Developing a Pan-Archival Linked Data Catalogue

Jone Garmendia[1[0000-0002-6532-2823]] and Adam Retter[2[0000-0001-9361-2126]]

[1] The National Archives of the United Kingdom, jone.garmendia@nationalarchives.gov.uk
[2] Evolved Binary, adam@evolvedbinary.com

Abstract. The UK National Archives has a large archival catalogue, available online since 2000. Our data is largely aligned to ISAD(G) and ISAAR(CPF) and is stored in a legacy relational database. We are now in a situation where the supporting infrastructure and the intrinsic data model are hindering our digital goals. We want to re-imagine archival practice by pioneering new approaches to description and access and by building a linked data catalogue. In this paper, we introduce Project Omega, which has evaluated standards, conceptual models and ontologies as most fit-for-purpose to underpin a new Pan-Archival Catalogue. This paper describes the early project findings and current implementation stage, providing an update on our five work streams: Data Modelling, Extract, Transform and Load of data, API, Catalogue Management System and Infrastructure.

Keywords: Archives, Catalogues, Data Models, Linked Data, Metadata, Ontologies.

1 Context

The National Archives of the United Kingdom (TNA) has a large and diverse archival catalogue. Our catalogue is itself an archival record while also a crucial business asset. The catalogue comprises several discrete database systems of varying technologies, although the data is largely aligned to ISAD(G) and ISAAR(CPF). The largest catalogue system (PROCat), which holds details of born-physical records, has been available online since 2000, and is backed by a relational database. Over the last 21 years, the infrastructure supporting our catalogue has expanded and diverged into an ecosystem of over 10 database systems. Separate systems were built, for example, to manage the legal conditions governing access to records, and to preserve digitised and born-digital archives.

In 2020, Project Omega started to explore the idea of replacing the aging archival catalogue systems. The project identified that a single system could be built to unify born-physical, digitised and born-digital catalogues by adopting a non-rigid (or schema-less) data model. It also recommended the use of a graph-based data model built with RDF (Resource Description Framework) technologies.

Our proposition is to move towards a single pan-archival linked data catalogue, taking a holistic view of an archive's assets (including all media, digital surrogates, and other record manifestations). To achieve this we need a sustainable data model and ontology that is flexible enough to support a second generation of complex born-digital accumulations as well as historical archives. We will gradually consolidate existing systems to introduce better workflows for accessions, data enhancement, enrichment, and for controlling access and publication of records. We imagine enhancing confidence in the integrity of the data, by introducing robust version management, provenance information, and audit trails. Through this project, we want to realize our ambition to reimagine archival practice and pioneer new approaches to description, data modelling and archival catalogue structures, delivering a new linked data catalogue.

This approach provides firm foundations for delivering our 'Archives for Everyone' strategy[1], enabling us to free our data and to unleash the power of an archival catalogue in a way that can support new forms of user engagement, participation, data re-use, and research. This data infrastructure project (Project Omega) is running in parallel with another project (Etna) that envisages what public interface we would create if we were to start completely anew with our website and vast catalogue[2].

2. Earlier Findings

In early 2020, Project Omega analysed the strengths and weaknesses of existing standards, to ascertain eligibility for expressing our conceptual data model. We focused our assessment on the following models:

- TNA-CS13 (aligned to ISAD(G))
- TNA-DRI (aligned to Dublin Core)
- EADv3
- DCATv2
- FRBR
- RDA
- BIBFRAME Lite + Archive
- Europeana Data Model
- RiC-CMv0.1 and draft RiC-O v0.2, including PIAAF project information
- The Matterhorn RDF Model approach.

In considering models for adoption, we have a strong preference to use open standards. The National Archives is committed to the use of open standards and is an active member of several international standards organisations. We adopt standards that align with the UK government’s open standards principles[3] and in particular that are developed through fair and transparent processes.

One model of particular interest was the Records in Contexts Conceptual Model (RiC-CM) and associated ontology RiC-O. These have been developed by the Expert Group

[1] https://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/archives-for-everyone

[2] The National Archives Discovery website presents over 35 million descriptions at https://discovery.nationalarchives.gov.uk/

[3] https://www.gov.uk/government/publications/open-standards-principles/open-standards-principle

on Archival Description under the auspices of the International Council on Archives (ICA). ICA provides a good home for archives and archivists to work together on standards. Unfortunately, we found it challenging to engage in the process for developing or contributing to RiC. We are looking forward to engaging with the Expert Group in an open and collaborative process. We would encourage both transparency and wider participation in RiC, to ensure RiC reflects the needs of ICA's membership and can benefit from the experience of prospective implementers like ourselves.

