

La gestió de les dades de la recerca

Ignasi Labastida i Juan
Oficina de Difusió del Coneixement
CRAI Universitat de Barcelona



Estructura

Dia 1: Sessió teòrica

- Què son les dades de la recerca
- El cicle de les dades
- Aspectes legals de les dades
- On emmagatzemar i publicar dades
- Polítiques sobre dades
- Com establir un servei de suport a la gestió

Estructura

Dia 2: Sessió pràctica

- Fem un pla de gestió de dades
- Compartim dades de recerca

Dades, conjunts de dades i bases de dades

Dada: L'element bàsic de la recerca que serveix per verificar-ne els resultats

Conjunt de dades: Una col·lecció de dades

Base de dades: Una col·lecció organitzada de conjunts de dades que s'ha agrupat seguint un criteri

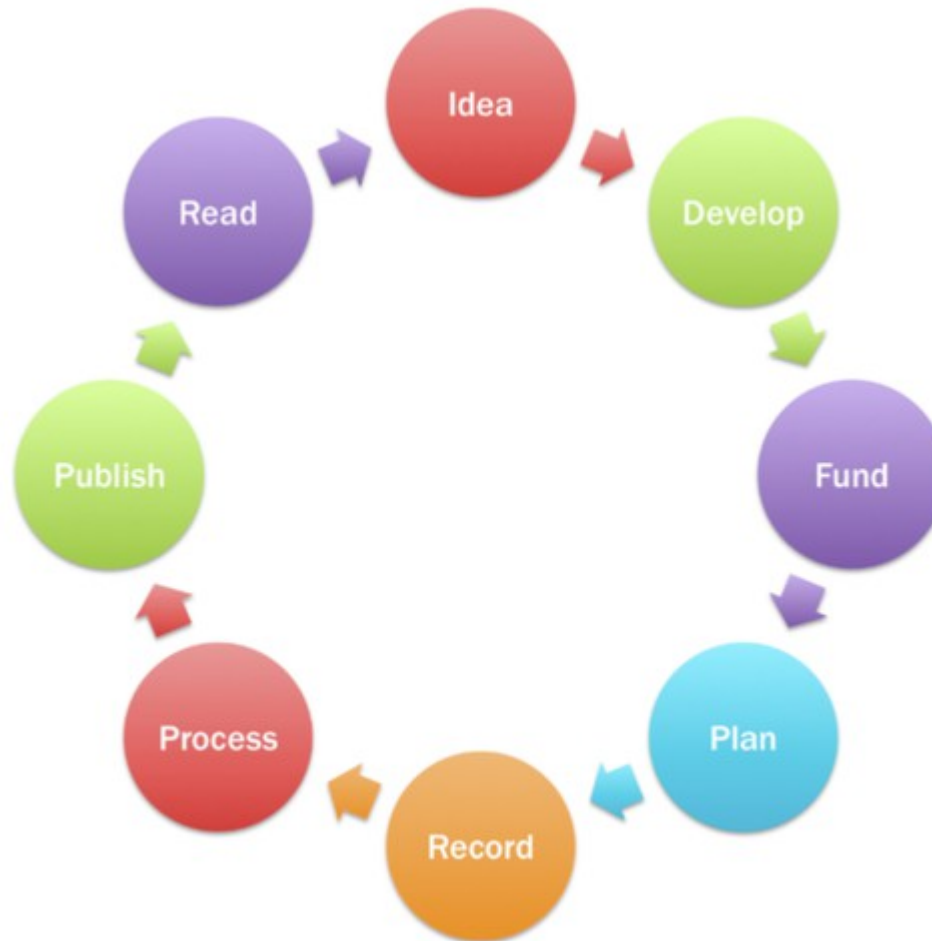
Tipus de dades segons obtenció

- Dades experimentals
- Dades simulades
- Dades observacionals
- Dades derivades
- Dades referencials

Tipus de dades segons format

- Textuals
- Numèriques
- Multimèdia
- Estructurades
- Programari
- Disciplinàries
- Instrumentals

