## Dataset Management as a Special Collection

### Introduction

University libraries face numerous issues regarding access to special collections, including datasets. Dataset licensing varies considerably, such as in terms of copyright or creative commons. These are issues that must be resolved in the coming years. Indeed, at present, datasets are scattered across different places and, depending on the topic in question, licences and retrieval can be difficult. Moreover, given challenges of access and manipulation, the inclusion of datasets in special collections can become a daunting task. Datasets are included in special collections if they are specialised, pertain to the collection's content, have research values or are old legacy formats that add value to the university library. Finally, they offer the possibility of generating new data if reused with adequate technology.

Special collections are characterised by their uniqueness, fragility, value, rarity and the difficulty of finding similar collections elsewhere. Typically, special collections were restricted to small portions of population, separated from the main library collection as a whole, and used for research or instruction (Tam 2017). Since the creation of the World Wide Web, libraries have used outreach projects to disseminate their special collections through their catalogues.

A virtual collection can offer numerous benefits to researchers, solving problems such as accessibility or findability, and providing innovative means of researching, teaching and learning, according to the Special Collections Working Group established by Association of Research Libraries (ARL) (Prochaska 2009). Virtual catalogues have removed geographical distance barriers and opening hours, and have expanded research and access to rare materials (Doi 2015), as well as presenting hidden material that no longer exists in the library catalogue (Tam 2017). Such materials have become digitised, with hard copies available in the library in case of an information technology (IT) failure. Nevertheless, digital special collections have always raised significant questions, affected by politics, legal requirements and technological solutions (Prochaska 2009).

A specific question arises when materials are not digitised but are instead digitally born and are to be included in special collections. In 2011, Goldman found that institutions had not created policies to protect digital materials and their properties (Goldman 2011). Garnett et al. also found that university libraries' special collection protection against disasters is under-researched (Garnett et al. 2018). In most cases, university libraries require IT support; special collections that are made available online must be treated carefully in order that the information presented is complete. Therefore, the very definition of a special digital collection becomes undefined, owing to the nature of digital material.

The definition of a special collection with born-digital material differs from a collection of analogue materials. In fact, Prochaska proposed how a born-digital special collection should be defined (Prochaska 2009). An approach to this definition is that a special collection referring to digitally born material is a collection of digital records. These records are digitally preserved – a primary source with research value

that may contain digital standalone material, collections of digital objects or datasets. Materials that belong to born-digital special collections can be computerised when adequate technology exists.

Significant improvements can be identified in the dissemination of digital special collections. Nevertheless, some elements, such as rare books, continue to require improvements, such as the importance of being fully searchable with a complete framework description using semantic technology. Moreover, the online dissemination of special collections does not necessarily mean that a special collection provides free access. Some institutions make profits because the digitising process is expensive (Tam 2017). This may be a barrier to researchers, who must instead visit the library to examine the non-digital material, increasing research costs as a result.

Although a research paper or journal may be easy to access and index through an online database, this is not necessarily the case when indexing datasets. In part, this is because datasets can be divided into different collections. Another factor is that some datasets may be indexed online but are not made visible due to permission licences or historical interest. This is especially true of special collections that are used to teach science, history, mathematics and other fields.

Libraries have potential means of offering datasets as special collections, especially when intending to do so for free. This paper presents a review of the literature that discusses datasets in special collections.

**Dataset locations and format representations**

When referring to digitally born material, a dataset is a set of data, represented in any digital format, that together have a meaning. These sets of data are available for computer processing in one or more digital formats. Datasets are not just limited to being a matrix of text or numbers; they can also contain collections of sounds, images or videos, combining different forms of information. Datasets have a data owner and must also have accompanying documentation which explains their use and processing systems. Datasets can be placed into a repository, the supplemental material of an article or any part of the Internet.

On university campuses, dataset location causes an issue for researchers, regardless of whether the data are digital or analogue (Farrell and Kelly 2018). Datasets are usually scattered around departments and need to be unearthed. Although a minority of scientists are aware of the importance of keeping dataset records for the long term, university policies should be enhanced not only to ensure that these records can be easily indexed and retrieved from library catalogues, but also so that research can be sustainable. Libraries can provide support in retaining research datasets as long as their staff are trained and in collaboration with IT support (Tenopir et al. 2017).

A dataset's format and representation would also indicate to a university library the type of dataset that can be treated and preserved, not only in terms of acquisition but also in the future. It must additionally be considered that datasets can exist in any format such as text, image, sound or video. The

format must be aligned with the digital preservation plan of the institution. Universities and researchers also create new types of datasets to be included in a special collection at a later date, and so they must adhere to the sustainability criteria (Library of Congress 2017). Therefore, datasets included in special digital collections should be readable by a computerised machine.

**The inclusion of datasets in a special collection**

The definition of a special collection is open to discussion. Materials vary from library to library and their value is not always recognized by budget-conscious academic administrators in some institutions (Hewitt and Panitch 2003). In 2013, a study found that there was a lack of consensus and precision concerning the definition of special collections (Dupont and Yakel 2013). According to this study, the key seems to be to define user-centric metrics and techniques. The dilemma also derives from the definition of what makes a collection special. Traditionally, rarity, location, and fragility, among other qualities, were reasons to include an object in a special collection. However, digitization and further online access have improved collection visibility (Cusworth et al. 2015). With regard to digital datasets, the challenge is determining the added value of the dataset to be included in a special collection.

