Digital Scriptorium 2.0 Environmental Scan
March 10, 2021

This environmental scan investigates current online cataloging projects to identify trends in description practices, controlled vocabularies, and the use of linked open data in the cultural heritage sector. Understanding how related projects manage their data collection and authority management will direct the Digital Scriptorium (DS) 2.0 implementation plan towards practices that will make the new platform more interoperable and sustainable. A variety of online catalogs, digital libraries, and research projects were included in the scan, all of which document the history of the book and most with a specific focus on manuscript objects:

15thc Book Trade - use material evidence of 15thc book trade to answer questions related to introduction of printing in the West (2014-2019)

Al-Furqan Digital Library - digital library of Al-Furqan Foundation, dedicated to preserving and studying Islamic written heritage (2013-present)

Biblissima - a virtual library of libraries: a discovery portal for the history of various texts and books that were written, translated, illuminated, collected and catalogued from Classical Antiquity through the 18th century (2017-present)

E-codices - virtual manuscript library of Switzerland (2005-present)

Europeana - Europe's digital library, museum, gallery and archive, providing online access to a vast store of cultural heritage material from across Europe (2008-present)

Firhist - union catalogue of manuscripts from the Islamicate world in UK institutions (2009-present)

Footprints - traces history and movement of Jewish books since inception of print (2014-present)

Handschriftenportal - central web portal for the manuscript collections of German institutions (2018-present)

Islamic Scientific Manuscripts Initiative (ISMI) - make accessible information on all Islamic manuscripts in the exact sciences (astronomy, mathematics, optics, mathematical geography, music, mechanics, and related disciplines), whether in Arabic, Persian, Turkish, or other languages (2018-present)

ManusOnline - database containing catalogue descriptions and digital images of manuscripts, private papers and archives held by Italian public, private and ecclesiastical libraries. (1988-present)

Medieval Manuscripts in Dutch Collections (MMDC) - contains descriptions of all medieval western manuscripts up to c. 1550 written in Latin script and preserved in public and semi-public collections in the Netherlands (2007-present)

Medieval Manuscripts in Flemish Collections (MMFC) - a complete, authoritative database of medieval manuscripts from the period 600 - 1600 that are kept in Flemish collections (2020-present)

Mapping Manuscript Migrations (MMM) - a semantic portal for finding and studying pre-modern manuscripts and their movements, based on linked collections of the Schoenberg Institute for Manuscript Studies, the Bodleian Libraries, and the Institut de recherche et d'histoire des textes. (2017-2020)

Material Evidence in Incunabula (MEI) - database of 15th-century printed books (2010-present)

Southeast Asia Digital Library - provides free access to archives of textual, still image, sound, and video resources, covering both historical and current information from Southeast Asia (2005-present)

Trends in Description Practices

At the Digital Scriptorium 2.0 Planning Meetings held in October 2020, a major point of discussion was the level of detail required for a manuscript description. How many data fields should a DS 2.0 record contain? Is it better to have briefer descriptions that are easier to manage, or to create as detailed descriptions as possible to capture the widest range of information for researchers?

To understand how other manuscript projects approach this question, this scan analyzed the data fields available in nine online manuscript cataloging projects: the Al-Furqan Digital Library, Biblissima, e-codices, Firhist, Handschriftenportal, ManusOnline, Medieval Manuscripts in Dutch Collections (MMDC), Medieval Manuscripts in Flemish Collections (MMFC), and the

Southeast Asia Digital Library. These projects all provide a single access point for manuscripts cataloged and housed at institutions across national or international institutions and include a range of material from the Middle East, Asia, and Western Europe. Each of these projects have a data model and schema that serves their own unique purposes. While many similar data fields appear across these projects, no one schema perfectly matches another. Still, common data fields did appear.

The majority of projects include roughly 20-30 data fields in their manuscript descriptions, though one very ambitious new project, MMFC, lists a potential set of over 150 data elements on their website. Every project describes the title, production date, and production place of the manuscripts in their dataset. Neary every project also includes information about the manuscript's current holding institution and shelfmark, the specific collection to which the manuscript belongs, subjects, human contributors such as authors and scribes, languages, provenance information, and general notes related to physical description. Most also managed specific fields related to genre, artist, material, extent, dimensions, binding, incipit/explicit, and references to the manuscript in other catalogs or scholarly literature.

No major differences in data fields appear between records for manuscripts produced in different geographic and cultural contexts. Firhist employs a TEI standard that is similar the one used by e-codices and HSP, though adapted for Islamic manuscript description by including instructions for encoding the various components of an Arabic name, for transliterating Arabic titles into Roman characters, and for the description of various scripts. The Al-Furqan Digital Library, though not using the same standard as Firhist, includes many similar fields dedicated to textual description. The Southeast Asia Digital Library has the fewest data fields among all of the projects in the scan, with a data model that aligns very closely with the Dublin Core. Its data includes manuscripts from a wide variety of geographies, cultures, and religions, which may necessitate a need for simplicity in the data model in order to accommodate this variability.

