Building an Integrated Digital Archives

Richard Lehane, State Records NSW

Introduction

I begin the talk by referring to Bearman's Archival Methods. *I confess that I read it hoping to find choice pearls of wisdom to decorate my account of State Records NSW's Digital Archives project. The punch line… not only had Bearman identified one of the key challenges with which we are grappling (*building an integrated digital archives*), but he had anticipated our solution, and rejected it, preferring instead his concept of intelligent artifices.*

Context

State Records NSW is approximately half-way through a three year project to build a whole of government digital archives for New South Wales (see http://futureproof.records.nsw.gov.au/digital-archive/). We started the project conventionally, following the pathways forged by established digital preservation programs (such as at the NAA and PROV). We absorbed the OAIS model. We considered *the* digital preservation problem (file format obsolescence) and which of the camps we should join (normalisation, migration or emulation). We looked at tools for processes like check-summing. Accepting that we would need to automate as much as possible (because of the 'digital deluge'), we looked at workflow solutions like *Archivematica* and NAA's *Digital Preservation Software Platform*.

But there are problems with this conventional framing of the digital preservation challenge and its related solutions:

1. underdeveloped solutions for metadata (PROV is a noteworthy exception here with its comprehensive schema but even it has areas of weakness, esp. managing unique agency-specific metadata)
2. 'black holes': the problem of file format obsolescence has been elevated to such an extent that it has crowded out many other digital preservation problems. In fact, file formats (like MS Word) are designed for sharing information and as such don't pose a major preservation problem. Programs that focus on this problem often ignore, or can't accommodate, digital records that exist in customised business systems, cloud environments, and even EDRMSs. These 'black holes' should be the focus.
3. Relevance. Government digital preservation programs in Australia aren't facing a 'digital deluge' but a digital drought. Government agencies, seeing little of benefit in the services offered by these programs, don't have strong incentives to transfer digital records as digital archives.

At State Records NSW, we have re-framed the digital preservation challenge. Instead of seeking the best way of – ingesting digital objects, running them through a magic workflow, and preserving them – we are taking a system migration approach. We're developing a methodology for assessing agency digital recordkeeping systems, for planning customised migration strategies for those systems, and for integrating those systems into a coherent digital archives. Our approach is described in this research paper http://futureproof.records.nsw.gov.au/system-migrations-to-archives-a-research-paper-from-the-digital-archives-team/ and in an article in the forthcoming (December 2012) issue of *Archifacts.*

We see these benefits in our approach:

1. it can be applied to any recordkeeping system (no black holes)
2. it endeavours to be relevant to government agencies by including them in preservation planning and by producing a methodology and tools that they are applicable to contemporary recordkeeping challenges (the migration of systems is one of the key challenges in our sector)
3. absent rigid workflows and with customised preservation strategies for individual systems, we can take a bespoke approach to metadata, authentically integrating agency systems into a coherent digital archives

Building an integrated digital archives

Let's unpack that last point: 'authentically integrating agency systems into a coherent digital archives'. What does this involve?

Integration is at the heart of the archival endeavour. You can read most archival processes/strategies (appraisal, arrangement and description, access) as being fundamentally about integration: the task of creating a coherent archive that incorporates disparate recordkeeping systems. Sue McKemmish contends that records are ever in a state of 'becoming'. The continuum model suggests that the same is true for archives, and indeed we are constantly re-creating archives as we integrate new records with them over time. With paper records, this integration happened above the 'item' layer through the documentation of ambient and provenance context (i.e. descriptions of series, functions etc.). With digital records, we have an opportunity to support much deeper integration. David Bearman devotes a chapter of *Archival Methods* to this opportunity (and challenge):

> *Over the past several years, the proliferation of on-line databases and machine readable information sources has made information scientists painfully aware that the problem of intellectual transportation across disciplinary perspectives is not resolved by making data available on-line or in full-text. Indeed it may be exacerbated, since in manual retrieval systems the human mind makes leaps across categories which are not supported by existing mechanisms in automated systems. As a practical matter, if we are to integrate a variety of externally developed databases into archival information systems in order to provide for retrieval without much in-house description of records, we need to determine how we can best make large machine readable data stores, consisting of a variety of sources, each collated for particular purposes and audiences, accessible through a single user interface.* (*Archival Methods*, Chapter V: Intelligent Artifices: Structures for Intellectual Control).