Once it has been decided to add a dataset to a special collection, another difficult challenge is the catalogue description, which will be found in a search. Librarians must compile digitised collections using a description of the various elements. The use of rich, adequate and consistent metadata with electronic collections is crucial to facilitate access to the dataset (Prochaska 2009). In the case of datasets that are going to be added to a special collection, it can become increasingly complex because datasets need a very detailed description. This detailed description would permit knowing how reuse it or interpret it. In the near future, there will likely be a need to create or enhance the current metadata schemes, thereby providing semantic richness to description facilities. Currently, there are initiatives such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and linked open data (LOD). In the case of OAI-PMH, this type of metadata permits maintaining the context as much as possible when metadata are shared with others (Han et al. 2009); this also supports interoperability by combining them with other standard metadata such as encoded archival description (EAD). LOD provides links to outside resources and maps metadata to linked data with friendly vocabulary such as the RDF vocabulary (Disambiguating descriptions 2016). Consequently, this permit combining distinct datasets together, giving the user more visibility into the collections.

Another challenge involves using assistive technologies to access datasets, which provide accessibility to all patrons. In 2012, it was found that 58% of special collections in 69 academic libraries were not screen-readable based on digitised textual documents (Southwell and Slater 2012). In a later 2016 study, it was found that assistive technology was still uncommon in rare books and manuscripts (Hardesty 2016). Textual documents have the possibility of being accessed via screen-readers. Instead,

datasets must be identified and later processed; hence, assistive technology must be enhanced to carry out these tasks.

**Datasets created in-house by the library**

Datasets can be generated by any means. Digitisation has been a solution over the past few years both to giving online visibility to special collections and also for digital preservation (Meger and Draper 2012). Nevertheless, although it is time-consuming and expensive, a physical copy always remains in the library to ensure access to materials in case digital damage occurs (Rink 2017). Hence, special collections are now findable online thanks to these technological processes. Because datasets can be part of a digitised collection, they can be created from both born-digital and digitised materials. Therefore, it is possible collect information for datasets from paper surveys, or even a self-administered method (Graves, Ball and Fraser 2007) or any other technique, such as mapping technologies to generate georeferenced datasets for geographic information systems (GIS) (Elliott 2014). A dataset composed of digital material must be interpreted and further processed by a software package, not only at the time of acquisition, but also in the long term. A library that decides to generate or create datasets needs to make many decisions concerning not only access, but also reading machine requirements, data processing, online access, and copyright statements.

**Acquisition of datasets for special collections**

In order to acquire research data, training and guidance could permit the establishment of deeper connections with libraries and research communities (Palumbo et al. 2015). University libraries will have to specialise in storing or indexing certain types of datasets. It will not be economically sustainable to index all generated datasets. This could be the differentiation between libraries regarding special collections. In this way, libraries will be specialised at attaining certain types of datasets so that they can later be included within a special collection. Therefore, partnering with other institutions may constitute a potential solution to ensuring that a special collection's catalogue includes a greater number of indexed datasets.

Finding and then including datasets as part of special collections can represent a daunting task, because it is necessary to avoid duplicate collections and redundant datasets. In order to locate and acquire datasets, university libraries will need to use crawlers such as Heritrix to locate data in digital preservation. However, certain questions should be considered when potential material is identified (Brunelle et al. 2016). A crawler should be selected based on the specific topic at hand. Rather than downloading, a summary of the data found should be sent and manual intervention should be used to guarantee quality and appropriateness in order to be part of a special collection.

**Security concerns when acquiring Datasets**

Another important issue in the future will be security. The guide to security theft in special collections must be updated to ensure its relevance in the new digital landscape (RBMS Security Committee 2009). In the literature vandalism, theft of special collections is a common term referring to stolen or damaged rare books, manuscripts or even DVDs -- often attracted by online catalogue descriptions (Higgins 2015),. In the digital landscape, security will become a more important issue because hacking or modifying datasets may threaten data accuracy, data reliability and their ability to be reused. Indeed, research experiments may become unrepeatable and confidence in the institution in question would be undermined. For instance, any modification to a dataset of the frequency sounds of whales could give a completely different classification of whales or leave the dataset completely unusable (Shamir et al. 2014). Creating policies and databases in the digital landscape for events similar to robberies would provide a clearer understanding of the most sensitive forms of datasets and their degree of vulnerability to digital attacks (Samuelson, Sare, and Coker 2012).

Also related to security, differentiating disaster and digital disaster have been discussed (Rachman and Afidhan 2018). A disaster affects access to collections or services, while a digital disaster means loss of and/or damage to data and digital collections, which directly impacts the continuous information access provided by the library. The same study also found that Indonesian libraries should first plan digital preservation strategies because the protection costs for digital collections are minimal compared to costs for the printed collections.

An important issue deserving consideration is the assurance that you are accessing accurate datasets as compared to the possibility that datasets may have been technologically altered. One possible solution is that libraries consider using digital rights management systems (DRM) as well as watermarking in the case of geographic information system (GIS) datasets in order to identify datasets. However, other authors argue that users prefer free data licensing (López 2003). In terms of managing DRM, the question of long-term access to the dataset is also critical. As discussed by Kaur et al. (2003), DRM can guarantee access by distributing DRM elements in different parts of the Internet. In this way, the authenticity of datasets, especially those that are protected, is guaranteed. Given the existence of false datasets and the use of rights management, other rights are also relevant.

