Structured Data Transcription with DataScribe for Omeka S

NEH-ODH Final Whitepaper

Hernán Adasme, Megan Brett, Daniel Howlett, Emily Meyers, Lincoln Mullen, Jessica Otis, and Greta Swain

Project Summary

From 2019 to 2022, a project team at the Roy Rosenzweig Center for History and New Media (RRCHNM) worked to create DataScribe: a structured data transcription module that extends the functionality of the Omeka S content management system. DataScribe generates an interface within the Omeka user dashboard that allows scholars to create projects and datasets from the images already uploaded to their Omeka installation. Users on each project can then create transcription records for each item in a dataset, employing a transcription interface linked to customizable structured data forms. This transcribed data is exportable for analysis and/or display in other platforms. DataScribe can be downloaded through the Omeka S module registry (https://omeka.org/s/modules/), on the DataScribe website (https://datascribe.tech), or directly from the DataScribe GitHub repository (https://github.com/chnm/Datascribe-module). Extensive documentation, including a user manual in English and Spanish, tutorials, case studies, and lesson plans are also available and can be found via the DataScribe website.
Project Origins and Goals

The Omeka Classic web publishing platform was first developed in 2008 by RRCHNM and has since become a standard for digital exhibits and collecting. In 2017, RRCHNM released a related platform, Omeka S, which was designed to make use of the affordances of linked data and allow multisite installations for more flexible use in universities as well as galleries, libraries, archives, and museums. The core functionalities of Omeka and Omeka S have been extended many times since their initial release, including the NEH-ODH-funded module Scripto (https://omeka.org/classic/docs/Plugins/Scripto/), which draws upon and adapts the same MediaWiki software that runs Wikipedia in order to support crowdsourced, freeform transcription of digitized primary source documents. While these extensions were and continue to be immensely valuable, as of 2019 there was no support within the Omeka universe for transcribing structured historical data.

At that time, the project co-directors Jessica Otis and Lincoln Mullen had each been working with structured data for several years. Otis began systematically digitizing copies of the early modern London Bills of Mortality in 2016 and spent much of 2016 and 2017 attempting a computational transcription solution using OCR before concluding that manual transcription was needed to reliably differentiate the numbers, given the bills’ low printing quality. Manual transcription began in 2017 and a spreadsheet-based template was soon implemented to speed up the transcription process; despite the speed gains in transcription, the spreadsheets still required considerable effort to clean before they could be imported into a SQL database for querying and analysis.

Mullen began collecting denominational statistics in 2013 and conducted trial transcriptions of data from yearbooks and other records compiled by Catholics, Methodists, Presbyterians, Congregationalists, and Episcopalians. In collaboration with his colleague John Turner, in 2018 he began digitizing congregation-level schedules from the 1926 U.S. Census of Religious Bodies, housed in the National Archives and Records Administration in Washington, DC. A subsequent grant from the National Endowment for the Humanities enabled the continued digitization as well as transcription and quantitative and spatial analysis of these denominational records.

The DataScribe module was thus conceived of as a way to extend the functionalities of the Omeka S platform to support the transcription of structured humanities data. While it was inspired by and formed a useful basis for further work on the project co-directors’ projects—which served as disparate motivating examples for development work—DataScribe was primarily intended to be a generalized solution to the problems facing scholars attempting to transcribe and analyze structured humanities data.

Humanities scholars working with digital methods are increasingly turning towards the quantitative analysis of data. Humanists have used methods such as mapping, network analysis, and statistical analysis of data to produce new interpretations of their fields, and data analysis undergirds research published in both customary prose forms, such as journals and
books, and in new forms, such as interactive visualizations and websites. These methods have a wide application across the humanities in general, including literary studies, classics, and art history. Within the discipline of history, these methods have been used for many topics including those in social, cultural, political, and religious history.

Doing data analysis presupposes that scholars have data to analyze. While there are large datasets that are commonly used by many scholars, the primary way most historians who work with data get their datasets is by creating the datasets for themselves. Whether in spreadsheets or CSV files, in Microsoft Access or Filemaker Pro, a common practice is for historians to gather a set of primary sources and transcribe them into a dataset which they then use for their personal research. This practice is, to some degree, inherent to the discipline of history: one of the fundamental ways that historians create new knowledge in their respective subfields is by looking at sources that other scholars have not yet examined. However, historians also create new knowledge by looking at familiar sources in a different way from their predecessors, more and more by treating them as data rather than as literary sources.

