Digital Description and Metadata at the National Archives. Digital Strategy

Jone Garmendia

DIGITAL DESCRIPTION AND METADATA AT THE NATIONAL ARCHIVES. DIGITAL STRATEGY

Jone GARMENDIA

Over the last eighteen years, The National Archives of the United Kingdom[1] has delivered a wide range of online catalogues and digital services and is now transforming to deliver an ambitious digital strategy. Our Digital Strategy[2] addresses both the challenge of digital records as well as our goal to become a digital archive by instinct and design. To achieve this goal, we must acknowledge that digital records disrupt archival practice, archival theory and the whole notion of what a professional archivist should be. This disruption represents an existential challenge for archives and archivists alike. To survive we must embrace the disruptive digital archive and the digital transformation of the physical archive, strengthening our digital capability and transforming our institutional culture.

Digital description

The National Archives has shared its approach to digital description through the publication of a position paper[3]. We are aware that the archive sector in the United Kingdom and perhaps some international archives may look at our practice as a benchmark, nevertheless, it is important to stress that our position paper is not a set of cataloguing guidelines for digital records. The position paper aims to record our practice and open the conversation widely, recognizing that both technology and archival practice will continue to evolve. In addition, we have engaged with the ICA Expert Group on Archival Description, contributing extensive feedback towards the development of the Records In Contexts Conceptual Model and with Richard Wallis and colleagues working behind the Schema.org initiative.

Descriptive practices and metadata have already been disrupted by a first generation of digital records. We have been able to handle and present online this

---

[1] The National Archives at Kew, London, is the official archive and publisher for the UK central government, and for England and Wales.

[2] J. Sheridan, *The National Archives Digital Strategy,* published at www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/digital-strategy/ and www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf, accessed on 5th September 2018.

[3] J. Garmendia et al., *Digital Cataloguing Practices at The National Archives,* published at www.nationalarchives.gov.uk/documents/digital-cataloguing-practices-march-2017.pdf, accessed on 5th September 2018.

first digital wave reasonably well by using an eccentric ISAD(G) framework and our own Discovery schema.

Our Discovery catalogue[4] can be described as a web portal and is the largest archival information resource in the United Kingdom. It provides access to well over 23 million descriptions for records held by The National Archives with 8.9 million being for digital assets; over a third is digital and this percentage is growing. Discovery also provides access to almost 11 million descriptive units for mostly analogue records held by other archives in England and Wales[5]. Our first generation of digital records has been made available online via Discovery and it covers the following three digital types:

Digital surrogates: copies of original paper and other analogue formats such as microfilm, slides, photographic negatives etc. The analogue record remains the archival original, the authentic record. Digital surrogates are analogue in essence and their descriptive metadata does not form part of the record.

Born-digital: records created in a variety of digital forms (for example Office files including Word and csv formats, pdfs, some audio, video and some early email). The born-digital object is inextricably bound to its metadata, which becomes part of the archival record and must be (and be seen to be) trustworthy[6].

Digitized records: in our jargon, these are the result of analogue material being digitised to a high standard, with additional provenance metadata captured and embedded within each file in order to become the official record for permanent preservation (instead of the original paper, which is not transferred to the archive). The provenance metadata allows us to assert the authenticity of the digitised record. Although technically complex to create, handle and preserve, these records have a paper heart, look very much like analogue material and share many diplomatic characteristics with the original that never reached the archive. They are not particularly challenging to present online either; users are familiar with their appearance and the creation of presentation versions is not problematic.

Second born-digital generation

An untamed second generation of born-digital content is already accumulating within government departments and other record creators. This scenario has been referred to as the 'digital iceberg', the 'digital heap' and even the 'digital wild west'.

---

[4] The National Archives online catalogue is available at discovery.nationalarchives.gov.uk.

[5] These figures were taken on 5th September 2018. We publish new catalogue descriptions daily. In the last financial year, an average of 53500 entries were released per month. They are the result of both new accessions and content enhancement projects, many carried out by volunteers.

[6] Colleagues in the Digital Archiving Department at The National Archives are currently working on a project called ARCHANGEL co-creating a novel prototype of Distributed Ledger Technology to ensure that record and metadata are authentic and trustworthy. Further information at blockchain.surrey.ac.uk/projects/archangel.html.