**Dataset curation**

Challenges will also materialise in curating datasets. Given that datasets will represent virtual collections with an enormous quantity of data, libraries will need to possess a clear framework policy and with an underlying technological structure that provides access to their special collections. Moreover, the curation process would differ, depending on how the dataset was acquired. A dataset acquired through a commercial or partnering process would depend on the source, which would need to be reviewed and audited according to the library's curation procedures. Additional economic and human effort would also be required to review the datasets according to the library's definition of its framework policy, as well as to ensure the quality of both the data and the source. As explained earlier, acquiring dataset descriptions or

complete collections will be seen as a technological issue. However, once the collection has been acquired, digital data curation will constitute the second step. Manual intervention will be required because digital datasets must be treated as part of a special collection. Software such as automatic labelling tools may also be used to curate hierarchical datasets (Liu et al. 2016). The creation of a standardised digital stamp or digital seal on the acquired source may be necessary to validate the source's quality and reliability in order to avoid a review of the curation process.

A dataset may be donated and hence constitute an in-house dataset that may not have been curated. This dataset would require a curation process to establish its appropriateness to be included as part of the special collection. In this phase, it is likely that skills in data science might be required to be included in the librarian or archivist training. Depending on the discipline or field, the curation concept may be difficult to implement. For example, earth science datasets would not only require a curation process, but the testing of the source's reliability and trustworthiness (Bugbee et al. 2018).

Finally, curated, online special collections of datasets would provide several advantages to users, including reusability, source authoring, reliability and the creation of new data. A question deserving of consideration is how the collection would be used, such as in terms of storing and promoting art objects or other physical items, as well as technological security issues, accessibility and storage in digital silos (Litchfield and Gilson 2013). Curators' skills will also need to evolve. Given that librarian training is not the same across countries, growing numbers of freelance professionals are assessing or cataloguing for special collections (Dupuigrenet 2004). Such a professional figure is likely to become more commonplace in most institutions. In the field of biology, distinguishing between expert curators, self-curators, community curators and automated curators has been proposed (Goble et al. 2008). Therefore, the training of curators is likely to represent an important issue to consider in a range of scientific fields in which special knowledge is necessary.

A 2009 study found that data curators would be expected to be responsible for metadata design to ensure the dataset's findability. Curators should also be working closer and more collaboratively with other institutions to ensure interoperability where datasets are shared by several institutions and the dataset is located in just one central location (Taranto 2009). Curators would also be responsible for defining the terms under which a dataset would be made visible online and how it would be reproduced or used. Some dataset collections will probably be more susceptible to digital injury and thus require greater care. The definition of digital injury would be a responsibility of the curator. Through the use of metrics, the curator may decide to select either similar or alternative dataset collections in order to enhance user engagement.

**Dataset auditing for quality assurance**

Given that curators would be responsible for questions regarding the quality of the dataset, auditing will be necessary not only to ensure standards but also to ascertain utility. In some cases, libraries would combine datasets with other institutions, and although some incompatibilities through interchange may arise, these should be rectifiable through quality assessment or data auditing. Datasets will most likely be subsumed as part of a large repository and, thus, continuous auditing processes will be necessary to guarantee quality, accessibility and the standard of authoring. This, in turn, will enable other individuals to use the collection for purposes other than those originally intended (Bugbee et al. 2018). In fields such as bioinformatics, data auditing is a requirement owing to a lack of accurate curation and data reliability, because curation is understood to be a process of cataloguing rather than representing a delivery (Goble et al. 2008).

Libraries with datasets such as a special collection will need to undertake digital preservation strategies to maintain their datasets. Even where these datasets are shared as part of a centralised repository that is maintained by several institutions, digital preservation would be necessary to ensure accessibility. Therefore, standardised dataset auditing would be required. A starting point may be the achievement of Trustworthy Repositories Audit & Certification, which is currently a standard (ISO 16363:2012) (The Center for Research Libraries, Online Computer Library Center, Inc. 2007), although this may need to be reviewed and amended in the future. In particular, the auditing strategy would need to ensure that the dataset does not exist under a proprietary format (Robertson and Borchert 2014). This would not only permit access to the data, but also the use of any proprietary or open-source software.

**Accessing distributed datasets from the special collection catalogue**

In most cases, libraries will have to provide access to other institutions' special collection materials within their catalogues, due to factors such as lacking sufficient storage in the repository. This situation affects libraries, because they have to index, tag and offer a degree of access, depending on whether the dataset is under copyright protection. In order to provide access to datasets and include them as part of a special collection following the curating process, technology must be used. If libraries are specialised in including certain types of datasets, they will need to use crawlers to index them. The use of semantic searchers as well as crawlers in-house is typically necessary. Crawled datasets must be accessible through a curation process that is not directly included in the special collection. In particular, rich metadata descriptions are required to render datasets findable.