The outcome of our assessment was published in a Catalogue Model Proposal paper[4], outlining findings and technology recommendations. The paper evaluated 35 test cases with sample catalogue data expressed using the various ontologies. We decided to adopt a graph linked data model adhering to the principles of the Records in Contexts Conceptual Model (RiC-CM) but using a combination of existing, mature, vocabularies (inspired by The Matterhorn RDF[5] Model approach), rather than adopting the RiC Ontology as available at the time.

The key RiC-O challenges for us were:

- its limited set of properties to model our current born-digital material,
- a lack of comprehensive facilities for describing and controlling access and availability conditions for descriptive metadata and digital records,
- our need for the model to handle revisions, redactions, manifestations, associated provenance metadata, access rights and mappings to other vocabularies.

We had to make some difficult trade-offs, balancing the convenience of using one single archival ontology against our existing data proposition, our catalogue business rules and the legal context surrounding access to public records in the UK. Our hybrid ontology makes use of matured and tested W3C vocabularies such as PROV-O and ODRL. This allows us to fulfil our business needs while modelling concepts in a wider multidisciplinary context, reaching to and beyond the world of archives, enhancing interoperability. We continue to review revisions to the RiC Conceptual Model and Ontology as they evolve, and believe that our approaches are similar, compatible, and travelling in the same direction.

Modelling metadata variation over time, in the context of increasing uncertainty, is a difficult challenge. Descriptive practice must be aware of temporal variation. To model metadata variation over time, we have separated the enduring form of a record from its transient descriptions. Therefore, in our new model, any changes to the description or arrangement of a record will generate a new description and/or arrangement. Any fact established in the past is immutable and fully transparent. We make use of a FRBR-like layering of entities to separate enduring concepts, temporal descriptions, and realisations. That, coupled with W3C (World Wide Web Consortium) Provenance Vocabulary (PROV) enables us to record how our records evolve. Additional properties in the data model are used to describe relationships between versions and their temporal extent.

[4] https://www.nationalarchives.gov.uk/documents/omega-catalogue-model-proposal.pdf

[5] https://fedora.phaidra.univie.ac.at/fedora/objects/o:1079685/methods/bdef:Content/download

PROV gives us the ability to store information about revisions, agents (people/organisations), and activities (the process of change).

In the UK public records system, early transfer of government files is encouraged. Records often reach The National Archives before they are 20 years old. Legal exemptions and instruments assert the types of information that need to remain closed for a particular period (under Data Protection and Freedom of Information legislation). In addition to providing intellectual control, describing our records and providing access, our catalogue manages closure metadata and the operational process of opening previously closed files (and vice versa). The W3C Open Digital Rights Language (ODRL) vocabulary has furnished us with an approach to model the legal conditions governing access to public records. In parallel, another project at TNA is investigating the concept of 'gradated access', which would allow different types of users (sometimes in different locations) varying degrees of access to records and their metadata depending on multiple conditions. Our research shows that we will likely be able to complement our ODRL policies for closure, with additional policies for 'gradated' access and online publication.

3. Implementation

The second phase of Project Omega started in January 2021. We have a small but cross-disciplinary team working on five parallel work streams: data modelling, ETL (Extract, Transform and Load of data), API, management system, and infrastructure. There are many intricate dependencies between tasks under each of the work streams. We use Agile methods and tools to overcome these difficulties and obtain quick insights and feedback from archival and metadata experts. During the first phase of the project, our data modelling focused on immutable record description and arrangement. In the second phase, while we are considering the detail of modelling authority files (corporate bodies, persons, etc.), we have learnt that it is more effective to run the data modelling and ETL work streams in parallel, as together they inform a better design in each other.

3. 1 Data Modelling Implementation

Our Project Omega Data Model[6] is built atop ideas from many existing approaches and papers. We started modelling with the International Council on Archives' Records in Context Conceptual Model (RiC-CM) and integrated many of the ideas around reuse from The Matterhorn RDF approach. Fundamentally, we derived two key axioms that underpin our work:

- A Record is not just the paper or the digital file. Its descriptive metadata is part of the record.

[6] https://www.nationalarchives.gov.uk/documents/omega-catalogue-data-model.pdf

When a catalogue description is subject to change or archival intervention, it must also be subject to preservation in the same manner as the paper or digital file.