The kinds of sources which historians have gathered into datasets are enormously varied. They include town registers, tax rolls, probate records, accounting books, missionary records, jail records, membership rolls for voluntary societies and religious congregations, library borrowing records, dishes from restaurant menus, ship logs and manifests, bills of mortality, and quantitative records of every kind. Records created by institutions and governments abound, and records kept as forms are particularly amenable to being transcribed into a dataset. Therefore, even as historians turn more towards quantitative methods, they are likely to continue to work as historians have always done by transcribing that huge array of historical sources into their own datasets that address their questions.

Though the historical sources which historians use are varied, sources within a specific study that contain data can often be transcribed into a similar structure. Structured data is data which has been organized in a way which permits the maximum flexibility in how it is queried and used. Put differently, structured data is data which can be used for the purpose for which it was originally created but also for other, originally unforeseen uses. In general, most datasets can be represented as tables of data organized according to the principles of data normalization. These principles reduce data redundancy and improve data integrity. These principles are well understood (though perhaps not by most humanists), and they undergird almost all relational databases. Generally speaking, the process of transcribing data from primary sources ought to involve creating a map from the elements of the source to the fields of a normalized data structure.

Except that is not how most historians do their transcriptions. Most historians simply reproduce the format of the source they are transcribing without critically engaging with its structure as data. Most historians who transcribe data do not understand the fundamental principles of data structure. Given the wide open canvas of spreadsheet programs like Microsoft Excel, they often make poor choices, leaving their datasets difficult to use for themselves and impossible to use by others. Besides these basic problems, creating datasets is very laborious. Appropriately
designed tools can reduce this burden and make the transcription of reusable data more feasible. Given the labor that goes into datasets, the ability to make them usable by others is essential.

The DataScribe module for structured data transcription was designed to address this issue. DataScribe built on the strengths of Omeka as a platform for gathering and distributing collections of sources. While this transcription module could have been built as stand-alone software, building this transcription tool on the Omeka platform enabled the images (and metadata) describing the historical sources to reside in the same place as the structured data transcribed from those sources. By utilizing the user accounts which are already a part of the Omeka S platform, scholars are able to collaborate on the transcription of sources into structured data. We were also able to leverage Omeka's existing user base. Many digital humanities projects are already built on top of Omeka S and Classic. Creating an Omeka S module allowed those users to gain the ability to transcribe structured data from their sources through the installation of a simple module.

The DataScribe module has a feature set and an interface built off existing Omeka S user categories to allow different types of access to the module based on user type, ranging from project owners who can define the data structures to be transcribed, reviewers who can review transcriptions, and transcribers who can transcribe sources. One key set of features revolves around the ability to define the data model for the transcription. A project owner is able to describe which fields should be transcribed and what types of data (for example, numeric, dates, short text, categorical, boolean) should be able to go into those fields. Another set of features revolves around the transcription of the data. Transcribers have the ability to view both the source to be transcribed and the transcription panel simultaneously, and the transcription panel is designed to maximize the transcriber’s productivity. Data as it is entered is automatically validated: for example, if a field is supposed to be a date, then the interface will check that it is a valid date within an optional range; if a field is supposed to be a positive integer, then that test will be enforced as well; categorical fields will allow the user to select from the allowed set of options. Transcribing users have the ability to mark fields as missing, illegible, or in need of review, thus allowing transcriptions to be flagged for review.

A third set of features relates to the management of a large scale transcription project. Project owners and reviewers are able to keep track of which documents have been transcribed. They are able to prioritize the transcription of specific documents: for example, documents that have a specific tag or metadata value can be presented to transcribers first. A final set of features relates to the export of the data forms and dataset. Project owners are able to validate and export both transcription forms and data from transcribed datasets. These exports are made available as JSON and CSV files, respectively, both of which are widely used formats.

The primary aim of this project was to create the module, not gather and transcribe any particular set of sources. Nevertheless, using specific yet varied kinds of sources was essential to the project’s success for two reasons. The first was that the Omeka S module needed to be able to abstract the details of many kinds of sources into a general pattern for transcribing
structured data. Tools in the digital humanities are successful when they address problems which can be well defined and which admit of a general solution. Bringing together very different kinds of sources, for which we already knew the potential pitfalls due to our experience working with them, was our way of making sure that the problem was well defined. We knew if the tool we built could transcribe both forms and tables, both early modern and twentieth-century sources, then there would be every reason to think that the tool would work for most other kinds of structured data that historians need to transcribe.

