Digital Scriptorium 2.0 Project Report II
February 17, 2021

Introduction

In July 2020, the IMLS awarded Penn Libraries a $100,000 National Leadership for Libraries Grant to oversee a one-year planning period for the future of Digital Scriptorium (DS), a consortium of US institutions who are dedicated to providing open access to their manuscript data and images. Since 1997, DS has hosted an online platform and database to facilitate this mission, but its purpose and technical infrastructure needs redevelopment in order to remain viable in the 21st century. The project team has been working throughout the grant period to refine the purpose and scope of the DS platform in order to ensure its technical and financial sustainability in the years to come. DS's ultimate goal is to become the national union catalogue of pre-modern manuscripts held in US collections.

Six months into the planning period, we have continued our weekly technical and advisory meetings for the project, completed an assessment of the current data within the DS platform, scanned related digital projects across the cultural heritage sector, and begun development of the DS 2.0 data model. Building upon the work completed during the initial months of the project, we are on track for a successful completion of the planning grant in July 2021. The project will culminate with a plan for DS 2.0 implementation presented to the DS Board of Directors.

Current Data Assessment

The DS platform has hosted manuscript data and images since 1997, building a database that now consists of 8543 manuscripts and 76101 images. Migrating the entirety of this dataset into the new DS platform may be an inefficient course of action, since (hypothetically) much of this legacy data has been superseded by more recent cataloging at the institutional level in the intervening years. Unfortunately, DS's current database platform does not store timestamps for any of the data or images within the system. This means that the database is unable to tell us when a manuscript record was created or when it was last updated. Instead, we contacted every DS member and associate institution with data in the platform to ask if they have any records about when they last updated their information. We also asked if these institutions publish their manuscript data in any other digital platform. 32 out of 33 current DS member or associate institutions responded to our request for information. We are grateful for the time and effort many of our consortium members spent to contribute information for this assessment. What follows is a brief summary of the information we received. A full report of the Current Data Assessment is available on the DS website.

Only 5 institutions have updated their data within the last 5 years. An additional 8 have definitely updated their data within the last 10 years. 3 institutions could not provide any estimate at all. This means that only about 40% of DS contributors have certainly updated their information within the last decade. 20 institutions do provide access to their manuscripts on other digital platforms, often via MARC records in their online public access catalogs. 13 institutions responded that they are dependent on DS as the sole access point for their manuscript data and images. This amounts to 1685 manuscript records (20% of the total manuscripts in DS) and 5497 individual images (7% of the total images in DS).

The 13 institutions who are reliant on DS as the sole digital access point for their manuscripts will need their data and images transferred from the current system into DS 2.0. While crosswalking the manuscript data should be a reasonably straightforward process, the images are a more complicated problem. The current DS platform hosts images on behalf of its consortium members, but this is not financially or technologically sustainable. The DS 2.0 project team is considering a variety of other options to facilitate a hosting solution for DS consortium members who cannot store their own images.

Environmental Scan

The environmental scan aims to survey current online cataloging projects to identify trends in description practices, controlled vocabularies, and linked open data in the cultural heritage sector. Understanding how related projects manage their data collection, authority management, and financial and technical infrastructure will help direct the DS 2.0 implementation plan towards practices and policies that will make our project more sustainable. By aligning with commonly used ontologies, for example, persistent manuscript identifiers managed within the DS 2.0 platform will be easier to integrate with other datasets. This type of interoperability will ensure that the knowledge produced by DS 2.0 will reach the widest audience possible.

Our scan initially focused on national online manuscript cataloging projects including e-codices (Switzerland), ManusOnline (Italy), Handschriftenportal (Germany), Firhist (UK), Medieval Manuscripts in Flemish Collections, and Medieval Manuscripts in Dutch Collections. These projects all aim to provide a union catalogue of manuscripts held by institutions within their nation. Projects with an international scope were also included, such as Biblissima, Europeana, Mapping Manuscript Migrations, the Southeast Asia Digital Library, and the International Standard Manuscript Identifier (ISMI) project. We have also investigated projects related to the history of printed books, including Footprints, the Incunabula Short Title Catalogue, and the 15c Book Trade project. Since Wikibase has been identified as a potential database platform for DS 2.0, we have also scanned other projects built using this software, including the Biblissima and GND authority files, the Factgrid database of historians, and the Enslaved project.

This research has helped to clarify the unique position DS 2.0's mission and technological framework will occupy in the cultural heritage sector. Where most national cataloging projects are focused on creating detailed and authoritative descriptions of manuscripts vetted by scholars, DS 2.0 will be more similar to a finding aid that offers streamlined records. This style of efficient description is more often seen at the international project level, such as in the Biblissima and Europeana manuscript collections. DS 2.0 will also prioritize authority controlled fields and the interoperability of its dataset with other projects, harnessing the power of linked open data to achieve these goals. Since linked open data is a relatively recent development in information technology, many of the projects examined in our scan were not designed according to its specifications. The environmental scan continues to progress along with the data model development, as we look for examples of ontology and vocabulary alignments that may help DS become more interoperable with other projects. A full report on the environmental scan will be published on the DS website.

Data Model Development

The development of the DS 2.0 data model began in January 2021. Mikko Koho, computer scientist and postdoctoral researcher at Aalto University, has been contracted to design the data model in consultation with the rest of the DS 2.0 project team. We have identified roughly 20 potential data fields that are integral to manuscript description and discoverability, which have arisen out of a comparison of existing manuscript catalog records across institutions and platforms. We are in the process of identifying ontology alignments that might be harnessed to ensure that the model is as interoperable as possible, such as Dublin Core, CIDOC-CRM, and the Getty Vocabularies. Since the data model is document-based and relatively simple, the project team hopes that its development will be concluded within the next few months.

Conclusion and Looking Ahead

The remaining months of the planning grant will focus on the completion of the DS 2.0 data model and the development of a DS 2.0 prototype platform. This prototype will help us test our conception of the DS 2.0 contribution workflow and demonstrate the viability of the project to potential funding agencies and DS consortium members. An additional stakeholder survey will be distributed to DS members and associates to gather feedback on the final implementation plans and the needs of the DS consortium going forward. The project team will host a stakeholder meeting in April or May to present the implementation plan and DS 2.0 data model, with a second meeting later this summer to demonstrate the DS 2.0 prototype. A final DS 2.0 Project Report will be published in May 2021.