- How a record changes through time, its physical form and description, provides valuable contextual information.
  Being able to understand the curation of the record through time provides valuable insight into record-keeping behaviours and their impact on records and users. Preserving and presenting this information allows the archive (and the government creator) to become fully transparent and accountable.

To achieve both axioms, all description of records must be preserved and become immutable. If an archivist wishes to change any element of the description or arrangement, a copy is made and the amended description becomes the live version. Our legacy system allows only for the live and one previous version. We will no longer replace a description with the amended one. Furthermore, we will record the provenance of each change, incorporating metadata about when, who, how, and why the description was amended. To implement this we decided to sub-divide the notion of the 'Description of a Record' as an entity into four distinct entities: Concept, Description, Realisation, and Digital File (when a digital file exists).

The Record Concept contains properties that are known to be permanently immutable (i.e. will never be amended). The Record Concept is a single anchor for each archival record, only asserting 'we know we have a record'. In practice, it is little more than an identifier. Each Record Concept may have as many Record Descriptions or Record Realisations as are required. It is entirely possible to have concurrent competing descriptions (e.g. curated vs. machine learning vs. public user contribution) and realisations of a record. The metadata properties, which have historically been considered the description of the record, are now split between the Record Description and the Record Realisation. By separating the description of a record into these four entities, we can easily create new descriptions, realisations, and arrangements without destroying any existing information. Serendipitously, the ability to have multiple competing descriptions and realisations of a record enables us to use these same constructs to manage redaction (when part of a record or description cannot be publicly accessed) and un-redaction (when the closed part can be reinstated).

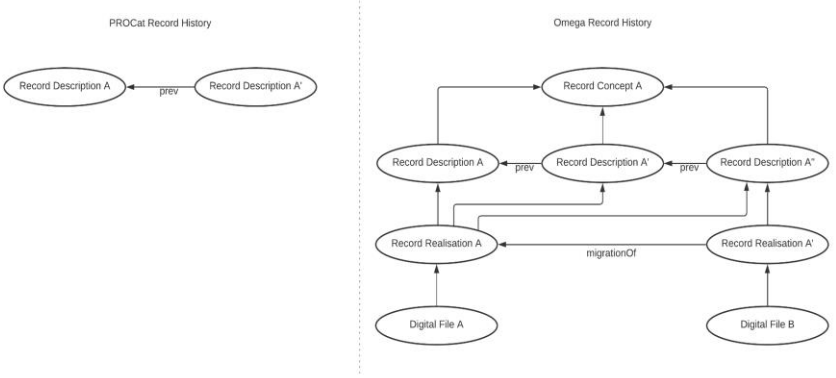

Figure 1. PROCat and Omega Record History

We have created a highly flexible data model (preserving changes and their purpose) from this FRBR-like separation of entities of a record description.

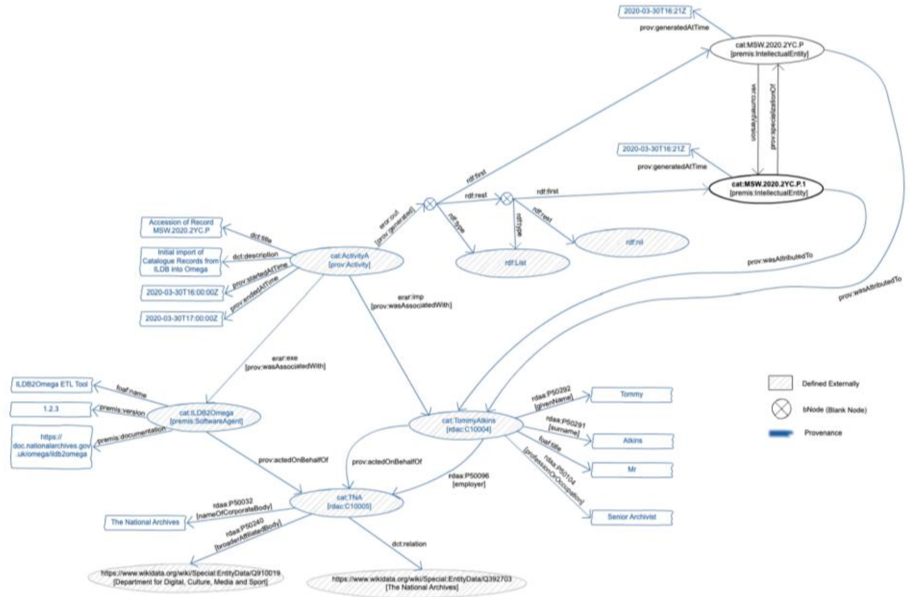

Figure 2. Example of Omega Record Provenance

Our data model documentation shows examples, advises on what properties and relationships to use and when, and provides mappings from the current data models (roughly aligned to ISAD(G)). We are adopting the RiC-CM concepts of ‘Record’ and ‘RecordSet’ to enable non-hierarchical relationships, moving away from the classic ISAD(G) hierarchical levels of description. Our approach for authority files (corporate bodies, persons, places, concepts) uses the same division into FRBR-like entities.