The second main reason that this step was necessary was that these kinds of historical sources provided the documentation necessary for the project's end users. Digital humanities tools are also successful when users can clearly understand how their problem is solved by the tool through clear documentation. In our case the most useful documentation was not only an abstract discussion of the properties of structured historical data, but also a set of examples of the kinds of sources which can be transcribed. This project created an interface which guides users into properly structuring their data and a robust set of documents explaining how to structure data using the examples we created.

In summary, historians were already actively engaged in collecting sources which could be transcribed into structured datasets. This process of collection and transcription is happening in many historical fields with all kinds of historical records, and the creation of such datasets is a basic way in which historical argumentation happens and new historical knowledge is created. But although there were a number of good options for the transcription of freeform text, there was no equivalent for the transcription of structured data which is intended to be computationally analyzed and visualized. In addition to maximizing the productivity of the people doing a transcription, it was essential that such a tool must be built on the best practices for data models and that it guide historians into how to see the underlying structure of their data—a step which databases and spreadsheets can never accomplish. Because transcription comes after collection, the Omeka platform for collecting and describing sources made an excellent platform for such a transcription tool. By being built on top of Omeka S, the DataScribe module gained features involved in collecting and describing data, as well as managing users and documents without having to build them from scratch. DataScribe was built to improve the ease of transcribing data, create better structure for data, and provide a useful platform for disseminating that data. The datasets created will have an impact on historical fields from many places, time periods, and subjects, by tapping into the energy in the field around data analysis and visualization.

**Project Activities, Team, & Participants**

In fall 2019, Otis and Mullen coordinated with other scholars and project advisors to determine what kinds of primary sources with structured data they were using and created mock workflows for transcription projects to inform the data model creation and design processes. The entire project team then worked to define the desired transcriptions of the sources and describe the key features of the module and user interface, ranking them by the degree to which they were essential.
In building the alpha module, Jim Safley focused on creating the administrative interface, with particular attention paid to generalizing the construction of data models in the administrative interface, while Kim Nguyen and Ken Albers focused on design and wireframing of the administrative, reviewer, and transcriber views for the module. Safley and Nguyen then began to implement the approved wireframes and conducted informal testing as they encountered bugs and other issues with the implementation.

Although the COVID-19 pandemic forced our team to leave our physical spaces in March 2020, the project team already had technological and workflow supports in place for work-from-home 2 days a week. The team leaned into those technologies and workflows in order to move our activities 100% online and thus managed to continue working to build, test, revise, and document the alpha module with minimal interruptions.

Otis, Mullen, Megan Brett, and Graduate Research Assistant (GRA) Dan Howlett alpha tested the module over the summer 2020, using six historical datasets, including the London Bills of Mortality and U.S. denominational data. Safely and Nguyen refined the alpha module based on feedback from the project team and issues that arose during alpha testing, with particular attention paid to user workflows and mouse-free navigation. Brett created and iteratively revised the module end-user documentation in support of our internal alpha testing. To create the foundation for future outreach, Nguyen created a wordmark for DataScribe and Albers created a website (https://datascribe.tech) for people to visit to learn more about the module. The team also created a sign-up form where scholars could opt-in for updates on the progress of the module’s development and release.

Alpha testing was completed and we officially launched in beta in November 2020. We collected feedback from beta testers on an ad hoc basis, as they used the module and reported back to us on their experiences. In order to ensure that we receive sufficiently robust feedback in this testing stage, we also ran a more formal process of user testing with project advisors and selected other volunteers.

Progress on development was temporarily halted in spring 2021 due to the RRCHNM programming team (including Albers, Nguyen, and Safley) being hired away en masse by an outside non-profit. GMU required a 6 month period of separation before team members were allowed to contract back to RRCHNM to continue revising the module. Further delays to production occurred during fall 2021 as the university was extremely slow to process the internal paperwork necessary for the contract. Safley and Nguyen’s deployment of bug fixes to the beta software did not begin until January 2022.

