Assessing the practicality of ARK identifier usage in a catalogue of medieval manuscripts

Halle Burns[1], Toby Burrows[2], J. Stephen Downie[1], David Lewis[2], Kevin Page[2], and Athanasios Velios[2]

[1] School of Information Sciences, University of Illinois at Urbana-Champaign, USA
[2] Oxford e-Research Centre, Department of Engineering Science, Oxford, UK
hburns2@illinois.edu

Abstract. In data management, the use of identifiers is essential for disambiguation and referencing. The scope of the use of identifiers varies. For example, disambiguation within an institution using integer identifiers may be sufficient for operational procedures, whereas digital scholarship using global resources relies on universally unique identifiers. In this paper we investigate practical routes to globally unique identifiers for the medieval manuscripts of the Bodleian Library. The Oxford Linked Open Data (OXLOD) and Mapping Manuscript Migrations (MMM) projects require unique identifiers for the transformation of the medieval manuscripts catalogue into linked data, in an effort to increase discoverability and consistency across platforms. We consider how Archival Resource Keys (ARKs), a type of URI, can be applied to the Medieval Manuscript catalog as well as determining how ARKs can support MMM's research goals. We begin with examining the Text Encoding Initiative (TEI) catalogue records to understand the data provided and identify and describe entities which do not presently have identifiers. Further, we evaluate ARKs for producing identifiers, prioritizing those which are required to answer common research questions.

Keywords: Archival Resource Keys, Linked Open Data, Metadata, Medieval Manuscripts.

1 Introduction

We are motivated by the Mapping Manuscript Migrations (MMM) Digging into Data Challenge, which seeks "to combine data from various sources to enable the large-scale analysis of the history and provenance of medieval and Renaissance manuscripts" [1]. MMM poses research questions drawing on the power of linked data as a method of enabling this analysis. Answering these requires disambiguating identifiers, particularly those related to changes in ownership and location of manuscripts. The Bodleian's Medieval Manuscript catalog is one of the key data sources for the MMM project.

Archival Resource Keys (ARKs) are considered by the Bodleian as a possible solution to tracking and management of manuscripts through different systems. This investigation looked at the suitability of ARKs for the broader set of requirements identified by MMM, and examined how they could be implemented in the data transformation and mapping processes of the MMM project.

2 Identifiers for Medieval Manuscripts

2.1 Tracking Provenance Information through Linked Data Mapping

The Bodleian catalogue files, encoded in TEI, can be transformed into a simpler, bespoke XML format, using XQuery. MMM (together with the related Oxford Linked Open Data project, OXLOD [2]) uses the mapping tool 3M [3], to transform the simpler XML into RDF triples, based primarily on two ontologies: CIDOC-CRM (for cultural heritage data) [4] and FRBRoo (for bibliographic information) [5].

This is where history- and provenance-based identifiers are needed. Historical information within the TEI documents is nested under a single "provenance" element. Limited disambiguation occurs within this element (see Fig. 1.).

```
<provenance>
   <persName role="formerOwner" key="person_37714110">
      William Greenwell
   </persName> (d. 1918)</provenance>
<provenance>Sotheby's 15 March 1907 and following day, lot 350, bought by
Leighton for £3 8s.</provenance>
```

Fig. 1. A provenance element from the Medieval Manuscript TEI metadata.

Examining the TEI record alone, it is impossible to disambiguate between separate provenance events. During the transformation from TEI to simplified XML, several elements are created within the provenance field, which are necessary for MMM to answer provenance questions and increases the granularity of the mapping in 3M. Fig. 2. shows the data from Fig. 1. represented in simplified XML. Additional elements are added to break the event into segments useable for 3M mapping. The person in "manuscript_8811_prov4" now has a specific URI associated with them, as well as a label, and a role. The event also possesses a note, which reads "William Greenwell (d. 1918)."

```
<provenance xml:id="manuscript_6811_prov4">
   <person>
      <uri>https://medieval.bodleian.ox.ac.uk/catalog/person_37714110</uri>
      <label>William Greenwell</label>
      <role>formerOwner</role>
   </person>
   <text>William Greenwell (d. 1918)</text>
</provenance>
```

Fig. 2. Simplified XML for the same provenance event.

2.2 What is an ARK?

An ARK is a persistent identifier meant to reference an "information object" and is made up of "a sequence of characters that contains the label, 'ark:'" [6]. Information objects can be physical, such as a book or intangible, such as a concept or a performance [7]. Unlike many other identifiers ARKs come with a promise of stewardship - an ARK must be maintained and should always resolve to the actual object being identified (or a copy of one) [7].

