Effective Stewardship of Humanities Data

Renaine Julian
Sarah Stanley

Overview

Defining research data as well as data in the humanities

Discussion of data management and associated issues

Funder requirements

Open Data

Data documentation and organization

What is Research Data?

“Data are outputs of research and inputs to scholarly publications and inputs to subsequent sharing and learning” (Borgman 2007)

What is Research Data?

"...The recorded factual material commonly accepted in the scientific community as necessary to **validate** research findings." (2 CFR 200.315(3))

What is data in the *humanities*?

Humanities objects of study that have been rendered machine readable, either through OCR, digitization, description, or other means

Christof Schöch on distinction between structured and unstructured data:

> Structured data is typically held in a database in which all key/value pairs have identifiers and clear relations and which follow an explicit data model. Plain text is a typical example of unstructured data, in which the boundaries of individual items, the relations between items, and the meaning of items, are mostly implicit. Data held in XML files is an example of semi-structured data, which can be more or less strictly constrained by the absence or presence of a more or less precise schema.

More on Humanities Data

See Joanna Drucker “Humanities Approaches to Graphical Display” in *Digital Humanities Quarterly*.

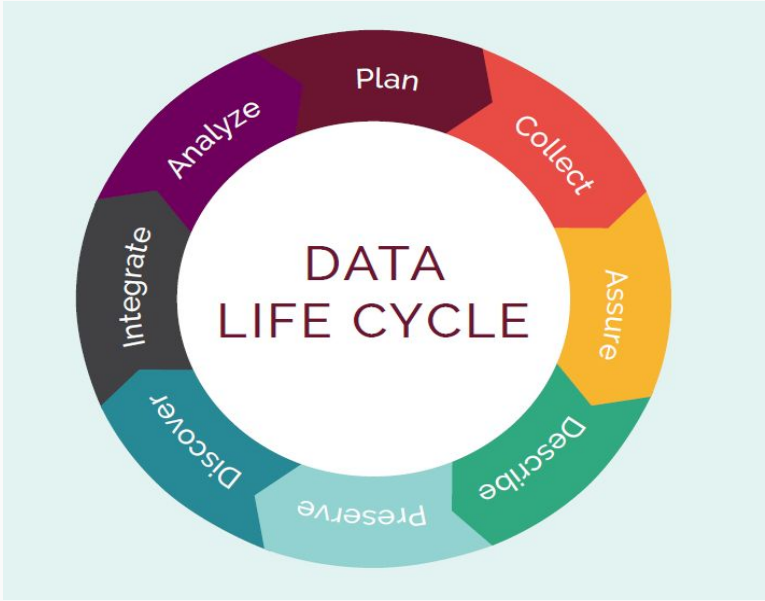
> *This requires first and foremost that we reconceive all data as capta*. Differences in the etymological roots of the terms data and capta make the distinction between constructivist and realist approaches clear. *Capta* is "taken" actively while *data* is assumed to be a "given" able to be recorded and observed. From this distinction, a world of differences arises. Humanistic inquiry acknowledges the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, not simply given as a natural representation of pre-existing fact.

What is Research Data Management?

"Research data management concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results.

It aims to ensure reliable verification of results, and permits new and innovative research built on existing information."

Whyte, A., Tedds, J. (2011)



Source: DataONE Redesigned by FSU Libraries

Why Manage Research Data?

“And yet, research data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine, it , reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself” (Gold 2007)

BILL & MELINDA GATES *foundation*

ALFRED P. SLOAN FOUNDATION

NATIONAL ENDOWMENT FOR THE
Humanities

In June 2011, The Office of Digital Humanities announced that their Digital Humanities Implementation Grant will require a DMP

DMP requirement modeled after NSF plan to “enable you to take advantage of data management resources that your institution may have already developed for applying to the NSF”

Requirement now extended to the majority of the grant programs administered by the Office of Digital Humanities

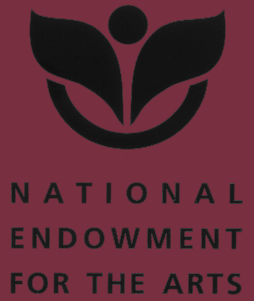
NATIONAL ENDOWMENT FOR THE
Humanities

Two main topics addressed in NEH DMPs:

1. What data are generated by your research?
2. What is your plan for managing the data?

The Office of Digital Humanities consulted the NSF Social, Behavioral and Economic Sciences Directorate when creating their own DMP requirements.

National Endowment for the Arts



September 2012, NEA's Office of Research and Analysis published *How Art Works*, which is the NEA's 5-year research agenda.

