüestionari, vegeu l'Annex C) a onze responsables de repositoris (vegeu la Taula 2) on es van abordar els fluxos de treball existents i les diferents limitacions de DSpace. Presentem els resultats de les entrevistes a la Taula 43 i els desenvolupem més endavant.

Recepció de suports físics

Hi ha dos casos d'institucions que reben suports físics. En el cas de la UB no tenen un flux de treball normalitzat, sinó que es "fa un procés *ad hoc* amb qualsevol suport que entri, adaptant els processos d'entrada[336]". Això seria en el cas de les tesis, que arriben en suport òptic o bé per memòria USB[337]. La UAB rep discs durs d'empreses externes de digitalització on "l'empresa externa es queda una còpia i nosaltres descarreguem els continguts als nostres servidors. Els discs durs es queden a la biblioteca com a segona còpia[338]".

[336] D. Iglesias, entrevista, 9 de gener de 2017

[337] I. Labastida; J. Casals, entrevista, 2 de febrer de 2017

[338] C. Azorín, entrevista, 30 de gener de 2017

Gestió i preservació de dades de recerca

Al Dipòsit Digital de la UB hi ha una col·lecció anomenada 'Dades de recerca' on es poden consultar alguns *datasets*. Per tal que l'investigador interessat els pugui dipositar al repositori, ha posar-se en contacte[336] amb els administradors. La UAB fa poc que ha començat a gestionar i preservar dades de recerca, els quals contenen plans de gestió de dades "degut sobretot a l'obligatorietat d'Horizon 2020[338]" i actualment estan treballant en el formulari per introduir les dades de recerca.

Perfil de Dublin Core per a dades de recerca

En cap institució vam trobar casos de perfils de DC realitzats especialment per a dades de recerca, però a la UB sí que s'utilitzen dos qualificadors per indicar on s'han obtingut les dades (coverage) i quan (temporal). "Quan un investigador envia dades, els hi demanem metadades bàsiques, però a més els demanem d'on són i quan les han recollit[337]".

Límit de mida per fitxer

Les respostes en aquest cas van ser variades. En el cas del Dipòsit Digital de la UB, la configuració permet la ingesta d'un fitxer amb una mida màxima de 500 GB[336, 337], mentre que al repositori O2 es pot ingestar un fitxer amb una mida d'1 GB[339]. Al Dipòsit Digital de Documents de la UAB es pot ingestar un fitxer amb un màxim de 2 GB; si un investigador realitzés alguna petició per ingressar un fitxer més gran requeriria una ampliació de maquinari o una solució a mida[338]. A UPCommons no s'han fet proves d'ingesta amb grans volums de dades[340] i al CSUC ens van informar de la problemàtica que ocasionaria posar un límit molt de mida molt gran per als fitxers, ja que hi ha un risc de caiguda dels repositoris si no s'escala el maquinari[341]; recordem que el CSUC manté repositoris col·laboratius amb 10 universitats.

[339] F. March; R. Padrós; C. Llueca, entrevista, 1 de febrer de 2017

[340] A. J. Prieto; J. Prats; A. Rovira, entrevista, 14 de febrer de 2017

[341] R. de la Vega, entrevista, 9 de març de 2017

Política de migració de formats

Dins la UB no hi ha una política explícita de migració de formats, però sí s'utilitza la funcionalitat de formats suportats per DSpace, així que hi ha una política implícita de formats acceptats, on s'ha produït algun cas de migració *ad hoc* com el d'una sèrie de pàgines web a PDF[336]. Una situació semblant succeeix a la UAB, on hi ha una sèrie de formats acceptats per a la producció científica[338]. A la UOC s'utilitzen a nivell institucional les eines en el núvol de Google, i per tant hi ha una política implícita de formats acceptats, que són aquells que accepta Google[339]. A la UPC no s'ha creat cap política, però sí s'ha fet alguna migració puntual[340] i finalment el CSUC encara no s'ha posicionat en aquest aspecte per manca d'"estàndards *de facto* per a la preservació", però sí que han programat *scripts* de migració de formats de vídeo i àudio[341].

Procediment per a la presència de virus

Les respostes en aquest cas es poden dividir en dues categories. En primer lloc, tenim les institucions on no hi ha cap tipus de procediment perquè "l'informàtic del repositori treballa amb Linux i segons ell, no hi ha virus en Linux[338]" o bé perquè confien en què els usuaris que gestionen els ítems digitals ja tenen controlat aquest aspecte[336]; en segon lloc, hi ha institucions en què es fan anàlisis antivirus amb comitès de seguretat de forma regular[339, 341] i per tant ja es poden detectar virus de forma proactiva abans que entrin al repositori.

Procediment tècnic per a dades confidencials

A la UAB existeixen procediments per anonimitzar els números de documents d'identitat[338], mentre que a la UOC es fan servir "processos d'ofuscació[339]". El CSUC no pot aplicar aquests procediments perquè no tenen res per anonimitzar; els continguts dels repositoris no són son del CSUC, sinó que són responsabilitat de les universitats[341]. A la UPC encara estan dissenyant polítiques al respecte[340], mentre que a la UB aquesta qüestió la deixen al criteri de l'autor que diposita els continguts[336], però si és necessari realitzar una anonimització per alguna queixa particular, sí que es realitzen les gestions pertinents per retirar dades confidencials[337].

Ús de BagIt

L'especificació BagIt es va utilitzar una vegada a la UAB[338] per intercanviar una col·lecció de manuscrits digitalitzats amb la Biblioteca de Catalunya, però en aquest cas es va fer a nivell intern i no mitjançant el repositori. A la resta d'institucions on hem realitzat les entrevistes no s'utilitza BagIt.

Taula 43. Resultats de les entrevistes a responsables de repositoris institucionals

	UB	UAB	UOC	UPC	CSUC
Recepció de suports físics?	Sí, però no es troba normalitzat	Discs durs	No	No	No
Gestió i preservació de dades de recerca?	Sí	Sí	No	No	No
Perfil de Dublin Core per a dades de recerca?	No, però s'utilitzen dues etiquetes (coverage i temporal)	No	No	No	No
Límit de mida de fitxer?	Fins a 500 MB	Fins a 2 GB	Fins a 1 GB	No s'indica	No s'indica
Política de migració de formats?	No, però sí hi ha una política de formats acceptats	No, però sí hi ha una política de formats acceptats	No, però sí hi ha una política de formats acceptats	No	No
Procediment per a la presència de virus?	No	No	Sí	No	Sí
Procediment tècnic per a dades confidencials?	No	Sí	Sí	No	No aplicable
Ús de BagIt?	No	Sí, però no amb al repositori	No	No	No

Fonts: D. Iglesias, entrevista, 9 de gener de 2017; C. Azorín, entrevista, 30 de gener de 2017; F. March; R. Padrós; C. Llueca, entrevista, 1 de febrer de 2017; I. Labastida; J. Casals, entrevista, 2 de febrer de 2017; A. J. Prieto; J. Prats; A. Rovira, entrevista, 14 de febrer de 2017; R. de la Vega, entrevista, 9 de març de 2017

5.4.2 Mida dels paquets/fitxers

La Taula 43 mostra que els límits de mida dels fitxers acceptats als repositoris són baixos. Això és degut a què no es tracten de repositoris exclusius de dades, ja que es reben diversos tipus de materials (com articles, treballs de l'alumnat o documentació institucional) que no requereixen en general una gran mida per a cada fitxer individual. Amb l'objectiu d'afirmar amb propietat els límits de mida de fitxer per realitzar la ingesta, vam fer diverses ingestes de fitxers ficticis generats amb el comanament de Windows 'fsutil.exe', el qual permet generar fitxers buits d'una mida en bytes determinada. Després de diverses proves hem comprovat que, efectivament, el límit màxim configurat per defecte és exactament 2.147.483.647 bytes o 2 GB tal i com mostrem a la Figura 58.

La raó d'aquest límit de mida a DSpace és que Cocoon, el marc de desenvolupament de l'entorn web de DSpace, es troba configurat per defecte amb aquest límit en bytes, tot i que es pot ampliar fins a 4 GB. Aquest paràmetre es pot canviar al fitxer 'core.properties' present al codi de DSpace, el qual mostrem a la Figura 59.

Evidentment, aquesta mida màxima de 4 GB no és acceptable per enviar AIPs més grans, així que en aquests casos s'haurà de descartar el sistema de formulari web. Com ja hem indicat al capítol 5.2.6, s'haurà de configurar l'enviament d'aquests paquets per altres vies com mitjançant el protocol FTP. Hi ha opcions alternatives, com en el cas de la University of Exeter, que per desenvolupar el seu repositori Open Research Exeter[342] va comptar amb la col·laboració de l'empresa Globus[343] per superar el problema de la transferència i emmagatzematge de fitxers grans (que poden arribar al TeraByte), ja que el seu repositori DSpace presentava una limitació d'1 GB. Gràcies al servei Globus Online implementat a Open Research Exeter (Evans; Lloyd-Jones; Cole, 2013), el repositori actualment té la capacitat de rebre *datasets* qualificats com a *big data* i la seva eficàcia ha estat demostrada pels investigadors de la University of Exeter[344].

[342] <https://ore.exeter.ac.uk/repository/>. [Consulta: 13/04/2017]

[343] <https://www.globus.org/>. [Consulta: 13/04/2017]

[344] L. Taylor, correu electrònic, 11 d'abril de 2017

Figura 58. Registre d'ingesta d'un fitxer de 2 GB a un repositori DSpace

Mostrar el registro sencillo del ítem

dc.contributor.author	Wilderbeek, Theo	
dc.date.accessioned	2017-04-02T09:30:12Z	
dc.date.available	2017-04-02T09:30:12Z	
dc.date.issued	2017	
dc.identifier.uri	http://localhost:8080/xmlui/handle/demo/12	
dc.description.provenance	Submitted by admin expire (theo.wilderbeek@gmail.com) on 2017-04-02T09:30:12Z No. of bitstreams: 1 test.txt: 2147483647 bytes, checksum: b3dc5e51b0698ddf18d48bbf16c1153f (MD5)	en
dc.description.provenance	Made available in DSpace on 2017-04-02T09:30:12Z (GMT). No. of bitstreams: 1 test.txt: 2147483647 bytes, checksum: b3dc5e51b0698ddf18d48bbf16c1153f (MD5) Previous issue date: 2017	en
dc.title	Arxiu test 1,99 GB (2147483647 bytes)	es_ES
dc.type	Other	es_ES

Ficheros en el ítem

Nombre: test.txt
Tamaño: 1.999Gb
Formato: Fichero de texto

Ver/

Este ítem aparece en la(s) siguiente(s) colección(ones)

- Test2
  Fitxers grans

Mostrar el registro sencillo del ítem

Font: L'autor

Figura 59. Paràmetre de mida màxima de fitxer a DSpace

```
42  # Specify maximum allowed size of the upload. Defaults to 10 Mb.
43  # DSpace Community: 2GB. The HTTP spec allows this to be increased to 4GB but
44  # Cocoon places a limit because the parameter is read into a signed integer.
45  org.apache.cocoon.uploads.maxsize=2147483647
```

Font: Codi de DSpace. Captura de l'autor

5.4.3 Ingesta de paquets BagIt

Com ja hem comentat al capítol 5.2.6, la ingesta de paquets BagIt requereix l'habilitació del complement Replication Task Suite, la qual cosa requereix certa perícia tècnica per programar-lo al repositori DSpace. Vam realitzar proves al DSpace de la nostra maquina local, però per habilitar Replication Task Suite calia tenir un espai

d'emmagatzematge a DuraCloud[345], que és un servei de *hosting* de DuraSpace, i per tant no vam poder fer progressos en aquest aspecte.

Un exemple d'ús de DSpace amb BagIt el trobem a Dryad, que fa servir un mòdul per compartir paquets de dades amb altres repositoris utilitzant BagIt com a protocol de transferència; no obstant, aquest mòdul encara no el fan servir per a la ingesta, sinó per a la difusió[346].

5.4.4 Metadades

Quant a la gestió d'esquemes de metadades, DSpace dóna suport a Dublin Core Qualificat per defecte (Castagné, 2013), mentre que amb DFXML no hi ha suport directe i amb PREMIS només fa un suport parcial, atès que DSpace només utilitza PREMIS quan importa o exporta AIPs en PREMIS que estiguin en format METS i només en el cas de l'Entitat Objecte[347]. Encara que DSpace permet la creació de metadades personalitzades (Castagné, 2013; Kurtz, 2010) hauríem de pensar, per una banda, quins serien els camps de metadades que necessiten els usuaris i per una altra, els camps que millor descriguin el contingut de cada *bag*. Sobre la primera qüestió, hem de recordar que el nostre model fa ús de quatre especificacions (a més d'algunes metadades de captura forense que genera Guymager) amb diversos tipus d'etiquetes, però no totes són útils per a la recerca d'informació. Per exemple, cap usuari farà una cerca d'un valor *hash* MD5 degut a què està produït per un algoritme i no pel llenguatge natural. Quant a la segona qüestió, considerem que és important l'ús de metadades que combinin tant la descripció de cada *bag* com aquelles pertinents a la preservació, altres que informin a nivell tècnic i per últim, necessitarem metadades que facin referència a l'adquisició de les dades dins la institució i al seu empaquetat.

Tot basant-nos en bones pràctiques en gestió de metadades dins repositoris de dades (Dietrich, 2010; Greenberg et al., 2009; Kurtz, 2010; Wira Alam, 2014) i en el model de

[345] <http://www.duracloud.org/>. [Consulta: 13/04/2017]

[346] Dryad wiki (2014). *BagIt Handshaking*. <http://wiki.datadryad.org/BagIt_Handshaking/>. [Consulta: 13/04/2017]

[347] DuraSpace wiki. *PREMIS schema*. <https://wiki.duraspace.org/display/DSDOC5x/DSpace+AIP+Format#DSpaceAIPFormat-PREMISSchema>. [Consulta: 07/09/2016]

metadades Dublin Core Qualificat que empra Dataverse[348], s'ha elaborat una taula prototipada, que mostrem a la Taula 44, de les diferents etiquetes que hauria de tenir mínimament el repositori de dades, si haurien de ser d'ús obligat o no, la seva cardinalitat i l'adequació de vocabularis controlats segons sigui necessari.

Segons Rice (2008) els *datasets* necessiten una descripció adequada per a la seva recuperació, preservació i reutilització. És per aquesta raó que hem triat Dublin Core Qualificat, que com ja hem indicat al capítol 3.3.1, té una bona funcionalitat com a estàndard de metadades descriptives amb l'avantatge que el programari DSpace està configurat per al seu ús. A més, és possible la creació d'extensions (Hodge; Templeton; Allen, 2005; Wira Alam, 2014) com per exemple, per a vocabularis controlats (Shaw et al., 2009).

Dins aquesta proposta no hem inclòs altres tipus de metadades com PREMIS degut a les dificultats tècniques d'integració amb DSpace. No obstant, és possible crear perfils d'aplicació que combinin elements de Dublin Core amb altres esquemes. Dryad fa servir un d'aquests perfils d'aplicació (Michener et al., 2011; Willis; Greenberg; White, 2012), que combina elements de Dublin Core, Darwin Core i Publishing Requirements for Industry Standard Metadata (Krause et al., 2015), però que en les seves primeres versions també incloïa DDI, EML i PREMIS (Carrier; Dube; Greenberg, 2007; White et al., 2008).

Taula 44. Proposta de metadades Dublin Core per al repositori de dades de la nostra proposta

Etiqueta	Metadada	Obligació R=Requerit O=Opcional	Cardinalitat R=Repetible NR=No rep	Comentaris
Títol	dc:title	R	NR	Títol del projecte de recerca
Autor(s) del *dataset*	dc:contributor. author	R	R	Es recomana l'ús de vocabularis controlats
Matèria	dc:subject	R	R	Es recomana l'ús de vocabularis controlats
Descripció	dc:description. abstract	R	NR	Descripció lliure dels continguts del *dataset*

[348] Dataverse (last updated 2017, Mar. 15). *Dublin Core Terms (DC Terms) Qualified Mapping - Dataverse DB Element Crosswalk.* <https://goo.gl/n05k9C>. [Consulta: 03/04/2017]

Etiqueta	Metadada	Obligació R=Requerit O=Opcional	Cardinalitat R=Repetible NR=No rep	Comentaris
Patrocinador / agència de finançament	dc:description. sponsorship	O	R	Informació sobre agències de finançament o persones que patrocinin l'ítem digital
Identificador del *dataset*	dc:identifier	R	NR	Es recomana l'ús de DOIs o *handles*
Data d'adquisició del *dataset*	dc:date.accessioned	R	NR	Data en què DSpace comença la custòdia del material
Data de disponibilitat del *dataset*	dc:date.available	R	NR	Data o interval de dates en què l'ítem es troba disponible al públic
Data de publicació del *dataset*	dc:date.issued	R	NR	Data en què el *dataset* es publica al DSpace
Tipus de fitxer	dc:type	R	NR	DSpace té un vocabulari controlat propi
Fonts de les dades	dcterms:source	O	R	Llista de llibres, articles, i altres fonts que hagin servit per elaborar el *dataset*
Material relacionat	dcterms:relation	O	R	Qualsevol material relacionat, excepte citació d'articles de revista
Cobertura geogràfica	dcterms:coverage.spatial	O	R	Informació sobre la cobertura geogràfica del *dataset*
Cobertura temporal	dcterms:coverage.temporal	O	R	Informació sobre la cobertura temporal del *dataset*
Declaració de drets	dc:rights	R	NR	Llicència de drets per a l'ús de les dades, com CCZero
Recurs que fa referència al *dataset*	dcterms:isReferencedBy	O	R	Publicació (article de revista, llibre o altre tipus) que utilitza aquest *dataset* (inclou citació, identificador permanent i URL permanent)
Mida	dc:format.extent	R	NR	Mida en MB o GB del recurs
Suport original	dc:format.medium	R	NR	Indicació del suport original on es trobava el *dataset* (memòria USB, disc dur, etc.). Es recomana l'ús de vocabularis controlats

Fonts: Wira Alam, 2014; Dataverse, last updated 2017, Mar. 15. *Dublin Core Terms (DC Terms) Qualified Mapping – Dataverse DB element crosswalk*. <https://goo.gl/n05k9C>. [Consulta: 30/04/2017]; Dietrich, 2010; Greenberg et al., 2009; Kurtz, 2010

6. Conclusions i línies futures de recerca

Aquesta tesi s'ha escrit amb l'objectiu d'aportar una solució de preservació a aquelles institucions que rebin o que rebran en el futur dades de recerca de gran mida (una circumstància que s'origina en part per les condicions de les agències de finançament als investigadors) en què hem emprat eines de codi obert perquè volem que qualsevol institució se'n pugui beneficiar sense haver de fer una inversió econòmica forta en quant a l'adquisició del programari. Per fer-ho, hem aplicat els mètodes de l'anàlisi forense digital, utilitzats habitualment per a la investigació criminal, a la preservació de dades de recerca i per donar-hi accés públic hem estudiat el programari de repositoris DSpace. Aquesta solució vol aportar una metodologia nova pel que fa a la preservació de col·leccions de dades de gran mida, amb un estudi i aplicació pràctica de la metodologia forense i una anàlisi teòrica per tal d'implementar la seva ingesta a un repositori DSpace.