El cicle de la recerca



CC BY Cameron Neylon

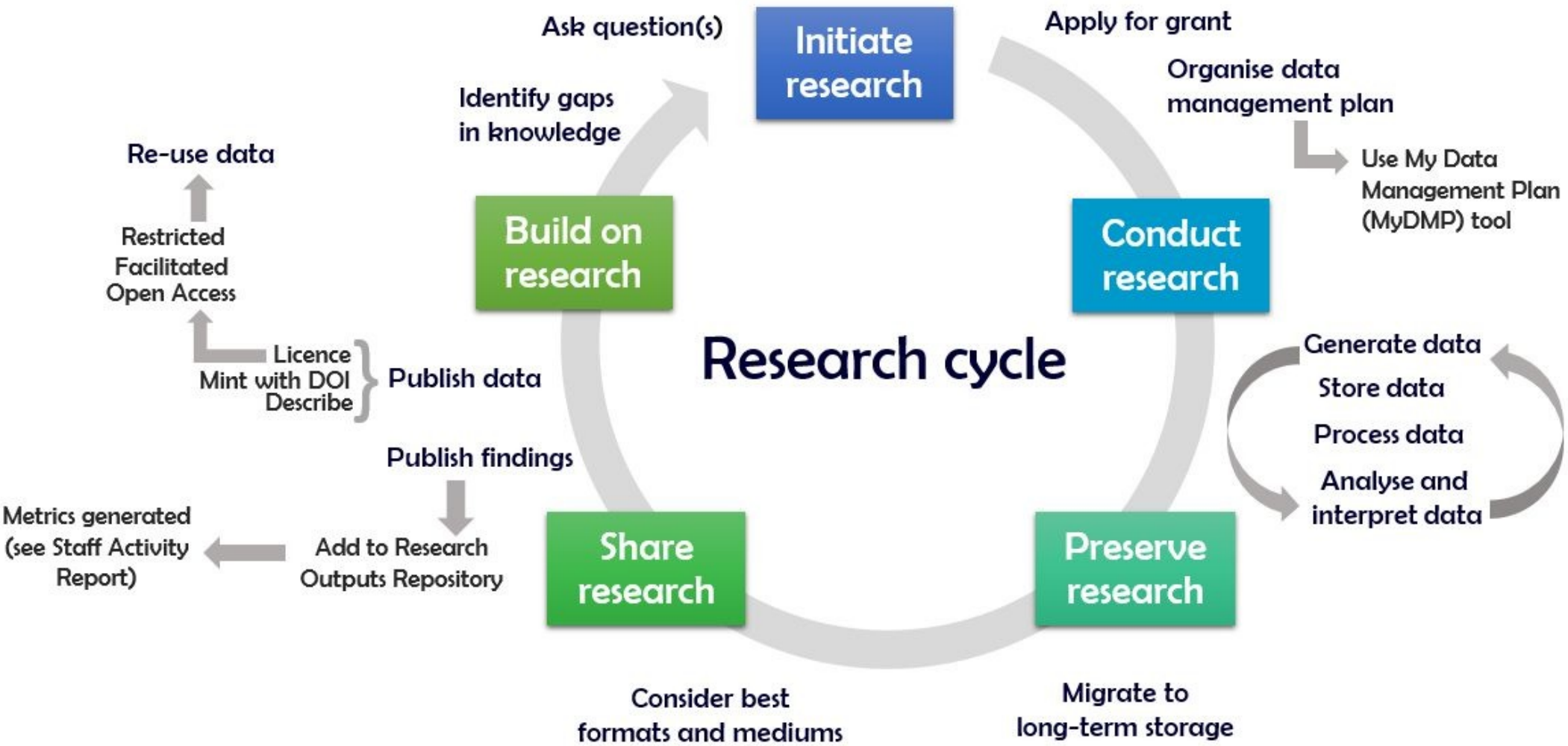
https://commons.wikimedia.org/wiki/File:Research_cycle.png

El cicle de les dades de la recerca



http://www.open.ac.uk/blogs/the_orb/?p=52

Els cicles combinats



<http://guides.library.unisa.edu.au/ResearchDataManagement>

Afirmacions per discutir

Les institucions haurien d'obligar a compartir dades de recerca

Compartir dades és més important que l'accés obert a les publicacions

Els investigadors només pensen sobre la gestió de dades si se'ls imposa

Les biblioteques haurien d'invertir en preservació de programari per facilitar la reutilització de les dades

Els investigadors estan més posicionats que les institucions per dissenyar i actualitzar una política de dades obertes adient

Les dades obertes hauria de ser una responsabilitat de la institució, no de l'investigador

La planificació

Els plans de gestió de dades responen a preguntes com:

- Quines dades es produiran, s'obtindran o s'observaran i en quin format
- Quin tractament tindran
- Com s'emmagatzemaran i es preservaran les dades
- Qui i quan tindrà accés a les dades
- Qui ostenta els possibles drets sobre les dades
- Es podran reutilitzar

La planificació

Els PGD de l'Horitzó 2020 demanen per a cada conjunt de dades

- Referència i nom
- Descripció
- Estàndards i metadades
- Intercanvi de dades
- Arxiu i preservació de les dades

Exemples de PGD

Descripcions

Raw data: Nearly all experiments under the proposal involve electrophysiological recordings in behaving rats. The primary (i.e., raw) forms of data are: 1) waveform recordings in the format of the .plx files (Plexon MAT recording system); 2) video and LED tracking files in the form of .avt and .dvt files (the latter is output of the animal's head position in x,y coordinates collected at 60 Hz by the Plexon Cineplex Studio program); 3) histological data in the form of Nissl-stained brain slices (to localize recording or microinjection sites). The applicant's laboratory houses a microscope with wide-field lenses and a digital camera and regularly practices archiving of relevant digital photographs of brain slices..

Exemples de PGD

Responsabilitats

The PI will be responsible for overall quality assurance, with lead country researchers and the UK research assistant undertaking specific activities to ensure quality control. Detailed protocols for extracting data from secondary sources will be developed, piloted, refined and agreed in Phase 2

Dades personals

A letter explaining the purpose, approach and dissemination strategy (including plans to share data) of the research, and an accompanying consent form (including to share data) will be prepared and translated into the relevant languages.

Exemples de PGD

Accés

All data, metadata, and analyses collected under the proposed experiments will be made publicly available as per NSF guidelines within 2 years of collection via published manuscripts, publicly available final reports to NSF, and/or from data archives at UCSD's Department of Cognitive Science.