It should be noted that this analysis is based on information that is openly available on each project's website. Many of these projects are in development, and some documentation may be unavailable to the public. In cases where a data dictionary or a description of the data schema was not available, sample records were examined to determine the available fields. It is entirely possible that some fields were missed due to this strategy for gathering information. This exercise is not meant to be exhaustive in its analysis, but rather to demonstrate general trends in description practices.

Authority Control

Authority control provides standardization within data fields, which allows for more reliable search results and indexing. Rather than searching for every variant spelling of an individual's name, for example, a user only needs to search with the standard form of the name to return all records associated with that person. Authority control can also improve standardization across platforms when different projects use the same controlled vocabularies to reference the same things. This is a great way to find connections between different manuscript catalogs and between different types of datasets.

All of the projects included this scan employed some type of authority control to certain data fields, though the degree to which they use the same controlled vocabularies is limited. Some of the older catalogs have analog authority files that don't correspond to standards set outside of their own datasets. This is the case in the ManusOnline, MMDC, and Islamic Scientific Manuscripts Initiative (ISMI), which provide searchable lists of standardized names used in their respective databases. These names can help human users as they navigate the databases, but they are not structured in a way that other projects can use or refer to them. The files are not available for download and do not contain unique identifiers to create stable references, therefore their computational use is limited.

Of the projects that do link to outside authorities, practices vary. Records produced in German-speaking areas such as e-codices and HSP favor the use of the German Integrated Authority File (GND). Firhist uses VIAF in their name authority and the Library of Congress Subject Headings for subject control, and also follows the Library of Congress's standards for transliteration of non-Roman scripts. MMFC also takes advantage of links to VIAF as well as the CERL Thesaurus, which harmonizes authority records from separate libraries in a similar way. Printed book projects like MEI, the 15thc Book Trade, and Footprints link to a range of similar controlled vocabularies as those seen in manuscript descriptions, including the Library of Congress Authorities, VIAF, and the CERL Thesaurus. Biblissima takes on the role of aggregation itself, managing authorities for persons, places, institutions, and even shelfmark references by harmonizing records across a variety of linked open data repositories including VIAF, CERL, Wikidata, GeoNames, and national libraries such as the Bibliothèque nationale de France, GND, and the Library of Congress.

As DS 2.0 considers how to best align its authority control with international practices, it is clear that there is no single set of standards to follow. Every project in this scan uses authorities that best suit its particular dataset and objectives. Only one of the four projects dealing explicitly with non-Western manuscripts attempt to link their authorized names to any outside authorities, pointing to a lack of representation for identities of people outside of Western European contexts in the controlled vocabularies used in other projects. Aggregated authority files such as VIAF, CERL, Wikidata, and Biblissima provide the greatest opportunities for linking, though they will certainly not offer comprehensive coverage for every identity or location referenced in DS 2.0's diverse dataset. Producing stable references for DS 2.0's internal authorities, and linking them to other authorities when possible, is the best way to ensure the accessibility and interoperability of DS 2.0. This can be accomplished by using linked open data.

Linked Open Data

Linked open data (LOD) is a set of specifications for publishing structured data on the internet, encoding the meaning of the information in a way that can be best utilized by computers. Data published as LOD uses uniform resource identifiers (URIs) to create permalinks for every data point, which other projects can link to in their datasets when they describe the same thing. LOD essentially turns the internet into a web of interconnected data, rather than a web of text files whose meaning can only be understood by human readers. The standardization and stability that LOD provides is especially useful in authority files, and it is no surprise that these specifications compose the technical infrastructure for Wikidata, GeoNames, and the Biblissima authority files.

Of the nine cataloging projects examined in depth for their current description practices, only Biblissima currently exposes their data as LOD, though HSP and MMFC plan to do so in later versions of their projects. Of the entirety of the manuscript projects and catalogs included in the whole scan, only Europeana, the Islamic Scientific Manuscripts Initiative, and Mapping Manuscript Migrations are added to this number. There are multiple reasons that LOD is not yet present in every project. LOD was first specified in 2006, and many of the datasets examined in this scan were established before that time. Many of these projects also encode their manuscript descriptions using TEI/XML, which though useful for text description is not an LOD standard. These records would have to be migrated to a structured data format such as RDF to be usable in the LOD environment.

It is telling that all of the projects in this scan that do or plan to utilize LOD are recipients of recent grant funding. LOD specifications and practices offer the best opportunity for accessible, linkable, and persistent data in our current technological environment. The DS 2.0 platform must produce LOD in order to fulfill its mission of providing open access to manuscript data.

Conclusion

This environmental scan has revealed some of the digital strategies that DS 2.0 should adopt to ensure that its new platform is as interoperable, sustainable, and useful as possible. Data fields that commonly appear in the manuscript descriptions of other projects will be included in the new DS 2.0 data model. This does not mean that DS 2.0's model will perfectly align with others, as the scan demonstrated that each project's schema will necessarily be different from others as it supports the specific needs of its own user community. DS 2.0 will prioritize authority control in its descriptions to enhance the interoperability of its dataset with other projects, harnessing the power of linked open data to achieve these goals. While no single controlled vocabulary appears to dominate the cultural heritage landscape, the aggregate files managed by services like Wikidata and Biblissima offer the greatest opportunities for reliable data linkage. Harnessing these latest technologies and policies will lay the groundwork for a robust new iteration of DS.