Digital Texts, Data Sources, and Analytics

This guide introduces you to the nuts and bolts of data. In today's world, it's important to understand data – what it is, how it's formatted, where to find it, and how to use it. Although far from exhaustive, this guide should give you some basics that can help you both to develop new research questions and to better understand what's possible in the wider world of your research field.

DATA FORMATS

- Plain text (e.g., Unicode UTF-8): the simplest, flattest, and most transferrable. *Publish your data-work in Unicode plain text for max transparency and uniformity*.
- CSV (comma-separated values) and TSV (tab-separated values) are two ways of formatting data in tables. *Hint: Use TSV if your data already has commas in it.*
- Other data formats: JSON, XML, HTML. Excel and FileMaker are examples of proprietary formats.
- Relational databases are used to link objects in different tables: examples are MySQL and Microsoft Access. (We are slowly leaving the period of relational databases! Watch out for "NoSQL" and "atomized" databases like OCHRE.)
- Queries: requests for data from databases and data streams.
- Media archives: maps, images, videos, etc.

ANALYTICS

- Analytics are tools that enable you to analyze data in ways that would be difficult without computer programming.
- You can write your own programs with Python, R, C/C++, Java, etc.
- *No programming? No problem*. Get your feet wet with free, online tools: Voyant, Tableau, Palladio. These resources allow you to input your data and run basic analytics/visualizations without programming knowledge.
- More: ArcGIS (for mapping + data), D3 (visualization; some JavaScript required).

DATA SOURCES

Digital Texts

- Online archives like Project Gutenberg offer a range of file formats. Be careful since these sources are rarely curated.
- HathiTrust: millions of titles.
- The UChicago Library: use "Find It!" to track down e-books.
- Turn paper books into digital data at the UChicago Visual Resources Center. Scan books quickly and use Optical Character Recognition (OCR) software to produce highly formatted plain text.
- The Institutional Repository (IR) and Digital Object Identifiers (DOI).

Qualitative and Quantitative Data

- Instruments and Sensors like Chicago's "The Array of Things"
- Other resources: USA Facts, census data, economic and statistical databases like Quandl, web surveys / polls and aggregators like 538.

WANT TO LEARN MORE?

- This handout is adapted from information and materials presented by Dr. Jeffrey Tharsen (Research Computing Center) as part of the UChicagoGRAD Workshop Series "Digital Literacy for Humanists." Dr. Tharsen welcomes questions related to these and other computing topics. Reach him at tharsen@uchicago.edu.
- Humanities Computing: Carmen Caswell (specialist in ArcGIS).