Project Omega: Modelling an Archive Catalogue to Support Future History

Dr K Faith Lawrence
The National Archives,
UK
faith.lawrence@nationalarchives.gov.uk

Adam Retter
Evolved Binary,
UK
adam@evolvedbinary.com

Jone Garmendia
The National Archives,
UK
jone.garmendia@nationalarchives.gov.uk

The National Archives has historically focused on physical records. Archival catalogues are structured according to the record within its surrounding context – not only its provenance but its place relative to other records in a collection. As a result physical records at TNA are categorized by the creating government department or body from which they came, and are organized based on the business function that generated them. Conversely, born-digital records are typically organised in one of two ways, 1) a heavily curated arrangement of records, whereby the arrangement of the records may not reflect the file plan of the digital files and associated metadata, or 2) a loosely curated arrangement whereby the arrangement is in fact the file plan of the digital files. In November 2019, TNA launched Project Omega with the aim of developing a proof-of-concept for a new system for managing catalogue data which could integrate data across the organisation - to do this we evaluated eleven current standards to find the model that could support a pan-archival catalogue and the complexities of the records themselves (physical, digital and digitised) and their associated metadata including spatial, temporal and agent relations.

The National Archives (TNA) of the UK is one of the world's leading archives. It publishes (https://discovery.nationalarchives.gov.uk/) more than 32 million catalogue descriptions for records held in i. our archive (which as well as physical records includes over 8 million digitised copies, born-digital records and the UK Government Web Archive); ii. more than 2500 archives across the UK (https://discovery.nationalarchives.gov.uk/find-an-archive). The catalogue is central to TNA and its role as the official archive and publisher for the UK Government, and for England and Wales. Our vision – 'Archives for Everyone' – sets out our ambition to be disruptive, inclusive and entrepreneurial. This includes redesigning our online offering and embracing our potential as a cultural heritage organisation.

Project Omega launched in November 2019 with the aim of developing a proof-of-concept for a new system for managing catalogue data. It quickly became apparent that the most value would be delivered by thinking beyond just recreating the catalogue and editorial systems to reimagining how catalogues at The National Archives (of which at least ten had been identified, some focused on born-physical and others on born-digital records) are envisioned and related, and how the published data could better interrelate with external resources within and outside the archives sector. To fully encapsulate the catalogue, it is necessary for the model to include temporal information, both in terms of the record itself and the metadata around the record which may change over time; agent information, in terms of the authorities related to the record, controlling access to the record, and the archivists creating and editing the record data; and spatial information about the record (physical and digital) and within the record.

This paper will discuss the first stages of the project: our evaluation and selection of data models from the standards identified as potentially relevant, and an analysis of the application of the selected model to our data. Currently TNA has implemented separate data models in use for physical records, (born) digital records, web archiving and front end record delivery among over ten models in current use. While all these models are, for the most part, informed by international archival standards (https://www.ica.org/en/public-resources/standards) they were developed to support specific use cases and as a result reflect only the different modalities of the records being archived and published making it difficult to fully exploit the range of data collected.

By comparing the strengths, weakness, similarities and differences of the existing models and the workflows that they were supporting, a list of eleven functional requirements was developed.

Using these requirements, twenty-three test cases encompassing both normal and complex situations in physical (twelve cases), digital records (ten case) and record access (one case) were identified and eleven current standards (TNA-CS13 Model, TNA DRI Catalogue Model, TNA Business Information, Architecture Model, Encoded Archival Description (EAD), Data Catalog Vocabulary (DCAT), Functional Requirements for Bibliographic Records (FRBR), Resource Description and Access (RDA), BIBFRAME Lite + Archive, Europeana Data Model, Records in Context (RiC), Matterhorn RDF Data Model) and models were evaluated against each.

Beyond the details of the evaluation itself, a number of interesting points were raised:

Almost all of the models separate the concept of a record from its realisation and many also allow multiple realisations. The later is key to supporting surrogates, redactions, preservation copies, digitisation and other variations on the original record. However, the majority of the models either support born-physical records or born-digital records with few having the needed flexibility to support the full range of physical, digitised, digital and potentially transdigital records that a forward looking, pan-archival system at TNA would require.

Whilst the older standards used a hierarchy model, the more modern standards have all adopted a graph model. This has potential implications for future development. While it is possible to support a graph model with a traditional relational or hierarchical database at the backend, to get the most benefit from such a model it is necessary to pair it with a graph database. A graph also allows for both expressing record provenance and, while not always explicit in the models, the provenance of the metadata surrounding the records. Further it supports both hierarchical

arrangement of records and more complex ad-hoc arrangements which supports our ability to describe the temporal and spatial relationships (physical and digital) between the records.

Following the evaluation, we have chosen to adopt the Matterhorn RDF model for Project Omega. While RiC might prove to be a better choice for adoption in the long term it currently does not support digital records to the level that we would require and the lack of interoperability with other widely used standards would limit our ability to fully benefit from data exchanges in the wider archival, cultural heritage and international communities. The model is still under active development so future iterations might be expected to rectify these issues. We will therefore revisit the issue when we move from pilot project to full development.

In looking at how we can create a model that can span physical and digital records we not only create the opportunity to bridge the gap between historical and current physical and digital records but acknowledge the future possibilities of transdigital and other, as yet unenvisioned, records which will challenge our understanding of both physical and digital archival catalogues. In undertaking this project, we were forced to reflect on the intersection of archival standards, web archiving, digital archiving and physical archiving and ask whether these practices can be brought together in a data model to support more efficient workflows, richer data and and better user experience both with our current archive and looking into the future where history that has not yet happened will become recorded and those resulting records, in whatever format, will need to be supported by our model.