Digital Scriptorium 2.0 Project Report IV
December 2, 2021

Introduction

In July 2020, the Institute of Museum and Library Services (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 worked 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.

Having completed the initial one-year planning period, the DS 2.0 project has now entered its its Bridge Year phase of implementation, which will run from August 2021-June 2022. The Bridge Year marks the interim between the planning grant funded exclusively by the IMLS from June 2020-June 2021, and the next round of grant funding which will hopefully begin in August 2022. The Bridge Year is funded in part with leftover grant funds from the original IMLS grant, which stayed under-budget due to the ongoing restrictions on travel and in-person meetings caused by the pandemic. Additionally, several DS member institutions generously contributed funds to support project management for DS 2.0 over the coming year. The bulk of the work over the Bridge Year will be dedicated to developing the prototype, beta, and alpha versions of the new DS 2.0 database.

DS 2.0 Prototype Database Progress

Technical problems relating to a security issue at Penn Libraries created significant delays in the DS 2.0 prototype database development over the summer. Originally scheduled for completion in September, prototype testing is on track to conclude by the end of 2021. The database is now up and running at Penn, developed using the DS 2.0 data model and schema produced during the planning year. A selection of sample records have been created within the prototype, resulting in ongoing changes to the data model and schema as they are adapted to the underlying Wikibase software that supports the database. Figure 1 shows one of these sample records. This is data that was harvested from a MARC record for UPenn LJS 102. The record includes information about the title of the manuscript, its language, and a test version of a unique DS identifier for this manuscript. The DS ID featured in this test record is merely a placeholder until actual DS IDs are generated. The DS 2.0 project team is currently investigating strategies for the creation of DS IDs. This process will be tested as a part of the DS 2.0 data transformation process.