3 Incorporating ARKs into the Mapping Process

We identified which CRM and FRBR classes were necessary for answering the MMM research questions. We then listed which elements were minted in 3M as part of the mapping process and which were already provided by the simplified XML. This list (see Table 1.) was used to develop potential solutions MMM could employ when minting identifiers.

Table 1. This table shows the elements required for MMM research questions tracked through the mapping process from TEI to 3M

TEI Element	3M Element	Example IDs from TEI
origDate/@notAfter	E12_Production	N/A
	E52_Time-Span	N/A
acquisition//date/@notBefore	E10_Transfer_of_Custody	N/A
	E52_Time-Span	N/A
TEI	E24_Physical_Man-Made_Thing	xml:id="manuscript_4165"
msPart	E24_Physical_Man-Made_Thing	xml:id="MS_Barocci_10-part1"
msItem	F22_Self-Contained_Expression	xml:id="MS_Barocci_10-part1-item1"
TEI/text/body/listBibl/bibl/author	F27_Work_Conception	N/A
provenance	E5_Event	N/A
titleStmt/title[@type='collection']	E78_Collection	N/A

3.1 Solution 1: ARKs for all necessary entities

The first solution involves creating ARKs for all possible entities needed for the transformation to RDF. This could be extremely powerful for researchers, since the large number of identifiers provided allows for a high degree of specificity in referencing these entities, however the thousands of event-based ARKs generated, would each need to be maintained and resolved. Additionally, the original XML documents would need to be modified to include the ARKs. As a result, implementation is unlikely due to the level of maintenance required.

3.2 Solution 2: Manuscript-only ARKs with TEI alteration

Solution two involves creating an ARK for the manuscript, with the rest of the identifiers created within the TEI document as built-in *xml:id* fields. This would require the TEI to be altered to support additional fields (example IDs for unmodified TEI records can be found in the third column of Table 1.). These new IDs would need to refer to another source, similar to the authority files that the Bodleian Libraries currently possess for people or place identifiers. This solution is the easiest from a development standpoint, as only one ARK would technically be maintained and resolved for each manuscript. However, this scenario does require reworking the current TEI metadata and the creation and maintenance of additional authority files, requiring extensive time and extra resources.

3.3 Solution 3: Manuscript-only ARKs

This solution involves creating ARKs only for the manuscript. Remaining identifiers needed for disambiguation will be produced by 3M, mirroring the current process. As with Solution 2, the only ARKs being maintained would be those for each manuscript. ARKs would not be a reasonable method for identifying provenance, as MMM cannot currently guarantee stewardship after the conclusion of the project. Additionally, ARKs for provenance events are unlikely because each identifier would need to resolve, requiring the creation of additional webpages that would also need to be maintained. While MMM could still use this data for the research questions, it would be otherwise inaccessible online, limiting its broader application.

4 Conclusions

Considering available resources, we conclude that Scenario 3 would work best in the short-term, though it is not ideal from a long-term linked data perspective. Moving forward, it is essential for MMM to determine what will happen to their data once the project is complete, which will help inform the type of identifier they might wish to use. Solution 3 would also meet the Bodleian Libraries' specific current requirements in the short term. But it would postpone addressing the wider issues relating to identifiers for TEI elements other than the manuscripts themselves.

It is important to note that this is a case study and not necessarily a blueprint for how other libraries should operate. If a library was looking to implement ARKs in the context of TEI-based metadata, the specific solutions described here could be adapted into an initial template. More generally our work exemplifies a methodology for assessing the practicality of adapting library identifiers to research studies, an approach which we hope to generalize in the future across the technologies used by all the MMM partners, not just the Bodleian.

References

1. Burrows, T., Hyvönen, E., Ransom, L., Wijsman, H.: Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts. University of Pennsylvania Press, Volume 3, Number 1, Spring 2018, 249–252 (2018).
2. OXLOD Homepage, https://www.glam.ox.ac.uk/oxford-linked-open-data-pilot, last accessed 2018/09/03.
3. 3M Homepage, http://139.91.183.3/3M/, last accessed 2018/08/09.
4. CIDOC-CRM Homepage, http://www.cidoc-crm.org/, last accessed 2018/07/28.
5. Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism, https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf, last accessed 2018/07/28.
6. N2T Archival Resource Key (ARK) Identifiers, http://n2t.net/e/ark_ids.html, last accessed 2018/08/08.

7. Kunze, J., & Rodgers, R.: The ARK Identifier Scheme. Berkley Planning Journal, https://escholarship.org/uc/item/9p9863nc, last accessed 2018/08/09.