http://doi.org/10.7207/twgn21-05

Born digital archive cataloguing and description

Jenny Bunn

DPC Technology Watch Guidance Note

May 2021

DigitalPreservationCoalition

1 Introduction

Archives, in common with other institutions such as libraries and museums, have traditionally worked on a curatorial or custodial basis. Such institutions have assumed responsibility for ensuring the ongoing existence and accessibility of vast collections of material in a wide variety of forms both physical and digital. To do this work and to fulfil this responsibility effectively, they have needed to research, gather and hold large amounts of information about their collections; a subset of which information has routinely been made publicly available, primarily to facilitate the discovery and understanding of those collections.

The task of collecting and/or generating this information has until recently been undertaken primarily through human labour and manual processes, but advances in technology mean that there is increasing scope for such processes to be automated. Then again advances in technology also mean that there is increasing scope for presenting and placing the information that has been collected into an interconnected global context - online. This briefing note will seek to provide guidance to support organisations and individuals in making sense of these developments.

2 Automating the collection and generation of information

One of the main sources from which archival institutions have traditionally gathered information about their collections, are the collections themselves. Folders will have titles and identifiers written on them, documents will be dated, the way in which the materials are structured or arranged will convey information about how they were connected and compiled for the purposes for which they were created and used. Now that these materials are arriving at archives in digital forms, some of the information archives have traditionally collected, can be extracted automatically.

Such extraction is easier when the required information is present in a form which is already explicitly structured as metadata, e.g. a pro-forma printed onto the front of a file for completion, or the convention of heading a letter in a certain way, now both translated into file format specifications and email protocols. Many tools have been developed to support metadata extraction of this kind and a lengthy list can be found in the Community Owned Digital Preservation Tool Registry (COPTR, 2021).

Automatic extraction of metadata may speed up the process of extracting information from collections, but it can only work with what is already there. If the required information appears only in a low quality form, e.g. an unhelpful title such as ‘Document 1’, the difference between the ability to extract what little descriptive and contextual detail may be directly available, and producing meaningful and informative metadata becomes very obvious. It is always important to consider the degree to which any original metadata is informative; email is one example where the original metadata can often be regarded as sufficiently informative (Stanford University, n.d.).

One type of information that archival institutions have long been interested in conveying, is the subject or content of the material they hold – what it is about. Pre-existing metadata (such as the titles assigned to documents or files) may provide some of this information, but much of it can only be gleaned by reading through or viewing/listening to that content. Advances in natural language processing (and other techniques more suited to non-textual forms such as audio, images and video) are now starting to make it possible for more of this kind of information to be automatically generated. For example, the extraction of entities, such as people, places or organisations, mentioned in textual content was previously a task for human indexers, but can now be carried out automatically using techniques such as Named Entity Recognition. Then again there are also promising advances towards both the automatic summarisation of text (Gonçalves, 2020) and its

automated classification (Anderson et al, 2019). Where metadata is the result of automatic generation, it is becoming accepted good practice to indicate this fact and, where appropriate, the degree of confidence with which a predicted classification, say, has been made. Some of these technical developments (in particular Named Entity Extraction) have started to be incorporated into archival workflows and tools, such as ePADD and BitCurator (ePADD, 2021; BitCurator, 2021), but others remain at a more experimental stage.

One area in which automated techniques remain particularly experimental is in the extraction/generation of information from the arrangement or structure of archival collections. Whilst not always the case, the ordering of material in an archival collection, particularly where that collection can be considered to be the result of a system of some kind, can also convey useful information for understanding how the collection fits together within an original/originating context. This idea is traditionally understood within archival science as the principle of original order (Zhang, 2012). Some existing metadata extraction tools can collect metadata such as a filepath, but automatic techniques cannot yet read or process filepaths to infer their underlying rationale or logic, let alone evaluate whether that logic provides any meaningful information in terms of contextualising an archival collection as embodying or representing a particular organisation or agency. Contextualising archival collections in this way is something archivists have traditionally prioritised through their conceptualisation of what they call the 'fonds' (Cook, 1992).