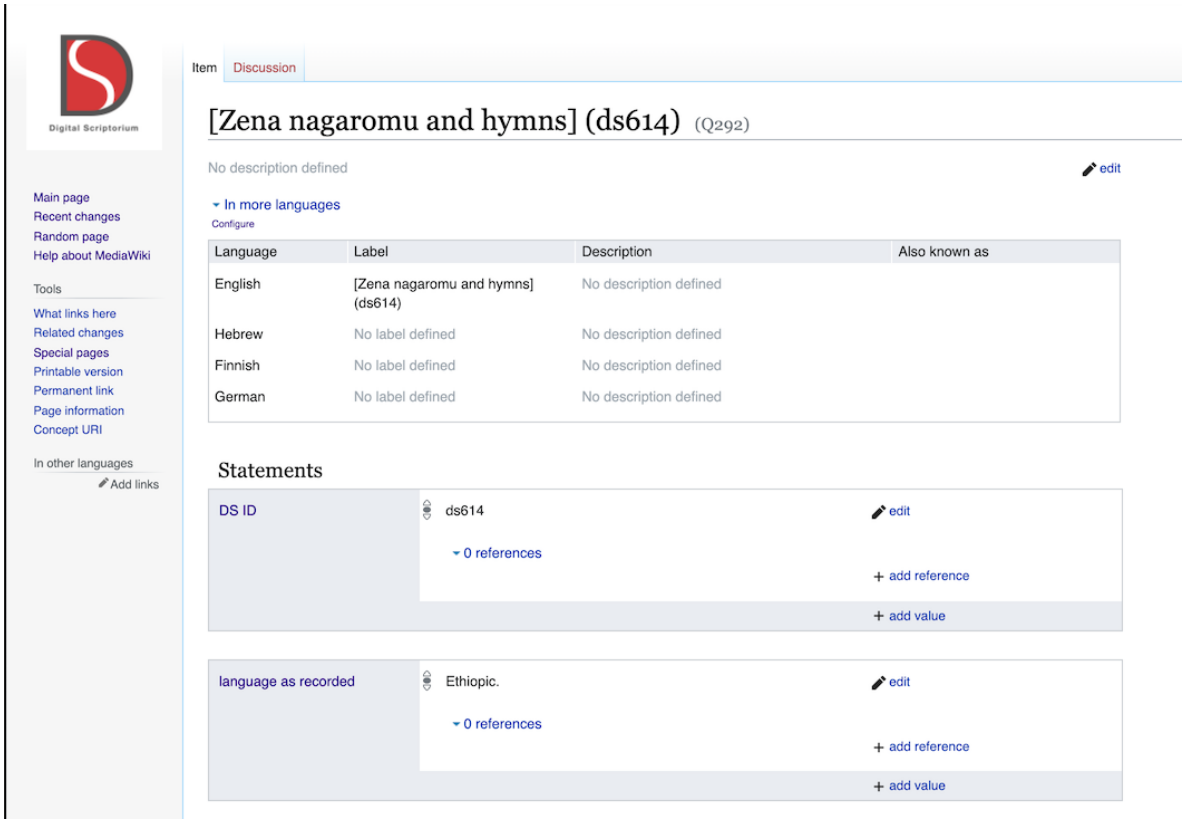
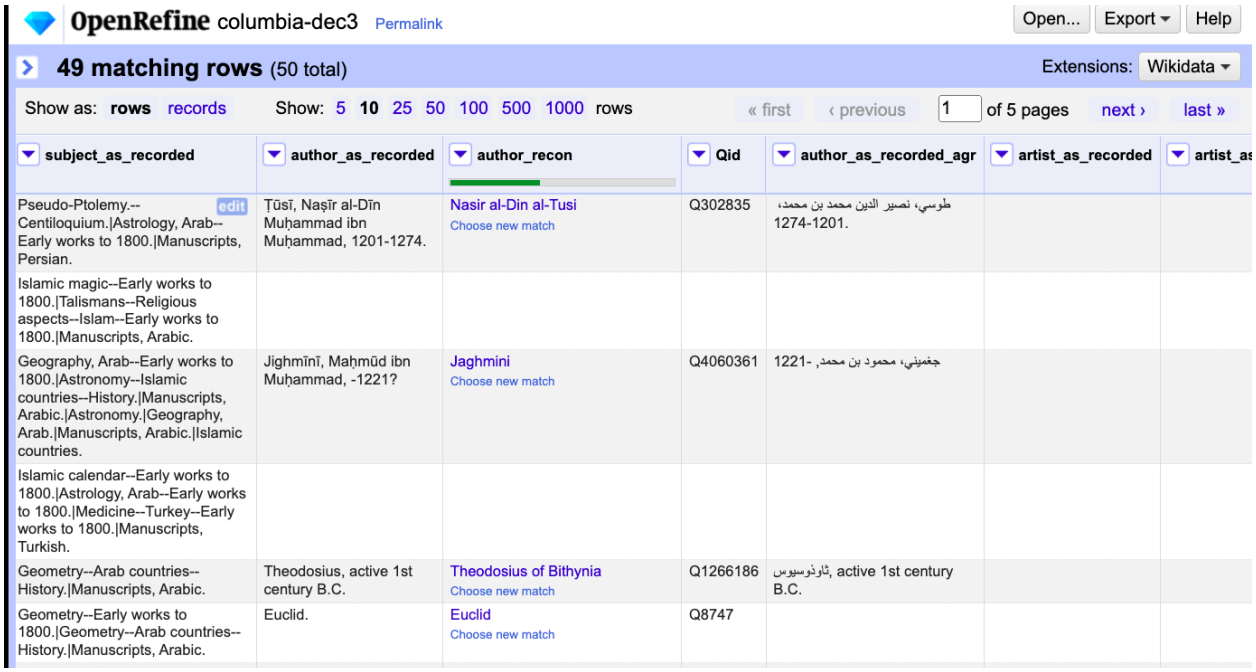


Figure 1: A test manuscript record in the DS 2.0 prototype database.

The Wikibase prototype record shown in Figure 1 demonstrates how the backend of the DS 2.0 database will appear to DS administrators. A separate user portal will begin development in 2022 to provide a more broadly accessible interface for users to browse and search the DS 2.0 database.