Reprenent la hipòtesi que vam formular al capítol 1.4, a continuació resumirem l'estudi que hem realitzat dels paràmetres i aportem les nostres conclusions i possibles línies de recerca per realitzar en el futur.

6.1 Requeriments de les agències de finançament

Al capítol 2.2 vam fer una anàlisi sobre polítiques de diferents tipus d'agències de finançament de la recerca; en alguns casos multidisciplinaris i en altres especialitzats en disciplines concretes. Exposarem a continuació els resultats de la nostra recerca, on es centrarem només en Horizon 2020, els Research Councils AHRC i ESRC i el Wellcome Trust, atès que la resta d'agències estan especialitzades en disciplines diferents de les ciències socials i les humanitats, i aportarem les nostres conclusions sobre el compliment o no del nostre model vers aquests requisits.

Horizon 2020

Com ja vam comentar, a partir de 2017 els projectes que participin a Horizon 2020 són d'accés obert per defecte. Segons estipula l'article 29.3 del *grant agreement* o acord de subvenció (European Commission, 2016d, p. 222), els investigadors estan obligats a, pel que respecta a les dades de recerca:

- Dipositar-les en un repositori de dades de recerca i prendre mesures per tal que terceres persones puguin accedir, extreure, explotar, reproduir i distribuir –lliure de càrrec per a qualsevol usuari– el següent:
    - Totes aquelles dades, incloses metadades associades, necessàries per validar els resultats presentats a publicacions científiques tan aviat com sigui possible;
    - Altres dades, incloses les seves metadades, tal com quedi especificat i dins els terminis de lliurament establerts en els plans de gestió de dades dels projectes de recerca
- Proporcionar informació –mitjançant el repositori– de les eines i els instruments que estiguin a disposició dels beneficiaris i que siguin necessaris per validar els resultats (i si és possible, subministrar les mateixes eines i els instruments)

Un apunt important és que el *grant agreement* no obliga a què les dades de recerca estiguin disponibles en accés obert de manera immediata (European Commission, 2016d, p. 228), ja que especifica que per a aquelles dades necessàries per validar resultats presentats a publicacions científiques, l'accés obert s'ha de fer "el més aviat possible", mentre que per a la resta de dades, els beneficiaris poden especificar períodes d'embargament per a les seves dades dins el pla de gestió de dades.

Podem concloure que el nostre model compleix amb tots els requeriments, atès que s'ha contemplat l'accés al DIP com obert a qualsevol usuari sense cap restricció i el dipòsit de les dades serà a un repositori DSpace dissenyat especialment per a dades de recerca. Quant a les metadades associades a les que fa referència l'article 29.3, entenem que són aquelles generades per l'investigador, i en tot cas també formarien part del DIP. Per altra banda, Horizon 2020 no especifica un temps concret per què les dades de recerca estiguin en accés obert, però sí que tenen prioritat les dades necessàries per validar els resultats científics. Si existeixen períodes d'embargament, l'administrador del repositori DSpace haurà de tenir-ho present durant la ingesta.

Research Councils UK

Recordem que el consorci RCUK té com a base el document *RCUK common principles on data policy* per a la gestió i compartició de dades de recerca. Dins aquests set principis, el segon és el que tracta de la preservació, que estableix el següent: "Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research". El mateix RCUK indica que els Research Councils esperen que la gestió de les dades de recerca es realitzi mitjançant un repositori.

Si ens centrem en els Research Councils especialitzats en ciències socials i humanitats, l'AHRC i l'ESRC, veurem que en el primer cas s'indica que les dades de recerca han d'estar disponibles en un repositori (no s'especifica cap, excepte en el cas de l'arqueologia) en un termini màxim de tres anys. En el segon cas, les dades de recerca estaran disponibles a un repositori en un termini màxim de tres mesos després de la finalització de subvencions. L'ESRC té el seu propi dipòsit de dades, l'UK Data Service, però és possible utilitzar altre repositori si les dades compleixen amb els principis FAIR (aquest requeriment també el trobem a Horizon 2020).

El nostre model compliria també aquests requeriments, ja que contempla el dipòsit de dades en un repositori institucional per a la seva preservació a llarg termini i que les dades siguin de lliure accés i reutilitzables, encara que això dependrà del tipus de llicència que utilitzi l'investigador. Com que el termini màxim que estipula l'ESRC per a l'accés es de tres mesos, els *datasets* subvencionats per aquest Council tindrien prioritat sobre aquells finançats per l'AHRC, que tenen un termini de tres anys.

Wellcome Trust

Entre les diferents disciplines que finança Wellcome Trust tenim les ciències socials i les humanitats. En aquest cas es requereix el dipòsit a un repositori de dades com per exemple l'UK Data Archive o el DANS per tal que es preservin a llarg termini sense especificar un termini màxim per fer el dipòsit. També en aquest cas el nostre model compleix amb aquests requeriments, ja que està dissenyat per a la preservació a llarg termini.

6.2 Tècniques d'anàlisi forense digital

Per tal de demostrar que les tècniques d'anàlisi forense digital són vàlides per a la captura de dades, la seva anàlisi profunda i l'edició i/o bloqueig de dades confidencials, es van estudiar casos d'ús rellevants d'aquestes tècniques a biblioteques i arxius (vegeu el capítol 4.2). Un cop establerta la utilitat a nivell teòric de les eines forenses dins els fluxos de treball de preservació, es van realitzar diverses proves amb l'entorn BitCurator (vegeu el capítol 4.3), dissenyat especialment per integrar les tècniques forenses a fluxos de treball a biblioteques i arxius. Reprendrem a continuació les nostres experiències amb el programari i exposarem les nostres conclusions.

Integritat de les dades

Com ja vam demostrar al capítol 4.3.1, BitCurator permet assegurar que no s'han produït modificacions a les dades gràcies a la incorporació en el programari d'un *write blocker* que bloqueja qualsevol escriptura sobre el contingut del suport que es munta al sistema. Amb les proves que es van realitzar amb la creació d'imatges forenses mitjançant Guymager es van fer diferents càlculs *checksum* (vegeu la Taula 20) que van servir per comprovar que no es va produir cap afectació, i per tant considerem que és suficient demostració. Una prova de la rellevància d'aquests càlculs *checksum* és que són vàlids davant un tribunal de justícia per demostrar que no s'han produït alteracions dins l'ítem digital original.

Anàlisi de les dades

Ha quedat demostrat que BitCurator permet realitzar diferents tipus d'anàlisi de dades (vegeu els capítols 4.3.2 i 4.3.3) que faciliten el control i la visualització del contingut del suport, que podem dividir en:

- Informes de metadades DFXML i PREMIS
- Informes de formats de fitxer
- Llistats de fitxers

Per fer aquesta anàlisi de dades, BitCurator fa servir la interfície gràfica BitCurator Reports, on es realitza en segon pla l'execució dels programaris fiwalk, bulk_extractor i altres *scripts* programats en Python. Farem esment a continuació del contingut de cada anàlisi i la seva utilitat.

Les metadades DFXML documenten les eines que s'han emprat per crear l'informe DFXML, informació sobre la pròpia imatge de disc, informació sobre les particions presents a la imatge, informació individual de cada fitxer present a la imatge de disc i l'extracció de dades privades i sensibles. Per tant, aquest fitxer ens presenta informació contextual, com el sistema o sistemes de fitxers (és a dir, si s'ha utilitzat Windows, Linux o altre sistema operatiu), els formats de fitxer o els valors *hash* de cada fitxer individual presents al suport original (vegeu el capítol 3.3.2). Dins els informes de metadades DFXML que hem utilitzat a les nostres proves, un és el que genera fiwalk, que aporta informació detallada de cada fitxer present al suport original, i un altre és el que genera bulk_extractor, on es documenten els tipus d'escàners que ha utilitzat per analitzar la imatge forense.

En el cas de l'informe de metadades PREMIS, es documenten tres actes de preservació dins el flux de treball de BitCurator: la creació de la imatge forense, l'anàlisi de les dades i l'extracció de dades privades i sensibles. Aquest informe serveix per acreditar que s'han realitzat de forma efectiva les operacions i amb quins programaris.

Quant als informes de formats de fitxers, BitCurator genera dos tipus: un llistat on s'informa dels formats detectats i el nombre d'ocurrències que ha localitzat. A les nostres proves s'ha demostrat que pot localitzar fitxers buits i amb presència de virus, a diferència del programari complementari DROID. No obstant això, les nostres proves amb DROID han demostrat que fa una anàlisi de formats de fitxer més acurada, ja que reconeix l'estàndard MIME i l'identificador PUID per realitzar el seus informes de formats de fitxer, mentre que BitCurator Reports utilitza fiwalk. La nostra conclusió en aquest aspecte, doncs, és que s'han de combinar els dos programaris per obtenir els millors resultats.

Quant al llistat de fitxers, BitCurator Reports aprofita l'anàlisi realitzada per fiwalk per generar un full de càlcul Excel d'onze columnes, on es poden consultar els noms de fitxer, l'extensió, la seva mida, el format, l'hora i data d'accés, creació i modificació i els valors *hash* MD5 i SHA1. Com ja hem comentat al capítol 5.2.4, aquest full de càlcul és útil per a la institució quant permet filtrar per exemple, tipus de formats o extensions, la qual cosa ajuda a crear llistats personalitzats per fer migracions de formats o edicions de dades confidencials.

Identificació i bloqueig d'informació privada i sensible

Tal com hem vist al capítol 4.3.2, l'eina bulk_extractor fa una anàlisi byte a byte del contingut de la imatge de disc i genera informes separats per categories en funció de com la seva configuració, ja que hi ha més de trenta possibilitats d'escanejat que permeten extreure números de telèfon, targetes de crèdit, adreces de correu electrònic, codi de Facebook, metadades Exif, etc. Després de les diferents proves podem concloure que és una eina potent d'extracció i més que suficient per al nostre model.

Com ja vam exposar al capítol 4.3.3, bitcurator_access_redaction permet redactar contingut, amb el bloqueig selectiu de fitxers (complet o parcial) mitjançant expressions regulars o bé indicant cadenes de text concretes. Això permet a un repositori oferir contingut lliure de dades privades i sensibles, ja sigui per requeriment legal o per una sol·licitud dels subjectes investigats per protegir la seva privacitat (vegeu el capítol 2.4.3) i exercir el seu dret a l'oblit (vegeu el capítol 2.4.4). Val a dir que bitcurator_access_redaction encara no està integrat dins l'entorn BitCurator, en el sentit que és un programari que cal descarregar i instal·lar per separat, a més que obliga a utilitzar línies de comanament, ja que no compta encara amb una interfície gràfica. Per altra banda, el seu ús no és senzill, ja que requereix editar amb cura les seves opcions de configuració per evitar l'edició i/o bloqueig no desitjat de fitxers. Per tant, aquesta funció presenta moltes possibilitats de desenvolupament per investigar.

6.3 Repositori DSpace

Tal com vam exposar al capítol 5.4, la implementació del nostre model a un repositori programat sota DSpace planteja diverses qüestions tècniques. A falta de poder validar totalment el model de forma pràctica, reprenem aquestes qüestions i indiquem fins a quin nivell les hem pogut resoldre i quines queden obertes per a la seva resolució en el futur.

Ingesta de gran volums de dades

La configuració per defecte de DSpace no permet una ingesta de grans volums de dades, però és perfectament possible habilitar aquesta funció amb una programació específica, així que aquest aspecte no representa un obstacle tècnic insalvable per a la implementació del model. Una prova d'això és que diverses institucions amb repositori DSpace utilitzen aquesta funció (vegeu el capítol 5.2.6).

Paquets BagIt

Com ja es va indicar al capítol 3.3.4, s'ha demostrat que l'especificació BagIt és una bona solució per crear, transferir i emmagatzemar grans col·leccions de dades. L'objectiu d'utilitzar aquesta especificació és la creació dels paquets AIP per a l'emmagatzematge dels datasets com a còpia d'arxiu de les imatges forenses i altres fitxers com informes i llistats. Per tal de poder implementar aquesta funció al repositori DSpace, caldria programar i posar a punt el mòdul Replication Task Suite (vegeu el capítol 5.2.6) o bé programar un de nou per realitzar la ingesta.

Metadades

Dins els diferents esquemes de metadades presents a la nostra proposta i que hem analitzat (vegeu el capítol 3.3), hem contemplat Dublin Core com l'esquema a utilitzar per descriure els *datasets* al repositori, mentre que la resta d'esquemes es fan servir per generar documents que acreditin les operacions tècniques i administratives que s'han realitzat sobre els mateixos.

Al llarg del capítol 3.3.1 hem deixat palès que Dublin Core és un esquema vàlid pel que respecta a la descripció de *datasets*, ja que DSpace està configurat per ser utilitzat amb Dublin Core i es pot adaptar a diferents necessitats, ja que permet l'ús de qualificadors i la creació de Perfils d'Aplicació per a casos específics d'ús, com en el cas del repositori Dryad.

En el cas de l'esquema DFXML, al capítol 3.3.2 hem exposat les seves capacitats per documentar de forma eficaç els processos tècnics forenses d'anàlisi de contingut (vegeu la Taula 36 per a més detalls), i d'extracció de dades confidencials (vegeu la Taula 35 per a més detalls). No ho fa, però, amb les operacions de bloqueig i/o redacció de dades confidencials sinó que és un procés accessori perquè el programari bitcurator_access_redaction no forma part integral de BitCurator. Això no vol dir que aquest procés no quedi documentat (vegeu el capítol 5.2.4), però seria desitjable poder tenir documentades totes les operacions forenses que s'han executat en un registre de metadades DFXML. S'ha de recordar que aquest esquema encara es troba en desenvolupament, així que no s'ha de descartar que els processos d'exportació de metadades DFXML millorin en el futur.

En el cas de la nostra proposta, el diccionari de metadades PREMIS (vegeu el capítol 3.3.3) ha demostrat ser eficaç per documentar els actes de preservació que executa BitCurator: la captura forense que realitza Guymager, l'anàlisi del sistema de fitxers que realitza fiwalk i l'extracció de dades confidencials que realitza bulk_extractor (vegeu la Figura 25). A més, també recupera l'Entitat Objecte (la imatge forense) que és la unitat subjecte a la preservació digital (vegeu la Figura 10). Tot i que BitCurator no utilitza PREMIS per gestionar els drets d'autor, no és un inconvenient per a la nostra proposta, ja que aquest aspecte el tractaria Dublin Core (vegeu la Taula 44).

Finalment, l'esquema BagIt (vegeu el capítol 3.3.4) permet gestionar de forma senzilla les metadades administratives d'un AIP, el qual pot arribar a la institució en forma d'un o més suports. La seva utilització és extensa en aquells repositoris que demanen una càrrega de fitxers de gran mida, així que la seva utilització quedaria validada per a la nostra proposta.

6.4 Línies futures de recerca

La nostra proposta es concentra en un flux de treball detallat (vegeu la Figura 56) on hem demostrat l'aplicació de les tècniques forenses per a la preservació de dades de recerca mitjançant el programari de codi obert BitCurator. Hem explicat amb detall els requeriments necessaris per engegar aquest flux de treball i facilitar posteriorment l'accés públic als *datasets*. Però com és lògic, durant el període de recerca s'han generat algunes qüestions que es poden explorar en el futur.

Per manca de recursos no ha estat possible fer proves directament amb maquinari especialitzat en anàlisi forense digital com l'estació FRED ni tampoc amb programari comercial com FTK. L'exploració directa d'aquestes eines pot ser interessant per descobrir noves funcionalitats que simplifiquin els processos del nostre flux de treball.

Per últim, una qüestió que es pot analitzar amb més profunditat és el de l'accés, que hem plantejat com obert però indirecte. Aquesta solució no ha de ser tancada necessàriament, sinó que es poden investigar opcions que permetin un accés més directe i senzill per a l'usuari final.

Annex A. Cas pràctic d'aplicació del flux de treball

A.1 Introducció

Un cop establerts els procediments del model teòric de preservació al capítol 5, dins aquest annex es presenta un cas pràctic fictici on es detalla pas per par cadascuna de les operacions. Amb aquest propòsit, s'ha utilitzat una memòria USB Lexar de 16 GB, però amb una preparació específica tenint en compte els procediments de treball.

La preparació ha consistit en seleccionar una mostra de formats de fitxers, contemplats específicament al nostre model, per demostrar la seva viabilitat en els casos que es poden presentar quan es preserven dades de recerca en ciències socials i humanitats. En la mesura del possible, aquests fitxers han estat creats pel propi autor durant el procés d'elaboració de la tesi, i en la resta de casos s'han fet servir fitxers de mostra de lliure descàrrega i ús a la xarxa. Mostrem la relació de fitxers i de formats de fitxers que hem utilitzat a la Taula 45.

Per altra banda, també s'ha elaborat un perfil fictici d'investigador de ciències socials que diposita les seves dades de recerca a un repositori de dades d'una institució també fictícia. Per aquesta raó hem generat dades personals de forma aleatòria, que consisteixen en un ID d'investigador i un correu electrònic relacionat.

Aquestes dades personals fictícies, juntament amb la mostra de formats de fitxer, s'han fet servir per al formulari de lliurament de dades que mostrarem al capítol A.2.