Increasingly though systems are being developed which automate the sorting, tagging and ordering of material to result in the building of a model which encodes the logic and implicit knowledge behind that ordering in the form of a machine readable knowledge graph (Microsoft, 2019). With the development and application of these (graph) technologies, the important connections and relationships (the ordering) which form or underpin specific aggregations of material may start to become more amenable to machine, as well as human, interpretation leading eventually perhaps to the automatic summarisation of context, as well as text.

3 Defining informational needs

Writing in 1993, Margaret Hedstrom presciently wrote that 'In the electronic era, the descriptive paradigm will shift from the current practice of augmenting scarce descriptive information to one of selecting from an abundance of metadata' (1993, p.59). In face of the current Climate Emergency, the importance of this selection becomes even more important, since energy and other resources should not be wasted on either extracting or subsequently storing and maintaining information unnecessarily. We need to define specific informational needs.

The question of what those needs are is highly dependent on the purposes from which those needs arise. Archival communities have, over time, evolved agreement on the information that they need for their purposes and have formalised this agreement in the form of metadata standards, schemas or specifications. So too have other communities, and the complexity of informational needs within the wider landscapes of cultural heritage and scholarly research have been visualised and listed by Jenn Riley and the Digital Curation Centre respectively (Riley, 2009-10; DCC, 2021).

Within the archival community some agreement (but not universal – divergent views were expressed particularly by those from the Australian recordkeeping tradition) as to the information that was required for their purposes was reached in the early 1990s. This agreement is formalised (and to some extent fossilised) in the General International Standard Archival Description (International Council on Archives, 2011). This initial agreement was reached on the basis of a relatively limited understanding of the informational requirements of born digital material. Since

that time, understanding has developed and it has become clear that information is required, not just to describe archival items in terms of what they are about and where they came from, but also to describe them in terms of instantiating and maintaining the integrity of what is now only a digital/informational object, rather than also a physical one. This evolving understanding can be seen in the conceptualisation, within the OAIS Information Model, of information, not just in terms of content and context, but also in terms of representation and fixity (International Standards Organisation, 2012). More practically it has also led to the need to reach agreement that increasing amounts of information are required for archival or, if you prefer, digital preservation purposes, and for the formalisation of this agreement into new standards or schemas, such as PREMIS (PREMIS, 2015).

One trend that is very noticeable when comparing these earlier and later definitions of informational needs (e.g. ISAD(G) in comparison with PREMIS) is that definition is being formalised at an increasingly granular and more detailed level. This trend is not just the result of the community's greater depth of understanding. It is also reflective of a need to accommodate both the relatively unsophisticated processing (in comparison with that of human beings) of the computerised information systems on/through which it is now being conveyed, and a growing expectation amongst some users that archival metadata should be amenable to their further processing using advanced computational techniques.

4 Conveying information held about collections.

Putting aside for now the new informational requirements around the instantiation and integrity of digital objects, archival institutions do still draw a distinction between those objects (their collections) and the information they hold about them. Information of this last sort has, since at least the 1980s, been stored in a digital format in a number of information systems reflective of the evolving state-of-the-art in terms of the wider information technologies landscape. The ICA Expert Group on Archival Description has categorised an evolution in this landscape from database and markup technologies to graph ones (Gueguen et al, 2013). Over this evolution the way in which information is structured as data to make it more amenable to known computational methods (that can produce the results and do the things we want them to do) has undergone a number of changes, with the most recent changes leading to a model (the Semantic Web) that seeks to atomise information into vast numbers of simple statements or assertions (triples) that can be processed 'automatically' to make the resulting complex and interconnected knowledge base more machine readable, navigable and understandable.

The full ramifications and possibilities of this last development are still being worked out in practice (archival, digital preservation, computer science, information systems, etc.) For example, within cultural heritage, individual communities are starting to redefine their informational needs at a more conceptual or abstract level in the form of conceptual models such as CIDOC-CRM and Records in Contexts (CIDOC CRM Special Interest Group, n.d.; International Council on Archives Expert Group on Archival Description, 2019). In this way, they are seeking to define both the things (entities) they want to make assertions about and the type of assertions they wish to make between them, and (sometimes also) additional relationships that allow for assumptions and inferences to be made about different kinds or classes of things and assertions. Work is already being undertaken in some institutions to convert the information they already hold into the more atomised format of such assertions, using for example the RDF standard (often serialised as either XML or JSON) to ensure the wider interoperability and tractability of the information so atomised (The National Archives, n.d., Europeana, n.d.).