Reutilització

User registration is required in order to access or download files. As part of the registration process, users must agree to the conditions of use governing access to the public release data, including restrictions against attempting to identify study participants, destruction of the data after analyses are completed, reporting responsibilities, restrictions on redistribution of the data to third parties, and proper acknowledgement of the data resource. Registered users will receive user support, as well as information related to errors in the data, future releases, workshops, and publication lists. The information provided to users will not be used for commercial purposes, and will not be redistributed to third parties

Eines per a fer PGD

DMP Online

<https://dmponline.dcc.ac.uk/>

PAGODA

<http://www.consorciomadrono.es/pagoda/>

DMP Tool

<https://dmptool.org/>

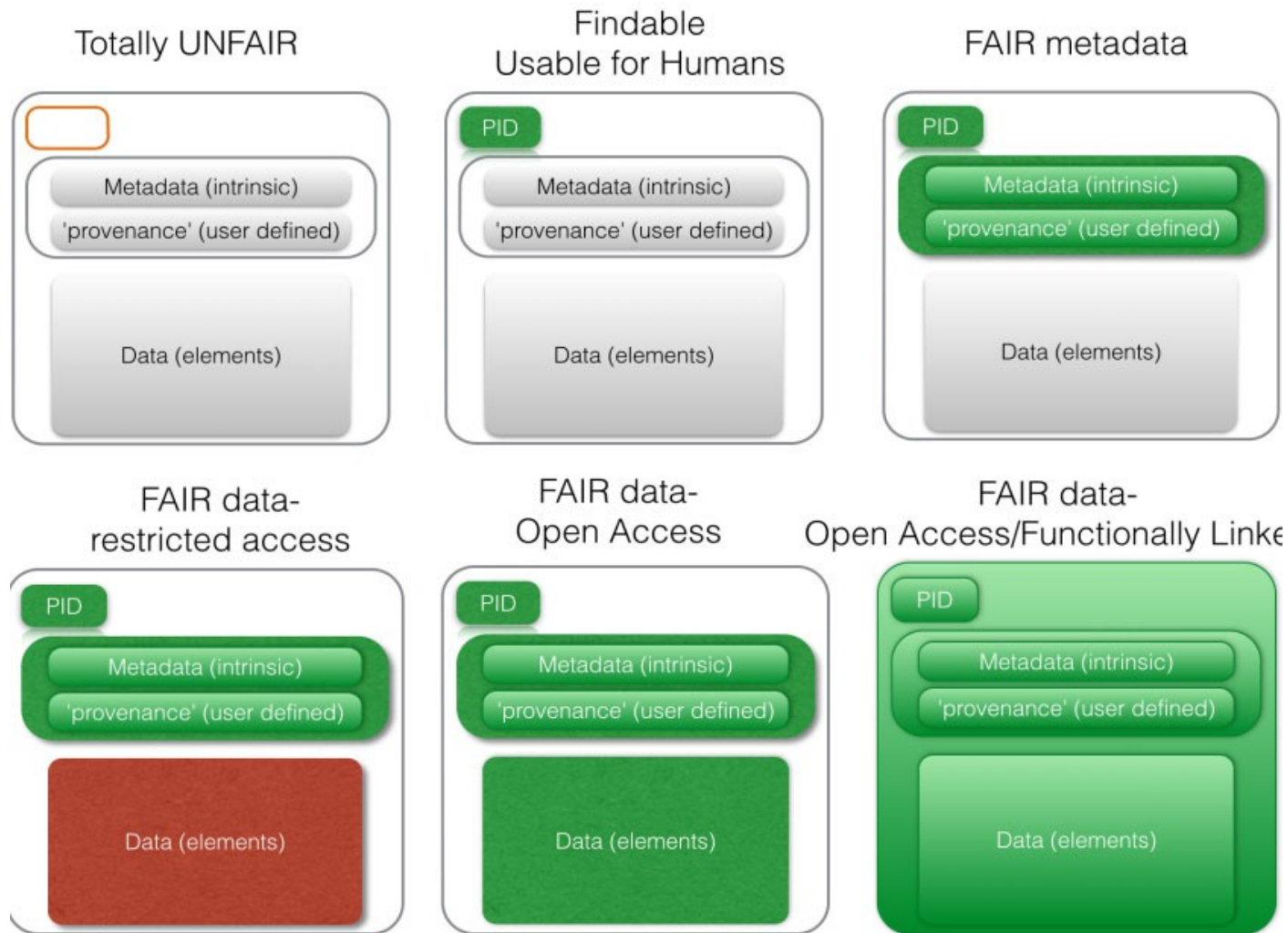
Treballant amb dades

Les dades haurien de ser FAIR

- Findable
- Accessible
- Interoperable
- Re-usable

Treballant amb dades

Data as increasingly FAIR Digital Objects



<https://www.force11.org/fairprinciples>

Treballant amb dades

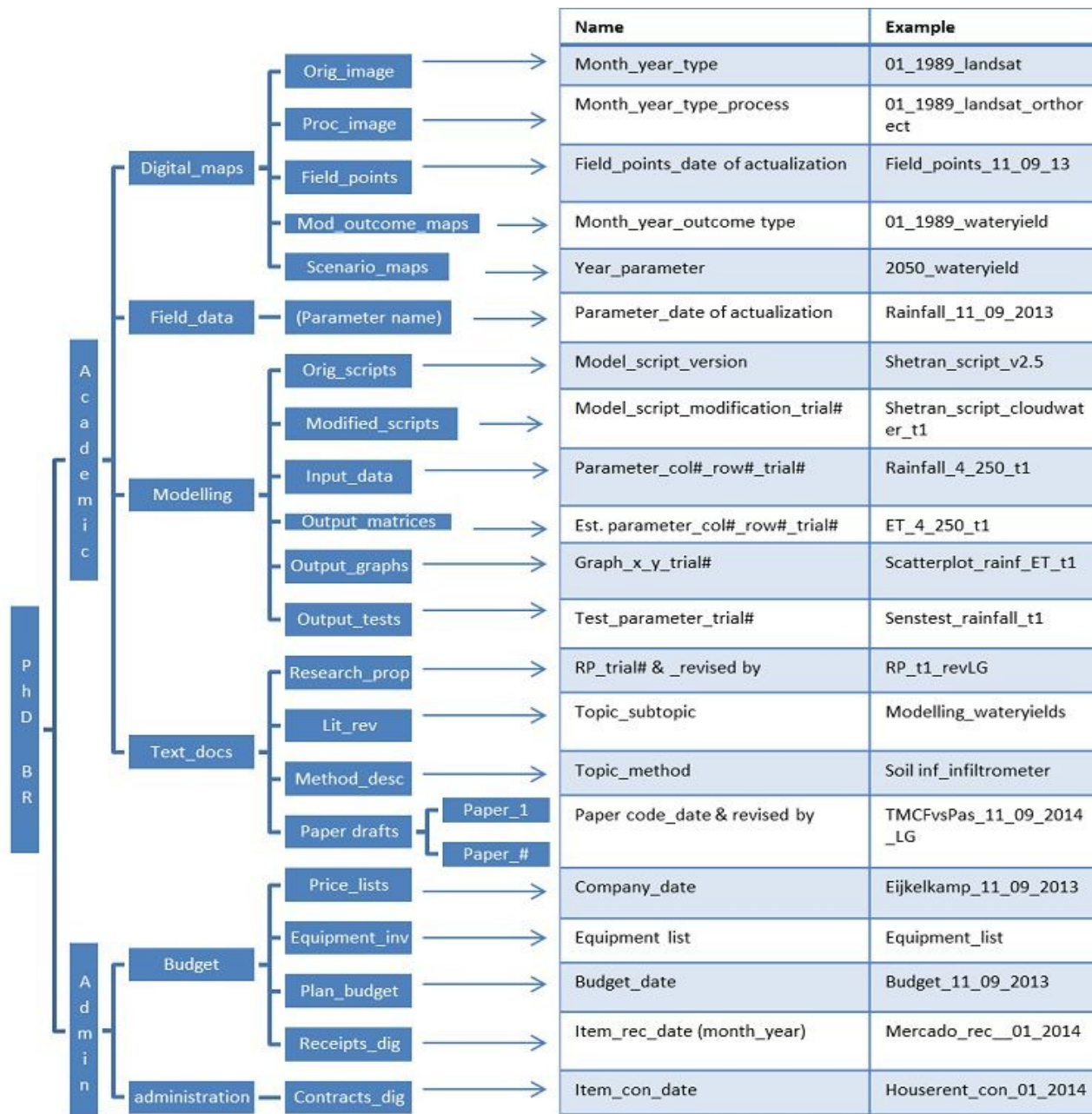
Formats estàndards i no propietaris, quan sigui possible

