HIST 2430 Digital Histories of Ethnic Boston
Simon Rabinovitch
Introduction to Data and Data Modeling

About

This handout is a complement to the module on working with data. It contains a list of important vocabulary, a description of the data preparation activity, and several links with more information relevant to the topic.

Key Words

- **Machine readability** - a set of properties of a file that enable it to be processed and operated on by computer programs, languages, and platforms
- **Human readability** - a set of properties of a file that enable its clear comprehension by humans
- **Digital accessibility** - clear, open, and user-friendly online access to documents and their content by a wide range of (human, rather than machine) users
- **Open format** - "one that is platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information." (US Gov. definition from 2009) - CSV fits into this definition
- **Delimiter-separated values** - two-dimensional arrays of data. Values in each row are separated with specific delimiter characters, while the separation of rows is indicated by a new line.
- **Delimiter (aka field separator)** - a specially chosen, **unique** character that specifies the boundaries between separate cells, for example - , (comma)
- **Rows** in delimited data are also sometimes referred to as "records"
- Comma-separated values (**CSV**) - the most commonly used type of delimited data format
- **Qualifier** - a special character which goes around the field in which a delimiter character appears, for cases when we don't want the program to read it as a delimiter - for example - " double quotes
- **Escape character** - a special character used to signify to the program that the character coming after escape character should be read as text, and not as a command to the program

Questions and Ideas to Consider

- How is machine-readability related to human-readability?
- Who and what is data created for?
- How is creation, delimitation, and organization of data conceptualized?

Step-by-Step Process

- Introduction into the purpose and basic vocabulary of data management and planning, with particular emphasis on CSV files
- Demonstration of CSV operation with material from the Hebrew Immigrant Aid Society (Boston) Digital Archives
- Planning, organization and initial implementation for turning that database into a CSV file and performing initial cleanup

Guide for year standardization:

- “c. year”, change to year (for example, change "c. 1950" to "1950")
- “<” or “> year”, change to year
- when there is a year range indicated or multiple years (e.g., 1896-97; 1893 or 1904), change to the earliest year indicated
- when the decade is indicated (e.g. 1980s), change to the middle of the decade (e.g., 1985)
- when there is a year with text (e.g., 1960s as Harvard Hillel Children’s School) change to the year indicated and remove the text (e.g., 1965)
- empty cells, or cases where the date is unknown, change to "unknown"

Helpful Resources

- A Primer on Machine Readable data - https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data
- Open Government Directive M-10-06 - https://obamawhitehouse.archives.gov/open/documents/open-government-directive
- Open Data Policy M13-13 - https://digital.gov/open-data-policy-m-13-13/
- More on the readme style of writing metadata: https://data.research.cornell.edu/content/readme