Google has recently built a means of searching scientific datasets (toolbox.google.com/datasearch), but it is not yet clear whether all datasets will be included. Similar tools will be necessary for libraries to distribute and share access to datasets with other institutions and partners, rather than making them all publicly available. The question may then be which kinds of datasets should be included as part of the collection and which should merely be indexed. Given that some datasets are distributed, a further question pertains to how dataset collections can be linked in

order to avoid issues of scattered or separate (yet similar) collections as music datasets (Raimond, Sutton, and Sandler 2008). It is also possible that dataset use in the future will be different from its use today. In order to facilitate the findings of datasets and to avoid issues of discoverability, it is necessary to utilise persistent identifiers (Woolcott, Payant, and Skindelien 2016).

**Dataset rights management and research uses**

Special collections are subject to copyright laws. However, special collection projects can be under the public domain once the copyright times out or the donor provides documents under a public domain licence. There are also exceptions to copyright infringement, including the fair use doctrine that is applied in contexts such as teaching scholarships or research (Buttler 2012). This means that materials from special collections such as rare books, manuscripts, or personal archives can be used as an alternative learning style in higher education, improving the pedagogical experience if an archivist and faculty member work together (Torre 2008). Hence, there is an opportunity to improve student learning through original primary sources (Horowitz 2015). This can also be applied to datasets, because libraries may have datasets generated in-house and other datasets that are acquired and managed in accordance with copyright or licencing agreements. From the researcher's point of view, a basic question will be to determine what type of access would permit access to the datasets. Special collections stimulate active learning. Having special collections of visual material is relevant to student training in collections related to science, technology, engineering, and mathematics (STEM) (Brown, Losoff, and Hollis 2014). Thus, it is an essential rights managements practice to provide access to special collections.

A study found that one top area regarding scientific dataset collections was digital management, and research data management librarians can provide good data practices and assist with digital rights management (Henderson and Knott 2015). Then, a university library would need to manage access to all kinds of datasets to be considered a primary information source and useful for researchers. For example, in a collection of audio records we can distinguish audio without music, a conference, a lecture, or audio with music. Datasets that are related to pieces of music and their relevant rights management schemes are complex because a piece of music can be represented by several parts (Reyes 2016). Hence, it is possible to find different rights holders inside a musical piece and it will be necessary to manage different types of access related to copyright issues.

Other questions include how audio can be reproduced and which player or players can guarantee not only playing but also distribution. In some cases, music datasets can be fixed to a precise player chosen by the copyright holder. The use of third-party software raises the acquisition cost, the digital preservation cost, and would also run into obsolescence problems in the long term. This situation needs to be resolved in the future. In the case of a copyright holder's disappearance or technological obsolescence, a dataset of this kind would not be reproduced anymore and consequently lose its value and data. It would be essential to organise datasets using data linked to semantic search in order to have a complete music dataset grouping data in only one distribution element.

In terms of video, visual elements of videos such as the format, video access, or format obsolescence are more complex. It is unclear what would happen if a dataset could permit access to videos in an obsolete format that cannot be reproduced. In any case, researchers have proposed a taxonomy categorising types of videos based on human action and activity recognition to video datasets (Chaquet, Carmona, and Fernández-Caballero 2013).

**Interoperability concerns and findability**

In addition to rights management, libraries with datasets will have to deal with issues such as interoperability or election adequacy for the end user, as well as findability. Interoperability would facilitate access to all kinds of formats, and small institutions may require their own dataset repositories (Schwartz et al. 2007).

Libraries will be able to generate database indexes that provide access to different datasets on various online sites under open access licensing. In most cases, datasets will need to be joined together in order to be useful. Thus, any dataset source must be flexible and capable of being integrated with other sources using appropriate metadata standards. One such standard is CERIF, a European standard for data formats and research information supported by the European institution euroCRIS (Biesenbender and Hornbostel 2016). University libraries that manage datasets should be able to manage standardised and non-standardised formats with different sources of data. One solution would be to create a worldwide standard to ensure interoperability, integration and digital preservation. Nevertheless, each scientific field tends to have different formats and standards.

**Digital preservation of datasets in a special collection**

Preservation of special collections has been covered in the literature for many years. A broad definition of the classic preservation concept is keeping things unchanged. However, this concept is not possible with digital preservation because technology and support evolve rapidly. In any archive where digital preservation is held, archivists need archival stability. This means that in any new project involving sorting IT into archives, archivists should participate in the project as embedded archivists (Chen 2007). This would permit to the IT teams to benefit from archivist experience, but also information flow.

University libraries can adapt existing digital preservation strategies to their special collections. Not all libraries have the same economic resources to apply to a digital preservation strategy. The application of a digital preservation plan would depend on librarian training and resources. There are several models of preserving collections recorded in the existing literature; for example, Lots of Copies Keep Stuff Safe (LOCKSS) and Open Archival Information System (OAIS). LOCKSS is a system that digitally preserves scientific journals (Reich and Rosenthal 2001). OAIS is a framework initially designed

to keep data space and was later used by libraries (Consultative Committee for Space Data Systems 2002). OAIS is probably the most universal model because it can be adapted to be used in different approaches such as archival sounds (Rodríguez 2016) or as a framework for data science management (Flathers, Kenyon, and Gessler 2017). However, the OAIS model is not an application model and questions about it remain unresolved (Cruz and Díez 2016). Further enhancements are needed in order to apply the model, because for some IT architectures it is difficult to find documentation about file formats (McKinney et al. 2014). However, this is a model that has been widely adopted by the archival community and developed by software companies.