In terms of RDF vocabularies, we specifically make use of the following:

- PREMISv3: for our basic entities and structure. RecordSets, Record Concepts, and Record Description are all premis:IntellectualEntity. Our Record Realisation is a premis:Representation
- W3C Provenance: to record who, when, and why, things were changed
- W3C Time Ontology: to express complex date information, e.g. Covering Dates of a Record or RecordSet
- Dublin Core: for many of our simpler data properties, e.g. identifier, title, creator, abstract, etc.
- RDA (Resource Description and Access) Ontology is used where we cannot find suitable properties of relationships from other preferred ontologies
- a very limited bespoke ontology when we absolutely need to add a property[7].

As we have no one comprehensive ontology to ensure the quality of the RDF data that we produce, we are using SHACL to validate our RDF data. At the time of writing, we have exported over 8 million item level records from the catalogue relational database into our new data model as RDF Turtle; the first step towards loading our data into a cloud-based graph database (Amazon Neptune).

Every entity (resource) in RDF must have a unique URI to identify it. During the first phase of Project Omega, we undertook a task to investigate and propose a new Catalogue Reference labelling scheme. The goal was to create a scheme which would be both friendly to human communication (verbal and written), be easy to generate computationally, and suitable for use in URIs. In addition, the scheme would have to be suitable for all media of records (born-physical, digitised, born-digital etc.) and scale to describe versions of description and realisation. This scheme is not designed to replace existing record identifiers (catalogue references) but rather to augment them. The new scheme is known as OCI (Omega Catalogue Identifier) and has the potential to become the canonical identifier of a record at TNA. We have published several articles on our URI and Identifiers research: Archival Catalogue Record Identifiers[8], Archival Identifiers for Digital Files[9], and Extreme Identifiers (for use in URIs)[10] and will fully document the scheme when it has been verified as fit for purpose.

3.2 Extract, Transform, and Load of Data

To reach the goals of Project Omega and replace the existing legacy catalogue systems, we must populate our new graph database with the data drawn from source databases. Before importing, data must be transformed to fit the data model. The existing systems

[7] See https://medium.com/the-national-archives-digital/reusing-standard-rdf-vocabularies-part-1-5a9bbfa58b85 and https://medium.com/the-national-archives-digital/reusing-standard-rdf-vocabularies-part-2-4e4a3ad0bbf5

[8] https://medium.com/the-national-archives-digital/archival-catalogue-record-identifiers-29b0a1fac9ba

[9] https://medium.com/the-national-archives-digital/archival-identifiers-for-digital-files-c448ff463c22

[10] https://medium.com/the-national-archives-digital/extreme-identifiers-for-use-in-uris-cae773b98cf7

are varied, and data must be drawn from SQL, JSON document databases (e.g. MongoDB), Microsoft Access, Excel spreadsheets, CSV files, XML files, RDF Data stores (e.g. Apache Jena), etc. Each data source offers new and unique technical and data challenges.

Rather than developing custom software for each data source, we decided to make use of an existing framework for building and executing ETL (Extract, Transform, and Load) processes. For this purpose, we chose Hitachi Vantara's PDI (Pentaho Data Integration) suite. PDI has the advantage of being an established product with good documentation and community support, Open Source, written in Java (which our team is familiar with), and extensible through authoring custom plugins (in Java).

PDI ships with many build-in steps, each of which performs a customisable action, such as extracting data from a SQL database, parsing emails, or finding and replacing text in strings. The benefit of using PDI is that many of the steps that we must perform to transform data from one system to our Omega model are already available. PDI allows you to connect visually these steps into custom transformations or jobs. Transformations can be re-used within other transformations/jobs as steps themselves, which enables developers to build up their own library of reusable components.

Much of the time taken in building transformations in PDI is about ensuring that the output data meets the standards of our data model. Data is often messy and inconsistent due to the long operating life and limited constraints of the original systems, which have been in use for over 20 years. We have had to build many steps within our workflows to clean up the data. For example, the catalogue relational database uses EAD XML to provide structured textual metadata for some properties of the records (e.g. Scope and Content), however there is little in the way of system constraints on how the EAD may be used, or even whether the XML is well formed! We have used existing Schema Validation and Regular Expression steps in PDI to clean-up the data and ensure the validity of the EAD XML.

