The Journal of Open Source Software

# DataScribe: An Omeka S module for structured data transcription

**Jessica M. Otis** [1], **James Safley** [2], **Megan Brett** [3], **and Lincoln Mullen** [1]

**1** Roy Rosenzweig Center for History and New Media, George Mason University, USA **2** Digital Scholar **3** Jefferson Library, Monticello

**DOI:** 10.21105/joss.05661

**Software**

- Review
- Repository
- Archive

**Editor:** Arfon Smith

**Reviewers:**

- @luxaritas
- @koenedaele

**Submitted:** 19 April 2023
**Published:** 07 January 2024



## Summary

DataScribe is a structured data transcription module that extends the functionality of the Omeka S content management system for digital exhibits. It creates an interface within the Omeka user dashboard that allows users to create projects and datasets from the images uploaded to their Omeka instance. Users then create transcription records for each item in a dataset using 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).

## Statement of Need

Scholars often collect sources, such as government forms or institutional records, intending to transcribe them into datasets which can be analyzed or visualized. Many transcription programs such as ABBYY FineReader (ABBYY, 2022), Scripto for Omeka S (Hamner et al., 2010), Tesseract (Weil et al., 2021), and Zooniverse Project Builder (Johnson et al., 2009) enable the manual or automated transcription into free-form text, but not into tables of data. The DataScribe module enables scholars to manually transcribe documents directly into a structured data format. Once scholars identify the structure of the data within their sources–such as numbers, dates, or controlled vocabularies–they can create forms that constrain and verify transcriptions done in the DataScribe interface. The transcriptions are then exported in tables of clean and tidy data that can be computationally analyzed or imported into a variety of analytical software programs. Because the module builds on Omeka S, scholars can also display transcriptions alongside the source images and metadata, crowdsource transcriptions, and publish their results on the web.

Projects using DataScribe include *Death by Numbers* (2016), which is transcribing the seventeenth- and eighteenth-century London Bills of Mortality, and *Mapping Religious Ecologies* (2018), which is transcribing the the 1926 United States Census of Religious Bodies. As part of the development of the module, the project team also created case study documentation for how DataScribe might be used to transcribe the London Bills of Mortality (Adasme et al., 2022), documentation on a 1903 plague outbreak in Chile in both Spanish and English (Adasme, 2022a, 2022b), the 1926 United States Census of Religious Bodies (Swain, 2022), and the 1950 United States Census (Brett, 2022).

## Acknowledgements

Development of this software was funded by the National Endowment for the Humanities (grant number HAA-266444-19).

## References

ABBYY. (2022). *ABBYY FineReader*. https://pdf.abbyy.com

Adasme, H. (2022a). *Peste bubónica en iquique, 1903: Transcripción de datos no tabulares usando DataScribe*. DataScribe.tech. https://datascribe.tech/casestudies/PesteBubonica_Iquique1903.pdf

Adasme, H. (2022b). *Plague in iquique, 1903: Transcribing non tabular data using DataScribe*. DataScribe.tech. https://datascribe.tech/casestudies/Plague_Iquique1903.pdf

Adasme, H., Howlett, D., & Meyers, E. (2022). *Death by numbers*. DataScribe.tech. https://datascribe.tech/casestudies/DataScribe_BillsOfMortality_CaseStudy.pdf

Brett, M. (2022). *An exercise in iteration: Transcribing the 1950 united states census with DataScribe*. DataScribe.tech. https://datascribe.tech/casestudies/DataScribeCaseStudy_1950CensusUS.pdf

*Death by numbers*. (2016). Roy Rosenzweig Center for History and New Media, George Mason University. https://deathbynumbers.org/

Hamner, C., Safley, J., Nguyen, K., Brett, M., Leon, S., Fahringer, A., Halabuk, J., Dauterive, J., Albers, K., Ghajar, L. A., & Brennan, S. (2010). *Scripto*. Roy Rosenzweig Center for History and New Media, George Mason University. https://scripto.org

Johnson, C., Chambers, C., Clement, D., Trouille, L., Bouslog, M., Yuen, M., Blickhan, S., Miller, S., Wolfenbarger, Z., Lintott, C., Noordin, S. A., Roberts, H., Mantha, K., Fortson, L., Huebner, S., Kiefer, T., Simmons, B., Krawczyk, C., Spiers, H., … Granger, W. (2009). *Zooniverse project builder*. https://www.zooniverse.org

*Mapping religious ecologies*. (2018). Roy Rosenzweig Center for History and New Media, George Mason University. https://religiousecologies.org

Swain, G. (2022). *American religious ecologies*. DataScribe.tech. https://datascribe.tech/casestudies/CaseStudy_AmericanReligiousEcologies.pdf

Weil, S., Smith, Ray, Podobny, Z., Abdulkader, A., Antonova, R., Beato, N., Breidenbach, J., Charron, S., Cheatle, P., Crouch, S., Eger, D., Huddleston, S., Johnson, D., Katikam, R., Kielbus, T., Lee, D.-S., Liu, Z., Moss, R., Newton, C., … Zaitsev, A. (2021). *Tesseract*. https://tesseract-ocr.github.io