Problems related to long-term digital preservation are diverse because the main goal is to preserve the bitstream (Rothenberg 1995). Information, physical support, and technology are the three cornerstones of the digital preservation field. To maintain some long-term digital objects, physical support can be avoided and technology can replace physical support; but for other items this is not the case. It remains unavoidable that information must be digitally preserved and this needs to be maintained in the long term. This means that information must be treated through different processes to ensure its access in the primary source. It is also necessary to include adequate metadata in its description in order to know both its meaning and its use. This issue does not arise with physical rare books or manuscripts, which can be observed and described in a catalogue that does not need to be extensive (Howell 2000).

Digital preservation has a broader meaning than these three cornerstones. It implies keeping raw material, its software and its physical medium, if necessary (Burrows 2000). This means that cost models and solutions are complex because there are no unique solutions in cost modelling for digital preservation (Bote, Fernandez-Feijoo, and Ruiz 2012). Datasets are groups of data that, when taken all together, have meaning. This implies that digital preservation strategies applied to datasets need to be carried out as a set instead of individual elements of a collection. Datasets have many different formats and it is possible that without adequate digital preservation strategies in the long term some of them will not be computer-processed. Issues such as format obsolescence or machine readability could impede processing or computer reading. Possible solutions in a digital preservation plan could be migration, which is one the common strategies in digital preservation. If datasets for any reason, as mentioned in a former section, need third-party software, it would be necessary to take into account acquiring new, updated software.

Geographical Information Systems (GIS) are mixed data with several formats that are continuously growing; there is a high risk of losing all this information. This means that this kind of data and geospatial data are not simple, textual data. Also, their storage models are inadequate for long-term preservation. These are very complex data delivered in dataset forms and their digital preservation needs should be planned when creating any research project (Clark 2016). Something similar happens with music.

Because an instrument can be kept for many years, electronically generated music has diverse supports that need to be maintained for content later. In most cases, the extraction of music from its

original format, known as migration, can result in losing some significant properties (Recker and Müller 2015). For instance, if the files of a dataset of bird sounds in tropical forest are altered (Ulloa et al. 2016), this means changes in significant properties. If this happens, consequently, the study could not be reused, unless the library had a robust digital preservation strategy that preserved the original datasets. This is a matter of special importance in keeping not only with this kind of dataset, but also local music history, where minorities can be affected, thereby losing an aspect of cultural heritage.

Other digital objects such as personal archives, digital literature, and institutional records exist only in physical forms. This means that in the future libraries should require policy plans, continuously train librarians who understand the urgency in keeping digital information in the long term, and have resources that are both technical and economically sustainable (Fisher 2017).

It is not possible to create a digital preservation strategy for a special collection without a strategic plan. This depends on whether they are born-digital or digitised materials, because datasets can be both types of digital objects. Procedures such as continuous auditing processes that permit deciding which steps to follow with regard to information should not be avoided. It is likely that in the future migration could be an option for datasets, but if digital preservation is planned in its initial phase, migrating support or technology could cost less than it does currently.

To preserve special collections, university libraries will have to account for many factors. If datasets are considered to be a set of data in any format, it is possible to classify them into text datasets, audio datasets, images datasets, video datasets, or mixed datasets. Therefore, datasets must be preserved as a set of digital objects, not as a simple element. Otherwise, libraries face the risk of misleading information or damaging the collection, which would render it unusable in the long term.

**Conclusion**

Libraries face numerous challenges when including datasets in special collections, including (but by no means limited to) information management and technological architecture. Regardless of whether the datasets are publicly available online or not, the inclusion criteria will help determine the cost to the institution. The roles of librarians, curators and researchers will continue to evolve. New skills in information literacy in terms of searching datasets, data auditing and rights management will also be developed as required. Where datasets are included in special collections that exist in partnership with other institutions, interoperability will become essential to ensuring access and reusability. Finally, a digital preservation plan will be necessary not only to ensure long-term access and usage, but also to guarantee quality, reliability, trustworthiness and the sustainability of the library as a whole.

**References**

Biesenbender, Sophie, and Stefan Hornbostel. 2016. 'The Research Core Dataset for the German Science System: Developing Standards for an Integrated Management of Research Information'. *Scientometrics* 108 (1): 401–12. https://doi.org/10.1007/s11192-016-1909-2.

Bote, Juanjo, Belen Fernandez-Feijoo, and Silvia Ruiz. 2012. 'The Cost of Digital Preservation: A Methodological Analysis'. *Procedia Technology*, 4th Conference of ENTERprise Information Systems – aligning technology, organizations and people (CENTERIS 2012), 5 (January): 103–11. https://doi.org/10.1016/j.protcy.2012.09.012.

Brown, Amanda, Losoff, Barbara and Hollis, Deborah. 2014. 'Science Instruction Through the Visual Arts in Special Collections'. *Portal: Libraries and the Academy* 14 (2): 197–216. https://doi.org/10.1353/pla.2014.0002.

Brunelle, Justin, Ferrante, Krista, Wilczek, Eliot, Weigle, Michel and Nelson, Michael. 2016. 'Leveraging Heritrix and the Wayback Machine on a Corporate Intranet: A Case Study on Improving Corporate Archives'. *D-Lib Magazine* 22 (1/2). https://doi.org/10.1045/january2016-brunelle.