Organització de carpetes i noms

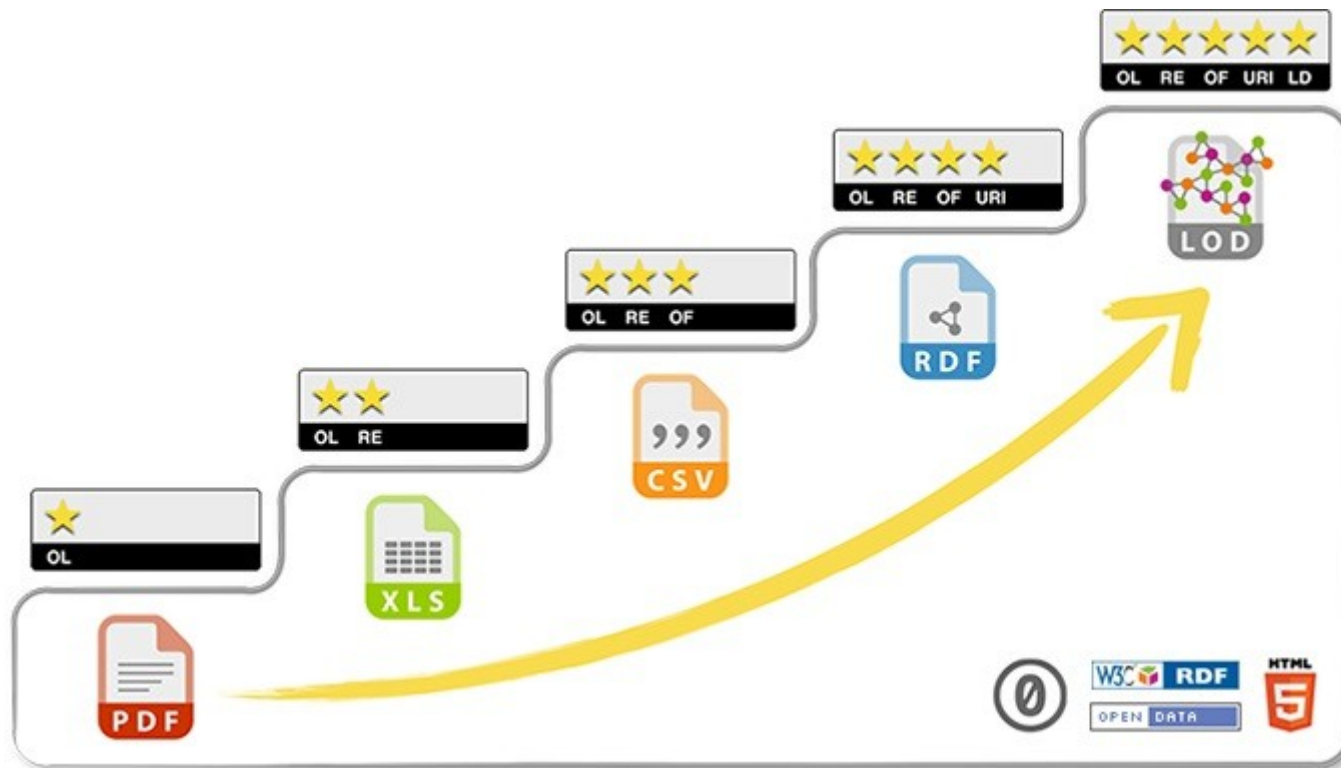
Noms amb sentit

Control de versions

Utilització de metadades (descriptives, estructurals, tècniques, administratives)



<http://www.wageningenur.nl/en/Expertise-Services/Data-Management-Support-Hub/Browse-by-Subject/Organising-files-and-folders.htm>



<http://5stardata.info/>

Desar les dades

Propi ordinador

Disc extern

USB

Servidor

Repositori intern

Núvol

Necessitats d'emmagatzematge

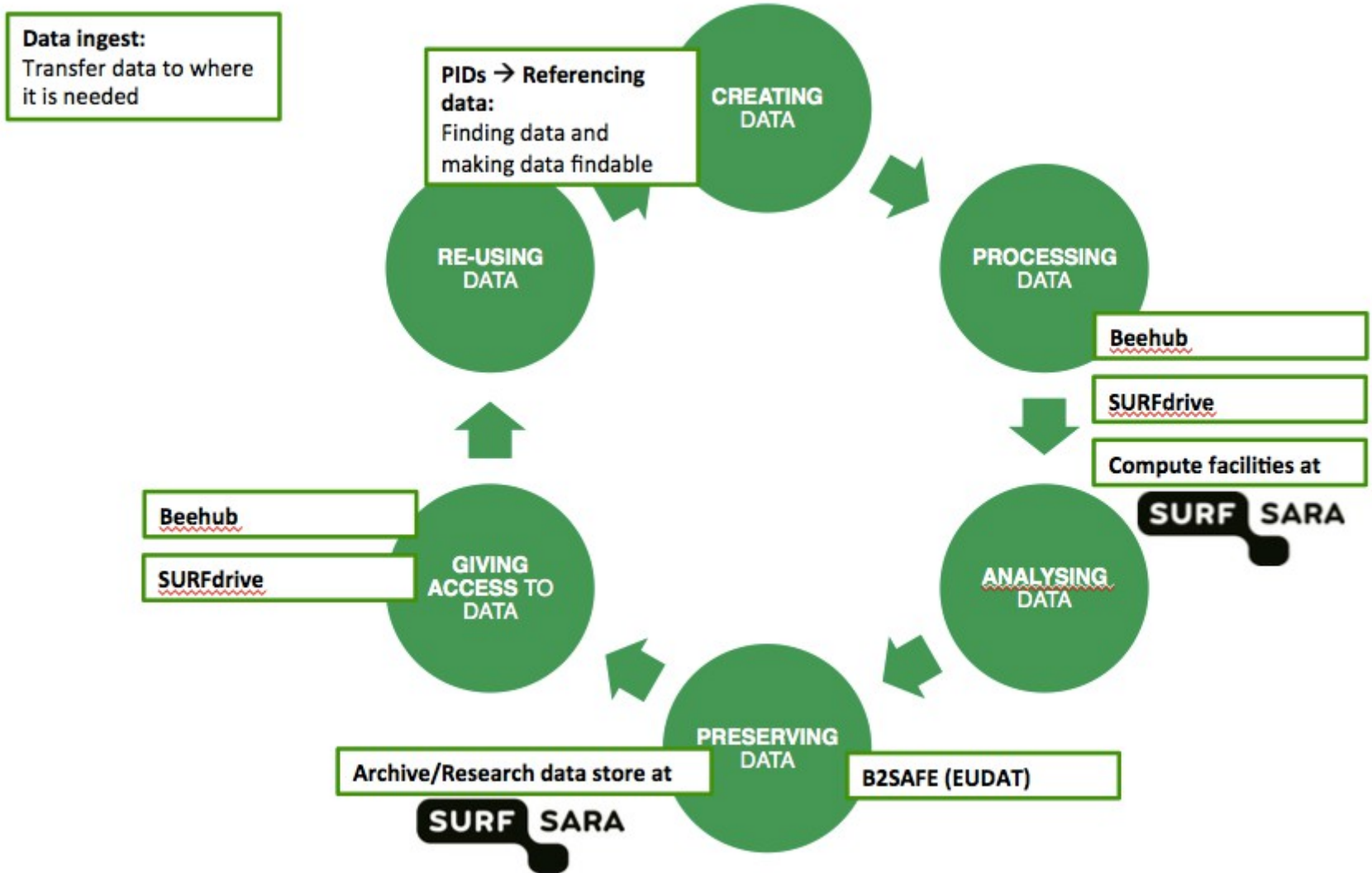
Seguretat contra pèrdues

Control d'accés

Espai

Possibles obligacions

Serveis d'emmagatzematge



<https://www.surf.nl/en/services-and-products/data-archive/portfolio-data-services/index.html>