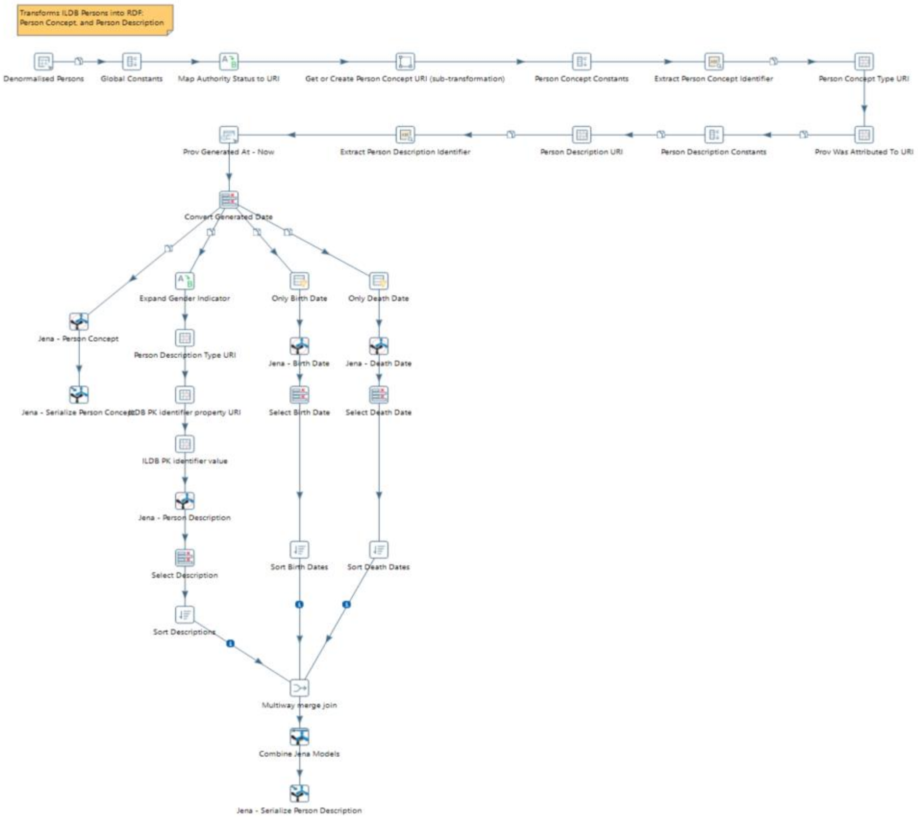
Transforms ILDB Persons into RDF:
Person Concept, and Person Description
Denormalised Persons
Global Constants
Map Authority Status to URI
Get or Create Person Concept URI (sub-transformation)
Person Concept Constants
Extract Person Concept Identifier
Person Concept Type URI
Prov Was Attributed To URI
Person Description Constants
Person Description URI
Extract Person Description Identifier
Prov Generated At - Now
Convert Generated Date
Jena - Person Concept
Jena - Serialize Person Concept
Expand Gender Indicator
Person Description Type URI
ILDB PK identifier property URI
ILDB PK identifier value
Jena - Person Description
Select Description
Sort Descriptions
Only Birth Date
Jena - Birth Date
Select Birth Date
Sort Birth Dates
Only Death Date
Jena - Death Date
Select Death Date
Sort Death Dates
Multiway merge join
Combine Jena Models
Jena - Serialize Person Description

Figure 3. PDI Transformation for transforming Person(s) from PROCat/ILDB to Omega RDF

To date we have faced many challenges with our ETL work. We will not dwell on these in detail here; instead, we would like to share how we overcame some of the obstacles:

- Building custom plugins for PDI to produce RDF Output. We released these as Open Source[11].
- Ensuring that we create a unique URI for each entity (Record, Person, Corporate Body, etc.) and that the same URI is reused for the same entity throughout the system. We developed further Open Source plugins for PDI[12] to ensure that a unique URI was created and stored only once.
- Processing and computing over historical dates from the (now) United Kingdom. This was ultimately related to how archival dates have been recorded, and the fact that the Julian/Gregorian Calendar switch-over was not undertaken

[11] https://github.com/nationalarchives/kettle-jena-plugins and https://blog.adamretter.org.uk/rdf-plugins-for-pentaho-kettle

[12] https://github.com/nationalarchives/kettle-atomic-plugins

universally at the same time. We investigated how to solve this programmatically[13] and also contributed an Open Source enhancement to PDI to fix the issue[14].