Bugbee, Kaylin, Ramachandran, Rahul, Maskey, Manil and Gatlin, Patrick. 2018. 'The Art and Science of Data Curation: Lessons Learned from Constructing a Virtual Collection'. *Computers & Geosciences* 112 (March): 76–82. https://doi.org/10.1016/j.cageo.2017.11.021.

Burrows, Toby. 2000. 'Preserving the Past, Conceptualising the Future: Research Libraries and Digital Preservation'. *Australian Academic & Research Libraries* 31 (4): 142–53. https://doi.org/10.1080/00048623.2000.10755131.

Buttler, Dwayne. 2012. 'Intimacy Gone Awry: Copyright and Special Collections'. *Journal of Library Administration* 52 (3–4): 279–29. https://doi.org/10.1080/01930826.2012.684506.

Chaquet, Jose, Carmona, Enrique and Fernández-Caballero, Antonio. 2013. 'A Survey of Video Datasets for Human Action and Activity Recognition'. *Computer Vision and Image Understanding.* 117 (6): 633–659. https://doi.org/10.1016/j.cviu.2013.01.013.

Chen, Su-Shing. 2007. 'Digital Preservation: Organizational Commitment, Archival Stability, and Technological Continuity'. *Journal of Organizational Computing and Electronic Commerce* 17 (3): 205–15. https://doi.org/10.1080/10919390701294012.

Clark, John. 2016. 'The Long-Term Preservation of Digital Historical Geospatial Data: A Review of Issues and Methods'. *Journal of Map & Geography Libraries* 12 (2): 187–201. https://doi.org/10.1080/15420353.2016.1185497.

Consultative Committee for Space Data Systems. 2002. 'Reference Model for an Open Archival Information System'. https://public.ccsds.org/publications/archive/650x0b1.pdf.

Cruz, José Ramón and Díez, Carmen. 2016. 'Open Archival Information System (OAIS): Lights and Shadows of a Reference Model'. *Investigación Bibliotecológica* 30 (70): 221–47. https://doi.org/10.1016/j.ibbai.2016.10.010.

Cusworth, Andrew, Hughes, Lorna, James, Rhian, Roberts, Owain and Lloyd, Gareth. 2015. 'What Makes the Digital "Special"? The Research Program in Digital Collections at the National Library of

Wales'. *New Review of Academic Librarianship* 21 (2): 241–48. https://doi.org/10.1080/13614533.2015.1034805.

Dupuigrenet, François. 2004. 'Enssib and the preservation of special collections in France'. *Conservation Science in Cultural Heritage* 4 (1): 209–14. https://doi.org/10.6092/issn.1973-9494/578.

Disambiguating descriptions. 2016. 'Disambiguating Descriptions: Mapping Digital Special Collections Metadata into Linked Open Data Formats - - 2016 - Proceedings of the Association for Information Science and Technology - Wiley Online Library'. *Proceedings of the Association for Information Science and Technology Banner*. https://doi.org/10.1002/pra2.2016.14505301096.

Doi, Carolyn. 2015. 'Local Music Collections: Strategies for Digital Access, Presentation, and Preservation—A Case Study'. *New Review of Academic Librarianship* 21 (2): 256–63. https://doi.org/10.1080/13614533.2015.1022663.

Dupont, Christian, and Elizabeth Yakel. 2013. '"What's So Special about Special Collections?" Or, Assessing the Value Special Collections Bring to Academic Libraries'. *Evidence Based Library and Information Practice* 8 (2): 9–21. https://doi.org/10.18438/B8690Q.

Elliott, Rory. 2014. 'Geographic Information Systems (GIS) and Libraries: Concepts, Services and Resources'. *Library Hi Tech News; Bradford* 31 (8): 8–11. http://dx.doi.org.sire.ub.edu/10.1108/LHTN-07-2014-0054.

Farrell, Shannon and Kelly, Julia. 2018. 'Identifying Potential Solutions to Increase Discoverability and Reuse of Analog Datasets in Various Campus Locations'. *Issues in Science and Technology Librarianship* 88. https://doi.org/10.5062/f4pc30nr.

Fisher, Katherine. 2017. 'Barriers to Digital Preservation in Special Collections Departments'. *Preservation, Digital Technology and Culture* 45 (4): 180–85. https://doi.org/10.1515/pdtc-2016-0027.

Flathers, Edward, Kenyon, Jeremy and Gessler, Paul. 2017. 'A Service-Based Framework for the OAIS Model for Earth Science Data Management'. *Earth Science Informatics* 10 (3): 383–393. https://doi.org/10.1007/s12145-017-0297-3.

Reyes, Artemisa. 2016. 'Los acervos de documentos musicales. ¿Libros raros, libros especiales?' *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 30 (70): 129–63. http://rev-ib.unam.mx/ib/index.php/ib/article/view/57609.

Garnett, Johanna, Arbon, Paul, Howard, David and Ingham, Valerie. 2018. 'Do University Libraries in Australia Actively Plan to Protect Special Collections from Disaster?' *Journal of the Australian Library and Information Association* 67 (4): 434–49. https://doi.org/10.1080/24750158.2018.1531678.