Difondre les dades

Repositoris temàtics (<http://www.re3data.org/>)

Repositoris institucionals

Repositoris nacionals (<http://easy.dans.knaw.nl>)

Zenodo (<http://zenodo.org>)

Figshare (<http://figshare.org>)

Mendeley Data (<https://data.mendeley.com/>)

Resum

http://ddd.uab.cat/pub/guibib/150829/repositorisdades_2016m5.pdf

La propietat intel·lectual de les dades

La dada

- Pot ser una obra: creació original
- Pot ser una prestació: enregistraments, meres fotografies
- Pot no ser protegible: fets, xifres

Diferents nivells de protecció:

- drets d'autor
- drets connexos o afins
- domini públic

La propietat intel·lectual de les dades

La base de dades o col·lecció

- Pot ser una obra: la selecció o disposició dels continguts constitueixen una creació intel·lectual original. Se'n protegeix l'estructura de la base de dades, no els continguts
- Pot ser una prestació: se'n protegeix la inversió substancial, avaluada qualitativament o quantitativa, que realitza el fabricant tan de mitjans financers, temps dedicat, esforç, energia o altres, per a l'obtenció, la verificació o la presentació del contingut

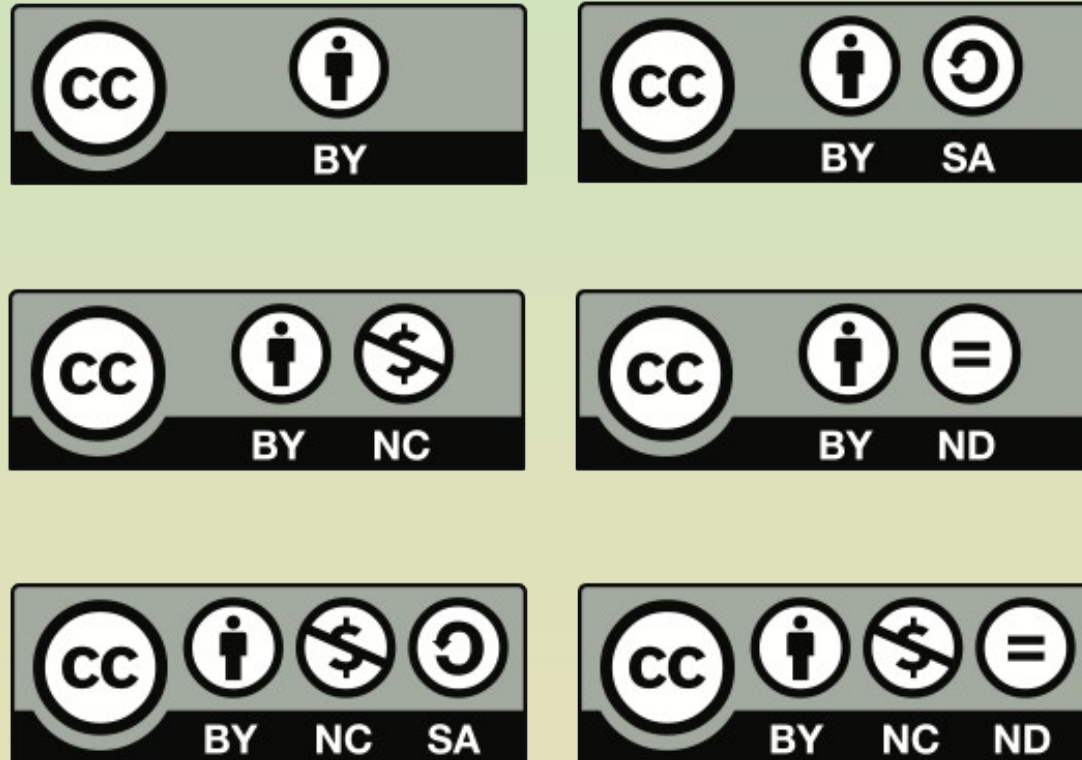
La propietat intel·lectual de les dades

Nivells de protecció de la base de dades o col·lecció

- Drets d'autor com a qualsevol obra
- Dret *sui generis*

L'usuari legítim només pot fer una extracció i/o reutilització d'una part no substancial de la base de dades sense autorització.

Llicències de Creative Commons



Llicències de Creative Commons

Fins a la versió 4.0 no estava clar com afectava a les bases de dades

S'hi incloïa un esment a obra i prestació però no hi havia claredat per al dret sui generis

Versió 4.0 s'especifica com afecten les condicions a les bases de dades

«Section 4 – Sui Generis Database Rights.»