Taula 45. Relació de fitxers i de formats de fitxer que s'han utilitzat al cas pràctic d'aplicació del flux de treball

Nom de fitxer	Format de fitxer	Font
Text		
Projecte de recerca	Word (.docx)	L'autor
Projecte de recerca	Text pla (.txt)	L'autor
Projecte de recerca	PDF/A (.pdf)	L'autor
Projecte de recerca	OpenDocument Text (.odt)	L'autor
sol_preinscripcio	Word 97/2003 (.doc)	L'autor
2016-2017 - Full Inscripció Projecte	PDF (.pdf)	L'autor
Text amb llenguatge de marques		
sample	Standard Generalized Markup Language (.sgm)	*The basic SGML document*. <https://goo.gl/INJ3q9>. [Consulta: 31/12/2016]
PANODEMO	Standard Generalized Markup Language (.sgm)	*The basic SGML document*. <https://goo.gl/INJ3q9>. [Consulta: 31/12/2016]
acronims	Extensible Markup Language (.xml)	L'autor
Text tabulat		
Glossari	Excel (.xlsx)	L'autor
acronims	CSV (.csv)	L'autor
acronims	Excel 97/2003 (.xls)	L'autor
acronims	OpenDocument Spreadsheet (.ods)	L'autor
Vídeo		
hale_bopp_1	MPEG2 (.mpg)	*HubbleSOURCE: MPEG Video Clips*. <http://hubblesource.stsci.edu/sources/video/clips>. [Consulta: 31/12/2016]
centaur_2	MPEG2 (.mpg)	*HubbleSOURCE: MPEG Video Clips*. <http://hubblesource.stsci.edu/sources/video/clips>. [Consulta: 31/12/2016]
VID_20161108_075347	MPEG4 (.mp4)	L'autor
Àudio		
a2002011001-e02	Waveform Audio File (.wav)	*Sound examples.* <http://www.music.helsinki.fi/tmt/opetus/uusmedia/esim/index-e.html>. [Consulta: 31/12/2016]
a2002011001-e02-128k	MP3 (.mp3)	*Sound examples.* <http://www.music.helsinki.fi/tmt/opetus/uusmedia/esim/index-e.html>. [Consulta: 31/12/2016]

Nom de fitxer	Format de fitxer	Font
a2002011001-e02-128k	OGG (.ogg)	*Sound examples.* <http://www.music.helsinki.fi/tmt/opetus/uusmedia/esim/index-e.html>. [Consulta: 31/12/2016]
example	Free Lossless Audio Codec (.flac)	*Sample Flac Sample Audio File (.flac) - free sample download*. <http://samplephotovideo.com/2015/12/sample-flac-audio-file>. [Consulta: 31/12/2016]
Imatge rasteritzada		
esquema de mostra	Photoshop (.psd)	L'autor
workflow	Portable Network Graphics (.png)	L'autor
workflow	Bitmap (.bmp)	L'autor
workflow	Tagged Image File Format (.tif)	L'autor
figura de mostra	JPEG (.jpg)	L'autor
Gràfic vectorial		
workflow	Scalable Vector Graphics (.svg)	L'autor
visualization_-_aerial	AutoCAD (.dwg)	*AutoCAD sample files*. <https://goo.gl/a02uNE>. [Consulta: 31/12/2016]
Gràfic 3D		
GeometryPrimitiveNodes	X3D (.x3d)	*X3D for web authors examples archive*. <https://goo.gl/KNeZ2S>. [Consulta: 31/12/2016]
Dades GIS		
TC_NG_Baghdad_IQ_Geo	GeoTIFF (.tfw, .tif, .txt)	*Earthstar geographics*. <http://www.terracolor.net/sample_imagery.html>. [Consulta: 31/12/2016]
Rivers_in_Southeast_Asia	Esri Shapefile (.dbf, .prj, .shp, .shx)	*Google Earth Solidario.* <https://www.google.es/earth/outreach/tutorials/importgis.html>. [Consulta: 31/12/2016]
Fitxers comprimits		
acronims	RAR (.rar)	L'autor
rivers_in_seasia_shapefile	ZIP (.zip)	*Google Earth Solidario.* <https://www.google.es/earth/outreach/tutorials/importgis.html>. [Consulta: 31/12/2016]
Altres		
workflow	Umlet (.uxf)	L'autor

Fonts: Indicats a la taula

A.2 Preparatius inicials

Els procediments s'han realitzat segons es van indicar al capítol 5.2.1, on hem treballat amb l'entorn BitCurator i una carpeta de base a la ruta '/home/bcadmin/Scripts', la qual conté tots els *scripts* necessaris per executar el flux de treball, a més d'altres fitxers de suport necessaris. A més, hem instal·lat el programari DROID a la ruta '/home/bcadmin/utils'.

Recepció del formulari del dipositant i del suport

En aquest cas fictici, el suport que hem utilitzat en aquest cas ha estat una memòria USB de 16 GB. Les dades del formulari han estat les següents:

- **Dades personals del dipositant**
  - Identificador: 37808285
  - Correu electrònic: pjimenez@uc.cat
- **Dades del pla de recerca**
  - Nom del pla de recerca: Social platform on cultural heritage in Catalonia
  - Codi identificador: 654612
  - Disciplina/àrea de coneixement: Humanitats
  - Descripció: Anàlisi del patrimoni cultural a Catalunya
  - Patrocinadors: Comissió Europea
  - Política de compartició de dades: Horizon 2020
  - Política de gestió de dades: Horizon 2020
  - Investigador principal: Enric Galván Álvarez
  - Data d'inici del projecte: 01/04/2015
  - Data de finalització del projecte: 30/03/2017
- **Organització tècnica de les dades**
  - Tipus de suport: Memòria USB
  - Sistemes operatius que s'han utilitzat: Windows
  - Tipologies de dades presents:
    - Text

        - Imatges
        - Fulls de càlcul/dades tabulars
        - Text amb llenguatge de marques
        - Gràfics vectorials
        - Imatges 3D
        - Àudio
        - Vídeo
        - Dades GIS
        - Altres: Diagrames UML, fitxers comprimits
    - Formats de fitxer utilitzats
        - Text: .pdf, .doc, .docx, .txt, .odt
        - Text amb llenguatge de marques: .sgm, .xml
        - Àudio: .wav, .mp3, .flac, .ogg
        - Vídeo: .mpg, .mp4
        - Imatges 3D: .x3d
        - Fulls de càlcul/dades tabulars: .xls, .xlsx, .ods, .csv
        - Dades GIS: ESRI (.dbf, .prj, .shp, .shx), GEOTIFF (.tfw, .tif, .txt)
        - Altres: .uxf, .rar, .zip
- **Heu utilitzat un o més esquemes de metadades?** No
- **Heu estructurat i/o normalitzat les vostres carpetes i fitxers?** No
- **Mida total de les dades:** 125 MB
- **Nombre total de fitxers presents al suport:** 38
- **Informació de drets d'autor**
    - *Les vostres dades de recerca tenen alguna llicència d'ús com Creative Commons o Open Data Commons?* Sí, Creative Commons Zero
- **Protecció de dades**
    - *Les vostres dades de recerca contenen informació personal i/o sensible que calgui que sigui bloquejada a l'accés públic?* Sí, els fitxers 'sol_preinscripcio.doc' i ' 2016-2017 - Full Inscripció Projecte.pdf'

Recordem que aquest és un cas fictici i per tant, encara que en aquest formulari s'hagi indicat que les dades de recerca es troben sota el programa Horizon 2020, no s'ha fet cap preparació específica quant a la seva organització ni estructura i per aquest motiu no

hem indicat cap estàndard de metadades ni tampoc cap estructura concreta d'organització d'arxius.

Creació de l'estructura de carpetes

Tal i com s'ha indicat al capítol 5.2.1, hem executat el primer *script* del nostre flux de treball, '0-Creació d'estructura de carpetes.sh', que ha generat totes les carpetes necessàries per treballar, amb tres carpetes principals: 'Arxiu intern', 'Ingesta AIP' i 'Workspace'. Com que l'ID de l'investigador en aquest cas pràctic és el 37808285, hem editat el *script* i hem reemplaçat 'ID investigador' per aquesta cadena numèrica.

Escanejar i arxivar el formulari

Hem escanejat el formulari en format PDF amb reconeixement de caràcters OCR i l'hem guardat a 'Workspace' amb el nom de fitxer 'Form.pdf'. En aquest cas pràctic, hem guardat el formulari en paper en una carpeta.

Assignar identificadors als suports

Com que només hem utilitzat un suport, que és una memòria USB, l'identificador del suport ha estat 37808285USB01, amb el qual hem fet un teixell i l'hem pegat al suport.

Fotografiar suport

Al no disposar d'una càmera rèflex, hem utilitzat una càmera compacte Nikon Coolpix S3300 amb una resolució de 4608 x 3456 píxels. S'han fet dues fotografies, una de la part anterior i una altre de la part posterior del suport, s'han copiat de la memòria interna de la càmera a la carpeta 'Workspace' i s'han canviat de nom a 'ID suport_01.jpg' i a 'ID suport_02.jpg'. Com que aquesta càmera només utilitza el format JPEG per a fotografies, no ha calgut fer cap canvi a la configuració de la càmera.

Figura 60. Fotografies de les parts anterior i posterior de la memòria USB utilitzada al cas pràctic

Font: L'autor

Examen físic del suport

Hem creat un full de càlcul dins 'Workspace' amb el nom 'Medialog.ods' on hem editat les següents cinc columnes:

- Tipus de suport. Hem indicat 'USB'
- Fabricant. Hem indicat 'Lexar' perquè és visible de forma externa
- ID suport. Hem indicat '37808285USB01'
- Observacions. No hem indicat res, ja que el suport no té etiquetes i les inscripcions originals són visibles
- Topogràfic. En aquest cas fictici, hem indicat 'USB01' per indicar una prestatgeria on es guarden memòries USB

A.3 Captura forense

Configurar maquinari i programari

Hem comprovat que BitCurator està configurat adequadament en mode només lectura per evitar qualsevol modificació en el suport. Un cop fet això, hem connectat directament la memòria USB a una de les ranures lliures de l'estació de treball.

Connectar *write blocker*

Aquest és un pas opcional. Com que no disposàvem de cap *write blocker*, hem passat al següent procés.

Crear i verificar imatge forense (Guymager)

Hem executat el *script* '1-Guymager.sh' per obrir la interfície gràfica de Guymager. El format d'imatge que hem utilitzat ha estat l'EnCase, el configurat per defecte, i hem utilitzat com a nom de fitxer 'imatge' a la ruta '/home/bcadmin/Workspace'. Hi ha diverses opcions de verificació d'imatge amb el càlcul de valors *hash*, però en aquest pràctic hem marcat totes: MD5, SHA1 i SHA256. Hem canviat l'opció de dividir la imatge en TB, per tal de generar la imatge forense en un sol fitxer. Per crear la imatge forense de 16 GB en el nostre sistema, s'ha necessitat un temps de 15 minuts i 37 segons.

Per altra banda, Guymager també ha generat de forma automàtica el fitxer de registre de captura amb el mateix nom de fitxer, 'imatge.info'. Hem consultat aquest fitxer de registrar per poder verificar que la imatge s'ha generat correctament.

Errors en el procés?

El registre de captura ens va donar "Source verification OK" i "Image verification OK". Com que no s'han produït errors, hem passat al procés següent de verificació de dades confidencials.

Presència de dades confidencials?

Hem revisat el formulari del dipositant per comprovar la presència o no de dades confidencials. En aquest cas, sí que s'han indicat dos fitxers, 'sol_preinscripcio.doc' i '2016-2017 - Full Inscripció Projecte.pdf'. Per tant, hem passat al següent pas corresponent al flux de treball, que és la creació i verificació de la imatge *raw*.

Crear i verificar imatge *raw* (Guymager)

Hem repetit els passos que vam realitzar per crear la imatge forense, amb la diferència d'haver de seleccionar l'opció d'imatge raw a Guymager i utilitzar el nom de fitxer 'imatge_raw', comprovant sempre que no es produeixi la divisió de la imatge en múltiples fitxers, tal i com hem creat la imatge forense. El directori per guardar aquesta imatge *raw* i el seu registre ha estat igualment 'Workspace'. Com que no s'han produït errors, hem passat al pas següent. Per crear la imatge *raw* s'ha requerit quasi el mateix temps que amb la imatge forense, que en aquest cas ha estat de 15 minuts i 33 segons.

Retirar suport

Hem desconnectat la memòria USB de l'estació de treball. Recomanem fer-ho de forma segura dins l'entorn BitCurator, amb l'opció 'Eject' marcant la icona del suport a la dreta de l'escriptori.

Guardar suport provisionalment

Hem desat la memòria USB en un lloc provisional. En el nostre cas, l'hem deixat a prop del lloc de treball.

A.4 Examen i anàlisi de contingut

Muntar imatge forense

Per muntar la imatge forense 'imatge.E01', hem fet ús del script '2-Muntar imatge forense.sh'.

Escanejar presència de virus (ClamAV)

Un cop la imatge es va muntar, hem executat el *script* '3-ClamAV.sh' per generar el registre d'escanejat 'clamav.log' a la ruta '/home/bcadmin/Workspace'.

Presència de virus?

Per comprovar la presència de virus a la imatge forense, hem obert el fitxer 'clamav.log' amb el programari gedit. Tal com hem vist a la Figura 37, si ClamAV detecta un virus ens ho mostrarà amb l'indicador FOUND al costat de la ruta del fitxer indicat. També es pot comprovar al sumari final, on es llisten els fitxers infectats. En aquest cas no s'ha localitzat cap virus, així que hem passat al següent pas corresponent.

Extreure dades confidencials (bulk_extractor)

En aquest cas hem executat el *script* '4-bulk_extractor.sh', el qual executa bulk_extractor que analitza la imatge forense i extreu dades confidencials en diversos fitxers de text pla, en funció de com s'ha configurat el *script*. Recordem que es generen a la ruta '/home/bcadmin/Workspace/bulk_extractor'. Hem cronometrat el temps que ha necessitat bulk_extractor per analitzar la imatge forense de 16 GB, que ha estat de 6 minuts i 1 segon.

Executar BitCurator Reports

Tal i com vam mostrar a la Figura 40, hem seleccionat la imatge forense 'imatge.E01', el directori '/bulk_extractor' i el directori 'bitcurator_reports' per poder generar els diferents informes i llistats de BitCurator Reports dins la casella 'Run All'. Un cop fet això, hem pitjat 'Run'. Els informes i llistats s'han generat al directori 'bitcurator_reports' dins 'Workspace'.

Generar informes de formats de fitxer (DROID)

Per executar DROID, hem utilitzat el *script* '6-DROID.sh' que en aquest cas busca el programari DROID a la ruta '/home/bcadmin/utils/DROID', crea un perfil en funció dels fitxers localitzats a la imatge forense muntada '/media' i exporta dos informes a '/home/Workspace', un en format XML (amb el nom de 'droid.xml') i un altre en format PDF (amb el nom de 'droid.pdf').

Afegir marca d'aigua als informes

El següent pas ha estat afegir una marca d'aigua amb l'identificador del suport (en aquest cas 37808285USB01) als informes 'droid.pdf', 'bc_format_bargraph.pdf', 'bulk_extractor_report.pdf', 'fiwalk_report.pdf' i 'format_table.pdf'. Hem executat el script '8-Watermark.sh' que executa Ghostscript amb el suport del fitxer PostScript 'stamp.ps' que teníem de base a la nostra ruta '/home/bcadmin/Scripts'.

A.5 Processat de continguts

Presència de dades confidencials?

Recordem que sí teníem presència de dades confidencials en dos fitxers ('sol_preinscripcio.doc' i ' 2016-2017 - Full Inscripció Projecte.pdf') segons s'indicà al formulari de l'investigador, així que hem passat al pas corresponent.

Identificar fitxers a redactar i bloquejar

Un cop hem identificat els fitxers al formulari, els hem contrastat amb els informes que ens ha donat bulk_extractor i BitCurator Reports. Per una banda, hem tret els valors *hash* dels fitxers del fitxer 'fiwalk-output.xml' de BitCurator Reports i per altra, hem consultat les extraccions de dades confidencials que, en aquest cas, han estat els fitxers amb números de telèfon (telephone.xls) i correus electrònics (email.xls). En aquest cas, hem comprovat que efectivament s'ha reconegut i extret el correu electrònic pjimenez@uc.cat, però no s'han reconegut la resta de dades confidencials presents, com el NIF, el telèfon o l'identificador de l'investigador. Per tant, hem consultat la imatge forense muntada i hem obert els fitxers amb l'editor hexadecimal per tal de comprovar que les cadenes de text es puguin trobar dins el codi màquina de cada fitxer. En el cas de '2016-2017 - Full Inscripció Projecte.pdf', amb dades com NIF, adreça o telèfon, no va ser possible trobar cap cadena de text coincident amb els valors de dades confidencials. Per tant, aquest fitxer s'ha hagut de restringir completament. En canvi, amb 'sol_preinscripcio.doc' sí que es van poder localitzar algunes cadenes de text i per tant es va poder redactar.

Per tant, ja teníem suficient informació per obrir un nou full de càlcul LibreOffice Calc, que hem guardat a 'Workspace' amb el nom 'Redaction.ods' on hem creat les sis columnes corresponents tal i com mostrem a la Taula 46:

Taula 46. Llista de fitxers a redactar i bloquejar al cas pràctic

Nom del fitxer original	Valoració	Acció	MD5 hash	SHA1 hash	Comentaris
2016-2017 – Full Matrícula UB.pdf	Restringit	SCRUB	4a0a0ccde6b33ded84d9ea9a6a44f2ec	99ff4ad948b9d090812ab8813c3c493a0f106f3c	No permet redacció, s'ha de bloquejar
sol_preinscripcio.doc	Per redactar	FILL	ba8bbffa1cb9e090e3d266edce444753	ddff4fdad4b8d08fcd79027421cc5856e8fdbdc4	S'han de redactar NIF, ID investigador, adreça, codi postal, telèfon, correu electrònic

Font: L'autor

Desmuntar imatge forense

Ja no necessitàvem que la imatge forense es trobi muntada, per tant l'hem desmuntat amb el *script* '7-Desmuntar imatge forense.sh'.

Configurar i executar bitcurator_access_redaction

El següent pas ha estat configurar el fitxer 'imatge_raw_config.txt' present a la ruta '/home/bcadmin/Scripts'. Segons el que hem indicat al full de càlcul, hem bloquejat completament el fitxer '2016-2017 - Full Inscripció Projecte.pdf' i hem fet redaccions concretes a 'sol_preinscripcio.doc' quant al NIF, l'identificador d'investigador, l'adreça, el codi postal i el correu electrònic. Mostrem aquest fitxer de configuració a la Figura 61.

Figura 61. Fitxer de configuració de redacció de continguts al cas pràctic

```
INPUT_FILE /home/bcadmin/Workspace/imatge_raw.dd
OUTPUT_FILE /home/bcadmin/Workspace/imatge_raw_redacted.dd
REPORT_FILE /home/bcadmin/Workspace/imatge_raw_redacted.json

# Busca el fitxer amb el valor hash indicat (2016-2017 - Full Inscripció
Projecte.pdf) i sobreescriu el contingut amb zeros
FILE_MD5 4a0a0ccde6b33ded84d9ea9a6a44f2ec SCRUB

# Omple seqüències que continguin les cadenes de text indicades amb el
codi ASCII 0x2a (caràcter "*")
SEQ_EQUAL pjimenezso@uc.cat FILL 0x2a
SEQ_EQUAL 654783665 FILL 0x2a
SEQ_EQUAL 37808285 FILL 0x2a
SEQ_EQUAL 70680790L FILL 0x2a
SEQ_EQUAL Bruc, FILL 0x2a
SEQ_EQUAL 08037 FILL 0x2a

# Executa la redacció (genera una imatge de disc raw redactada)
COMMIT
```

Font: L'autor

La cadena de text corresponent a l'adreça, que originalment era "Bruc, 250", no es va poder editar degut a què bitcurator_access_redaction no funciona amb cadenes de text que continguin espais en blanc. Per tant, hem editat només el nom del carrer, ja que considerem que és suficient per ocultar l'adreça de l'investigador. Si afegim una seqüència que redacti "250", això crearia una edició de tots els fitxers dins la imatge que continguin aquesta seqüència i per tant, hem decidit no fer-la. El pas següent ha estat executar el *script* '9-Redacció de fitxers.sh' que ha creat el fitxer 'imatge_raw_redacted.dd'.