DS 2.0 Data Transformation Process Testing

With prototype development well underway, the DS 2.0 project team has begun testing the data transformation process that will eventually crosswalk DS member data into the DS 2.0 database. Cornell University, Columbia University, the Free Library of Philadelphia, the Huntington Library, Princeton University, and the University of Pennsylvania have contributed sample manuscript records from their library catalogs to use for testing. Data from these records are extracted from their original format, including MARC and TEI, and then migrated into the DS 2.0 transition spreadsheet following a schema that corresponds to the DS 2.0 data model. This spreadsheet is then uploaded into OpenRefine, a data cleaning and reconciliation program that facilitates mass editing of spreadsheet data. Within OpenRefine, the data is enhanced with links to authority files when possible, such as the Getty Vocabularies, VIAF, and Wikidata. Figure 2 shows the transformation process in action, displaying a selection of data taken from Columbia University's MARC records.



OpenRefine columbia-dec3 Permalink — Open... Export Help

49 matching rows (50 total) — Extensions: Wikidata

Show as: **rows** records — Show: 5 **10** 25 50 100 500 1000 rows — « first ‹ previous 1 of 5 pages next › last »

subject_as_recorded	author_as_recorded	author_recon	Qid	author_as_recorded_agr	artist_as_recorded	artist_as
Pseudo-Ptolemy.--Centiloquium.|Astrology, Arab--Early works to 1800.|Manuscripts, Persian. edit	Ṭūsī, Naṣīr al-Dīn Muḥammad ibn Muḥammad, 1201-1274.	Nasir al-Din al-Tusi Choose new match	Q302835	طوسي، نصير الدين محمد بن محمد، 1274-1201.		
Islamic magic--Early works to 1800.|Talismans--Religious aspects--Islam--Early works to 1800.|Manuscripts, Arabic.						
Geography, Arab--Early works to 1800.|Astronomy--Islamic countries--History.|Manuscripts, Arabic.|Astronomy.|Geography, Arab.|Manuscripts, Arabic.|Islamic countries.	Jighmīnī, Maḥmūd ibn Muḥammad, -1221?	Jaghmini Choose new match	Q4060361	جغميني، محمود بن محمد, -1221		
Islamic calendar--Early works to 1800.|Astrology, Arab--Early works to 1800.|Medicine--Turkey--Early works to 1800.|Manuscripts, Turkish.						
Geometry--Arab countries--History.|Manuscripts, Arabic.	Theodosius, active 1st century B.C.	Theodosius of Bithynia Choose new match	Q1266186	ثاوذوسيوس, active 1st century B.C.		
Geometry--Early works to 1800.|Geometry--Arab countries--History.|Manuscripts, Arabic.	Euclid.	Euclid Choose new match	Q8747			

Figure 2: A screenshot taken from the OpenRefine program showing the DS 2.0 transition spreadsheet with data harvested from Columbia University's MARC records.

The screenshot shows the *author_as_recorded* column with the author name as it appears in Columbia's MARC records, the *author_recon* column with a link to the Wikidata item for that author, the *Qid* column with Wikidata's unique identifier for that item, and the *author_as_recorded_agr* column with the author name in Arabic as provided in Columbia's records. DS will eventually produce its own name authority records and unique identifiers that will supersede the links to Wikidata. For now, these reconciliations with Wikidata assist the DS 2.0 project team in testing the data transformation process in OpenRefine.

Upcoming Goals

By the end of this year, the DS 2.0 project team hopes to complete the OpenRefine data transformation process for a selection of the submitted sample records and upload the transformed data into the DS 2.0 prototype database. This will complete a key phase of prototype testing and pave the way for beta development to begin in the new year. In February 2022, the IMLS will announce whether they have accepted the pre-proposal for DS 2.0 implementation funding. If accepted, the full proposal will be due in March 2022, with a notification of award the following July. With or without additional funding, the DS 2.0 database is on track to launch next year, thanks to the Bridge Year funding. A stakeholder meeting will be held next spring to update the DS community on the project's progress. In the meantime, DS administrators have begun meeting with DS members to discuss the DS 2.0 project and answer questions about how each institution will contribute data. DS member representatives should be on the lookout for an invitation to schedule a member meeting with their institution over the coming months. The next Project Report will be published in April 2022.