Llicències de Creative Commons

Quan hi ha drets sui generis:

- Es permet extraure, reutilitzar i compartir tots els continguts o una part substancial de la base de dades (amb finalitat comercial o no, depèn de la llicència)
- Si aquests continguts s'inclouen en una base de dades de la qual s'ostenten drets sui generis, serà considerada una obra derivada
- Quan es comparteixin els continguts caldrà complir amb les condicions de la llicència

Una altra eina de Creative Commons

Eina legal per renunciar a tots els drets de propietat intel·lectual que permet la llei vigent aplicable

Si la renúncia no és prou efectiva, l'eina actua com a llicència per permetre qualsevol ús



Altres eines legals per a bases de dades

Open Data Commons

- Public Domain Dedication and License (PDDL)
- Attribution License (ODC-By)
- Open Database License (ODC-OdbL)

UK Open Government Licence

Licence Ouverte/Open Licence

Llicències Open Data Commons

Aplicables a dades i bases de dades

Public Domain Dedication and License (PDDL)

- Funciona com CC0: renúncia i llicència, si cal

Attribution License (ODC-By)

- Autorització sense restriccions sempre que se'n reconegui l'autoria

Open Database License (ODC-OdbL)

- Cal reconèixer-ne l'autoria i mantenir les mateixes condicions en una base de dades «derivada»

Requeriment de Llicències

...as far as possible, projects must then take measures to enable for third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data. One straightforward and effective way of doing this is to attach Creative Commons Licence (CC-BY or CC0 tool) to the data deposited

Pilot on Open Research Data in Horizon 2020

Data should be covered by a CC BY license or a less restrictive license

PLoS Open Data Policy

Polítiques

Institucionals: Regne Unit, Països Baixos, Alemanya...

Finançadors: Comissió Europea

Editorials: PLoS

Polítiques editorials: PLoS

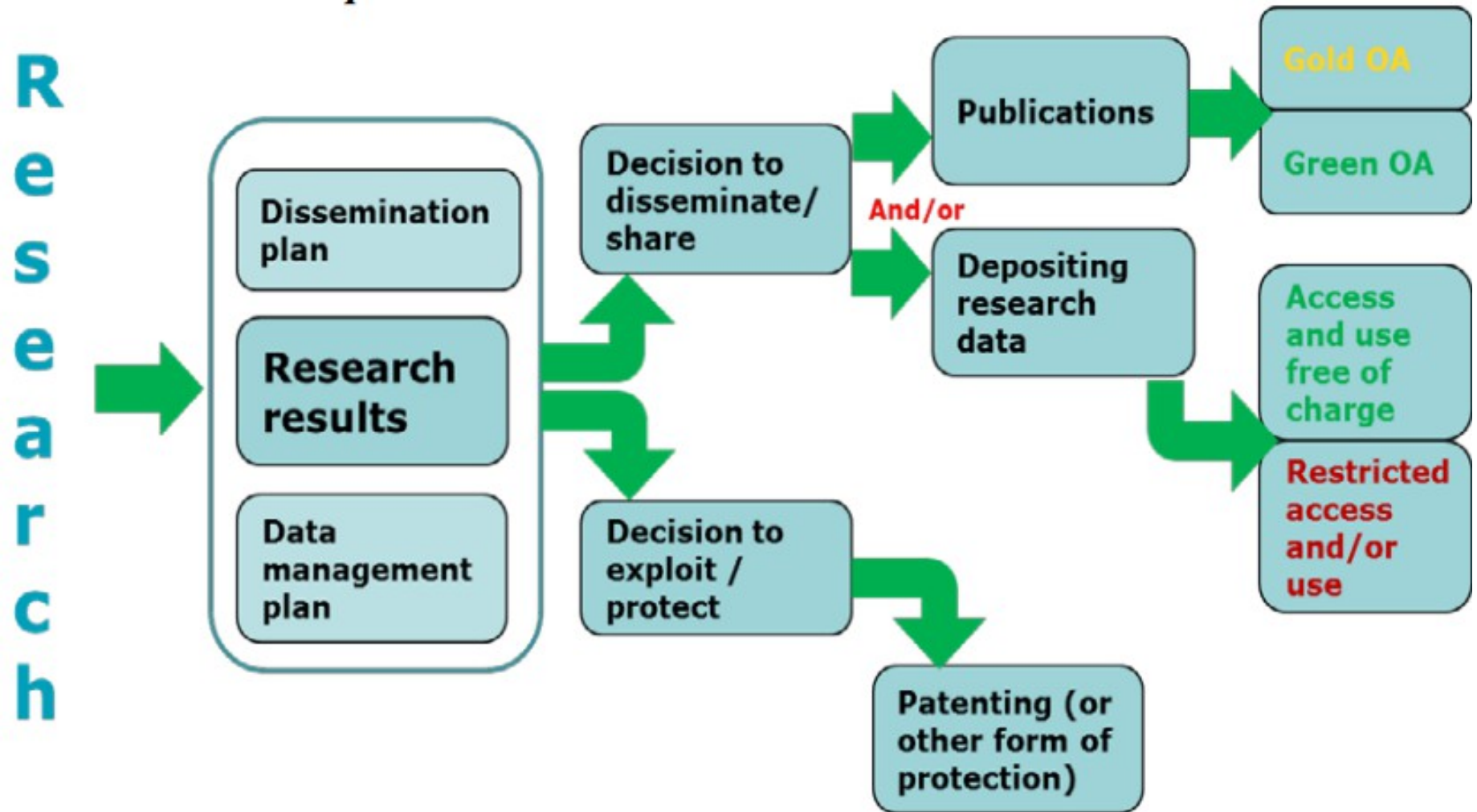
PLoS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.