The realisation that information about collections is now being viewed in a global accessible context (online) is also leading to a need for those undertaking born digital archive cataloguing and description to keep abreast of the norms, vocabularies and standards that are gaining the most currency in that environment. For example, the Dublin Core Metadata Element Set is a widely used vocabulary for fifteen properties or forms of assertion that it is particularly useful to make in order to facilitate resource discovery (Dublin Core Metadata Initiative, n.d.). Then again Schema.org is a common vocabulary and mechanism by which information about the content or what is discussed within a web page can be surfaced to major search engines, again facilitating resource discovery (Schema.org, n.d.; Schema Architypes Community Group, 2017). Institutions are also starting to reconsider how they uniquely (at this global level) and persistently identify and support referencing (and de-referencing) to the things they are making assertions about (Towards an National Collection – Heritage PIDS, n.d.). Online publication of any data (metadata included) brings with it the need to consider relevant legislation (such as Data Protection) and the clear indication of terms for re-use. Where appropriate, commonly understood licensing schemes, such as Creative Commons, and emerging rights vocabularies, such as the Open Digital Rights Vocabulary, should be preferred. (Creative Commons, n.d.; W3C, 2018).

5 Conclusion

Archive cataloguing and description now takes place in a born-digital environment, even if the material it deals with is both analogue and digital in nature. Those practising it are adapting and evolving all the time in response. This guidance note highlights a number of environmental shifts in order to discuss the possibilities and challenges they have brought, as follows:

- The material which is being described is increasingly being digitally realised. On the plus side this offers up the opportunity for the 'automatic' extraction from it of the sort of information that is required for archive cataloguing and description. On the minus side those practising archival description now need to grapple with the idea that they need to do more actual describing (that is to say instantiating and maintaining the integrity of objects) than they did when the objects being described already had a more directly corresponding physical form.

- The information which is being extracted/generated as archive description is now itself born-digital. On the plus side this means it can be (indeed has been) contributed to the global information superhighway that is the internet and become (at least potentially) available to countless millions. On the minus side this means that those practising archival description now need to grapple not just with the rules of that road, but also with the other communities and individual interest groups on it. The rules of the road consist of both the standards such as XML and RDF that enable the road to exist, but also a willingness to constantly re-define and re-encode information and the knowledge needed to understand it at ever more granular levels of detail. The presence of other communities and interest groups, means that those practicing archive description must also learn to take into account the norms, standards and vocabularies they adopt and use, learning to compromise in the pursuit of common interests (such as facilitating resource discovery).

It is not necessary to grapple with all the challenges identified above at once or alone. Indeed as has been highlighted in the second point above, some of them are about overcoming the tendency (which arguably has been present within the archival if not the digital preservation community) to work in isolation. No specific guidance has been offered, but a number of directions for further

exploration have been indicated and the references will point you to sources to follow up on those that seem most relevant. The horizons for born digital archive cataloguing and description are expanding all the time, those who wish to practice it must therefore constantly work to expand their own.

6 References

Anderson, B. G., Prom, C .J., Hutchinson, J. A., Chandrashekhar, A., Michael, B., Udhani, S., Sammons, M., Dolski, A., Hamilton, K., Kaushik, S., and Shrivastava, M. (2019). The Cybernetics Thought Collective: A History of Science and Technology Project Portal White Paper. Available at: http://hdl.handle.net/2142/106050 [accessed 10 May 2021]

BitCurator. (2021). *BitCurator*. Available at https://github.com/bitcurator/bitcurator-nlp/wiki

CIDOC CRM Special Interest Group. n.d. *CIDOC Conceptual Reference Model*. Available at: http://www.cidoc-crm.org/ [accessed 4 May 2021]

Cook, T. (1992). *The Concept of the Archival Fonds in the Post-Custodial Era: Theory, Problems and Solutions*. Archivaria 35, pp.24-37. Available at: https://archivaria.ca/index.php/archivaria/article/view/11882 [accessed 4 May 2021]