During this time period, the remaining members of the project team focused on documentation and the creation of test datasets to support the full public release of the module. Since we had a stable beta release, the documentation was revised and expanded to ensure that public users had sufficient support to employ the module. Brett, Mullen, and Otis stress-tested the module extensively in projects that were supported by other funding sources, and came up with several
additional key feature requests that included the ability to export form data in JSON format and re-import it as the form for other datasets. The case study documentation on Religious Ecologies was completed by GRA Greta Swain and used as a model for other documentation created over the course of 2022. Brett and Otis worked with Howlett and Swain to propose an in-person workshop at the German Historical Institute's December 2021 digital humanities conference, which was accepted but delayed by the Omicron wave of COVID-19 and rescheduled to summer 2022. The project team also onboarded RRCHNM’s newest programmer, Jason Heppler, who served as an in-house backup and was particularly active in assisting Brett with updating the datascribe.tech website as new content was created and needed to be uploaded.

DataScribe was fully released in March 2022 and is available to download through both the DataScribe github repository (currently: https://github.com/chnm/Datascribe-module) and the Omeka.org module list (https://omeka.org/s/modules/). Upon its release, the project team conducted a publicity campaign to promote the module. This included outreach through social media and listservs, with the assistance of advisory board members, as well as presentations at the Renaissance Society of America annual meeting in Dublin; an in-person workshop at the German Historical Institute in summer 2022; and numerous online webinars and workshops which were well attended by both scholars from within academia and by members of the wider GLAM community. We fixed several bugs in the module discovered after the public release, and added key features—including the ability to move datasets between projects—that we discovered we needed during additional testing.

With the assistance of GRAs Hernán Adasme and Emily Meyers, we spent the remainder of 2022 completing the project documentation, case studies, tutorials, and other user-oriented documentation. RRCHNM’s new Community Engagement Coordinator, Bridget Bukovitch, joined the team to consult on outreach. To ensure adequate time to finish our publicity campaign and documentation, we asked for and received a 6-month no-cost extension which we used to wrap up project activities in the fall of 2022.

Project Outcomes

The primary outcome of this project was the creation of DataScribe, an Omeka S module for structured data transcription. Alongside the module, the project produced extensive documentation both on the module and its use cases, in English and Spanish. Supporting materials include case studies, tutorials, and lesson plans, which are available on the project website and which formed the basis for a series of workshops held in 2022.

The module and documentation have been disseminated under a GNU General Public License v3.0 through the Omeka website (https://omeka.org/s/modules/), the CHNM GitHub (https://github.com/chnm/datascribe-module), and project website (https://datascribe.tech). A copy of these materials and the final whitepaper will also be deposited together in George Mason University’s MARS institutional repository after the formal end of the project.
Project Evaluation and Impact

The DataScribe module was downloaded over 200 times in the first eight months after its official release in March 2022, showing it already is having a significant impact in enabling scholarly research. Over 100 people signed up for our beta release email and approximately 85 people—including academic faculty, students, librarians, and other public scholars—attended our 2022 in-person and online workshops.

While DataScribe was created within the Anglosphere, and was designed as natively English software, the project team—primarily Adasme—took steps to translate some of our materials into Spanish as part of an effort to build a multilingual community of DataScribe users that goes beyond the Anglosphere. This effort began with the translation of the DataScribe interface and the translation of the documentation and continued with the hosting of a Spanish language workshop which covered transcription, reviewing of transcriptions, and project creation. The outreach team personally invited people working in the DH field in Chile and México to attend. One of the highlights of the Spanish workshop was the presence of Leonor Riesco, Director of the Digital Humanities Certificate of the Universidad Finis Terrae and one of the leading figures of DH in Chile. Riesco and the attendees were very interested in using both DataScribe and Omeka for their own research projects.

Both our user testing and the results of our outreach within the DH community made clear that our primary audience for the module will be scholars within academia and the GLAM (Galleries, Libraries, Archives and Museums) communities, particularly historians and other humanists working with historical sources. The level of interest and questions from participants in our varied workshops demonstrated a clear need for our module, and working with participants also allowed us to evaluate and produce the kind of support resources some scholars would need to help them understand how to think about structured data.

While it is still too soon to know what the long-term impact of the DataScribe module will be for its user base, it is possible to evaluate its impact on the three active RRCHNM projects that are currently using the module and which formed the basis of four of the case studies (two in English, one in both English and Spanish) available on the DataScribe website. Two of these are projects that were part of the original motivation for developing the module, while the third is the work of a GRA who used his own research interests to develop a case study using the module. An additional case study grew out of the use of the newly publicly available 1950 Census returns as a demonstration dataset for workshops.