Muntar imatge *raw* redactada

Hem muntat la imatge *raw* redactada amb el nostre *script* '10-Muntar imatge redactada.sh'.

Comprovar redaccions

Hem consultat el fitxer JSON que es va generar amb el nom 'imatge_raw_redacted.json', amb especial atenció a l'etiqueta "filename", que és la que indica el fitxer que ha estat redactat.

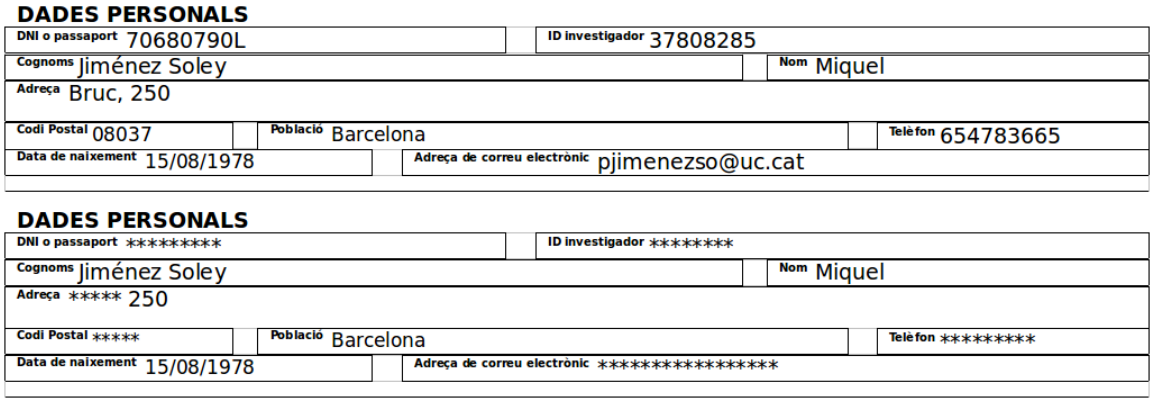
Hem comprovat que efectivament els fitxers que hem seleccionat per ser redactats s'han modificat correctament navegant en els continguts de la imatge muntada. En el cas de

'sol_preinscripcio.doc', els caràcters que hem indicat s'haurien d'haver omplert de asteriscs, i en el cas de '2016-2017 - Full Inscripció Projecte.pdf', tot el contingut hexadecimal del fitxers s'hauria d'haver sobreescrit a zeros. Per comprovar aquest últim cas, hem obert aquest fitxer amb un editor hexadecimal.

Errors a la redacció?

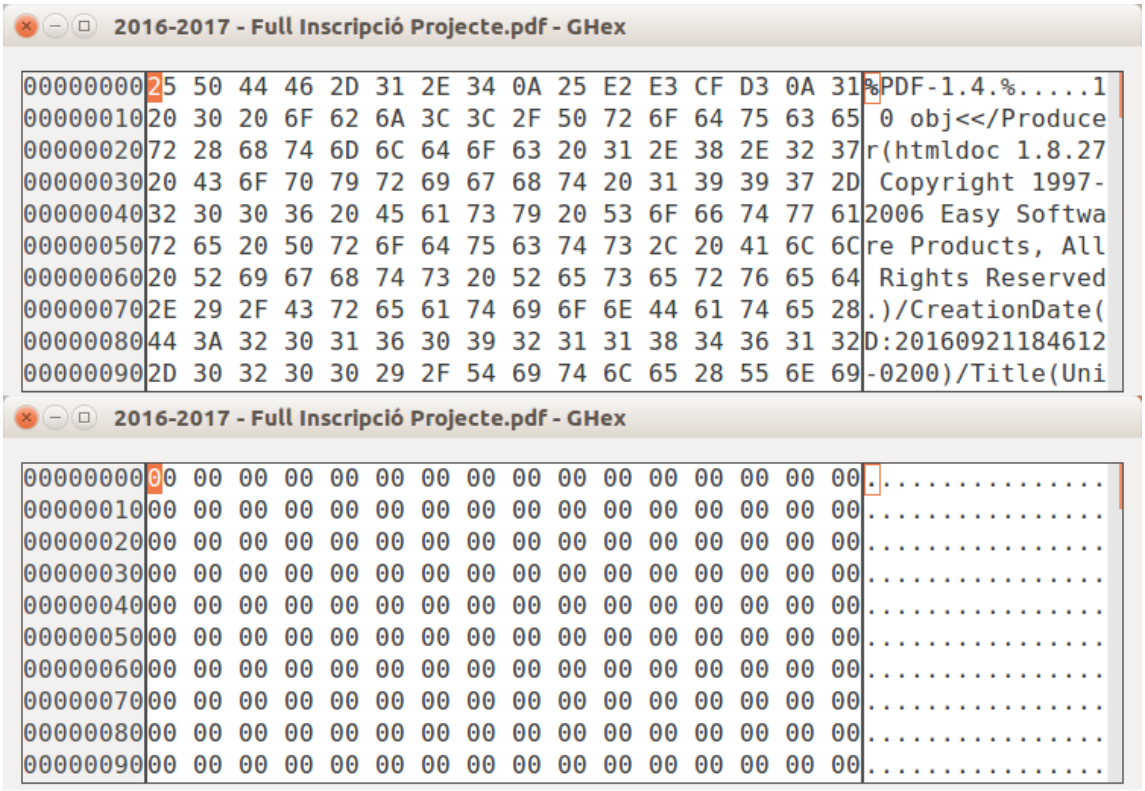
Al fitxer JSON no hem trobat que s'hagin redactat fitxers diferents als seleccionats al nostre full de càlcul 'Redaction.ods'. Un cop comprovat els fitxers de la imatge muntada, podem dir que la redacció ha estat un èxit, tal i com mostrem a la Figura 62 i a la Figura 63. En el primer cas, s'han canviat les cadenes de text indicades amb el codi ASCII 0x2A (l'asterisc) i en el segon cas, s'ha substituït tot el contingut hexadecimal del fitxer per zeros, la qual cosa fa que el fitxer estigui totalment buit i per tant no es pot consultar.

Figura 62. Text de dades personals abans i després de la seva redacció al cas pràctic



Font: L'autor

Figura 63. Contingut hexadecimal abans i després del seu bloqueig al cas pràctic



Font: L'autor

Identificar fitxers amb formats no preferits

En aquest flux de treball fem només migració de formats d'imatge rasteritzada, que en el nostre cas hem considerat el TIF com l'adient per a la preservació. Tal i com hem vist al capítol 5.2.4, tenim un script que migra els formats PNG i Photoshop a TIF i per tant són aquests dos formats els que hem de localitzar. Per fer-ho, hem consultat els informes de format de fitxer generats per DROID i per BitCurator Reports, que en el primer cas ens ha localitzat un fitxer en PNG i un altre en Photoshop. En el cas de l'informe de BitCurator Reports en Excel, tenim la possibilitat de filtrar columnes i la utilitzarem amb la columna 'File Format', on hem marcat que ens filtri aquelles columnes que continguin el valor 'Adobe Photoshop Image' i 'PNG image data'. Els resultats han coincidit amb DROID, ja que teníem un fitxer de cada format, però en aquest cas tenim els noms de fitxer, que en el primer cas ha estat 'esquema de mostra.psd' i en el segon cas, 'workflow.png'.

A continuació hem creat un nou full de càlcul amb LibreOffice Calc que ha rebut el nom de 'Migration.ods' i l'hem editat tal i com mostrem a la Taula 47.

Taula 47. Llista de fitxers a migrar al cas pràctic

Nom del fitxer original	MIME type	Programari	Nom del fitxer migrat	MIME type
esquema de mostra.psd	image/vnd.adobe.photoshop	ImageMagick	esquema de mostra_migrated.psd	image/tiff
workflow.png	image/png	ImageMagick	workflow_migrated.tif	image/tiff

Font: L'autor

Recordem que hem tret la informació de la primera columna del fitxer Excel de BitCurator Reports, la segona i la cinquena columna fan referència al format de fitxer segons la identificació MIME, la tercera columna al programari que s'utilitza per migrar, i la quarta columna fa referència al nom que rebrà el fitxer migrat.

Executar Disk Image Access Interface

Ara que ja hem identificat els fitxers que volem migrar, hem executat el *script* '12-Disk Image Access Interface.sh' que ens ha obert la interfície gràfica d'aquest programari.

Extreure fitxers a normalitzar

Un cop obert Disk Image Access Interface, hem seleccionat la imatge forense 'imatge.E01' i s'ha mostrat la seva estructura de fitxers i carpetes tal i com mostra la Figura 49. Hem seleccionat els fitxers que hem llistat anteriorment i els hem exportat a 'Workspace' dins el subdirectori 'Pre-migration'.

Normalitzar formats

El següent pas a fer ha estat executar el *script* '13-Migration.sh' que ha creat els nous fitxers migrats al subdirectori 'Post-migration' dins 'Workspace'. En aquest cas, teníem 'esquema de mostra_migrated.tif' i 'workflow.tif'.

Guardar fitxers generats

Per tal de guardar els fitxers que hem generat fins ara a les rutes corresponents dins '/home/bcadmin/Arxiu intern' i '/home/bcadmin/Ingesta AIP' hem editat el *script* '14-

Reanomenar i moure fitxers.sh' i hem canviat les cadenes de text 'ID investigador' pel número identificador de l'investigador que hem indicat en el nostre cas pràctic (37808285) i 'ID suport' per l'identificador que hem assignat al suport en aquest cas (37808285USB01). Un cop fet això, els fitxers s'han canviat de nom a les formes normalitzades que hem indicat al Taula 40 dins 'Arxiu intern' i s'ha generat un fitxer ZIP dins 'Ingesta AIP' amb tots els fitxers seleccionats per a ser preservats al repositori tal com hem indicat a la Taula 41. En aquest cas hem cronometrat el temps que s'ha trigat en comprimir el fitxer ZIP, que entre altres fitxers contenia dues imatges de disc (la imatge forense i la imatge *raw* redactada), cadascuna de 16 GB. El temps total que el nostre sistema va requerir per comprimir el fitxer ZIP fou de 16 minuts i 1 segon.

Guardar suport

Un cop hem acabat amb el flux de treball, el pas següent seria guardar el suport a l'espai físic definitiu dins al full de càlcul 'Medialog.ods'. En el nostre cas, l'hem guardat a una prestatgeria per als suports que posseeix l'autor.

A.6 Preparació dels paquets AIP

Com que ja tenim el fitxer ZIP comprimit dins 'Ingesta AIP', hem executat el *script* que ha creat l'estructura segons l'especificació BagIt. Abans, però, l'hem editat en funció de les característiques de l'AIP. Com que al nostre cas pràctic s'ha tractat d'un investigador que vol lliurar les seves dades de recerca a una institució fictícia, els camps de metadades BagIt que havíem d'editar al nostre *script* '15-BagIt.sh' els hem deixat de la següent manera:

- Source organization: Universitat de Catalunya
- Organization address: Carrer Balmes, 500, 08022 Barcelona
- Contact name: Dídac Reguant
- Contact phone: +34 934345578
- Contact mail: dreguant@uc.cat

- External description: Dades de recerca del projecte "Social platform on cultural heritage in Catalonia". Imatge forense del suport original, imatge *raw* amb redacció de dades confidencials, fotografies del suport original, informes i llistats i fitxers migrats
- External identifier: 37808285USB01
- Bag group identifier: No l'hem utilitzat en aquest cas, ja que només teníem un *bag*
- Bag count: Hem indicat '1 de 1'

S'ha dipositat més d'un suport?

En aquest cas només tenim un suport de l'investigador, així que el pas següent seria el d'ingesta d'AIP. Per poder executar aquest procés, caldria l'ús d'un repositori DSpace amb un servidor real, així que hem aturat el cas pràctic en aquest procés.

Annex B. Formulari de lliurament de dades de recerca

Dades personals del dipositant

Identificador

Correu electrònic

Dades del pla de recerca

Nom del pla de recerca

Codi identificador

Disciplina/àrea de coneixement

Descripció

Patrocinadors

Política de compartició de dades

Política de gestió de dades

Investigador principal

Data d'inici del projecte

Data de finalització del projecte

Organització tècnica de les dades

Tipus de suport:

☐ Discs durs interns ☐ Discs òptics (CD-ROM i/o DVD-ROM)

☐ Discs durs externs USB ☐ Memòries USB

Sistemes operatius que s'han utilitzat:

☐ Windows ☐ OS X ☐ Linux

☐ Altres (especificar):

Tipologies de dades presents:

☐ Text ☐ Text amb llenguatge de marques ☐ Àudio

☐ Imatges ☐ Gràfics vectorials ☐ Vídeo

☐ Fulls de càlcul/dades tabulars ☐ Imatges 3D ☐ Dades GIS

☐ Altres (especificar):

Formats de fitxer utilitzats:

Text (pdf, doc...)

Text amb llenguatge de marques (xml, sgml...)

Àudio (wav, flac...)

Imatges (jpeg, tiff...)

Gràfics vectorials (svg, dwg...)

Vídeo (mpeg, avi...)

Fulls de càlcul/dades tabulars (csv, xls, ods,...)

Imatges 3D (x3d...)

Dades GIS (ESRI, GEOTIFF...)

Altres (especificar)

Heu utilitzat un o més esquemes de metadades? En cas afirmatiu, indiqueu-lo(s).

Heu estructurat i/o normalitzat les vostres carpetes i fitxers? En cas afirmatiu, indiqueu quin patró heu fet servir.

Mida total de les dades (indiqueu-la en MB, GB o TB):	**Nombre total de fitxers presents al suport:**

Informació de drets d'autor

Les vostres dades de recerca tenen alguna llicència d'ús com Creative Commons o Open Data Commons? En cas afirmatiu, indiqueu-lo.

Protecció de dades

Les vostres dades de recerca contenen informació personal i/o sensible que calgui que sigui bloquejada a l'accés públic? En cas afirmatiu, indiqueu els fitxers afectats.

Lloc i data **Signatura del dipositant**

Annex C. Qüestionari de les entrevistes realitzades a responsables de repositoris

1. El nostre model contempla la preservació de grans volums de dades mitjançant la recepció de suports com memòries USB, discs òptics i discs durs. La vostra institució té un flux de treball normalitzat per a l'emmagatzematge d'aquests suports?
2. Per construir el model ens hem basat en l'estàndard OAIS. La vostra institució també el fa servir com a model de referència per a la preservació?
3. El repositori de la vostra institució ja està gestionant i preservant dades de recerca? En cas afirmatiu, de quina manera ho feu?
4. Hem elaborat un model de metadades descriptives en Dublin Core en el que seria necessària la creació de qualificadors per a camps especialitzats. Està fent servir el vostre repositori algun perfil de metadades per a dades de recerca? En cas afirmatiu, utilitzeu algun d'específic per a dades en ciències socials i humanitats?
5. És possible la ingesta de paquets de gran mida (al voltant d'1 TB) sense dividir els paquets en unitats més petites?
6. Una de les funcionalitats del model és la migració d'un cert nombre de formats de fitxer a altres aptes per a la preservació. La vostra institució té actualment polítiques al respecte? En cas afirmatiu, feu servir algun procés automatitzat? Ho feu abans o després del procés d'ingesta?
7. Penseu que el model proposat podria ser assumit amb els coneixements actuals del vostre personal o amb un cert nivell de formació addicional?
8. Teniu algun procediment en els casos que aparegui un virus o un programari maliciós dins els continguts que accepta el vostre repositori?
9. Hi ha casos en què es pot presentar contingut amb dades privades o confidencials que no es pot difondre en accés obert. Teniu algun procediment per a aquests casos?
10. Els paquets d'ingesta del nostre model s'han creat segons l'especificació BagIt. Coneixeu aquesta especificació? En cas afirmatiu, el vostre repositori accepta aquest tipus de paquets? Utilitzeu aquesta especificació per a algun altre servei de la biblioteca?
11. Penseu que la proposta de flux de treball és viable dins els processos actuals del repositori?

Annex D. Glossari

Arxiu híbrid: Arxiu que conté material en paper i material nascut digital.

Bash: Intèrpret i llenguatge de comanaments d'Unix.

Checksum: Funció *hash* que serveix per detectar canvis accidentals en una seqüència de dades per protegir la seva integritat.

CP/M: Sistema operatiu desenvolupat per al microprocessador Intel 8080.

Disc dur d'estat sòlid: Dispositiu que utilitza memòria flash per emmagatzemar les dades, sense parts mecàniques. Més ràpid i més costós que el disc dur magnètic.

Docking station: Perifèric dissenyat per poder connectar un altre dispositiu a una estació de treball, com un disc dur.

EAD: Estàndard XML per codificar eines d'ajuda arxivístiques.

ELF: Format de fitxer utilitzat per a executables, codi objecte, biblioteques compartides i abocaments de memòria.

FC5025: Dispositiu que permet connectar una unitat de disquets de 5 ¼ polzades a un port USB.

Fedora: Sistema de repositori de codi obert per a la gestió i difusió de contingut digital. No s'ha de confondre amb la distribució Linux del mateix nom.

Fez: Projecte de codi obert per produir i mantenir una interfície web flexible dins Fedora per a qualsevol biblioteca o institució que permeti configurar i publicar o arxivar documents.

FinalDraft: Programari assistent per a l'escriptura de guions per al cinema i la televisió.

Flask: Infraestructura minimalista escrita en Python que permet crear aplicacions web ràpidament i amb un nombre mínim de línies de codi.

FRED: Estació de treball especialment equipada per fer anàlisis digitals forenses.

Garmin: Empresa especialitzada en tecnologia GPS.

IETF: Organització que desenvolupa i promou estàndards d'Internet.

Imatge forense: Còpia bit per bit d'un suport o una partició d'un suport amb objectes digitals, com un disc dur. Sovint incorpora metadades per validar que es tracta d'una còpia idèntica a la font original.

Imatge lògica: Còpia de tots o part dels fitxers referenciats a un sistema de fitxers.

Imatge *raw*: Còpia bit per bit de les dades d'un suport o una partició d'un suport sense addicions com la incorporació de metadades.

ISO: Organització internacional que promou estàndards propietaris, industrials i comercials.

Jiffy: Unitat de temps definida per la freqüència base del rellotge del sistema. Per convenció, 0,01 segons.

JSON: Estàndard obert basat en text dissenyat per a l'intercanvi de dades de lectura directa.

Karen's Directory Printer: Programari que permet imprimir el nom de cada fitxer en una unitat, juntament amb la mida del fitxer i la seva data i hora de darrera modificació, així com els seus atributs.

KryoFlux: Dispositiu que permet connectar unitats de disquet per USB i recuperar els continguts amb programari especialitzat.

LNK: Extensió de fitxer per a un fitxer drecera utilitzat per Windows per apuntar a un fitxer executable.

Maquinari: Col·lecció d'elements físics dins un sistema informàtic. S'inclouen discs durs, plaques base, memòries USB, etc.

MIME: Estàndard d'Internet que permet identificar els formats de fitxer que s'envien a clients Web.

NeXus: Format comú de dades per a ciència de muons, neutrons i raigs X.

OAI-PMH: Protocol dissenyat per a la transmissió de metadades en format Dublin Core.

Objecte digital: Seqüència estructurada de bytes.

Programari: Conjunt de components lògics necessaris per realitzar tasques específiques dins un sistema informàtic.

PUID: Esquema d'identificadors persistents, únics i unívocs per identificar formats d'objectes digitals.

Python: Llenguatge de programació orientat a objectes.

RFC: Memoràndum que descriu mètodes, recerques o innovacions aplicables a Internet i sistemes connectats a Internet.

***Scriptorium* digital**: Lloc on es preserven i repliquen materials nascuts digitals mitjançant tècniques d'anàlisi forense digital.

Servidor: Programari o maquinari capaç d'atendre peticions d'altres màquines i donar-hi les respostes adequades.

Servlet: Programa Java que amplia les capacitats d'un servidor.

Sistema de fitxers: Mètode mitjançant el qual es controla i recupera la informació. Als sistemes informàtics existeixen diversos com FAT, NTFS, HFS, etc.

Suport fugitiu: Suport amb material nascut digital que es troba dins una col·lecció de documents en paper i que no ha rebut cap avaluació ni estudi abans de la seva adquisició.

UUID: Identificador únic universal de 128 bits que s'utilitza per identificar un objecte o entitat de forma unívoca.

Valor *hash*: Valor que permet verificar la integritat de les dades d'un objecte digital i detectar possibles errors durant la seva transmissió i emmagatzematge.

Video-CD: Format estàndard per a l'emmagatzematge de vídeo en un disc compacte.

Write blocker: Dispositiu que bloqueja l'escriptura d'un suport per tal d'evitar qualsevol alteració de la informació original.

X.509: Estàndard per gestionar certificats digitals i encriptació de claus públiques.

XOR: Tècnica per ofuscar dades; s'utilitza sovint per ocultar dades i codi sensibles dins fitxers i programes amb *malware*.

Bibliografia

AIMS Work Group (2012). *AIMS born-digital collections: an inter-institutional model for stewardship*. <http://dcs.library.virginia.edu/aims/white-paper/>. [Consulta: 26/03/2017].

Akers, Katherine G.; Green, Jennifer A. (2014). "Towards a symbiotic relationship between academic libraries and disciplinary data repositories: a Dryad and University of Michigan case study". *International Journal of Digital Curation*, vol. 9, n. 1, p. 119-131. <http://dx.doi.org/10.2218/ijdc.v9i1.306>. [Consulta: 05/06/2016].

Aleixandre-Benavent, R. *et al.* (2013). "Disponibilidad en abierto de los artículos y de los datos brutos de investigación en las revistas pediátricas españolas". *Anales de Pediatría*, vol. 82, n. 1, p. e90-e94. <http://dx.doi.org/10.1016/j.anpedi.2013.11.014>. [Consulta: 08/08/2015].

Altman, Micah; Crosas, Mercè (2013). "The evolution of data citation: from principles to implementation". *IASSIST Quarterly*, vol. 37, n. 1-4, p. 62-70. <http://www.iassistdata.org/sites/default/files/iqvol371_4_altman.pdf>. [Consulta: 02/04/2016].

Andreoli-Versbach, Patrick; Mueller-Langer, Frank (2014). "Open access to data: an ideal professed but not practised". *Research Policy*, vol. 43, n. 9, p. 1621-1633. <http://dx.doi.org/10.1016/j.respol.2014.04.008>. [Consulta: 27/03/2016].

Ariza López, Francisco Javier *et al.* (2012). "Preservación de la información geográfica: perspectivas y situación en España". *GeoFocus (Artículos)*, n. 12, p. 171-200. <http://geofocus.rediris.es/2012/Articulo8_2012.pdf>. [Consulta: 02/05/2016].

Armbruster, Chris; Romary, Laurent (2010). "Comparing repository types: challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication". *International Journal of Digital Library Systems*, vol. 1, n. 4, p. 61-73. <https://arxiv.org/abs/1005.0839>. [Consulta: 26/03/2017].

Ashcroft, John; Daniels, Deborah J.; Hart, Sarah V. (2004). *Forensic examination of digital evidence: a guide for law enforcement*. <https://www.ncjrs.gov/pdffiles1/nij/199408.pdf>. [Consulta: 26/03/2017].

Assante, Massimiliano *et al.* (2016). "Are scientific data repositories coping with research data publishing?". *Data Science Journal*, vol. 15, p. 6. <http://dx.doi.org/10.5334/dsj-2016-006>. [Consulta: 08/12/2016].

Ayers, Rick; Brothers, Sam; Jansen, Wayne (2014). *Guidelines on mobile device forensics (NIST Special Publication 800-101 Revision 1)*. <http://dx.doi.org/10.6028/NIST.SP.800-101r1>. [Consulta: 17/07/2016].

Barrera-Gomez, Julianna; Erway, Ricky (2013). *Walk this way: detailed steps for transferring born-digital content from media you can read in-house*. Dublin, Ohio: OCLC Research. ISBN 9781556534546. <http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf>. [Consulta: 20/07/2014].

Bengtson, Jason (2012). "Preparing for the age of the digital palimpsest". *Library Hi Tech*, vol. 30, n. 3, p. 513-522. <http://dx.doi.org/10.1108/07378831211266636>. [Consulta: 26/10/2014].

Bercovitz Rodríguez-Cano, Rodrigo (2006). *Manual de propiedad intelectual*. Valencia: Tirant lo Blanch. ISBN 848456679X.

Bermúdez, Óscar; Barragán, Antonio; Alonso, F. (2011). "La gestión de los datos polares en España: una aproximación a la contribución de las ciencias de la vida". *Ecosistemas*, vol. 20, n. 1, p. 94-103. <http://www.revistaecosistemas.net/index.php/ecosistemas/article/view/16>. [Consulta: 05/04/2015].

Borgman, Christine L. (2012). "The conundrum of sharing research data". *Journal of the American Society for Information Science and Technology*, vol. 63, n. 6, p. 1059-1078. <http://dx.doi.org/10.1002/asi.22634>. [Consulta: 08/08/2015].

Bradley, Jessica R.; Garfinkel, Simson L. (2015). *bulk extractor 1.4: user manual*. <http://digitalcorpora.org/downloads/bulk_extractor/BEUsersManual.pdf>. [Consulta: 16/11/2016].

Breeding, Marshall (2012). "From disaster recovery to digital preservation". *Computers in Libraries*, vol. 32, n. 4, p. 22-24. <https://librarytechnology.org/repository/item.pl?id=16821>. [Consulta: 31/03/2017].

Byard, Roger W. *et al.* (2016). "Locard's principle of exchange, dental examination and fragments of skin". *Journal of Forensic Sciences*, vol. 61, n. 2, p. 545-547. <http://dx.doi.org/10.1111/1556-4029.12964>. [Consulta: 30/08/2016].

Carlson, Jake (2012). "Demystifying the data interview: developing a foundation for reference librarians to talk with researchers about their data". *Reference Services Review*, vol. 40, n. 1, p. 7-23. <http://dx.doi.org/10.1108/00907321211203603>. [Consulta: 08/12/2016].

Carlson, Jake; Stowell-Bracke, Marianne (2013). "Data management and sharing from the perspective of graduate students: an examination of the culture and practice at the Water Quality Field Station". *Portal: Libraries and the Academy*, vol. 13, n. 4, p. 343-361. <http://dx.doi.org/10.1353/pla.2013.0034>. [Consulta: 26/03/2017].

Carrier, Sarah; Dube, Jed; Greenberg, Jane (2007). "The DRIADE project: phased application profile development in support of open science". *Proceedings of the International Conference on Dublin Core and Metadata Applications*, p. 35-42. <http://dcpapers.dublincore.org/pubs/article/view/875>. [Consulta: 08/04/2017].

Carroll, Laura *et al.* (2011). "A comprehensive approach to born-digital archives". *Archivaria*, vol. 72, p. 61-92. <http://pid.emory.edu/ark:/25593/cksgv>. [Consulta: 13/07/2016].

Castagné, Michael (2013). *Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra*. <http://dx.doi.org/10.14288/1.0075768>. [Consulta: 26/03/2017].

CCSDS (2011). *Audit and Certification of Trustworthy Digital Repositories: recommended practice (CCSDS 652.0-M-1: magenta book)*. Washington, DC: CCSDS. <https://public.ccsds.org/pubs/652x0m1.pdf>. [Consulta: 26/03/2017].

CCSDS (2012). *Reference model for an Open Archival Information System (OAIS): recommended practice (CCSDS 650.0-M-2: magenta Book)*. Washington, DC: CCSDS. <https://public.ccsds.org/pubs/650x0m2.pdf>. [Consulta: 26/03/2017].

Chassanoff, Alexandra; Woods, Kam; Lee, Christopher A. (2016). "Digital preservation metadata practice for disk image access". *Digital preservation metadata for practitioners*. Cham, Switzerland: Springer International Publishing, p. 99-109. ISBN 9783319437613.

Chen, Hsin-liang; Zhang, Yin (2014). "Functionality analysis of an open source repository system: current practices and implications". *The Journal of Academic Librarianship*, vol. 40, n. 6, p. 558-564. <http://dx.doi.org/10.1016/j.acalib.2014.09.012>. [Consulta: 31/05/2015].

Childs, Sue *et al.* (2014). "Opening research data: issues and opportunities". *Records Management Journal*, vol. 24, n. 2, p. 142-162. <http://dx.doi.org/10.1108/RMJ-01-2014-0005>. [Consulta: 02/06/2015].

Chisum, W. Jerry; Turvey, Brent E. (2006). "A history of crime reconstruction". *Crime reconstruction*. Waltham, MA: Academic Press, p. 1-35. ISBN 9780123693754.

Cohen, Michael; Garfinkel, Simson; Schatz, Bradley (2009). "Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow". *Digital Investigation*, vol. 6, n. SUPPL., p. S57-S68. <http://dx.doi.org/10.1016/j.diin.2009.06.010>. [Consulta: 04/09/2016].

Cohen, Michael; Schatz, Bradley (2010). "Hash based disk imaging using AFF4". *Digital Investigation*, vol. 7, n. Supplement, p. S121-S128. <http://dx.doi.org/10.1016/j.diin.2010.05.015>. [Consulta: 17/07/2016].

Cohn, Jeffrey P. (2012). "DataONE opens doors to scientists across disciplines". *BioScience*, vol. 62, n. 11, p. 1004. <http://dx.doi.org/10.1525/bio.2012.62.11.16>. [Consulta: 19/04/2016].

Cramer, Tom; Kott, Katherine (2010). "Designing and implementing second generation digital preservation services: a scalable model for the Stanford Digital Repository". *D-Lib Magazine*, vol. 16, n. 9/10. <http://dx.doi.org/10.1045/september2010-cramer>. [Consulta: 13/04/2017].

Crispino, Frank (2008). "Nature and place of crime scene management within forensic sciences". *Science and Justice*, vol. 48, n. 1, p. 24-28. <http://dx.doi.org/10.1016/j.scijus.2007.09.009>. [Consulta: 30/08/2016].

CRL; OCLC (2007). *Trustworthy Repositories Audit & Certification: criteria and checklist*. Chicago, Illinois; Dublin, Ohio: CRL; OCLC. <http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf>. [Consulta: 24/03/2016].

Crosas, Mercè (2011). "The Dataverse network®: an open-source application for sharing, discovering and preserving data". *D-Lib Magazine*, vol. 17, n. 1/2. <http://dx.doi.org/10.1045/january2011-crosas>. [Consulta: 05/04/2015].

Crosas, Mercè (2012). "A data sharing story". *Journal of eScience Librarianship*, vol. 1, n. 3, p. 173-179. <http://dx.doi.org/10.7191/jeslib.2012.1020>. [Consulta: 16/11/2016].

Dallmeier-Tiessen, Suenje *et al.* (2014). "Enabling sharing and reuse of scientific data". *New Review of Information Networking*, vol. 19, n. 1, p. 16-43. <http://dx.doi.org/10.1080/13614576.2014.883936>. [Consulta: 20/07/2014].

Dappert, Angela; Enders, Markus (2010). "Digital preservation metadata standards". *Information Standards Quarterly*, vol. 22, n. 2, p. 4-13. <http://dx.doi.org/10.3789/isqv22n2.2010.01>. [Consulta: 03/07/2016].

Dearborn, Carly C.; Barton, Amy J.; Harmeyer, Neal A. (2014). "The Purdue University Research Repository: HUBzero customization for dataset publication and digital preservation". *OCLC Systems & Services*, vol. 30, n. 1, p. 15-27. <http://dx.doi.org/10.1108/OCLC-07-2013-0022>. [Consulta: 20/07/2014].

Del Pozo, Nicholas; Elford, Douglas; Pearson, David (2009a). "Invited demo: Mediapedia: managing the identification of media carriers". *DigCCurr2009*, p. 76-78. <http://www.slideshare.net/natlibraryofaustralia/mediapedia>. [Consulta: 24/07/2016].

Del Pozo, Nicholas; Elford, Douglas; Pearson, David (2009b). "Invited demo: Prometheus: managing the ingest of media carriers". *DigCCurr2009*, p. 73-75. <http://www.slideshare.net/natlibraryofaustralia/prometheus-13399586>. [Consulta: 24/07/2016].

Dietrich, Dianne (2010). "Metadata management in a data staging repository". *Journal of Library Metadata*, vol. 10, n. 2-3, p. 79-98. <http://dx.doi.org/10.1080/19386389.2010.506376>. [Consulta: 16/11/2014].

Dietrich, Dianne *et al.* (2016). "How to party like it's 1999: emulation for everyone". *Code4Lib Journal*, n. 32. <http://journal.code4lib.org/articles/11386>. [Consulta: 26/07/2016].

Dietrich, Dianne; Adelstein, Frank (2015). "Archival science, digital forensics, and new media art". *Digital Investigation*, vol. 14, n. S1, p. S137-S145. <http://dx.doi.org/10.1016/j.diin.2015.05.004>. [Consulta: 25/07/2016].

Dijk, Elly; Doorn, Peter (2014). "Providing access to research data, publications and current research information at Data Archiving and Networked Services - DANS". *7th Conference on Grey Literature and Repositories*. <http://hdl.handle.net/20.500.11755/00acd75e-c765-4230-b0a6-fb4699832c07>. [Consulta: 26/03/2017].

Dollar, Charles M.; Ashley, Lory J. (2013). "Long-term digital preservation". *Managing electronic records: methods, best practices, and technologies*. Hoboken, New Jersey: John Wiley & Sons, p. 285-315. ISBN 9781118218297.

Donaldson, Devan Ray; Yakel, Elizabeth (2013). "Secondary adoption of technology standards: the case of PREMIS". *Archival Science*, vol. 13, n. 1, p. 55-83. <http://dx.doi.org/10.1007/s10502-012-9179-0>. [Consulta: 03/07/2016].

Downs, Robert R.; Chen, Robert S. (2004). "Cooperative design, development, and management of interdisciplinary data to support the global environmental change research community". *Science and Technology Libraries*, vol. 23, n. 4, p. 5-19. <http://dx.doi.org/10.1300/J122v23n04_02>. [Consulta: 01/05/2016].

Downs, Robert R.; Chen, Robert S. (2010a). "Designing submission and workflow services for preserving interdisciplinary scientific data". *Earth Science Informatics*, vol. 3, n. 1, p. 101-110. <http://dx.doi.org/10.1007/s12145-010-0051-6>. [Consulta: 25/04/2016].

Downs, Robert R.; Chen, Robert S. (2010b). "Self-assessment of a long-term archive for interdisciplinary scientific data as a trustworthy digital repository". *Journal of Digital Information*, vol. 11, n. 1. <https://journals.tdl.org/jodi/index.php/jodi/article/view/753>. [Consulta: 26/04/2016].

Dube, Jed; Carrier, Sarah; Greenberg, Jane (2007). "DRIADE: a data repository for evolutionary biology". *Proceedings of the 7th ACM/IEE Joint Conference on Digital Libraries*, p. 481. <http://dx.doi.org/10.1145/1255175.1255280>. [Consulta: 26/03/2017].

Duranti, Luciana (2009). "From digital diplomatics to digital records forensics". *Archivaria*, vol. 68, n. Fall, p. 39-66. <http://archivaria.ca/index.php/archivaria/article/view/13229/14548>. [Consulta: 24/03/2017].

Eisenberg, Rebecca S. (2006). "Patents and data-sharing in public science". *Industrial and Corporate Change*, vol. 15, n. 6, p. 1013-1031. <http://dx.doi.org/10.1093/icc/dtl025>. [Consulta: 19/04/2016].

Eitken, Brian *et al.* (2008). "The Planets Testbed for digital preservation". *Code4Lib Journal*, n. 3. <http://journal.code4lib.org/articles/83>. [Consulta: 16/07/2016].

Elford, Douglas *et al.* (2008). "Media matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia". *World Library and Information Congress: 74th IFLA General Conference and Council*. <http://archive.ifla.org/IV/ifla74/papers/084-Webb-en.pdf>. [Consulta: 24/07/2016].

España (1978). "Constitución española. Última modificación: 27 de septiembre de 2011". *Boletín Oficial del Estado*, n. 311. <https://www.boe.es/legislacion/codigos/codigo.php?id=151&modo=1&nota=0>. [Consulta: 28/05/2017].

España (1996). "Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual. Última modificación: 5 de noviembre de 2014". *Boletín Oficial del Estado*, n. 97. <https://www.boe.es/buscar/pdf/1996/BOE-A-1996-8930-consolidado.pdf>. [Consulta: 12/06/2016].

España (1998). "Ley 5/1998, de 6 de marzo, de incorporación al Derecho español de la Directiva 96/9/CE, del Parlamento Europeo y del Consejo, de 11 de marzo de 1996, sobre la protección jurídica de bases de datos". *Boletín Oficial del Estado*, n. 57, p. 7935-7940. <https://www.boe.es/boe/dias/1998/03/07/pdfs/A07935-07940.pdf>. [Consulta: 11/09/2016].

España (1999). "Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.". *Boletín Oficial del Estado*, n. 298, p. 43088-43099. <https://www.boe.es/buscar/doc.php?id=BOE-A-1999-23750>. [Consulta: 11/09/2016].

España (2008). "Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal". *Boletín Oficial del Estado*, p. 4103-4136. <https://www.boe.es/buscar/doc.php?id=BOE-A-2008-979>. [Consulta: 11/09/2016].

European Commission (2007). *COM(2007) 56 final: Scientific information in the digital age: access, dissemination and preservation*. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2007:0056:FIN:EN:PDF>. [Consulta: 25/04/2015].

European Commission (2011a). *COM(2011) 882 final: Open data: an engine for innovation, growth and transparent governance*. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF>. [Consulta: 25/04/2015].

European Commission (2011b). *Survey on open access in FP7*. Luxembourg: Publications Office of the European Union. <http://dx.doi.org/10.2777/81083>. [Consulta: 26/03/2017].

