The Shock of the New: Testing the Pan-Archival Linked Data Catalogue with Users (short paper)

Alex Green[1] and Dr K Faith Lawrence[2]

[1, 2] The National Archives, Kew Surrey, TW9 4DU, United Kingdom

Abstract

The UK National Archives' goal is to re-imagine archival practice, pioneer new approaches to description and build a new linked data catalogue. The Pan-Archival Catalogue will bring together into one management system descriptions of both physical and digital records from a variety of sources within the organization. This report briefly describes the users' feedback on aspects of the new data model when first shown in the new editorial interface and as part of business processes.

Keywords

Archives, Catalogues, User Research, Data Model, Linked Data.

1. Introduction

Archives are changing. New ways of preserving, describing and presenting records are emerging and archivists are rising to the challenge. At last year's conference, our colleagues presented a paper on the development of The National Archives' Pan-Archival linked data catalogue and our need to replace the ageing system with a new catalogue to manage the metadata for all types of records. [1]

We have started our exploration of how the model fits with the needs of users and their processes. In this brief report, we reflect on the initial user responses to the implementation of the new model in the first iterations of the editorial user interface and related editorial and accessioning workflows. This highlighted a number of assumptions we had made about the application of linked data in general, and our data model in particular, to the machinery of the editorial process within the creation and management of The National Archives' catalogue.

Proceedings TPDL2022: 26th International Conference on Theory and Practice of Digital Libraries, 20-23 September 2022, Padua, Italy

EMAIL: alex.green@nationalarchives.gov.uk (A. 1); faith.lawrence@nationalarchives.gov.uk (A. 2)
ORCID: 0000-0002-6993-4157 (A. 1); 0000-0001-9200-1921 (A. 2)

2. Enter the Users

The first phase of the project produced a draft of our conceptual data model. A living document, it described our approach to ‘replacing legacy systems, reducing duplication and creating new opportunities through unlocking the unrealised potential in The National Archives’ data’ [2]. This model embraced new thinking in archival description (notably the ICA’s Records in Contexts (RiC) [3]) to enable us to meet our strategic goal of reimagining archival practice for the 21st century. A significant change introduced by the model is the division of a record (an intellectual entity) into four entities, each with their own properties: an unchanging Concept, its associated temporal Description(s), Realisations (specific physical or digital instance(s)) and individual Digital Files (i.e. computer files in the case of digital records).

In parallel, we carried out some initial user research on the existing editorial interface to give us an understanding of the current processes, issues and new requirements using an early prototype based on the existing physical records model. As the new model is a marked departure, and because the catalogue is a business critical system, a more rigorous approach to our user research was needed. A team of user experience (UX) researchers and service designers carried out formal user research to ensure that, as well as adhering to accessibility and editorial standards, we were following UK Government Digital Service Standard [4] best practice to put user needs at the centre of new product development. As we explained the new model to this team and provided them with sample data, we realised how significant the impact of the new data model would be on our colleagues' day-to-day work.

3. Respecting the fonds?

The principle of provenance forms the basis of most institutional archival description. Within the profession however, there has been a growing acknowledgement of the multiplicity of perspectives on archival records beyond the institutional. The RiC-Conceptual Model (RiC-CM) incorporates this wider perspective, encompassing the single hierarchy resulting from the *Respect des fonds* approach but taking it further. It defines a record set as both ‘one or more records that are grouped together by an agent based on the records sharing one or more attributes or relations’ and ‘some other selection and grouping that fulfils a particular purpose or purposes (for example, a classification that reflects or supports the purposes of a researcher)’ [5]. We decided to adopt RiC’s record set and its associated definitions but its application remained in question. Our existing catalogue describes the records and their arrangement when in use in their creating department therefore the types of record sets necessary to support the current catalogue data are: Fonds (Department), Division, Series, Subseries, Subsubseries and File. This list of types is expected to expand as born-digital records are brought into the model with the addition of Item (where child ‘Sub-items’ exist) and Directory (for a folder containing individual digital files). However, with the wider definition of a record set, we could support alternative groupings or arrangements (as well as the principle of original order). This would be a significant departure from the current process, even discounting what we have termed ‘catalogue-adjacent record sets’, such as those records grouped as a result of research or presentation which are beyond the scope of the project. We can however, foresee cases where records could belong to more than one record set e.g. a set of records relating to an event created originally by the Home Office and subsequently sent to an inquiry, could be reflected in two arrangements according to their differing uses by the two bodies.