Goble, Carole, Robert Stevens, Duncan Hull, Katy Wolstencroft, and Rodrigo Lopez. 2008. 'Data Curation + Process Curation=data Integration + Science'. *Briefings in Bioinformatics* 9 (6): 506–17. https://doi.org/10.1093/bib/bbn034.

Goldman, Ben. 2011. 'Bridging the Gap: Taking Practical Steps Toward Managing Born-Digital

Collections in Manuscript Repositories'. *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 12 (1): 11–24. https://doi.org/10.5860/rbm.12.1.343.

Graves, Anna, Jean Ball, and Eliza Fraser. 2007. 'Data Management: The Building Blocks of Clean, Accurate and Reliable Longitudinal Datasets'. *International Journal of Multiple Research Approaches* 1 (2): 156–74. https://doi.org/10.5172/mra.455.1.2.156.

Han, Myung-Ja, Cho, Christine, Cole, Timothy and Jackson, Amy. 2009. 'Metadata for Special Collections in CONTENTdm: How to Improve Interoperability of Unique Fields Through OAI-PMH'. *Journal of Library Metadata* 9 (3–4): 213–38. https://www.tandfonline.com/doi/abs/10.1080/19386380903405124.

Hardesty, Emily. 2016. 'Accessibility and Special Collections Libraries: Using Technology to Close the Digital Divide'. *Public Services Quarterly* 12 (4): 329–33. https://doi.org/10.1080/15228959.2016.1222757.

Henderson, Margaret, and Knott, Teresa. 2015. 'Starting a Research Data Management Program Based in a University Library'. *Medical Reference Services Quarterly* 34 (1): 47–59. https://doi.org/10.1080/02763869.2015.986783.

Hewitt, Joe, and Panitch, Judith. 2003. 'The ARL Special Collections Initiative'. *Library Trends* 52 (1): 157–171.

Higgins, Silke. 2015. 'Theft and Vandalism of Books, Manuscripts, and Related Materials in Public and Academic Libraries, Archives, and Special Collections'. *Library Philosophy and Practice; Lincoln*, 0_1,1-23. http://search.proquest.com/lisa/docview/1739062808/abstract/BFA3CBFE0AD4423PQ/7.

Horowitz, Sarah. 2015. 'Hands-On Learning in Special Collections: A Pilot Assessment Project'. *Journal of Archival Organization* 12 (3–4): 216–29. https://doi.org/10.1080/15332748.2015.1118948.

Howell, Alan. 2000. 'Perfect One Day—Digital The Next: Challenges in Preserving Digital Information'. *Australian Academic & Research Libraries* 31 (4): 121–41. https://doi.org/10.1080/00048623.2000.10755130.

Kaur, Kirn, Hein, Stefan, Schrimpf, Sabien, Ras, Marcel and Holzmayer, Manuela. 2003. 'Report on DRM Preservation.' http://www.alliancepermanentaccess.org/wp-content/uploads/sites/7/downloads/2014/06/APARSEN-REP-D31_1-01-1_4_incURN.pdf.

Library of Congress. 2017. 'Sustainability Factors'. 2017. https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml.

Litchfield, Robert and Gilson, Lucy. 2013. 'Curating Collections of Ideas: Museum as Metaphor in the Management of Creativity'. *Industrial Marketing Management*, B2B Service Networks and Managing creativity in business market relationships, 42 (1): 106–12. https://doi.org/10.1016/j.indmarman.2012.11.010.

Liu, Ruoqian, Palsetia, Diana, Paul, Arindam, Al-Bahrani, Reda, Jha, Dipendra, Liao, Wei-keng, Agrawal, Ankit and Choudhary, Alok. 2016. 'PinterNet: A Thematic Label Curation Tool for Large Image

Datasets'. In *2016 IEEE International Conference on Big Data (Big Data)*, 2353–62. https://doi.org/10.1109/BigData.2016.7840868.

López, Carlos. 2003. 'Digital Rights Management of Geo-Datasets: Protection against Map Piracy in the Digital Era'. *GIM International* 17 (2): 51–53. http://www.thedigitalmap.com/~carlos/papers/rep03_1/RightsManagementForDigitalCartography.pdf.

McKinney, Peter, Steve Knight, Jay Gattuso, David Pearson, Libor Coufal, David Anderson, Janet Delve, Kevin De Vorsey, Ross Spencer, and Jan Hutař. 2014. 'Reimagining the Format Model: Introducing the Work of the NSLA Digital Preservation Technical Registry'. *New Review of Information Networking* 19 (2): 96–123. https://doi.org/10.1080/13614576.2014.972718.

Meger, Amy Lowe, and Daniel Draper. 2012. 'Digital Preservation and Access of Agricultural Materials'. *Journal of Agricultural & Food Information* 13 (1): 45–63. https://doi.org/10.1080/10496505.2012.637437.

Palumbo, Laura, Jantz, Ron, Lin, Yu-Hung, Morgan, Aletia, Wang, Minglu, White, Krysta, Womack, Ryan, Zhang, Yingting and Zhu, Yini. 2015. 'Preparing to Accept Research Data: Creating Guidelines for Librarians'. *Journal of EScience Librarianship* 4 (2). https://doi.org/10.7191/jeslib.2015.1080.