Okay. So what we don't want is a digital archives that is merely a container for siloed systems that must each be approached/interrogated individually. From an access perspective, our goals in building a digital archives are to:

1. create simple, intuitive interfaces that can be used by general users to explore the digital archives as a coherent whole
2. ensure that government agencies can continue to rely on their records post-transfer and continue to use them, ideally seamlessly (by connecting their current business systems to the digital archives-as-backend)
3. integrate across, and not just above, agency recordkeeping systems by developing a capacity for structured querying of the contents of the digital archives.

Bearman identified two potential solutions for creating integrated digital archives and thereby obtaining these goals. Bearman's two solutions are metadata mapping (an approach he didn't like, but the one which we are pursuing at State Records NSW) and his own concept of intelligent artifices.

Metadata mapping (State Records NSW's approach)

I note in this part of the talk that, while State Records NSW has bedded down much of its methodology for migrating digital recordkeeping systems, this part of our approach is still largely conceptual – we've got much work to do and our metadata mapping strategy is by no means settled.

The obvious integration strategy for a digital archives is to enforce a common vocabulary by mapping metadata/structured data in digital recordkeeping systems. Bearman doesn't like this approach because it is labour intensive and semantically messy (mappings can never be precise and you can easily end up distorting meaning).

At State Records NSW, we can answer Bearman's first charge by admitting that yes, mapping is labour intensive, but in fact our whole approach to digital preservation is labour intensive (developing customised migration plans for individual systems) and, since we propose taking this case-by-case approach anyway, we can incorporate the mapping of metatdata/structured data into the migration methodology.

To Bearman's second charge we can offer only a partial defence. A mapping strategy that would indeed be semantically brutal is mapping to a fixed vocabulary (defining a preferred schema at the outset, and then forcing agency created metadata and structured data to conform to that schema). We're not taking this approach at State Records. What we propose instead is to create a *metadata registry* (follow our progress at https://github.com/srnsw/Metadata-Registry).

This metadata registry is a growing vocabulary of preferred terms with defined constraints. We'll record preferred terms in this metadata registry and, when we encounter metadata and structured data that matches those preferred terms we will map and transform it accordingly. But it is an open vocabulary so that when we discover metadata and structured data for which there are no appropriate mappings, we can register new terms (preferably based on existing vocabularies but, where necessary, coining new terms). The metadata registry will be an online resource that users can consult when constructing queries for the digital archives. It will also be available as a best practice guide for NSW agencies to consult when they are considering what metadata terms to use in new recordkeeping systems. The metadata registry makes heavy use of linked data technologies: the terms themselves will comply with the RDF data model, we anticipate supporting SPARQL queries over the digital archives, and the registry itself will be exportable (at any moment in time) as a Dublin Core Application Profile.

The flexibility offered by a linked data approach means that we can likely avoid gross distortions of meaning. Indeed if, in the future, agencies themselves use linked data to manage their own metadata and structured data, no transformations would be required (we could just assert mappings to the registry). In the near term, however it must be admitted that we can't entirely evade Bearman's critique. So what does he propose instead?

Bearman approach: intelligent artifices

With his concept of intelligent artifices, Bearman proposes a front-end solution to the integration challenge. Instead of changing the underlying data, Bearman suggests preserving the data as-is but presenting it through an intelligent interface that would allow users to search across seven dimensions (time, space etc.). The system would use a kind of fuzzy matching to guess answers to queries based on the raw data. Crucially, users could choose to override the system's matches and those decisions would be available as hints to later users following similar lines of enquiry. So, rather than a universal map, you achieve integration by allowing users to forge trails through the archive, assisted by an intelligent system, and benefitting from the discoveries of earlier users.

Conclusion

Two decades after *Archival Methods*, and it is still too early for conclusions. Bearman's concept of intelligent artifices is an intriguing one that deserves more exploration. At State Records, we're taking a different tack, the path that Bearman did not recommend, but with the advent of linked data technologies we think it has the most potential. In any case, if we can move the digital preservation discourse on from "emulation vs file format migration" to "intelligent artifices vs linked data mappings"… well that will be an end worth pursuing!