In this heap, the identification of unique, authentic records becomes blurred; and rules around expected record-keeping behaviours no longer apply. Second generation born-digital content possesses at least one of the following characteristics:

- Unreliable provenance
    - Lack of clear creator(s) or hybrid creation including non-government parties
    - Erratic or broken link between government function and digital accumulation
    - Amorphous accumulation, stored in unstructured shared drives.
- Include corrupted files, embedded files or objects composed of multiple file formats
- Suspected duplication with other digital and printed records (that might be annotated)
- Not fixed in time (the end date of the record may be uncertain or set in the future as re-use may bring the record back into activity)
- Dubious authenticity

Examples would include shared drives without meaningful metadata structures, unstructured email servers created in environments where basic record-keeping and information management practices have not been applied, a variety of datasets, software applications, computer code and even algorithmic records (when machine learning techniques may have been used to develop government policies or decision making processes). We do not believe that we can render these accumulations within a traditional online catalogue such as our own Discovery system or the Archives Portal Europe, at least as they currently stand. Archives will need to develop and adopt an entirely new approach to description and metadata for this second generation of born-digital content, moving away from the international standard ISAD(G). We are looking at what our future data model might be, using metadata to develop a different style of archival digital description. For example, we are exploring probabilistic, contextual and temporally aware description.

Probabilistic description

Probabilistic description is about acknowledging in a transparent manner that data is imperfect and about embracing uncertainty. We are introduction confidence ratings about the trustworthiness of our metadata using computational methods. Our Traces Through Time Project[7] explored the use of probabilistic techniques to disambiguate names and link catalogue descriptions that referred to the same

7 Further information about the project, including links to its source code and schema are available at www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-projects/traces-through-time/ , accessed on 5th September 2018.

person. These personal names existed as free text within the scope and content metadata; they were just citizens in the records, not authority-controlled entities for prominent people. Our researchers used big data, contextual metadata and algorithms to build connections and create links labelled as 'other possible matches'. Transparency is extremely important; therefore, data linked computationally in our catalogue displays a warning and a probabilistic confidence rating (strong, medium or weak match). We could apply a similar approach to reveal how much we trust the metadata accompanying second-generation born-digital records.

Contextual description

Contextual description covers the creation, administrative and custodial history of both record and metadata; it also includes a variety of relationships, migrations, transformations and related events. The context of each record is vital in understanding its value as historical or legal evidence, and as a cultural asset. It is useful to think of the well-known assertion 'if content is king, then context is God.'[8] This mantra from the digital marketing and social media domains translates really well to the archival environment. The value of the content in our records and metadata is greatly diminished without context. Providing context is a powerful way to exercise intellectual control over an archive. Traditionally we have curated contextual information through the description of creators, the role, function and administrative history of creating bodies, the selection criteria, the custodial history and other circumstances surrounding the transfer to the historical archive. These contextual elements will remain relevant at accumulation or collection level for the second digital generation but we need to consider what other types of contextual metadata we need to acquire from the people creating born-digital material. Furthermore, we need to explore how much of this metadata[9] we should curate editorially as archivists; and exactly what (and how much) should be derived or extracted by processing born-digital records and their metadata computationally. Should we also attempt to derive any of this new contextual metadata from external and non-archival sources such as DBpedia?

There are also some fairly easy practical steps that we can take within a first generation online catalogue to improve context and findability for born-digital records. Digital file names are often meaningless without context. Some extreme examples in our catalogue include: 'Next week (2).msg', 'RE 2 random thoughts.msg', 'ClseUp Extra swg slip[A204460].jpg' or even '561,B4 OvVw Lwr Rat eaten[A206843].jpg'. We have used metadata from the multiple parent folders within the original file structures to provide some meaningful and searchable

[8] G. Vaynerchuk, various online sources, including www.garyvaynerchuk.com/content-is-king-but-context-is-god/, accessed on 5th September 2018.

[9] Several blogs by staff at the National Archives provide additional insights into our understanding of context, metadata and digital archiving, for example, blog.nationalarchives.gov.uk/?s=context+metadata.

information for born-digital files, which would have otherwise been extremely difficult to find. In our Discovery catalogue, the names of parent folders are mapped within the 'arrangement' field for each digital asset, offering a provenance trail, exposing the original filing structure and enabling search retrieval. As a result, a search for 'archival conservation' or 'limp vellum conservation' would find the file entitled 'ClseUp Extra swg slip[A204460].jpg'; and a search for 'London bombings' or 'solicitor emails' would find, among others, the file called 'RE 2 random thoughts.msg'.

Temporally aware description

Temporally aware description is a difficult concept. Space and time permeate paper and born-digital records differently; and we are just beginning to feel the impact. The records life cycle model identified boundaries between current and historical records. This intellectual model is not helpful in the new digital environment, as born-digital records are not necessarily fixed in time. We have handled paper records from the records life cycle perspective. This model is anchored on a vision of the archive as the final destination[10], after the record has reached the end of its life, has died administratively and has moved from the current, legal, time into the historical time.

The Records Continuum premise that records are always in a process of becoming[11] offers a more compelling model for born-digital material. Temporal variation and temporal uncertainty are becoming a new reality for both metadata and digital objects.

For example, digital records generally offer a variety of date metadata, which introduces a degree of uncertainty. Archivists handling a digital accession have to decide what date(s) should be presented to our users as the record creation date in our catalogue. We would like to publish additional date metadata but the crucial point here is that, at present, an archivist makes an intellectual choice for each digital accession in order to ensure that we identify accurate start and end dates for digital files. This is not a straightforward or scalable process. Different Electronic Records Management Systems store many automatically generated dates. Record creation or last modified dates can be overwritten when migrating digital records between systems in government offices, and then again, later, when exporting or preparing data for transfer to the national archive. There is a real risk that an inaccurate 'last modified' date could generate a later opening date for viewing than legally required. To avoid this scenario The National Archives asks government departments to use software (e.g., Teracopy) to preserve time stamps and minimise risk.

---

[10] H. Jenkinson, *A manual of Archives Administration*, Urbana-Champaign, 1937, p. 112.

[11] S. McKemmish, "Placing Records Continuum Theory and Practice", in *Archival Science,* 1. 4, 2001, p. 339-359.

With digital records and the new technologies used to process and re-use them over time uncertainty has increased and changed in nature. This has led to another complex archival problem around the time continuum: zombie records. We found a paper file from 1967 that had been transferred to The National Archives in 2003, taken on loan by the creating body and lost while on loan (with our online catalogue amended to record that fact). 13 years later, the file resurrected in digitised form, within a transfer of records that had been digitised after a damaging fire in 2008. The borrowed file (that could have already been accessed in the reading rooms and cited by researchers) had been scanned for business purposes, subsequently used digitally for a few years and later transferred (again) to The National Archives as a new record. It would not be particularly difficult for an archivist to curate a description for this file individually, providing adequate accumulation dates, custodial history, archivist's note and other contextual information. The challenge is whether the actions on the re-used record in digital form have actually brought back the file into the legal time; and whether this would merit a change to the last modified date and the legal record opening date for the public to view it (or for full online publication). Archival records are being stretched into new shapes and structures through digital aggregation: 'Records can even have multiple lives in spacetime as the contexts that surround their use and control alter and open up new threads of action, involving re-shaping and renewing the cycles of creation and disposition[12].

There is also a second challenge: we will not have the time or the resources to make these decisions individually as we need to manage digital transfer at great scale and pace. Data uncertainty can be managed through probabilistic description; however, born-digital records that are not fixed in time, that return to the legal time or go on a loop without reaching their final destination in the historical archive, are extremely difficult to control. New ways of exploiting metadata and technology will help us to provide multiple dated and traceable instances of a born-digital record. Versioning control software may help along the way but what I would like to figure out is how this temporal multiplicity could be presented to the public in a clear and usable manner. For purely born-digital zombies (disregarding hybrids) we could opt for the presentation of several manifestations of a uniquely referenced asset, but there will be many other perfectly valid approaches to represent digital records in a constant state of becoming.

Parting thoughts

We do not have solutions for all of these problems. In 2017-2018, The National Archives used Agile methodologies to explore: a) improvements to the digital transfer process, carrying out user research with government creators, and b) a new presentation system for the second generation of born digital archives.

---

[12] F. Upward, "Modelling the Continuum as Paradigm Shift in Recordkeeping and Archiving Processes, and Beyond – A Personal Reflection", in *Records Management Journal*, 10. 3, 2000, p. 120.

During the DARIAH-EU workshop on 15th May 2018, I briefly described how Agile methodologies were being used at The National Archives and encouraged attendees to use them beyond software development, for example devising time boxes to create time for thinking and innovation. I also shared our ambition to make records available for computational analysis, as an aggregate, and our commitment to research how archives could exploit metadata to provide enhanced context and content.

A final reflection: I often find the following behaviour unsettling but difficult to avoid. Sometimes we immerse ourselves in a long series of 'digital' projects believing that just by delivering the project objectives we are becoming a digital archive, or delivering significant change to our profession. For those with strategic responsibilities it is also very hard to formulate the right vision for digital change: a vision that reaches beyond the operational provision of digital preservation or IT training for the archives workforce. To achieve digital transformation and remain relevant in the future we need to revolutionise our culture. How this could be done extends beyond the remit of this piece and the original workshop presentation. We should look at ourselves with a critical eye but I feel optimistic that, with the help of a new generation of archivists from diverse educational backgrounds, we will embrace real digital and cultural transformation.