From a staff user’s perspective, how should we present our records arranged in different ways by the people who used them? How would we show the contexts of the different arrangements? Due to the urgency of replacing our ageing system, finding the answers to these questions can be deferred for the moment.

4. Splitting the Record

The division of a record into four entities (see Section 2) is a fundamental shift from our current ISAD(G)-based data model[1]. Each entity has its own properties, some of which exist at multiple levels, for example, both the Description and the Realisation have Scope and Content but each contain different information. Other properties are unique to a specific level, for example, only Realisations have Physical Extent and Form.

When we considered how to present these different entities and their properties in the new user interface, it became clear that the data could not be split easily into these entities and properties using automation. To explore this, we mapped an existing catalogue description to the new model.[2] The current Scope and Content describes the record as ‘Middlesex: Westminster (now in London Borough of Westminster). Plan of Buckingham House and grounds abutting on Green Park and St James's Park. Shows garden layout, with trees in elevation. Reference table to plots marked EFGHI and KLM on plan. Scale: 1 inch to 60 feet. Compass indicator. [By] Charles Evans. This plan, annotated 'No 150' at the top of the sheet, is similar to MPE 1/378, but refers to different portions of the site. A copy of this plan, made in May 1760, is MFQ 1/450’

[1] See https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

[2] See the catalogue entry on Discovery: https://discovery.nationalarchives.gov.uk/details/r/C4048574

Table 1
Record Scope and Content Description Reworked into the New Data Model

Scope and Content: Middlesex: Westminster (now in London Borough of Westminster). Plan of Buckingham House and grounds abutting on Green Park and St James's Park. Shows garden layout, with trees in elevation. Reference table to plots marked EFGHI and KLM on plan. Scale: 1 inch to 60 feet. Compass indicator.	**Copies Information**: A copy of this plan, made in May 1760 is MFQ 1/450. **Map Scale**: 1:720 **Related Material**: This plan is similar to MPE 1/378, but refers to different portions of the site. **Places**: Westminster, Green Park and St James's Park **Creator**: Charles Evans
Realisation 1 (the physical record held at TNA):	**Scope and Content**: This record is hand drawn
Realisation 2 (a digitised copy held by the Image Library at TNA):	**Scope and Content**: This record is a digitised copy

Staff acknowledged that our mapping, augmentation and arrangement of the description (see Table 1) were valid but were concerned about how this would be achieved without extensive re-cataloguing. Some of this metadata is not captured in this structure by the existing accessioning process so it would necessitate changes not just internally at The National Archives but also government departments who are responsible for describing the records they transfer to us. Clearly, this approach will need further consultation.

5. From the Specific to the General

Our existing data model is based on the accepted archival principle that 'archival description proceeds from the general to the specific' i.e. data is normalized so that it is held at the highest point possible in the hierarchy [6]. With the new data model, we are revisiting this principle. Denormalizing the data, i.e. moving the information from the upper levels down to the level to which it applies, might be more accurate in some cases. It could also simplify the queries needed to return relevant records. However, this approach is not without its issues from both practical and archival perspectives.

Some properties can safely be moved down to the record level: if there is only one organisation in Immediate Source of Acquisition and the series is no longer accruing, this value could be denormalised. Archivists are reluctant to make inaccurate statements and even when it seems simple to denormalize the data, it is not always sound to do so. For example, where the creator information is currently held at series rather than at file or item level. If a series only has one creator then the logical assumption would be that the information could be propagated down to any records within that series. Conceptually however, in recognizing that the catalogue data is both a work in progress and exists in an open world, the assignment of the creator at series level only indicates that at least some of the records in that series came from that creator rather than being a statement about all of them. From a functional perspective, this nuance may not be obvious to a researcher using the public catalogue, so denormalizing substantiates an assumed, but unknown, association.

Where there are multiple creators listed at series level (see Fig. 1) denormalization is more risky as dates, the most obvious means of disambiguation, are not granular enough to separate many of the edge-cases, and, as the individual records themselves may have more than one creator, there cannot be a clear and automatically applied delimitation. While these edge cases are only a small percentage of the total, the number is still significant enough to require manual checking, and this cannot be achieved during the initial data migration exercise given the project's deadline.