All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository, unless already provided as part of the submitted article.

If authors use repositories with stated licensing policies the policies should not be more restrictive than the Creative Commons Attribution (CC BY) license.

<http://journals.plos.org/plosone/s/data-availability>

Polítiques finançadors: Horitzó 2020



Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Polítiques finançadors: Horitzó 2020

Fins 2017, pilot restringit a:

- Future and Emerging Technologies
- Research infrastructures – part e-Infrastructures
- Leadership in enabling and industrial technologies – Information and Communication Technologies
- Societal Challenge: 'Secure, Clean and Efficient Energy' – part Smart cities and communities
- Societal Challenge: 'Climate Action, Environment, Resource Efficiency and Raw materials' – except raw materials
- Societal Challenge: 'Europe in a changing world – inclusive, innovative and reflective Societies'
- Science with and for Society

Polítiques finançadors: Horitzó 2020

Excepcions:

- Incompatibilitat amb l'obligació de protegir els resultats si poden ser explotats comercialment o industrialment
- Incompatibilitat amb temes de confidencialitat o de seguretat
- Incompatibilitat amb regulacions sobre dades personals
- Pot posar en perill el principal objectiu del projecte
- No es generaran o recolliran dades
- Si hi ha altres raons legítimes

Polítiques finançadors: Horitzó 2020

- Dades i metadades associades necessàries per validar els resultats publicats
- Altres dades i metadades associades especificades en el pla de gestió de dades

Pla de gestió de dades:

- Lliurament en els sis primers mesos

Mandat en dos passos: dipositar les dades en un repositori i facilitar-ne la reutilització mitjançant llicències (CC BY/CC0)

Costos elegibles per a la implementació del pilot

Polítiques institucionals

Accés Obert per defecte: requeriment estàndard. Ha de ser d'obligat compliment amb possibilitat de tancar dades en casos en què es puguin demostrar temes ètics, legals, de confidencialitat, de seguretat

Responsabilitats: definició de la responsabilitat de tots els agents implicats dins la institució: els òrgans de govern, els investigadors i tot el personal de suport a la gestió i publicació de les dades de recerca.

Lloc de dipòsit: determinació del lloc de dipòsit de les dades, podent ser el repositori institucional o altres repositoris de confiança.

Termini de dipòsit: publicació en obert tan aviat com sigui possible. La política contemplarà períodes de temps raonables entre l'obtenció de les dades i la seva publicació en obert.

Pla de gestió de dades: elaboració d'un pla de gestió de dades com a requeriment estàndard, en les activitats de recerca de la institució. Aquest requeriment es podrà implementar de manera progressiva.

Polítiques institucionals

Costos: establiment de mecanismes oportuns per repercutir els costos de la infraestructura de la gestió de les dades. Per exemple, garantint un espai estàndard sense cost a tots els investigadors i establint tarifes per a requeriments específics.

Preservació i conservació: compromís de la institució per preservar i conservar les dades dipositades durant un període de temps determinat (10-20 anys, per exemple), establint mecanismes per garantir-ne l'accés i la seva integritat.

Acompanyament i seguiment: compromís de la institució de dotar d'un servei d'assessorament i suport als investigadors en la gestió de dades de recerca i l'establiment d'un pla de formació.

La política inclourà mecanismes de seguiment i d'avaluació de la seva eficàcia i possibles incentius per als investigadors.

Reutilització de les dades

Dependrà de les condicions d'ús

Caldrà citar-ne la font:

- Títol
- Autors
- Editor/Repositori
- Data
- Versió
- ID

Citació de dades

Deschenes, Elizabeth Piper, Susan Turner, and Joan Petersilia. Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release [Computer file]. ICPSR06849-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000. doi:10.3886/ICPSR06849

*Yannic G, Pellissier L, Dubey S, Vega R, Basset P, Mazzotti S, Pecchioli E, Vernesi C, Hauffe HC, Searle JB, Hausser J (2012) Data from: Multiple refugia and barriers explain the phylogeography of the Valais shrew, *Sorex antinorii* (Mammalia: Soricomorpha). Dryad Digital Repository. <http://dx.doi.org/10.5061/dryad.2jj36325>*

Rodriguez, Tommy (2013): 17,170 Base Pair Alignment of Thirteen Time-Extended Lineages [data: (complete) mtDNA; format: ClustalW]. figshare. <https://dx.doi.org/10.6084/m9.figshare.815894> Retrieved: 16 26, Jan 04, 2016 (GMT)

Frederico Girosi; Gary King, 2006, 'Cause of Death Data', <http://hdl.handle.net/1902.1/UOVMCPSWOL> UNF:3:9JU+SmVyHgwRhAKclQ85Cg== IQSS Dataverse Network [Distributor] V3 [Version].

Cool, H. E. M. and Bell, M. (2011), Excavations at St Peter's Church, Barton-upon-Humber [dataset] (York: Archaeology Data Service), doi: 10.5284/1000389

Servei de suport a la gestió

Transversal: «Starting the Conversation: University-wide Research Data Management Policy»

<http://www.oclc.org/research/publications/library/2013/2013-08r.html>

Un exemple, Tampere University of Technology

<http://scienceport.tut.fi/c.php?g=216291&p=1427737>

Debat

ilabastida@ub.edu

@ignasi

Oficina de Difusió del Coneixement

CRAI Universitat de Barcelona

<http://www.ub.edu/odc>