Shaping Humanities Data: Use, Reuse, and Paths Toward Computationally Amenable Cultural Heritage Collections

Thomas Padilla
thomaspadilla@ucsb.edu
UC Santa Barbara, United States of America

Sarah Potvin
spotvin@library.tamu.edu
Texas A&M University, United States of America

Laurie Allen
laallen@upenn.edu
University of Pennsylvania, United States of America

Stewart Varner
svarner@sas.upenn.edu
University of Pennsylvania, United States of America

Galleries, libraries, archives, and museums (GLAMs) increasingly seek to make digitized and born-digital collections accessible as data optimized for computational methods and tools common to the Digital Humanities. Preparation and publication of collections as data extends possible collection use beyond the analog object interactions that collection interfaces tend to try and emulate. In line with open data efforts, libraries, archives, and museums typically work to assign open licenses to these data. Current access methods are widely divergent, spanning simple provision of compressed collection objects in ZIP files, exposing static collection websites that can be crawled using a tool like rsync, leveraging Github for text collection access, provisioning an API, enabling FTP access to collections, mediating computational processes performed on collection data through a platform, to facilitating data access through use of torrent technology. Concurrently, in response to researcher requests for data-mining, commercial publishers have developed a range of processes for delivering proprietary corpuses with terms and conditions that significantly limit or expressly forbid data sharing, including providing libraries with physical hard drives loaded with the data. There are no consensus-driven best practices that guide the generation, description, and provisioning of computationally amenable GLAM collections for the range of communities that fall within the Digital Humanities. Without best practices in this space, institutions run the risk of misplaced investment of resources that foster the creation of irregular, ultimately disorienting data access environments. Indeed, the panoply of institutional approaches poses a challenge to GLAM institutions seeking best practices and clear guidelines for publishing collections as data.

One major barrier to the development of consensus-driven best practice is an incomplete understanding of *how* digital humanists, among others, are using and reusing cultural heritage data. This workshop aims to make progress towards bridging that gap. Research indicates that types of use exhibited by digital humanists include but are not limited to text analysis, image analysis, mapping, sound analysis, and network analysis. Orientation to the full scope of *academic* use types can be gained through in-depth analysis of data use practices across disciplines as represented in core Digital Humanities journals (Padilla and Higgins 2016), by reviewing works at the annual global Digital Humanities conference (Weingart 2016), and by studying edited volumes that have to this point effectively compiled a broad range of research in this space (Gold 2012; Gold and Klein 2016; Burdick, Drucker, Lunenfeld et al 2012; Schreibman, Siemens, Unsworth 2016).

This workshop will build upon this orientation by engaging directly with digital humanists' existing and projected research and pedagogical practices that draw upon ever growing GLAM collections. Blending short talks by practitioners, guided discussion, and workshopping of the organizers' draft framework (further described below), the workshop will focus on how researchers and educators use GLAM collections that have been made accessible as data, and will extend to consider how these uses should inform collection creation and access.

The organizers of this workshop are members of the project team for "Always Already Computational: Library Collections as Data," an effort sponsored by the Institute of Museum and Library Services in the United States of America through their National Forum grant program. The organizers have observed that GLAM approaches to the preparation of collections as data are often heavily influenced by national or regional priorities and associated infrastructures. Yet the use and reuse of these open data is necessarily international. While the organizers of the workshop are US-based, the workshop aims to surface geographically-diverse praxis. The short talks in the workshop have been selected through an open CFP facilitated by an international program committee.

The workshop may be structured thematically, based on talks and demos solicited via the CFP. Participants will be encouraged to consider how efforts to develop computationally amenable collections, which run the risk of recreating and reinforcing long standing biases inherent in cultural heritage collection practice, provide an opportunity to reframe, enrich, and/or contextualize collections in a manner that seeks to avoid replication of bias.

Potential themes may include:

- **Use and Reuse:** How are data used and reused? What methods and tools are commonly employed? Do these differ by disciplinary community? What types of data are used? What types of data are desired but are difficult to use for reasons included but not limited to copyright status, content type (e.g. video, audio, web, software), size? How, when, and where are data reused? What factors enhance or inhibit the likelihood of data reuse?
- **Access:** What can we learn from our collective experiences working to access data from within and outside of the cultural heritage community? What are preferred methods of data access? What factors should be considered when deciding among access methods? When is simple click and download of bulk collections appropriate? What characteristics define an optimally useful API (application programming interface) for a wide range of users with varying technical expertise? Is an API always the best route to go? Are there a mix of options that should be considered? What considerations inform development of those access options?
- **Description and Discovery:** How do digital humanists locate appropriate data? What tools are used to search for data? What information about the data is necessary to enable use? When compiling meta-collections of data, how are digital humanists maintaining provenance and merging disparate metadata?

Bibliography

Burdick, A., Drucker, J. and Lunenfeld, P., Presner, T. and Schnapp, J. (Eds) (2012). *Digital_Humanities*. Cambridge: MIT Press. https://mitpress.mit.edu/books/digitalhumanities

Gold, M. (Ed) (2012). *Debates in the Digital Humanities.* Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/1

Gold, M. and Klein, L. (Eds) (2016). *Debates in the Digital Humanities.* Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/2

Padilla, T., Higgins, D. (2016). Data Praxis in the Digital Humanities: Use, Production, Access. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 644-646.

Schreibman, S., Siemens, R. and Unsworth, J. (Eds). (2016). *A New Companion to Digital Humanities*. 2nd edition. Wiley-Blackwell.

Weingart, S. (2016). “Submissions to DH2016 (pt. 1).” On *the scottbot irregular*. http://scottbot.net/submissions-to-dh2016-pt-1/