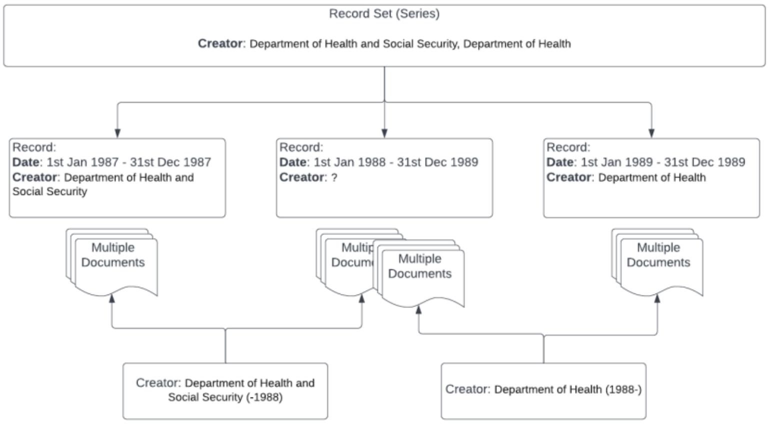
Record Set (Series)
Creator: Department of Health and Social Security, Department of Health
Record:
Date: 1st Jan 1987 - 31st Dec 1987
Creator: Department of Health and Social Security
Record:
Date: 1st Jan 1988 - 31st Dec 1989
Creator: ?
Record:
Date: 1st Jan 1989 - 31st Dec 1989
Creator: Department of Health
Multiple Documents
Multip Docum
Multiple Documents
Multiple Documents
Creator: Department of Health and Social Security (-1988)
Creator: Department of Health (1988-)

Figure 1: A record set with multiple creators listed at series level

If we cannot denormalize the existing data, either as a whole or in part, could we look to the future data to improve the accuracy of the catalogue? Could the data better reflect reality at the level at which the information is held? One example could be holding the creator of a record at the level of that record. The ideal is to represent the truth but, as with the additional requirements around the capture of more structured metadata described in Section 4, we are not receiving creator data for individual records now, so these changes would need to be discussed with both internal staff and those in government departments responsible for describing and transferring the records. From a user-centric perspective, the aim of the new system is to streamline the editorial process and reduce the workload by removing inefficiencies in the interface rather than generate additional work. Some information could be captured automatically but, returning to the example of creator, the transferring department and the creating department are not necessarily the same and the transferring department may not know the creating department if the records are older, and/or inherited from elsewhere in government. This leaves us in a position where conceptually it would be valuable to denormalize the data, and the data model supports us doing so, but it may not be feasible in reality.

6. Conclusion: Challenging our Assumptions

Our work with the UX team challenged some key assumptions that we had made in the early stages of the project. While we had shared our data model publicly and sent it for review by members of our core user group who were familiar with conceptual models, it was not until we began incorporating parts of the model in the wireframes for the initial interface that the impact for the archivists and their working practices became clear.

The work on the new catalogue system looks both inwards, to improving the interface for the editorial team, and outwards to the data contributed by other teams to the system or those supplied with data sent from the system. Moving to a linked data catalogue offers many advantages when searching, processing and exploring the data but for the staff managing the data, the benefits are less clear. While the staff are enthusiastic and engaged, change is never an easy proposition and a key component of successful change management is showing the direct benefits to the people affected by the change. We have more questions than answers, but we have a better idea of what the questions are. It is the users rather than the technology that should drive the change, especially where it has implications for the editorial process. As we reach the stage of the project where technology and users meet, and under the pressure of delivery deadlines, we are seeing more points of negotiation and re-evaluation emerge. We will continue to learn as we start to build the user interface iteratively: testing our assumptions about how staff will work with the new model to ensure it meets their needs and allows us to make the best use of the model's potential.

7. References

[1] J. Garmendia, A, Retter, Developing a Pan-Archival Linked Data Catalogue, in Proceedings of Linked Archives International Workshop 2021, CEU-WS.org, pp. 93-10393-103. URL: http://ceur-ws.org/Vol-3019/LinkedArchives_2021_paper_7.pdf.

[2] J. Garmendia, A, Retter, Developing a Pan-Archival Linked Data Catalogue, in Proceedings of Linked Archives International Workshop 2021, CEU-WS.org, p. 103. URL: http://ceur-ws.org/Vol-3019/LinkedArchives_2021_paper_7.pdf.

[3] ICA, Records in Contexts Conceptual Model, Consultation Draft v0.2 July 2021. URL: https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf.

[4] Government Digital Service, Service Standard, 2019. URL: https://www.gov.uk/service-manual/service-standard.

[5] ICA, Records in Contexts Conceptual Model, Consultation Draft v0.2 July 2021, p. 23. URL: https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf.

[6] The General International Standard for Archival Description, September 201, p. 8. URL: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition.