Overview (first section): 150 words (150 limit)
Overall abstract (including first section): 1634 (1500 limit)

A layered digital library for cataloguing and research: practical experiences with medieval manuscripts, from TEI to Linked Data

In this paper we report our experiences developing and applying a set of digital infrastructure elements which, in combination, realise a layered digital library (Page et al 2017) for the investigation of manuscript provenance.

We describe several related technical contributions: (i) encoding of manuscript catalogue and local authority records as TEI; (ii) using Github for version control, issue tracking, and collaboration; (iii) automated production of catalogue user interfaces derived from the TEI; (iv) an XML processing workflow identifying, extracting, and processing TEI elements for reuse in research; (v) mapping workflow output into a CIDOC-CRM RDF export; (vi) reconciliation of RDF entities with external authorities enabling the creation and use of Linked Data bridging multiple datasets.

We contextualise the co-evolution of these components and exemplify their use in studies of the provenance of medieval manuscripts. We reflect on the flexibility and extensibility provided by our layered approach, and the independent benefits for catalogers and scholars.

Catalogue implementation and Linked Data workflow

The foundation layer of the approach described herein is the TEI encoding of manuscript metadata undertaken by the University of [redacted] Library. TEI has previously been used to encode text-based catalogues of medieval and renaissance manuscripts[1], and we briefly reference the particular problems and solutions posed for the [redacted] previously described elsewhere ([rref1]).

The digital catalogue records are mostly derived from earlier printed catalogues, especially the Quarto and Summary Catalogues published between 1853 and 1924, though in many cases they have been enhanced and updated for the digital catalogue. More than 9,200 Western medieval manuscripts are described.

[1] https://wiki.tei-c.org/index.php/TEI_manuscript_catalogues

The TEI XML format was chosen for this detailed cataloguing of manuscripts because of its rich and flexible syntax: it can encode a complete retrospective conversion of existing catalogue description texts, adding structured markup of specific concepts and identifiers adapted to the various formats of historical catalogues, while allowing a variable degree of comprehensive or selective markup as required or desired.

The [redacted] catalogue records are implemented using a customisation of the TEI P5 manuscript description module[2], schema with minor variations for Western, Islamic and Oriental manuscripts. Significant effort has been invested in the creation of local authority files for works, people and places, also using TEI. These local authorities have been, in turn, manually reconciled with URIs of records in external authorities such as VIAF, Library of Congress, Bibliothèque nationale de France, Système Universitaire de Documentation, Gemeinsame Normdatei, and WikiData.

TEI records are created and edited in the Oxygen editor, and stored in repositories[3] using the Git version control system. These repositories are hosted on GitHub, which also provides for issue tracking and collaboration - requests for modifications or additional markup in support of researcher investigations, such as that described in this paper, can be added, trialled, reverted, or otherwise properly and consistently managed without negatively impacting on the traditional library functions of the catalogue.

This TEI layer is, therefore, focussed purely on the creation and maintenance of the XML record files, with appropriate support tools and functionality, and which could easily be transferred to an equivalent or improved file repository systems should the need arise.

While the TEI records are freely and openly available via GitHub, the primary interface for library users is a Medieval Manuscripts collections website[4], where the full gamut of traditional searching, browsing and viewing functionalities are provided. The website is built using open source technologies including XSLT, XQuery, Solr and Blacklight providing a user interface. Since the website is generated atop a version controlled check out of the TEI catalogue layer, it too can be developed, improved independently of the other layers.

A further benefit flowing from this separation of concerns is the ability to create parallel specialised data layers targeted towards distinct areas of research, which supplement rather than supplant the canonical TEI catalogue and its core digital library functions. Here the flexibility and adaptability offered by TEI is excessive, even potentially counterproductive in its complexity for computational processing of the catalogue metadata.

Instead, to answer specific research questions, we desire reasoning over logically consistent relationships such as those defined by the CIDOC Conceptual Reference Model[5], and the

[2] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html
[3] [redacted]
[4] [redacted]
[5] The CIDOC-CRM http://cidoc.ics.forth.gr/

ability to cross-reference multiple corpora and authorities using Linked Data, for which we create a selective RDF layer derived from the catalogue metadata. While in theory the RDF conversion could be comprehensive, including all available TEI elements and attributes, in practice the mapping process is detailed and complex, and so scoping the conversion according to the investigation at hand engenders progress; RDF complements this approach by providing data structures well suited to future extensions retaining consistency.

The first stage of processing simplifies the TEI records, extracting the pertinent information (manuscripts, parts, works, authors and other people, places) in a more rigid XML structure, and implemented as an XQuery[6]. This transforms the source records into a single file, conforming to a structured list of desired metadata fields, normalising some data (e.g. languages), referencing authority files, and building URIs.

The second stage of processing ingests the simplified XML into the the 3M mapping tool (Oldman, Theodoridou and Samaritakis, 2010) for transformation to a data model combining entities and relationships from CIDOC-CRM and FRBRoo[7]. Here entities are also reconciled with local and external authorities, and RDF is exported ready for querying against research questions.

In creating these two alternate versions, or layers, atop the TEI encoded catalogue, we can serve several distinct but complementary motivations: a robust, maintainable and consistent record system for cataloguing; a visible and discoverable interface for browsing and searching the catalogue; and a malleable data structure for detailed scholarly investigation. These parallel the affordances offered by the encodings used in each layers, deploying TEI and RDF (and, similarly Solr/Blacklight) according to their strengths. In the remainder of the paper we focus on the last of these motivations, detailing the use of the digital library for research into manuscript provenance.

Application to manuscript provenance research

Having described the infrastructural components and overall workflow, we spend the remainder of the paper demonstrating the use of novel digital library for research into the provenance of medieval manuscripts: their origins and movements, and the collectors and owners involved in their history.

As the result of changes in ownership over the centuries, European manuscripts are now spread all over the world in diverse library, museum and gallery collections. Information relating to their often complicated histories is dispersed and fragmented across numerous sources, compelling historians and other researchers to make painstaking and time-consuming searches of printed and online catalogues. Digital tools which can assist in these searches and record their outcomes are of great benefit; cross-referencing and reconciliation across catalogues even moreso.

[6] [redacted]

[7] An object-oriented version of FRBR harmonized with CIDOC-CRM

As such, our ultimate aim is to search across multiple distributed catalogues[8,9], using ontologies to describe conceptual equivalencies and indirect relationships, overcoming differences in underlying catalogue structures, and so enabling unified searching and interrogation. Here, however, we constrain discussion to the completed implementation at the University of [redacted], noting it provides a template for creating equivalent layers over other catalogue systems, and that the queries below would be equally applicable to a combined search across multiple collections[10].

There has been little previous work transforming TEI manuscript catalogues into RDF suitable for the combined data explorations described here. The Medieval Electronic Scholarly Alliance (MESA) has published samples of its transformations from the Walters Art Museum into the Dublin Core based schema[11] of the Advanced Research Consortium (ARC); while Compton and Schwartz (2019) outline the general motivations and benefits of TEI to RDF conversion., based primarily on Dublin Core.

For our modelling, we began by considering the TEI markup for the manuscripts records themselves, which can be complex and hierarchical, often describing a manuscript divided into several parts, each with its own history and containing works-within-works (e.g. a collection of poetry and individual poems). Information about the provenance of a manuscript is sometimes encoded with a single XML element describing the entire history of the manuscript, and sometimes as multiple elements each recounting one event in that history. Dates might be encoded with ‘date’ tags or attributes on the ‘provenance’ element, and so on.

Given this inherent complexity within the data, we needed to identify a more limited number of ‘frames of reference’ to practically scope our RDF conversion. Consulting with other manuscript scholars identified archetypal questions which would be required of any data investigation. We include an illustrative selection of these queries here, which will be referenced to the TEI, simplified XML, 3M mapping, RDF. and SPARQL queries implemented to achieve them:

- How many manuscripts survive that contain Spanish texts written in gothic rotunda were produced in Castile for an abbey or convent?
    - Of these, which were owned during the nineteenth century by English private collectors?
    - Which are now owned by an institution in North America?
- What French collectors purchased manuscripts since the end of the Wars of Religion (after 1598)?

[8] [redacted]

[9] Furthermore, the RDF described here for our manuscript catalogues has already formed part of Linked Data network combining records from the gardens, libraries and museums of the University of [redacted] as part of the [redacted] project, including the [redacted].

[10] Indeed this reusability and extensibility is one of our primary motivations for using Linked Data.

[11] http://wiki.collex.org/index.php/Submitting_RDF

Our focus on manuscript provenance and associated research questions scoped our choice of elements and attributes to include in the intermediate simplified XML, selecting necessary entities, cross-referencing information from the local authority files, and creating the URIs required for the next stage of processing. The authority files themselves, being essentially flat lists, could be mapped in the 3M tool directly.

Within the 3M tool, we give examples of mapping from the TEI customisation to CIDOC-CRM and FRBRoo[12], taking care to separate evidence derived directly from the text from that which embeds institutional knowledge (i.e. inscriptions require interpretation).

Finally, we provide examples of SPARQL resolving the research questions above, paying attention to how a researcher can take advantage of data semantics to overcome complexities not immediately apparent in natural language statements of the query. For example, breaking apart a query to retrieve “manuscripts from 1550-1600 produced in European countries” entails reasoning variable temporal constructs, mapping and resolution of external spatial definitions (Getty, wikidata).

References

[rref1] [redacted]

Compton, C. and Schwartz, M. 2018. ‘More Than “Nice to Have”: TEI-to-Linked Data Conversion’. Digital Humanities 2018.

Oldman, D., Theodoridou, M., and Samaritakis, G. 2010. Using Mapping Memory Manager (3M) with CIDOC CRM. Version 4g.
http://83.212.168.219/DariahCrete/sites/default/files/mapping_manual_version_4g.pdf

Page, K.R., Bechhofer, S., Fazekas, G, Weigl D.M, and Wilmering, T. 2017. ‘Realising a layered digital library: exploration and analysis of the live music archive through linked data’. Proc. 17th ACM/IEEE Joint Conference on Digital Libraries, pp.89-98.

Acknowledgments

[redacted]

[12] For example, from the catalogue work item (TEI `bibl`) to F1 Work; `msItem` to `F22 Self-contained Expression`; and so on.