The case study, Religious Ecologies, written by Greta Swain, demonstrates how DataScribe can be used for structured historical sources with hundreds of thousands of forms involving a variety of data types from text to numbers to complex geographical data. This project began transcribing before the DataScribe module was available but their spreadsheet-based method of transcription was cumbersome and confusing. Transcribers found it very easy to make mistakes, forget the transcription rules, or include typos. Creating (and then using) DataScribe’s custom transcription form helped the project standardize data, kept transcription
errors to a minimum, and made each field easier to analyze or visualize. DataScribe also made it easy for the team to collaborate on the transcription process. The project manager could assign tasks to team members (transcribers or reviewers) and review transcription progress, while all team members could communicate about the transcription process asynchronously, directly in the DataScribe interface. Ultimately, in addition to being a transcription interface, DataScribe proved itself to also be a platform that enabled the team to plan and carry out a project workflow. This project is ongoing with funding from the National Endowment for the Humanities, and datasets created as part of this project are available at https://religiousecologies.org/

The case study, *Death by Numbers*, written by Hernán Adasme, Dan Howlett, and Emily Meyers, demonstrates the utility of DataScribe for structured historical sources with a more uniform type of data—numbers—but with hundreds of data points and several different subsets of data to be transcribed from each document. Moreover, the documents for this project were generated over 150 years, leading to subtle shifts in the form over time. DataScribe allowed the project team to chunk these documents into a large number of datasets that are tracked and managed through a dashboard. Each dataset has its own transcription guidelines, especially around missing and illegible data, which can be easily called up from within the transcription interface. New datasets can be quickly and easily created using the form export and import features, when unexpected variants in the data are discovered during the transcription process. DataScribe also enabled the expansion of the transcription team as it makes it easy to assign, review, and track the work of multiple transcribers within the interface. Finally, DataScribe allows the data to be easily downloadable in a CSV that can then be imported into the project database, posted on the project’s website, and used in team members’ analyses of the data. This project is ongoing with funding from the National Science Foundation, and datasets created as part of this project are available at https://deathbynumbers.org/ and the project GitHub at https://github.com/chnm/bom/tree/main/bom-data/data-csvs

The case study *Plague in Iquique*, written by Hernán Adasme, demonstrates the viability of using DataScribe to extract structured data from a non-tabular or semi-structured historical source. The case study explores the Informe Sobre la Epidemia de Peste Bubónica en Iquique in 1903, a document presented to the government of Chile by a medical commission investigating a 1903 outbreak of the bubonic plague in the port of Iquique. The report provides detailed and heterogeneous data about the 167 first cases of the outbreak. This project demonstrates how DataScribe can provide a template for the systematic capture of quantifiable data even when the historical source is only partially structured. The case study walks the reader through the process of creating a transcription form that captures information from these paragraphs—ranging from dates, spatial, numeric, and text strings—that suits the historical questions being asked by scholars. *Plague in Iquique* continues to be an ongoing project that relies heavily on DataScribe, which allows for the creation of datasets for the study of the outbreak, using unstructured textual data.

Last, but certainly not least, this project has had a significant impact on the professional development of the GRAs who worked on the module: our co-authors Hernán Adasme, Dan
Howlett, Emily Meyers, and Greta Swain. All four gained experience in working on a collaborative DH software development project and presenting the results of the team’s work to a public audience in presentations and workshops. Howlett was integral to the alpha testing and initial documentation processes, while Swain focused on creating the model for our subsequent case studies. Adasme and Meyers helped create our tutorials and several case studies, and grew comfortable enough with the module to eventually take over running our online webinars and workshops. Adasme also suggested—and implemented—the translation of our documentation into Spanish, which he followed up on by creating a Spanish-language case study and running a Spanish-language workshop to expand our outreach efforts to a more linguistically diverse audience. All four grew in experience and confidence over the course of the project, and took on public-facing roles that helped raise their profile and provide them valuable connections in the DH community that we expect to help them as they continue their careers.

**Project Continuation and Long-Term Impact**

The DataScribe module is currently in a stable place and there are no plans for major new feature development. As various active projects at RRCHNM continue to use the module, it will continue to be assessed for bugs and checked for compatibility with new versions of the Omeka S core as they are released. Additional supporting documentation may be added to the DataScribe website as they are developed for class projects, etc. However, the module’s major long-term impact will be best judged by the success of current and future digital humanities projects that use it as the basis for the structured data transcription and subsequent analysis.