European Commission (2012a). *C(2012) 4890 final: Access to and preservation of scientific information*. <http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf>. [Consulta: 06/06/2016].

European Commission (2012b). *COM(2012) 401 final: Towards better access to scientific information: boosting the benefits of public investments in research*. <http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf>. [Consulta: 20/04/2014].

European Commission (2012c). *Online survey on scientific information in the digital age*. Luxembourg: Publications Office of the European Union. <http://dx.doi.org/10.2777/7549>. [Consulta: 31/01/2016].

European Commission (2015). *Horizon 2020: first results*. Brussels: European Commission <http://dx.doi.org/10.2777/420545>. [Consulta: 24/03/2016].

European Commission (2016a). *COM(2016) 178 final: European Cloud Initiative - Building a competitive data and knowledge economy in Europe*. <http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266>. [Consulta: 15/09/2016].

European Commission (2016b). *Guidelines on FAIR data management in Horizon 2020 (version 3.0)*. <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>. [Consulta: 06/09/2016].

European Commission (2016c). *Guidelines on open access to scientific publications and research data in Horizon 2020 (version 3.1)*. <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf>. [Consulta: 06/09/2016].

European Commission (2016d). *Horizon 2020 Annotated Model Grant Agreement: version 2.2 (25 November 2016)*. <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf>. [Consulta: 09/04/2017].

Evans, Jill; Lloyd-Jones, Hannah; Cole, Gareth (2013). *Final report on the Open Exeter project to JISC*. <http://hdl.handle.net/10871/14845>. [Consulta: 11/04/2017].

Farquhar, Adam; Hockx-Yu, Helen (2007). "Planets: integrated services for digital preservation". *International Journal of Digital Curation*, vol. 2, n. 2, p. 88-99. <http://dx.doi.org/10.2218/ijdc.v2i2.31>. [Consulta: 16/07/2016].

Fernández Molina, Juan Carlos (2010). "Preservación digital y derechos de autor: ¿un conflicto sin solución?". *V Congreso Nacional de Bibliotecas Públicas*. Gijón: Subdirección General de Coordinación Bibliotecaria. <http://hdl.handle.net/10421/4902>. [Consulta: 26/03/2016].

Ferrer-Sapena, Antonia; Peset, Fernanda; Aleixandre-Benavent, Rafael (2011). "Acceso a los datos públicos y su reutilización: open data y open government". *El Profesional de la Información*, vol. 20, n. 3, p. 260-269. <http://dx.doi.org/10.3145/epi.2011.may.03>. [Consulta: 18/07/2014].

Forstrom, Michael (2009). "Managing electronic records in manuscript collections: a case study from the Beinecke Rare Book and Manuscript Library". *American Archivist*, vol. 72, n. 2, p. 460-477. <http://dx.doi.org/10.17723/aarc.72.2.b82533tvr7713471>. [Consulta: 22/07/2016].

García-García, Alicia; López-Borrull, Alexandre; Peset, Fernanda (2015). "Data journals: eclosión de nuevas revistas especializadas en datos". *El Profesional de la Información*, vol. 24, n. 6, p. 845-854. <http://dx.doi.org/10.3145/epi.2015.nov.17>. [Consulta: 28/06/2016].

Garfinkel, S. *et al.* (2006). "Advanced Forensic Format: an open, extensible format for disk imaging". *Advances in Digital Forensics II*, vol. 222, p. 14-27. <http://dx.doi.org/10.1007/0-387-36891-4_2>. [Consulta: 06/08/2016].

Garfinkel, Simson (2012). "Digital Forensics XML and the DFXML toolset". *Digital Investigation*, vol. 8, n. 3-4, p. 161-174. <http://dx.doi.org/10.1016/j.diin.2011.11.002>. [Consulta: 18/08/2016].

Garfinkel, Simson; Cox, David (2009). "Finding and archiving the Internet Footprint". *Digital Lives Research Conference: Personal Digital Archives for the 21st Century*. <http://hdl.handle.net/10945/44446>. [Consulta: 28/05/2017].

Garfinkel, Simson L. (2006). "AFF: a new format for storing hard drive images". *Communications of the ACM*, vol. 49, n. 2, p. 85-87. <http://dx.doi.org/10.1145/1113034.1113076>. [Consulta: 26/12/2016].

Garfinkel, Simson L. (2009a). "Automating disk forensic processing with SleuthKit, XML and Python". *Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, p. 73-84. <http://dx.doi.org/10.1109/SADFE.2009.12>. [Consulta: 14/08/2016].

Garfinkel, Simson L. (2009b). "Providing cryptographic security and evidentiary chain-of-custody with the Advanced Forensic Format, Library, and Tools". *International Journal of Digital Crime and Forensics*, vol. 1, n. 1, p. 1-28. <http://dx.doi.org/10.4018/jdcf.2009010101>. [Consulta: 28/05/2017].

Garfinkel, Simson L. (2013). "Digital media triage with bulk data analysis and bulk_extractor". *Computers and Security*, vol. 32, p. 56-72. <http://dx.doi.org/10.1016/j.cose.2012.09.011>. [Consulta: 26/12/2016].

Gengenbach, Martin J. (2012). *«The way we do it here»: mapping digital forensics workflows in collecting institutions*. <http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf>. [Consulta: 14/04/2014].

Gengenbach, Martin J.; Chassanoff, Alexandra; Olsen, Porter (2012). "Integrating digital forensics into born-digital workflows: the BitCurator project". *Proceedings of the American Society for Information Science and Technology*, vol. 49, n. 1. <http://dx.doi.org/10.1002/meet.14504901343>. [Consulta: 26/10/2014].

Giménez Chornet, Vicent (2014). "Criterios ISO para la preservación digital de los documentos de archivo". *Códices*, vol. 10, n. 2, p. 135-150. <https://revistas.lasalle.edu.co/index.php/co/article/view/3267>. [Consulta: 02/05/2015].

Gómez, Nancy-Diana; Méndez, Eva; Hernández-Pérez, Tony (2016). "Datos y metadatos de investigación en ciencias sociales y humanidades: una aproximación desde los repositorios temáticos de datos". *El Profesional de la Información*, vol. 25, n. 4, p. 545-555. <http://dx.doi.org/10.3145/epi.2016.jul.04>. [Consulta: 14/09/2016].

Great Britain. Cabinet Office (2007). *Public Bodies 2007*. [London]: Cabinet Office. <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/266224/PublicBodies2007.pdf>. [Consulta: 26/03/2017].

Great Britain. Cabinet Office (2012). *Open data white paper: unleashing the potential*. [London]: Cabinet Office. <http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf>. [Consulta: 21/04/2015].

Greenberg, Jane *et al.* (2009). "A metadata best practice for a scientific data repository". *Journal of Library Metadata*, vol. 9, n. 3-4, p. 194-212. <http://dx.doi.org/10.1080/19386380903405090>. [Consulta: 26/06/2016].

Groves, Trish (2009). "Managing UK research data for future use". *BMJ*, vol. 338, p. b1252. <http://dx.doi.org/10.1136/bmj.b1252>. [Consulta: 22/07/2014].

Guarino, Alessandro (2013). "Digital forensics as a big data challenge". *ISSE 2013 Securing Electronic Business Processes*, p. 197-203. <http://dx.doi.org/10.1007/978-3-658-03371-2_17>. [Consulta: 13/02/2016].

Guy, Marieke; Donnelly, Martin; Molloy, Laura (2013). "Pinning it down: towards a practical definition of “research data” for creative arts institutions". *International Journal of Digital Curation*, vol. 8, n. 2, p. 99-110. <http://dx.doi.org/10.2218/ijdc.v8i2.275>. [Consulta: 30/08/2015].

Heidorn, P. Bryan (2008). "Shedding light on the dark data in the long tail of science". *Library Trends*, vol. 57, n. 2, p. 280-299. <http://dx.doi.org/10.1353/lib.0.0036>. [Consulta: 08/12/2016].

Hernández-Pérez, Tony; García-Moreno, María-Antonia (2013). "Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios". *El Profesional de la Información*, vol. 22, n. 3, p. 259-263. <http://dx.doi.org/10.3145/epi.2013.may.10>. [Consulta: 21/07/2014].

High level Expert Group on Scientific Data (2010). *Riding the wave: how Europe can gain from the rising tide of scientific data*. <https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data>. [Consulta: 28/05/2017].

Higman, Rosie; Pinfield, Stephen (2015). "Research data management and openness: the role of data sharing in developing institutional policies and practices". *Program: electronic library and information systems*, vol. 49, n. 4, p. 364-381. <http://dx.doi.org/10.1108/PROG-01-2015-0005>. [Consulta: 27/03/2016].

Hodge, Gail; Templeton, Clay; Allen, Robert (2005). "A metadata element set for project documentation". *Science & Technology Libraries*, vol. 25, n. 4, p. 5-23. <http://dx.doi.org/10.1300/J122v25n04_02>. [Consulta: 02/08/2016].

Houghton, Bernadette (2015). "Trustworthiness: self-assessment of an institutional repository against ISO 16363-2012". *D-Lib Magazine*, vol. 21, n. 3/4. <http://dx.doi.org/10.1045/march2015-houghton>. [Consulta: 26/04/2016].

ICPSR (2012). *Guide to social science data preparation and archiving: best practice through the data life cycle* (5th ed.). Ann Arbor, MI: ICPSR. ISBN 9780891388005. <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>. [Consulta: 26/03/2016].

John, Jeremy Leighton (2008). "Adapting existing technologies for digitally archiving personal lives: digital forensics, ancestral computing, and evolutionary perspectives and tools". *5th International Conference on Preservation of Digital Objects (iPRES)*. <http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf>. [Consulta: 20/01/2015].

John, Jeremy Leighton (2009). "The future of saving our past". *Nature*, n. 459, p. 775-776. <http://dx.doi.org/10.1038/459775a>. [Consulta: 13/07/2016].

John, Jeremy Leighton *et al.* (2010). *Digital Lives. Personal digital archives for the 21st century: an initial synthesis (version 0.2)*. <http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf>. [Consulta: 13/07/2016].

John, Jeremy Leighton (2012). *Digital forensics and preservation*. <http://dx.doi.org/10.7207/twr12-03>. [Consulta: 13/07/2016].

Johnston, Leslie (2010). "Releasing open source at the Library of Congress". *OCLC Systems & Services: International digital library perspectives*, vol. 26, n. 2, p. 94-102. <http://dx.doi.org/10.1108/10650751011048461>. [Consulta: 23/10/2016].

Johnston, Wayne (2012). "Digital preservation initiatives in Ontario: trusted digital repositories and research data repositories". *Partnership: The Canadian Journal of Library & Information Practice & Research*, vol. 7, n. 2. <http://dx.doi.org/10.21083/partnership.v7i2.2014>. [Consulta: 23/08/2015].

Khayyat, Mashael; Bannister, Frank (2015). "Open data licensing: more than meets the eye". *Information Polity*, vol. 20, n. 4, p. 231-252. <http://dx.doi.org/10.3233/IP-150357>. [Consulta: 11/09/2016].

Kim, Youngseek; Stanton, Jeffrey M. (2012). "Institutional and individual influences on scientists' data sharing practices". *Journal of Computational Science Education*, vol. 3, n. 1, p. 47-56. <http://dx.doi.org/10.22369/issn.2153-4136/3/1/6>. [Consulta: 08/12/2016].

Kim, Yunhyong; Ross, Seamus (2012). "Digital forensics formats: seeking a digital preservation storage container format for web archiving". *International Journal of Digital Curation*, vol. 7, n. 2, p. 21-39. <http://dx.doi.org/10.2218/ijdc.v7i2.227>. [Consulta: 18/07/2014].

King, Gary (2007). "An introduction to the Dataverse Network as an infrastructure for data sharing". *Sociological Methods & Research*, vol. 36, n. 2, p. 173-199. <http://dx.doi.org/10.1177/0049124107306660>. [Consulta: 23/08/2016].

Kirschenbaum, Matthew G.; Ovenden, Richard; Redwine, Gabriela (2010). *Digital forensics and born-digital content in cultural heritage collections*. Washington, DC: Council on Library and Information Resources. <http://www.clir.org/pubs/reports/pub149/pub149.pdf>. [Consulta: 07/02/2016].

Knight, Gareth (2012). "The Forensic Curator: digital forensics as a solution to addressing the curatorial challenges posed by personal digital archives". *International Journal of Digital Curation*, vol. 7, n. 2, p. 40-63. <http://dx.doi.org/10.2218/ijdc.v7i2.228>. [Consulta: 18/07/2014].

Kousha, Kayvan; Thelwall, Mike (2014). "Disseminating research with web CV hyperlinks". *Journal of the Association for Information Science and Technology*, vol. 65, n. 8, p. 1615-1626. <http://dx.doi.org/10.1002/asi.23070>. [Consulta: 29/05/2016].

Kowalczyk, Stacy; Shankar, Kalpana (2011). "Data sharing in the sciences". *Annual Review of Information Science and Technology*, vol. 45, p. 247-294. <http://dx.doi.org/10.1002/aris.2011.1440450113>. [Consulta: 05/04/2017].

Krause, Edward M. *et al.* (2015). "Evolution of an application profile: advancing metadata best practices through the Dryad data repository". *International Conference on Dublin Core and Metadata Applications*, p. 63-75. <http://dcevents.dublincore.org/IntConf/dc-2015/paper/view/338>. [Consulta: 28/05/2017].

Kruse, Warren G.; Heiser, Jay G. (2001). *Computer forensics: incident response essentials*. Boston, Massachusetts: Addison-Wesley. ISBN 9780201707199.

Kurtz, Mary (2010). "Dublin Core, DSpace, and a brief analysis of three university repositories". *Information Technology and Libraries*, vol. 29, n. 1, p. 40-47. <http://dx.doi.org/10.6017/ital.v29i1.3157>. [Consulta: 31/05/2015].

Lavoie, Brian (2014). *The Open Archival Information System (OAIS) reference model: introductory guide* (2nd ed.). <http://dx.doi.org/10.7207/twr14-02>. [Consulta: 22/02/2016].

Lawrence, Bryan *et al.* (2011). "Citation and peer review of data: moving towards formal data publication". *International Journal of Digital Curation*, vol. 6, n. 2, p. 4-37. <http://dx.doi.org/10.2218/ijdc.v6i2.205>. [Consulta: 08/12/2016].

Layne, R. *et al.* (2012). "Long term preservation of scientific data: lessons from jet and other domains". *Fusion Engineering and Design*, vol. 87, n. 12, p. 2209-2212. <http://dx.doi.org/10.1016/j.fusengdes.2012.07.004>. [Consulta: 31/03/2017].

Leach-Murray, Susan (2016). "Figshare—Get credit for your research". *Technical Services Quarterly*, vol. 33, n. 1, p. 98-99. <http://dx.doi.org/10.1080/07317131.2015.1093855>. [Consulta: 29/05/2016].

Lee, Christopher A. (2005). *Defining digital preservation work: a case study of the development of the reference model for an Open Archival Information System*. University of Michigan. <http://hdl.handle.net/2027.42/39372>. [Consulta: 16/09/2016].

Lee, Christopher A. (2012a). "Archival application of digital forensics methods for authencity, description and access provision". *Comma*, vol. 2012, n. 2, p. 133-140. <http://dx.doi.org/10.3828/comma.2012.2.14>. [Consulta: 28/01/2015].

Lee, Christopher A. *et al.* (2012). "BitCurator: tools and techniques for digital forensics in collecting institutions". *D-Lib Magazine*, vol. 18, n. 5/6. <http://dx.doi.org/10.1045/may2012-lee>. [Consulta: 26/10/2014].

Lee, Christopher A. (2012b). "Digital forensics meets the archivist (and they seem to like each other)". *Provenance, Journal of the Society of Georgia Archivists*, vol. 30, n. 1, p. 3-7. <http://digitalcommons.kennesaw.edu/provenance/vol30/iss1/2>. [Consulta: 07/02/2016].

Lee, Christopher A. *et al.* (2013). *From bitstreams to heritage: putting digital forensics into practice in collecting institutions*. <http://www.bitcurator.net/wp-content/uploads/2013/11/bitstreams-to-heritage.pdf>. [Consulta: 13/07/2016].

Lee, Christopher A. (2014). "Up close and personal: individual digital traces as cultural heritage and discovery through forensics tools". *Personalized Access to Cultural Heritage (PATCH)*. <http://patch2014.files.wordpress.com/2012/07/submission-4-version-of-dec-13-22_45.pdf>. [Consulta: 20/01/2015].

Lee, Christopher A. *et al.* (2014). *From code to community: building and sustaining BitCurator through community engagement*. <http://www.bitcurator.net/wp-content/uploads/2014/11/code-to-community.pdf>. [Consulta: 20/01/2015].

Lee, Christopher A.; Tibbo, Helen R. (2007). "Digital curation and trusted repositories: steps toward success". *Journal of Digital Information*, vol. 8, n. 2. <https://journals.tdl.org/jodi/index.php/jodi/article/view/229>. [Consulta: 23/03/2017].

Lee, Christopher A.; Woods, Kam (2011). *Digital Acquisition Learning Laboratory: a white paper*. <http://www.digpres.com/publications/dall-white-paper.pdf>. [Consulta: 24/07/2016].

Lee, Christopher A.; Woods, Kam (2012). "Automated redaction of private and personal data in collections: toward responsible stewardship of digital heritage". *Proceedings of Memory of the World in the Digital Age: Digitization and Preservation: Documentary Heritage*, p. 298-313. <http://ils.unc.edu/callee/p298-lee.pdf>. [Consulta: 28/01/2015].

Lee, Dong Joon; Stvilia, Besiki (2014). "Developing a data identifier taxonomy". *Cataloging & Classification Quarterly*, vol. 52, n. 3, p. 303-336. <http://dx.doi.org/10.1080/01639374.2014.880166>. [Consulta: 22/08/2016].

Lee, Dong Joon; Stvilia, Besiki (2017). "Practices of research data curation in institutional repositories: a qualitative view from repository staff". *Plos One*, vol. 12, n. 3, p. e0173987. <http://dx.doi.org/10.1371/journal.pone.0173987>. [Consulta: 31/03/2017].

Levelt Committee; Noort Committee; Drenth Committee (Eds.) (2012). *Flawed science: the fraudulent research practices of social psychologist Diederik Stapel*. <http://www.mpi.nl/publications/escidoc-1569964>. [Consulta: 27/03/2016].

Loftus, Mary J. (2010). "The author's desktop". *Emory Magazine*, n. Winter, p. 23-27. <http://www.emory.edu/EMORY_MAGAZINE/2010/winter/winter-2010.pdf>. [Consulta: 31/05/2017].

Lyle, Jared *et al.* (2013). "An applied approach to data curation training at the Inter-University Consortium for Political and Social Research (ICPSR)". *DigCurv*. <http://ceur-ws.org/Vol-1016/paper14.pdf>. [Consulta: 31/05/2016].