Prochaska, Alice. 2009. 'Digital Special Collections: The Big Picture'. *RBM: A Journal of Rare Books, Manuscripts and Cultural Heritage* 10 (1): 13–24. https://doi.org/10.5860/rbm.10.1.313.

Rachman, Yeni Budi, and Saiful Afidhan. 2018. 'Digital Disaster Preparedness of Indonesian Special Libraries'. *Preservation, Digital Technology & Culture; Berlin* 47 (2): 54–59. http://dx.doi.org.sire.ub.edu/10.1515/pdtc-2018-0009.

Raimond, Yves, Christopher Sutton, and Mark Sandler. 2008. 'Automatic Interlinking of Music Datasets on the Semantic Web'. In *CEUR Workshop Proceedings*, 369:8. Beijing, China. http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper18.pdf.

RBMS Security Committee. 2009. 'ACRL/RBMS Guidelines Regarding Security and Theft in Special Collections: Approved by the ACRL Board of Directors, September 2009'. *College & Research Libraries News* 70 (10). https://doi.org/10.5860/crln.70.10.8273.

Recker, Astrid, and Müller, Stefan. 2015. 'Preserving the Essence: Identifying the Significant Properties of Social Science Research Data'. *New Review of Information Networking; London* 20 (1–2): 229–35. http://dx.doi.org.sire.ub.edu/10.1080/13614576.2015.1110404.

Reich, Vicky, and Rosenthal, David. 2001. 'LOCKSS: A Permanent Web Publishing and Access System'. *D-Lib Magazine* 7 (6). http://mirror.dlib.org/dlib/june01/reich/06reich.html.

Rink, Katrina. 2017. 'Displaying Special Collections Online'. *The Serials Librarian* 73 (2): 170–78. https://doi.org/10.1080/0361526X.2017.1291462.

Robertson, Wendy, and Borchert, Carol Ann. 2014. 'Preserving Content from Your Institutional Repository'. *The Serials Librarian* 66 (1–4): 278–88.

https://doi.org/10.1080/0361526X.2014.881209.

Rodríguez, Perla Olivia. 2016. 'OAIS in the Preservation of Digital Audio Objects'. *Investigación Bibliotecológica* 30 (70). https://doi.org/10.1016/j.ibbai.2016.10.009.

Rothenberg, Jeff. 1995. 'Ensuring the Longevity of Digital Information'. *Scientific American* 272 (1): 42–47. https://doi.org/10.1038/scientificamerican0195-42.

Samuelson, Todd, Laura Sare, and Catherine Coker. 2012. 'Unusual Suspects: The Case of Insider Theft in Research Libraries and Special Collections'. *College & Research Libraries* 73 (6): 536–68. https://doi.org/10.5860/crl-307.

Schwartz, Scott, Prom, Christopher, Rishel, Christopher and Fox, Kyle. 2007. 'Archon: A Unified Information Storage and Retrieval System for Lone Archivists, Special Collections Librarians and Curators'. *Partnership: The Canadian Journal of Library and Information Practice and Research* 2 (2). https://doi.org/10.21083/partnership.v2i2.246.

Shamir, Lior, Carol Yerby, Robert Simpson, Alexander M. von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller, and John Wallin. 2014. 'Classification of Large Acoustic Datasets Using Machine Learning and Crowdsourcing: Application to Whale Calls'. *The Journal of the Acoustical Society of America* 135 (2): 953–62. https://doi.org/10.1121/1.4861348.

Southwell, Kristina, and Slater, Jacquelyn. 2012. 'Accessibility of Digital Special Collections Using Screen Readers'. *Library Hi Tech* 30 (3): 457–71. https://doi.org/10.1108/07378831211266609.

Tam, Marcella. 2017. 'Improving Access and "Unhiding" the Special Collections'. *The Serials Librarian* 73 (2): 179–85. https://doi.org/10.1080/0361526X.2017.1329178.

Taranto, Barbara. 2009. 'It's Not Just about Curators Anymore: Special Collections in the Digital Age'. *RBM: A Journal of Rare Books, Manuscripts and Cultural Heritage* 10 (1): 30–36. https://doi.org/10.5860/rbm.10.1.315.

The Center for Research Libraries, Online Computer Library Center, Inc. 2007. 'Trustworthy Repositories Audit and Certification: Criteria and Checklist'. http://www.crl.edu/PDF/trac.pdf.

Torre, Meredith. 2008. 'Why Should Not They Benefit from Rare Books?: Special Collections and Shaping the Learning Experience in Higher Education'. *Library Review* 57 (1): 36–41. https://doi.org/10.1108/00242530810845044.

Ulloa, Juan Sebastian, Amandine Gasc, Phillipe Gaucher, Thierry Aubin, Maxime Réjou-Méchain, and Jérôme Sueur. 2016. 'Screening Large Audio Datasets to Determine the Time and Space Distribution of Screaming Piha Birds in a Tropical Forest'. *Ecological Informatics* 31 (January): 91–99. https://doi.org/10.1016/j.ecoinf.2015.11.012.

Woolcott, Liz, Andrea Payant, and Sara Skindelien. 2016. 'Partnering for Discoverability: Knitting Archival Finding Aids to Digitized Material Using a Low Tech Digital Content Linking Process'. *Code{4}lib*, no. 34. https://works.bepress.com/andrea-payant/4/.