3.3 API and Catalogue Management System

Building a new Data Model for our Pan-Archival Catalogue and importing the existing data is a first step. However, to deliver a successful and usable product, The National Archives must be able to use the new linked data catalogue to exert intellectual control over its holdings. The main legacy editorial system has over 100 active users who need to manage the accessioning of new records into the archive, curate individual descriptions, ingest and enhance data in bulk while also providing quality assurance for a constant flow of data generated by many cataloguing projects. To this end, we will be building a new web-based, catalogue management and editorial application. So far, we have identified key technologies for our application that will allow users to interact with and manage our data graphs. We have also undertaken some preliminary research into user experience and user interface design.

As well as allowing our staff and volunteers to interact with our new catalogue, we want to open opportunities to discover new relationships between its content and to create new ways to present and contextualise our records. To this end, our catalogue management and editorial system will access an API (Application Programming Interface) to communicate with the database, instead of interacting directly with it. This API will enable the new catalogue management system and other applications run by other parts of TNA to deliver their services in a joined up manner. We envision an ecosystem of applications that both consume and contribute to our graph to deliver, for example, catalogue, preservation, gradated access and public online services.

3.4 Infrastructure

Our Infrastructure work stream has barely started; so far, all development has occurred in local environments. To date, we have established resources and procured contracts to set up a Virtual Private Cloud within Amazon's Web Services. We have just started to move all development into the Cloud, and to set up a Proof of Concept product utilising Amazon Neptune (Graph Store) and EC2 (Elastic Compute Cloud).

4. Conclusion

Devising and developing a linked data catalogue with the ambition to become a Pan-Archival model is hard. The National Archives is devoting a very significant amount of intellectual effort, technical expertise and financial investment to transform our archival catalogue infrastructure. The pan-archival approach has led us to collaborate with many teams and domain experts, as we have to consider much more than just our main catalogue. Project Omega is not a green-field project, instead we are having to

[13] https://medium.com/the-national-archives-digital/processing-historical-dates-d7ddb5814de8

[14] https://github.com/pentaho/pentaho-kettle/pull/8006

reverse-engineer existing systems, data, and legacy processes. Furthermore, the catalogue has to keep functioning while we migrate into our new model and processes. This is a dynamic catalogue that, in spite of COVID restrictions, made available over 560,000 new or enhanced catalogue descriptions in the financial year that ended on 31 March 2021.

A crucial and challenging part during the first phase of the project was the need to achieve buy-in and resources from our internal leadership. We worked tirelessly to sell the idea and communicate the advantages (and potential future cost savings) to the business; e.g. replacing existing legacy and unsupported software, reducing duplication and creating new opportunities through unlocking the unrealised potential in TNA's data. It is also fair to acknowledge that this would not have been achieved without the vision and endorsement of our Digital Director, who championed the project from the outset.

Although we face technical, conceptual and data challenges every week, we keep iterating, making continuous improvements and learning. There is a sense of professional gratification each time we are able to tackle, document and share our approaches to the resolution of an issue. We are committed to open source development and the sharing of our work through blogs and The National Archives public GitHub. Our modelling and implementation experience should hopefully aid others embarking on linked data transformation projects.

For other archival institutions that are looking to develop or improve their catalogues, we hope that our research can prove helpful to inform their own decisions. Our advice would be the following:

- consider carefully the scope of your project and any legacy, technical or human constraints
- think about what data models are most appropriate for your data
- tackle upfront your provenance and transparency requirements (e.g. do you wish to preserve all changes and versions?)
- re-use existing vocabularies to facilitate linking with the wider world
- agree an identifiers scheme that strikes a balance between human communication and the ability of a machine to compute over it
- be ready to get your hands dirty fixing data to make progress, historic data is wonderfully inconsistent.

We must stress the benefits and long-lasting value of linked data initiatives. Being part of the Semantic Web, using the tools and knowledge developed by others and collaborating to make them more usable for archives is a very worthy cause. We are excited by the possibilities that linked data will bring, for example, by using external relationships to enrich our own records via links to resources such as Legislation.gov.uk, Office for National Statistics, government datasets, Wikidata, etc. Finally, we would like to make it easier for other institutions and individuals to reference and use our data, placing records and descriptions in the larger context, reaching beyond the archival community.