Macdonald, Stuart; Martinez-Uribe, Luis (2010). "Collaboration to data curation: harnessing institutional expertise". *New Review of Academic Librarianship*, vol. 16, n. sup1, p. 4-16. <http://dx.doi.org/10.1080/13614533.2010.505823>. [Consulta: 05/06/2016].

Mannheimer, Sara *et al.* (2014). "A balancing act: the ideal and the realistic in developing Dryad's preservation policy". *First Monday*, vol. 19, n. 8. <http://dx.doi.org/10.5210/fm.v19i8.5415>. [Consulta: 05/04/2017].

Meister, Sam; Chassanoff, Alexandra (2014). "Integrating digital forensics techniques into curatorial tasks: a case study". *International Journal of Digital Curation*, vol. 9, n. 2, p. 6-16. <http://dx.doi.org/10.2218/ijdc.v9i2.325>. [Consulta: 14/02/2016].

Mennes, Maarten *et al.* (2013). "Making data sharing work: the FCP/INDI experience". *NeuroImage*, vol. 82, p. 683-691. <http://dx.doi.org/10.1016/j.neuroimage.2012.10.064>. [Consulta: 19/04/2016].

Michener, William *et al.* (2011). "DataONE: Data Observation Network for Earth - preserving data and enabling innovation in the biological and environmental sciences". *D-Lib Magazine*, vol. 17, n. 1/2. <http://dx.doi.org/10.1045/january2011-michener>. [Consulta: 08/04/2017].

Minor, David *et al.* (2009). "Chronopolis: preserving our digital heritage". *6th International Conference on the Preservation of Digital Objects (iPRES)*, p. 141-147. <http://escholarship.org/uc/item/0t2075qc>. [Consulta: 31/05/2017].

Minor, David *et al.* (2010). "Chronopolis digital preservation network". *International Journal of Digital Curation*, vol. 5, n. 1, p. 119-133. <http://dx.doi.org/10.2218/ijdc.v5i1.147>. [Consulta: 31/03/2017].

Misra, Sunitha; Lee, Christopher A.; Woods, Kam (2014). "A web service for file-level access to disk images". *Code4Lib Journal*, n. 25. <http://journal.code4lib.org/articles/9773>. [Consulta: 12/01/2015].

Molloy, Jennifer C. (2011). "The Open Knowledge foundation: open data means better science". *PLoS Biology*, vol. 9, n. 12, p. e1001195. <http://dx.doi.org/10.1371/journal.pbio.1001195>. [Consulta: 21/04/2016].

Murray-Rust, Peter (2008). "Open data in science". *Serials Review*, vol. 34, n. 1, p. 52-64. <http://dx.doi.org/10.1016/j.serrev.2008.01.001>. [Consulta: 14/06/2015].

Nash, Jacob L.; Wheeler, Jonathan (2016). "Desktop batch import workflow for ingesting heterogeneous collections: a case study with DSpace 5". *D-Lib Magazine*, vol. 22, n. 1/2. <http://dx.doi.org/10.1045/january2016-nash>. [Consulta: 06/09/2016].

National Archives of Australia (2010). *AGLS Metadata Standard part 2: usage guide. Version 2.0*. <http://www.agls.gov.au/pdf/AGLS%20Metadata%20Standard%20Part%202%20Usage Guide.PDF>. [Consulta: 20/06/2016].

National Institutes of Health (2015). *NIH grants policy statement*. <http://grants.nih.gov/grants/policy/nihgps/nihgps.pdf>. [Consulta: 06/03/2016].

National Research Council (2008). *Earth observations from space: the first 50 years of scientific achievements*. Washington, DC: The National Academies Press. ISBN 9780309110952. <http://www.nap.edu/openbook.php?record_id=11991>. [Consulta: 02/05/2015].

National Research Council. Committee on Geophysical and Environmental Data (2001). *Resolving conflicts arising from the privatization of environmental data*. Washington, DC: The National Academy Press. ISBN 9780309075831. <http://dx.doi.org/10.17226/10237>. [Consulta: 26/03/2017].

National Science Foundation (2015). *Today's data, tomorrow's discoveries: increasing access to the results of research funded by the National Science Foundation*. <http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>. [Consulta: 04/04/2015].

Nesmith, Tom (1999). "Still fuzzy, but more accurate: some thoughts of the «ghosts» of archival theory". *Archivaria*, vol. 47, p. 136-150. <http://archivaria.ca/index.php/archivaria/article/view/12701/13875>. [Consulta: 23/03/2017].

nestor (2006). *Catalogue of criteria for trusted digital repositories*. Frankfurt am Main: nestor Working Group. <http://nbn-resolving.de/urn:nbn:de:0008-2006060703>. [Consulta: 26/04/2016].

nestor (2014). *Guidelines for the creation of an institutional policy on digital preservation*. [Frankfurt am Main]: nestor Working Group. <http://nbn-resolving.de/urn:nbn:de:0008-2014111006>. [Consulta: 24/03/2016].

Nina-Alcocer, Víctor; Blasco-Gil, Yolanda; Peset, Fernanda (2013). "Datasharing: guía práctica para compartir datos de investigación". *El Profesional de la Información*, vol. 22, n. 6, p. 562-568. <http://dx.doi.org/10.3145/epi.2013.nov.09>. [Consulta: 21/07/2014].

OECD (2007). *OECD principles and guidelines for access to research data from public funding*. Paris: OECD Publications. ISBN 9789264034020. <http://dx.doi.org/10.1787/9789264034020-en-fr>. [Consulta: 03/05/2015].

Palmer, Gary (2001). "A road map for digital forensic research". *Proceedings of the 2001 Digital Forensics Research Workshop (DFRWS 2004)*. <http://www.dfrws.org/2001/dfrws-rm-final.pdf>. [Consulta: 14/02/2016].

Pampel, Heinz *et al.* (2013). "Making research data repositories visible: the re3data.org registry". *PloS ONE*, vol. 8, n. 11, p. e78080. <http://dx.doi.org/10.1371/journal.pone.0078080>. [Consulta: 26/03/2017].

Pappalardo, Kylie; Fitzgerald, Anne (2007). *A guide to developing open access through your digital repository*. Queensland: Queensland University of Technology. ISBN 9780980298840. <http://eprints.qut.edu.au/9671/1/9671.pdf>. [Consulta: 26/03/2016].

Pearce-Moses, Richard (2005). *A glossary of archival and records terminology*. Chicago, IL: The Society of American Archivists. ISBN 1931666148. <http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf>. [Consulta: 11/10/2016].

Peset, Fernanda; Ferrer-Sapena, Antonia; Subirats-Coll, Imma (2011). "Open data y linked open data: su impacto en el área de bibliotecas y documentación". *El Profesional de la Información*, vol. 20, n. 2, p. 165-174. <http://dx.doi.org/10.3145/epi.2011.mar.06>. [Consulta: 16/07/2014].

Pirounakis, George; Nikolaidou, Mara (2009). "Comparing open source digital library software". Theng, Y. (ed.). et al. (ed.). *Handbook of research on digital libraries: design, development, and impact*. Hershey, PA: IGI Global, p. 51-60. ISBN 9781599048796. <http://dx.doi.org/10.4018/978-1-59904-879-6.ch006>. [Consulta: 03/06/2015].

Pitt, Mark A.; Tang, Yun (2013). "What should be the data sharing policy of cognitive science?". *Topics in Cognitive Science*, vol. 5, n. 1, p. 214-221. <http://dx.doi.org/10.1111/tops.12006>. [Consulta: 20/07/2014].

Pollitt, Mark (2010). "A history of digital forensics". *Advances in Digital Forensics VI: Sixth IFIP WG 11.9 International Conference on Digital Forensics*, vol. 337, p. 3-15. <http://dx.doi.org/10.1007/978-3-642-15506-2_1>. [Consulta: 25/03/2017].

Poole, Alex H.; Lee, Christopher A.; Murillo, Angela P. (2012). "CurateGear: enabling the curation of digital collections". *D-Lib Magazine*, vol. 18, n. 11/12. <http://dx.doi.org/10.1045/november2012-poole>. [Consulta: 16/02/2017].

PREMIS Editorial Committee (2008). *Data Dictionary for preservation metadata: version 2.0*. <http://www.loc.gov/standards/premis/v2/premis-2-0.pdf>. [Consulta: 03/07/2016].

PREMIS Editorial Committee (2015). *Data Dictionary for preservation metadata: version 3.0*. <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>. [Consulta: 03/07/2016].

PREMIS Working Group (2005). *Data Dictionary for preservation metadata: final report*. <https://www.loc.gov/standards/premis/v1/premis-dd_1.0_2005_May.pdf>. [Consulta: 03/07/2016].

Ramjoué, Celina (2015). "Towards open science: the vision of the European Commission". *Information Services & Use*, vol. 35, n. 3, p. 167-170. <http://dx.doi.org/10.3233/ISU-150777>. [Consulta: 26/03/2016].

Redwine, Gabriela *et al.* (2013). *Born digital: guidance for donors, dealers, and archival repositories*. Washington, DC: Council on Library and Information Resources <http://www.clir.org/pubs/reports/pub159/pub159.pdf>. [Consulta: 16/02/2016].

Reilly Jr., Bernard; Waltz, Marie E. (2014). "Trustworthy data repositories: the value and benefits of auditing and certification". *Research data management: practical strategies for information professionals*. West Lafayette, Indiana: Purdue University Press, p. 109-126. ISBN 9781557536648.

Reith, Mark; Carr, Clint; Gunsch, Gregg (2002). "An examination of digital forensic models". *International Journal of Digital Evidence*, vol. 1, n. 3. <http://www.just.edu.jo/~Tawalbeh/nyit/incs712/digital_forensic.pdf>. [Consulta: 21/06/2016].

Renear, Allen H.; Sacchi, Simone; Wickett, Karen M. (2010). "Definitions of dataset in the scientific and technical literature". *Proceedings of the ASIST Annual Meeting*, vol. 47, n. 1, p. 1-4. <http://dx.doi.org/10.1002/meet.14504701240>. [Consulta: 26/03/2016].

Rice, Robin (2008). "Applying DC to institutional data repositories". *Proceedings of the International Conference on Dublin Core and Metadata Applications*, p. 212. <http://dcpapers.dublincore.org/pubs/article/view/945>. [Consulta: 08/04/2017].

Ridge, Enda (2014). *Guerrilla analytics: a practical approach to working with data*. Burlington, Massachusetts: Morgan Kaufmann Publishers. ISBN 978-0128002186.

RLG; NARA (2005). *An audit checklist for the certification of Trusted Digital Repositories: draft for public comment*. Mountain View, CA: RLG. <http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file2416.pdf>. [Consulta: 31/05/2017].

RLG; OCLC (2002). *Trusted Digital Repositories: attributes and responsibilities*. Mountain View, California: RLG. <https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf>. [Consulta: 25/04/2016].

Rodríguez Gairín, Josep Manuel; Sulé Duesa, Andreu (2008). "DSpace: un manual específico para gestores de la información y la documentación". *BiD: textos universitaris de biblioteconomia i documentació*, n. 20. <http://www2.ub.edu/bid/consulta_articulos.php?fichero=20rodri2.htm>. [Consulta: 03/06/2015].

Ross, Seamus; Gow, Ann (1999). *Digital archaeology: rescuing neglected and damaged data resources*. London: Library Information Technology Centre. ISBN 1900508516. <http://eprints.gla.ac.uk/100304/1/100304.pdf>. [Consulta: 23/03/2017].

Royal Society (2012). *Science as an open enterprise*. London: The Royal Society. <https://royalsociety.org/policy/projects/science-public-enterprise/Report/>. [Consulta: 03/05/2015].

Schatz, Bradley L. (2015). "Wirespeed: extending the AFF4 forensic container format for scalable acquisition and live analysis". *Digital Investigation*, vol. 14, n. S1, p. S45-S54. <http://dx.doi.org/10.1016/j.diin.2015.05.016>. [Consulta: 04/09/2016].

Schumacher, Jaime; VandeCreek, Drew (2015). "Intellectual capital at risk: data management practices and data loss by faculty members at five american universities". *International Journal of Digital Curation*, vol. 10, n. 2, p. 96-109. <http://dx.doi.org/10.2218/ijdc.v10i2.321>. [Consulta: 31/03/2017].

Schumann, Natascha; Recker, Astrid (2013). "De-mystifying OAIS compliance: benefits and challenges of mapping the OAIS reference model to the GESIS Data Archive". *IASSIST Quarterly*, vol. 36, n. 2, p. 6-11. <http://www.iassistdata.org/sites/default/files/iqvol36_2_recker.pdf>. [Consulta: 22/06/2016].

Shaw, Robert; Corns, Anthony; McAuley, John (2009). "Archiving archaeological spatial data: standards and metadata". *Computer Applications and Quantitative Methods in Archaeology*. <http://archive.caaconference.org/2009/articles/Shaw_Contribution187_c%20(1).pdf>. [Consulta: 25/06/2016].

Simón Castellano, Pere (2012). "El derecho al olvido en el universo 2.0". *BiD: textos universitaris de biblioteconomia i documentació*, n. 28. <http://dx.doi.org/10.1344/105.000001808>. [Consulta: 11/09/2016].

Smallwood, Robert F. (2013). *Managing electronic records: methods, best practices, and technologies*. Hoboken, New Jersey: John Wiley & Sons. ISBN 9781118218297.

Smit, Eefke (2011). "Abelard and Héloise: why data and publications belong together". *D-Lib Magazine*, vol. 17, n. 1/2. <http://dx.doi.org/10.1045/january2011-smit>. [Consulta: 26/03/2016].

Smith, MacKenzie *et al.* (2003). "DSpace: an open source dynamic digital repository". *D-Lib Magazine*, vol. 9, n. 1. <http://dx.doi.org/10.1045/january2003-smith>. [Consulta: 02/04/2017].

Steinhart, Gail; Dietrich, Dianne; Green, Ann (2009). "Establishing trust in a chain of preservation: the TRAC checklist applied to a Data Staging Repository (DataStaR)". *D-Lib Magazine*, vol. 15, n. 9/10. <http://dx.doi.org/10.1045/september2009-steinhart>. [Consulta: 24/08/2015].

Strodl, Stephan *et al.* (2007). "How to choose a digital preservation strategy: evaluating a preservation planning procedure". *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, p. 29-38. ISBN 9781595936448. <http://dx.doi.org/10.1145/1255175.1255181>. [Consulta: 16/07/2016].

Sugimoto, Shigeo; Baker, Thomas; Weibel, Stuart L. (2002). "Dublin Core: process and principles". *Digital libraries: people, knowledge, and technology*. Springer Berlin Heidelberg, p. 25-35. ISBN 9783540002611. <http://dx.doi.org/10.1007/3-540-36227-4_3>. [Consulta: 15/08/2016].

Tansley, Robert *et al.* (2003). "The DSpace institutional digital repository system: current functionality". *2003 Joint Conference on Digital Libraries*, p. 87-97. ISBN 0-7695-1939-3. <http://dx.doi.org/10.1109/JCDL.2003.1204846>. [Consulta: 31/05/2015].

Tenopir, Carol *et al.* (2011). "Data sharing by scientists: practices and perceptions". *PloS ONE*, vol. 6, n. 6, p. e21101. <http://dx.doi.org/10.1371/journal.pone.0021101>. [Consulta: 13/07/2014].

Tenopir, Carol *et al.* (2015). "Changes in data sharing and data reusue practices and perceptions among scientistis worldwide". *PLoS One*, vol. 10, n. 8, p. e0134826. <http://dx.doi.org/10.1371/journal.pone.0134826>. [Consulta: 08/12/2016].

Thelwall, Mike; Kousha, Kayvan (2016). "Figshare: a universal repository for academic resource sharing?". *Online Information Review*, vol. 40, n. 3, p. 333-346. <http://dx.doi.org/10.1108/OIR-06-2015-0190>. [Consulta: 29/05/2016].

Thomas, Susan (2011). "Curating the I, digital: experiences at the Bodleian Library". Lee, Christopher A. (ed.). *I, digital: personal collections in the digital era*. Chicago: Society of American Archivists, p. 280-301. ISBN 1931666385.

Thomas, Susan; Martin, Janette (2006). "Using the papers of contemporary British politicians as a testbed for the preservation of digital personal archives". *Journal of the Society of Archivists*, vol. 27, n. 1, p. 29-56. <http://dx.doi.org/10.1080/00039810600691254>. [Consulta: 17/07/2016].

Unión Europea (2013). "Reglamento (UE) nº 1291/2013 del Parlamento Europeo y del Consejo de 11 de diciembre de 2013 por el que se establece Horizonte 2020, Programa Marco de Investigación e Innovación (2014-2020)". *Diario Oficial de la Unión Europea*, n. 347, p. 104-173. <http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/fp/h2020-eu-establact_es.pdf>. [Consulta: 27/03/2017].

Unión Europea (2016). "Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos". *Diario Oficial de la Unión Europea*, n. 119, p. 1-88. <https://www.boe.es/doue/2016/119/L00001-00088.pdf>. [Consulta: 11/09/2016].

Van Garderen, Peter (2010). "Archivematica: using micro-services and open-source software to deliver a comprehensive digital curation solution". *7th International Conference on Preservation of Digital Objects (iPRES)*, p. 145-149. <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf>. [Consulta: 01/09/2016].

Van Garderen, Peter; Mumma, Courtney C. (2013). "Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation". *10th International Conference on Preservation of Digital Objects (iPRES)*. <http://purl.pt/24107/1/>. [Consulta: 26/10/2016].

Vandeven, Sally (2014). "Forensic images: for your viewing pleasure". *SANS Institute*. <https://www.sans.org/reading-room/whitepapers/forensics/forensic-images-viewing-pleasure-35447>. [Consulta: 13/08/2016].

Vardigan, Mary; Lyle, Jared (2014). "The Inter-university Consortium for Political and Social Research and the Data Seal of Approval: accreditation experiences, challenges, and opportunities". *Data Science Journal*, vol. 13, p. PDA83-PDA87. <http://dx.doi.org/10.2481/dsj.IFPDA-14>. [Consulta: 12/06/2016].

Vardigan, Mary; Whiteman, Cole (2007). "ICPSR meets OAIS: applying the OAIS reference model to the social science archive context". *Archival Science*, vol. 7, n. 1, p. 73-87. <http://dx.doi.org/10.1007/s10502-006-9037-z>. [Consulta: 05/04/2016].

Vines, Timothy H. *et al.* (2014). "The availability of research data declines rapidly with article age". *Current Biology*, vol. 24, n. 1, p. 94-97. <http://dx.doi.org/10.1016/j.cub.2013.11.014>. [Consulta: 21/07/2014].

Vogel, Dustin (2014). "Qualified Dublin Core and the Scholarly Works Application Profile: a practical comparison". *Library Philosophy and Practice*, vol. 1085. <http://digitalcommons.unl.edu/libphilprac/1085>. [Consulta: 02/08/2016].

Wallis, Jillian C.; Rolando, Elizabeth; Borgman, Christine L. (2013). "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology". *PLoS ONE*, vol. 8, n. 7, p. e67332. <http://dx.doi.org/10.1371/journal.pone.0067332>. [Consulta: 23/08/2015].