> Offers a framework for studying research topics critical to a broader public understanding of the arts' value and/or impact for individuals and communities

NEA requires applicants to submit a data management plan documenting how any **raw data** and **metadata** resulting from the proposed project will be maintained during and beyond the life of the grant

Research projects that offer plans to make data available to researchers and the public will be given special consideration in the application review process

Common Data Management Issues

Obtaining original & complete file set

Lack of data documentation

Instrument-specific and proprietary file formats

Incomplete and/or incoherent data

What can we do to make things better?

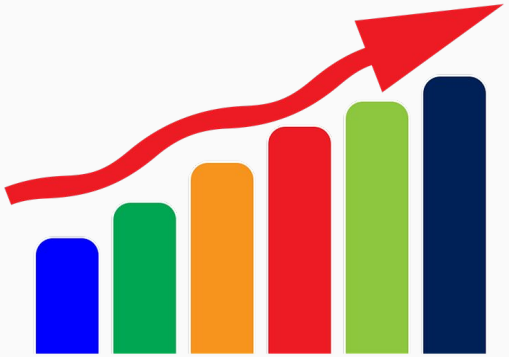
Making data openly available whenever possible

Documentation of datasets and research methodologies

Organization of datasets and documentation

Consistent file versioning

Plans for data storage and backup



What is open data?

Open Data

Available on the web

Free to download and use

Non proprietary file formats (.txt, .csv)

Copyright

Non-Open Data

Not available on the web

Data behind a paywall

Proprietary file formats (SPSS, .xls)

Copyright

What are “methods”?

Description of things like:

- ➔ How was a dataset acquired or assembled
- ➔ How to prepare raw datasets for analysis
- ➔ How to interpret Data
  - ◆ Data groupings
  - ◆ Coding
  - ◆ Units of analysis

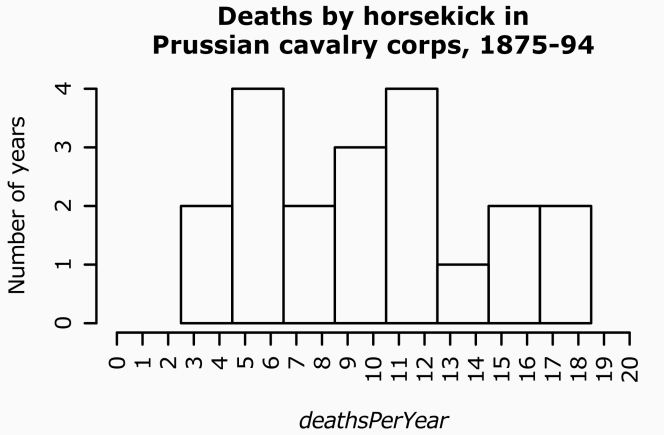


Photo source: https://commons.wikimedia.org/wiki/File:R-horsekick_totals-histogram.svg

Curating Humanities Data: An example

Text Creation Partnership

- ➔ TEI-encoded texts
- ➔ Early English Books Online
- ➔ Eighteenth Century Collections Online
- ➔ Evans Early American Imprints
- ➔ As of 2014, most texts are available to use for free!

TCP and Data Curation

- Uses TEI P3 (1999-2002)
- TEI P3 guidelines use SGML, rather than XML
- Several efforts to update (and curate) the data
  - University of Michigan now has XML versions of the texts
  - https://github.com/textcreationpartnership has TEI P5 versions of all of the phase I texts

README.txt

README.txt originates from computer science

Plain text file intended to accompany dataset or project folder

First place to look in order to understand what files are and how they should be used

Contents of README.txt

Project name

Project summary (abstract)

Location & description of previous work on project

Funding/Financial information

Primary contact information

Other people working on the project

Data Dictionaries

Ideal for documenting the structure for spreadsheets and datasets with several variables

More elaborate than README.txt usually

Allows new researcher to quickly understand contents of spreadsheets/datasets

Provide space to define variables & provide context while keeping raw data streamlined and computable

