CDH Dataset Curation | Resources

Tools

Data Cleaning / Editing

- OpenRefine
  - power tool for cleaning tabular data and some XML
  - supports regular expressions
- Atom
  - free and open-source text and code editor
  - powerful search across files
  - supports regular expressions
  - robust community of developers who contribute "packages" that extend Atom's functionality, including data transforms, specialized syntax highlighting, easy GitHub integration, etc
- Recogito
  - collaborative data annotation platform for text and images
  - includes syntax for places, people, events
  - multiple export formats
  - runs auto Named-Entity Recognition
- Breve
  - web based visual tool for seeing data errors in tabular data
  - NEH-funded project under development at Stanford's Center for Spatial and Textual Analysis
- WTFcsv
  - web based visual tool for a quick snapshot of the data in a csv file

Databases / Repositories

- Airtable: collaborative database platform
  - allows you to embed a browsable copy of your database in a webpage
  - super user friendly, with tutorials that explain features like pivot tables
  - free account allows for 2GB server space and revision history 2 weeks old, but further features cost $$
- Mukurtu: content management system supporting Indigenous knowledge systems and values
  - grassroots platform currently used by six hundred different groups to "curate their own Web sites and regulate access in accordance with custom"
  - multiple records can be generated for single digital heritage items, allowing for overlapping cultural narratives
  - "There is rarely just one story, one set of information, or one way of knowing cultural heritage materials."

Developed by Grant Wythoff

- Zenodo: open access repository maintained by CERN
  - automatically assigns DOIs to all files
  - a great alternative to the for-profit academia.edu
  - syncs with and preserves GitHub repositories: "Just Log in with your GitHub account and click here to start preserving your repositories."
    - Used by CDH for making project code citable
      Derrida's Margins codebase: https://doi.org/10.5281/zenodo.1453447
      PPA codebase https://doi.org/10.5281/zenodo.2400705
  - "your research output is stored safely for the future in the same cloud infrastructure as CERN's own LHC research data."
- Figshare: another data repository used by CDH
  - Automatically sets metadata and datasets as CC0 (because authorship cannot be claimed on factual data). If this is a concern, can be published as filesets instead so we can specify CC-BY. See https://knowledge.figshare.com/articles/item/what-is-the-most-appropriate-licence-for-my-data
  - example project: Derrida's Margins https://figshare.com/collections/Derrida_s_Margins_datasets/4256927
  - Automatically generates DOIs, which can be reserved before you publish a dataset if you need something to reference for publication. Click "cite" on the Figshare website to see the DOI

Project Management

- Asana
  - online project management platform with shared to-do lists
- Trello
  - team communications app in a message board format
- Slack
  - group communications with topic-based channels

Tutorials

"Cleaning Data with Open Refine,"*The Programming Historian*

"Cleaning Data with OpenRefine for Ecologists" and "OpenRefine for Social Science Data", *Data Carpentry: Building Communities Teaching Universal Data Literacy*

Checklist for Digital Humanities Projects, La Red de Humanidades Digitales (RedHD), English and Spanish versions available

Zed A. Shaw, *Learn SQL The Hard Way* [book]

Methods and Best Practices

- Research Data Management at Princeton

Developed by Grant Wythoff

  - Provides general and Princeton specific guidance and information on managing research data
- DH Curation Guide
  - Asks, "How do we align the care for digital materials with the methods/goals of traditional humanities disciplines?"
  - Introductory essays on different aspects of data curation in digital humanities, with links to relevant readings
  - produced by NEH-funded workshops in 2014 at Maryland Institute for Technology in the Humanities and University of Illinois Center for Informatics Research in Science and Scholarship
- UCLA Library: Data Management for the Humanities
  - extensive research guide
- PM4DH | Project Management for the Digital Humanities
  - developed by Emory Center for Digital Scholarship
  - "curriculum for managing digital projects in academic libraries and other settings"
- Managing and Sharing Data: Best Practices for Researchers [PDF]
  - Created by the UK Data Archive, "the UK's largest collection of digital research data in the social sciences and humanities."
  - produced in 2011, a slightly outdated but thorough rundown of best practices for sharing, management, documenting, formatting, storing, and ethics
- Kristin Briney, *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success* (Exeter, UK: Pelagic Publishing, 2015).

Example Datasets

- browse projects featured in *Journal of Open Humanities Data*
  - "features peer reviewed publications describing humanities data or techniques with high potential for reuse"
- Data Refuge
  - "a community-driven, collaborative project to preserve public climate and environmental data"
  - currently building a "Storybank", or map of data use cases and "life stories"
  - includes a number of toolkits for the rescue and protection of public data
  - spearheaded by UPenn's Program in Environmental History Lab
- Early African American Film
  - wonderful example of thorough documentation
  - networks of producers/actors/directors in early twentieth century "race film"
- Collections as Data: Part to Whole
  - UNLV / University of Iowa / U Penn led Mellon grant, supports a number of project applicants
  - "Collections as data produced by project activity will exhibit high research value, demonstrate the capacity to serve underrepresented communities, represent a

diversity of content types, languages, and descriptive practices, and arise from a range of institutional contexts."

- NYPL's "What's on the Menu?"
  - crowdsourced project that has garnered lots of public interest
  - interesting method of organically generating their data model
- Black Anthology Project
  - "information related to over 600 African American short stories that appeared in 100 African American and American anthologies published between 1925 and 2017."
  - tabular data on underrepresented authors and circulation histories
- ToposText
  - "an indexed collection of ancient texts and mapped places relevant the the history and mythology of the ancient Greeks from the Neolithic period up through the 2nd century CE"
- Quill Project
  - marking up "negotiated texts" written/decided by committee: constitutions, legislative proceedings, statements, etc.
  - "legibility to the general public only of secondary concern" – an archive primarily for scholars
  - example: https://www.quillproject.net/event_visualize/493

Readings

Data Cleaning

Katie Rawson & Trevor Muñoz, "Against Cleaning" (2016)

Mia Ridge, "Mia Ridge explores the shape of Cooper-Hewitt collections", *Cooper-Hewitt Labs* (2012)

Lauren F. Klein, "The Image of Absence: Archival Silence, Data Visualization, and James Hemings," *American Literature* 85, no. 4 (2013)

Data and Method

Tanya E. Clement, "Where Is Methodology in Digital Humanities?", *Debates in the Digital Humanities 2016*

Ryan Cordell, "Teaching Humanistic Data Analysis" (2019)

Luke Stark and Anna Lauren Hoffmann, "Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture," *Cultural Analytics* (2019)

Daniel Rosenberg, "Data Before the Fact," in *"Raw Data" Is an Oxymoron,* ed. Lisa Gitelman (MIT Press, 2013)

Johanna Drucker, "HTML and Structured Data" (2013)

Michael Hancher, "Re: Search and Close Reading," *Debates in the Digital Humanities 2016*

Ricardo L. Punzalan, Diana E. Marsh, Kyla Cools, "Beyond Clicks, Likes, and Downloads: Identifying Meaningful Impacts for Digitized Ethnographic Archives," *Archivaria* 84 (Fall 2017)