COPTR. (2021). *Community Owned Digital Preservation Tool Registry*. Available at: https://coptr.digipres.org/index.php/Category:Metadata_Extraction

Creative Commons. (n.d.). *Creative Commons*. Available at: https://creativecommons.org/ [accessed 17 May 2021]

Digital Curation Centre. (2021). *Disciplinary Metadata*. Available at: https://www.dcc.ac.uk/guidance/standards/metadata [accessed 4 May 2021]

Dublin Core Metadata Initiative. n.d. *Dublin Core Metadata Initiative*. Available at: https://dublincore.org/ [accessed 4 May 2021]

ePADD. (2021). *ePADD*. Available at https://library.stanford.edu/projects/epadd

Europeana. (n.d.). Linked Open Data. Available at: https://pro.europeana.eu/page/linked-open-data [accessed 17 May 2021]

Gonçalves, L. (2020) *Automatic Text Summarization with Machine Learning – An overview*. Available at: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25 [accessed 4 May 2021]

Gueguen, G., da Fonesca, V. M. M., Pitti, D. V., and Sibille-de Grimoüard. (2013). *Toward an International Conceptual Model for Archival Description: A Preliminary Report from the International Council on Archives' Experts Group on Archival Description*. The American Archivist 76 (2), pp.567–584. Available at: https://doi.org/10.17723/aarc.76.2.p071x02401282qx2 [accessed 27 April 2021]

Hedstrom, M. (1993). *Descriptive Practices for Electronic Records: Deciding What Is Essential and Imagining What Is Possible*. Archivaria 36, pp.53-63. Available at: https://archivaria.ca/index.php/archivaria/article/view/11934 [accessed 4 May 2021]

International Council on Archives. (2011). *ISAD(G): General International Standard Archival Description*. 2nd edn. Available at: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition [accessed 4 May 2021]

International Council on Archives Expert Group on Archival Description. (2019). *Preview – Records in Contexts Conceptual Model Version 0.2*. Available at: https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf [accessed 4 May 2021]

International Standards Organisation. (2012). *ISO 14721: 2012 Space data and information transfer systems – Open archival information system (OAIS) – Reference model*. Available at: https://www.iso.org/standard/57284.html

Microsoft. (2019). *Introducing Project Cortex*. Available at: https://techcommunity.microsoft.com/t5/microsoft-365-blog/introducing-project-cortex/ba-p/966091 [accessed 4 May 2021]

PREMIS. (2015). *PREMIS Data Dictionary for Preservation Metadata*. 3rd edn. Available at: https://www.loc.gov/standards/premis/v3/ [accessed 4 May 2021]

Riley, J. (2009-10). *Seeing Standards: A Visualization of the Metadata Universe*. Available at: http://jennriley.com/metadatamap/ -:~:text=A%20small%20set%20of%20the,standards%20for%20cultural%20heritage%20metadata. [accessed 4 May 2021]

Schema Architypes Community Group. (2017). Schema Architypes Community Group. Available at: https://www.w3.org/community/architypes/ [accessed 4 May 2021]

Schema.org. (n.d.). *Schema.org*. Available at: https://schema.org/ [accessed 4 May 2021]

Stanford University. (n.d.). Email Collections. Available at https://epadd.stanford.edu/epadd/collections [accessed 17 May 2021]

The National Archives. (n.d.). *Project Omega*. Available at: https://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/our-digital-cataloguing-practices/project-omega/ [accessed 4 May 2021]

Towards a National Collection – Heritage PIDS. (n.d.). Heritage PIDs – Resources. Available at: https://tanc-ahrc.github.io/HeritagePIDs/resources.html [accessed 4 May 2021]

W3C. (2018). ODRL Vocabulary & Expression 2.2. Available at: ODRL Vocabulary & Expression 2.2 (w3.org) [accessed 17 May 2021]

Zhang, J. (2012). *Original Order in Digital Archives*. Archivaria 74, pp.167-93. Available at: https://archivaria.ca/index.php/archivaria/article/view/13410 [accessed 4 May 2021]