Westra, Brian *et al.* (2010). "Science and technology resources on the Internet: selected Internet resources on digital research data curation". *Science and Technology Librarianship*, n. 63. <http://dx.doi.org/10.5062/F46D5QXS>. [Consulta: 05/06/2016].

White, Hollie C. *et al.* (2008). "The Dryad data repository: a Singapore framework metadata architecture in a DSpace environment". *International Conference on Dublin Core and Metadata Applications*, p. 157-164. <http://dcpapers.dublincore.org/pubs/article/view/928>. [Consulta: 05/04/2017].

Whyte, Angus (2015). *Where to keep research data: DCC checklist for evaluating data repositories*. Edinburgh: Digital Curation Centre <http://www.dcc.ac.uk/sites/default/files/documents/publications/Where%20to%20keep%20research%20data.pdf>. [Consulta: 27/05/2016].

Wicherts, Jelte M.; Bakker, Marjan (2012). "Publish (your data) or (let the data) perish! Why not publish your data too?". *Intelligence*, vol. 40, n. 2, p. 73-76. <http://dx.doi.org/10.1016/j.intell.2012.01.004>. [Consulta: 21/04/2016].

Wilderbeek, Theo (2013). *Disseny d'una unitat d'anàlisi forense digital en una biblioteca*. Universitat de Barcelona; Universitat Pompeu Fabra. <http://hdl.handle.net/2445/100441>. [Consulta: 16/07/2016].

Wilderbeek, Theo; Térmens, Miquel (2015). "Creación de unidades de análisis forense en bibliotecas". *El Profesional de la Información*, vol. 24, n. 1, p. 44-54. <http://dx.doi.org/10.3145/epi.2015.ene.06>. [Consulta: 01/04/2015].

Wilkinson, Mark D. *et al.* (2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*, vol. 3, p. 160018. <http://dx.doi.org/10.1038/sdata.2016.18>. [Consulta: 22/12/2016].

Williams, Peter *et al.* (2008). "Digital Lives: report of interviews with the creators of personal digital collections". *Ariadne*, n. 55. <http://www.ariadne.ac.uk/issue55/williams-et-al/>. [Consulta: 13/07/2016].

Willis, Craig; Greenberg, Jane; White, Hollie (2012). "Analysis and synthesis of metadata goals for scientific data". *Journal of the American Society for Information Science and Technology*, vol. 14, n. 4, p. 1505-1520. <http://dx.doi.org/10.1002/asi.22683>. [Consulta: 08/04/2017].

Wilsey, Laura *et al.* (2013). "Capturing and processing born-digital files in the STOP AIDS Project Records: a case study". *Journal of Western Archives*, vol. 4, n. 1. <http://digitalcommons.usu.edu/westernarchives/vol4/iss1/1/>. [Consulta: 13/08/2016].

Wira Alam, Andias (2014). "Dublin Core metadata for research data - lessons learned in a real-world scenario with datorium". *Proceedings of the International Conference on Dublin Core and Metadata Applications*, p. 64-73. <http://dcpapers.dublincore.org/pubs/article/view/3703>. [Consulta: 25/09/2016].

Wolverton, Michael (2016). "Digital forensics in the library". *Nature*, vol. 534, p. 139-140. <http://dx.doi.org/10.1038/534139a>. [Consulta: 13/07/2016].

Wong, Pak Chung; Wong, Kwong-Kwok; Foote, Harlan (2003). "Organic data memory using the DNA approach". *Communications of the ACM*, vol. 46, n. 1, p. 95-98. <http://dx.doi.org/10.1145/602421.602426>. [Consulta: 16/07/2016].

Woods, Kam *et al.* (2015). "Functional access to forensic disk images in a web". *12th International Conference on Digital Preservation (iPRES)*, p. 191-195. <http://phaidra.univie.ac.at/o:429564>. [Consulta: 26/12/2016].

Woods, Kam; Brown, Geoffrey (2009). "From imaging to access: effective preservation of legacy removable media". *Archiving 2009*, n. 1, p. 213-218. <http://www.digpres.com/publications/woodsbrownarch09.pdf>. [Consulta: 08/03/2015].

Woods, Kam; Chassanoff, Alexandra; Lee, Christopher A. (2013). "Managing and transforming digital forensics metadata for digital collections". *10th International Conference on Preservation of Digital Objects (iPRES)*. <http://purl.pt/24107/1/>. [Consulta: 03/08/2016].

Woods, Kam; Lee, Christopher A. (2012). "Acquisition and processing of disk images to further archival goals". *Archiving 2012*, p. 147-152. <http://ils.unc.edu/callee/archiving-2012-woods-lee.pdf>. [Consulta: 02/02/2015].

Woods, Kam; Lee, Christopher A. (2015). "Redacting private and sensitive information in born-digital collections". *Archiving 2015*, n. 6, p. 2-7.

Woods, Kam; Lee, Christopher A.; Garfinkel, Simson (2011). "Extending digital repository architectures to support disk image preservation and access". *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries - JCDL '11*, p. 57-66. New York, NY: ACM. ISBN 9781450307444. <http://dx.doi.org/10.1145/1998076.1998088>. [Consulta: 13/07/2016].

Woods, Kam; Lee, Christopher A.; Misra, Sunitha (2013). "Automated analysis and visualization of disk images and file systems for preservation". *Archiving 2013*, p. 239-244. <http://ils.unc.edu/callee/p239-woods.pdf>. [Consulta: 14/02/2016].

Working Group on Expanding Access to Published Research Findings (2012). *Accessibility, sustainability, excellence: how to expand access to research publications*. <https://www.acu.ac.uk/research-information-network/finch-report-final>. [Consulta: 21/04/2015].

Wright, Sarah J. *et al.* (2013). "Using data curation profiles to design the Datastar dataset registry". *D-Lib Magazine*, vol. 19, n. 7/8. <http://dx.doi.org/10.1045/july2013-wright>. [Consulta: 25/08/2015].

Yoon, Ayoung (2014). "End users' trust in data repositories: definition and influences on trust development". *Archival Science*, vol. 14, n. 1, p. 17-34. <http://dx.doi.org/10.1007/s10502-013-9207-8>. [Consulta: 25/04/2016].

Índex de taules

Índex de figures

Llista d'acrònims

3FR	Flexible File Format Raw
AAT	Art & Architecture Thesaurus
AC3	Audio Compression - 3
ADNI	Alzheimer's Disease Neuroimaging Initiative
ADS	Archaeology Data Service
AEPD	Agencia Española de Protección de Datos
AFF	Advanced Forensic Format
AFFLIB	Advanced Forensic Format Library and Tools
AHRC	Arts and Humanities Research Council
AI	Adobe Illustrator
AIC	Archival Information Collection
AIFF	Audio Interchange File Format
AIMS	An Inter-institutional Model for Stewardship
AIP	Archival Information Package
AIU	Archival Information Unit
API	Application Programming Interface
ARCO	Accés, Rectificació, Cancel·lació i Oposició
ARW	Alpha RaW
ASCII	American Standard Code for Information Interchange
ATA	AT Attachment
ATAPI	ATA Packet Interface
AVI	Audio Video Interleave
BBSRC	Biotechnology and Biological Sciences Research Council
BEAM	Bodleian Electronic Archives and Manuscripts
BioLINCC	Biologic Specimen and Data Repository Information Coordinating Center
BMP	BitMaP
BoDAR	Born-Digital Archives Working Group
BSD	Berkeley Software Distribution
BWF	Broadcast Wave Format

CAD	Computer-Aided Design
CALM	Computerisation for Archives, Libraries and Museums
CC	Creative Commons
CCSDS	Consultative Committee for Space Data Systems
CCZero	Creative Commons Zero
CDR	Carolina Digital Repository
CD-RW	Compact Disc-ReWritable
CERP	Collaborative Electronic Records Project
CIESIN	Center for International Earth Science Information Network
CLOCKSS	Controlled Lots of Copies Keep Stuff Safe
CP/M	Control Program for Microcomputers
CR2	Canon Raw version 2
CRAI	Centre de Recursos per a l'Aprenentatge i la Investigació
CRL	Center for Research Libraries
CRW	Canon RaW
CSS	Cascading Style Sheets
CSUC	Consorci de Serveis Universitaris de Catalunya
CSV	Comma Separated Values
CTIC	Centro Tecnológico de la Información y Comunicación
CTN	Clinical Trials Network
DANS	Data Archiving and Networked Services
DataONE	Data Observation Network for Earth
dBASE	data Base
dbGaP	database of Genotypes and Phenotypes
DC	Dublin Core
DCC	Digital Curation Centre
DCMI	Dublin Core Metadata Initiative
DCR	Digital Camera Raw
DDC	Dewey Decimal Classification
DDI	Data Document Initiative
DDL	Data Definition Language
DFXML	Digital Forensics XML
DICOM	Digital Imaging and Communication in Medicine

DIMAC	Disk Image Access for the Web
DIP	Dissemination Information Package
DMID	Division of Microbiology and Infectious Diseases
DMP	Data Management Plan
DNG	Digital Negative Graphics
DOAR	Directory of Open Access Repositories
DOI	Digital Object Identifier
DOSBox	Disk Operating System Box
DPE	DigitalPreservationEurope
DPN	Digital Preservation Network
DRAMBORA	Digital Repository Audit Method Based On Risk Assessment
DRIADE	Digital Repository of Information and Data for Evolution
DRO	Deakin Research Online
DROID	Digital Record and Object IDentification
DSA	Data Seal of Approval
DTD	Document Type Definition
DVD	Digital Versatile Disc
EAD	Encoded Archival Description
EaaS	Emulation as a Service
EASE	Edinburgh Authentication SErvice
EASY	Electronic Archiving System
EIDE	Enhanced Integrated Drive Electronics
ELF	Executable and Linkable Format
EML	Ecological Metadata Language
eMSS	eMANUSCRIPTS
ENCODE	ENCyclopedia Of DNA Elements
ePADD	email: Process, Appraise, Discover, Deliver
EPI	El Profesional de la Información
EPS	Encapsulated PostScript
EPSRC	Engineering and Physical Sciences Research Council
ERF	Epson Raw Format
ESRC	Economic and Social Research Council
ESRI	Environmental Systems Research Institute

EUA	Estats Units d'Amèrica
Exif	Exchangeable image file format
FAIR	Findable, Accessible, Interoperable and Re-usable
FAT	File Allocation Table
Fedora	Flexible Extensible Digital Object Repository Architecture
FCP	Functional Connectomes Project
FFmpeg	Fast Forward MPEG
FFV1	Fast Forward Video codec 1
FIDO	Forensic Information in Digital Objects
FITBIR	Federal Interagency Traumatic Brain Injury Research
FLAC	Free Lossless Audio Codec
FLV	Flash Video
FP7	Seventh Framework Programme for Research
FPR	Format Policy Registry
FRED	Forensic Recovery of Evidence Device
FTK	Forensic ToolKit
FTP	File Transfer Protocol
futureArch	Future of Archives
GB	GigaByte
GCC	GNU Compiler Collection
GCDIS	Global Change Data and Information System
GeoTIFF	Geospatial Tag Image File Format
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen
GHz	GigaHertz
GIF	Graphics Interchange Format
GIS	Geographic Information System
GML	Geography Markup Language
GNU	GNU's Not Unix
GPL	General Public License
HAROLD	Hull Archives Recovery of Legacy Data
HDF5	Hierarchical Data Format 5
heiDATA	Heidelberg Research Data Repository
HFS	Hierarchical File System

HIPC	Human Immunology Project Consortium
HTTP	Hypertext Transfer Protocol
IASA	International Association of Sound and Audiovisual Archives
IBM	International Business Machines
ICPSR	Interuniversity Consortium for Political and Social Research
IDE	Integrated Drive Electronics
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IGY	International Geophysical Year
ImmPort	Immunology Database and Analysis Portal
IMT	Internet Media Type
INDI	International Neuroimaging Data-sharing Initiative
INSPIRE	Infrastructure for Spatial Information in the European Community
IQSS	Institute for Quantitative Social Science
ISO	International Organization for Standardization
ISSN	International Standard Serial Number
JHOVE	JSTOR/Harvard Object Validation Environment
JISC	Joint Information Systems Committee
JPEG	Joint Photographic Experts Group
JP2	JPEG 2000
JSON	JavaScript Object Notation
JSTOR	Journal STORage
KDC	Kodak Digital Camera
KML	Keyhole Markup Language
KNAW	Koninklijke Nederlandse Akademie van Wetenschappen
LCC	Library of Congress Classification
LCSH	Library of Congress Subject Headings
LGPL	Lesser General Public License
libewf	Library of Expert Witness Compression Format
LOCKSS	Lots of Copies Keep Stuff Safe
LOPD	Llei Orgànica de Protecció de Dades
LPCM	Linear Pulse-Code Modulation
LPI	Llei de Propietat Intel·lectual

MARBL	Manuscripts, Archives, and Rare Book Library
MARC	MAchine-Readable Cataloging
MB	MegaByte
Mbit	Megabit
MBOX	Mailbox
MeSH	Medical Subject Headings
METS	Metadata Encoding and Transmission Standard
MHz	MegaHertz
MID	MapInfo Interchange Data
MIF	MapInfo Interchange Format
MIME	Multipurpose Internet Mail Extensions
MIT	Massachusetts Institute of Technology
modENCODE	Model Organism ENCyclopedia Of DNA Elements
MOV	QuickTime Movie
MP3	Moving Picture Experts Group Layer-3 Audio
MP4	Moving Picture Expert Group-4
MPEG	Moving Picture Experts Group
MRC	Medical Research Council
MRI	Magnetic Resonance Imaging
MRW	Minolta RaW
MS	MicroSoft
NARA	National Archives and Records Administration
NARCIS	National Academic Research and Collaborations Information System
NASA	National Aeronautics and Space Administration
NCRAD	National Cell Repository for Alzheimer's Disease
NDA	NIMH Data Archive
NDAR	National Database for Autism Research
NDCT	National Database for Clinical Trials Related to Mental Illness
NDSA	National Digital Stewardship Alliance
NEF	Nikon Electronic Format
NERC	Natural Environment Research Council
nestor	Network of Expertise in long-term STOrage and long-term availability of digital Resources

NHGRI	National Human Genome Research Institute
NHLBI	National Heart, Lung and Blood Institute
NHPRC	National Historical Publications and Records Commission
NIA	National Institute on Aging
NIAGADS	NIA Genetics of Alzheimer's Disease Data Storage
NIAID	National Institute of Allergy and Infectious Diseases
NICHD	National Institute of Child Health and Human Development
NIDA	National Institute on Drug Abuse
NIDDK	National Institute of Diabetes and Digestive and Kidney Diseases
NIH	National Institutes of Health
NIMH	National Institute of Mental Health
NLA	National Library of Australia
NLNZ MET	National Library of New Zealand Metadata Extraction Tool
NLP	Neuro Linguistic Programming
NSF	National Science Foundation
NTFS	New Technology File System
NWO	Nederlandse Organisatie voor Wetenschappelijk Onderzoek
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OASIS	Organization for the Advancement of Structured Information Standards
OCLC	Online Computer Library Center
ODbL	Open Database License
ODC	Open Database Connectivity
ODS	OpenDocument Spreadsheet
OECD	Organisation for Economic Co-operation and Development
OGC	Open Geospatial Consortium
OKF	Open Knowledge Foundation
OOXML	Office Open XML
OpenAIRE	Open Access Infrastructure for Research in Europe
ORCID	Open Researcher and Contributor ID
ORF	Olympus Raw Format
Paradigm	Personal Archives Accessible in Digital Media
PCI	Peripheral Component Interconnect

PCW	Personal Computer Word processor
PDDL	Public Domain Dedication and License
PDF	Portable Document Format
PDF/A	Portable Document Format Archive
PDI	Preservation Description Information
PedsMRI	Pediatric MRI Repository
PEF	Pentax Format
Planets	Preservation and Long-term Access through Networked Services
Plato	Planning Tool
PREMIS	PREservation Metadata: Implementation Strategies
PTAB	Primary Trustworthy Digital Repository Authorization Board
PUID	PRONOM's Persistent Unique Identifier
QTVR	QuickTime Virtual Reality
RAF	Raw, Fuji
RAM	Random Access Memory
RCHME	Royal Commission on the Historical Monuments of England
RCUK	Research Councils UK
RDF	Resource Description Framework
RDoCdb	Research Domain Criteria Database
RFC	Request For Comments
RLG	Research Libraries Group
ROAR	Registry of Open Access Repositories
ROM	Read Only Memory
RTF	Rich Text Format
SAA	Society of American Archivists
SAS	Serial Attached SCSI
SATA	Serial AT Attachment
SBP	Systems Biology Program
SCP	Secure copy
SCSI	Small Computer System Interface
SDMX	Statistical Data and Metadata Exchange
SDR	Stanford Digital Repository
SEDAC	SocioEconomic Data and Applications Center

SFTP	Secure File Transfer Protocol
SGML	Standard Generalized Markup Language
SHARE IT	Spatial Heritage & Archaeological Research Environment IT
SIARD	Software Independent Archiving of Relational Databases
SIP	Submission Information Package
SPSS	Statistical Package for the Social Sciences
STFC	Science and Technology Facilities Council
SUL	Stanford University Libraries
SUSE	Software und Systementwicklung
SVG	Scalable Vector Graphics
SWF	Shockwave Flash
SWORD	Simple Web-service Offering Repository Deposit
TB	TeraByte
TBI	Traumatic Brain Injury
TDR	Trusted Digital Repository
TEDDY	The Environmental Determinants of Diabetes in the Young
TFW	Tiff World File
TGA	Truevision Graphics Adapter
TGN	Thesaurus of Geographic Names
TIFF	Tag Image File Format
TIFF/IT	Tagged Image File Format/Image Technology
TRAC	Trustworthy Repositories Audit and Certification
UAB	Universitat Autònoma de Barcelona
UB	Universitat de Barcelona
UBC	University of British Columbia
UDC	Universal Decimal Classification
UiT	Universitetet i Tromsø
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNF	Universal Numerical Fingerprint
UOC	Universitat Oberta de Catalunya
UPC	Universitat Politècnica de Catalunya
URI	Uniform Resource Identifier

USB	Universal Serial Bus
UUID	Universally Unique IDentifier
UVA	University of Virginia
VCEG	Video Coding Experts Group
VML	Vector Markup Language
W3C	World Wide Web Consortium
W3C-DTF	World Wide Web Consortium Data Time Format
WARC	Web ARChive
WAV/WAVE	WAVE form Audio File Format
WebGIS	Web-based Geographical Information System
Winpe	Windows Preinstallation Environment
WinUAE	Windows Unix Amiga Emulator
WMA	Windows Media Audio
WMV	Windows Media Video
WPD	WordPerfect Document
X3D	eXtensible 3 Dimensions
XML	eXtensible Markup Language
XOR	Exclusive or
ZIRC	Zebrafish International Resource Center