Contents of Data Dictionaries

Variable name

Variable definition

How variable are measured

Data units

Data format

Minimum and maximum values

Coded values and their meanings

Representation of null values

Precision of measurement

Ways to organize project files

By project

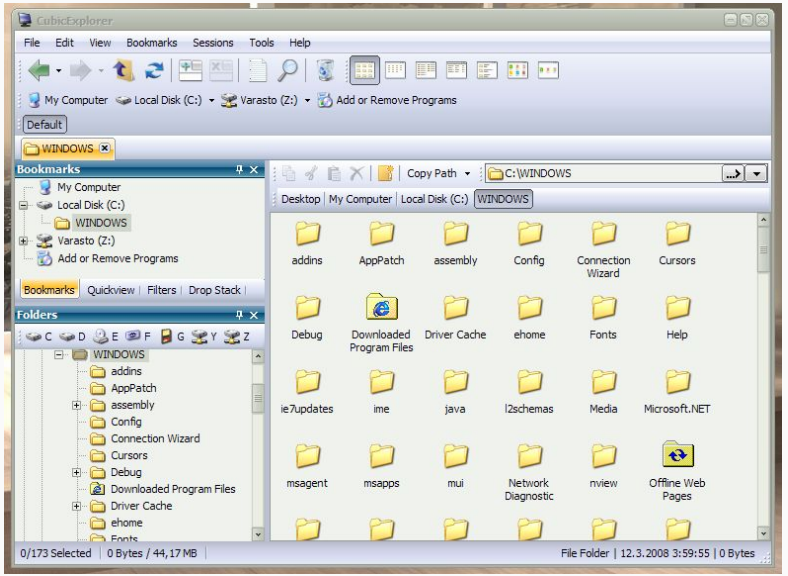
By researcher

By date

By analysis type

By data type

Any combination of the above



https://commons.wikimedia.org/wiki/File:Cubicexplorer.png

File naming conventions

Makes organizing data easier

Helps to avoid duplicate information

Use different naming conventions

Good names convey context about what the file contains by stating information like:

- Analysis type
- researcher name/initials
- Data type
- Version number

File naming best practices

Be descriptive

Be consistent

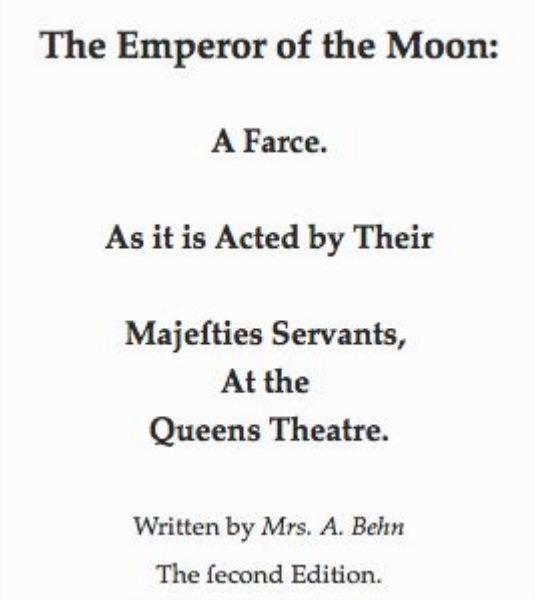
Keep it short (<25 characters)

Use underscores or dashes instead of spaces (CamelCase vs. pothole_case)

Avoid special characters such as “ / ; ! * ? {}

Follow the date conventions: YYYY-MM-DD or YYYYMMDD

Women Writers Online - Using File Names to Organize Files



The Emperor of the Moon:

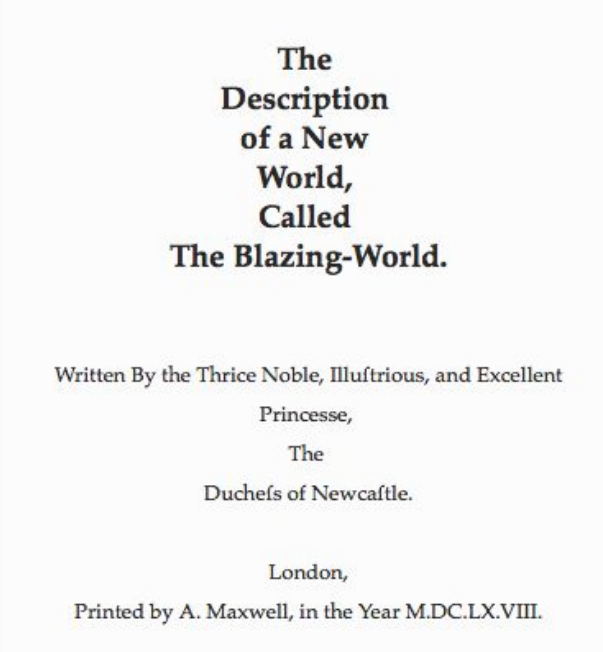
A Farce.

As it is Acted by Their

Majeſties Servants,
At the
Queens Theatre.

Written by *Mrs. A. Behn*

The ſecond Edition.



The
Description
of a New
World,
Called
The Blazing-World.

Written By the Thrice Noble, Illuſtrious, and Excellent

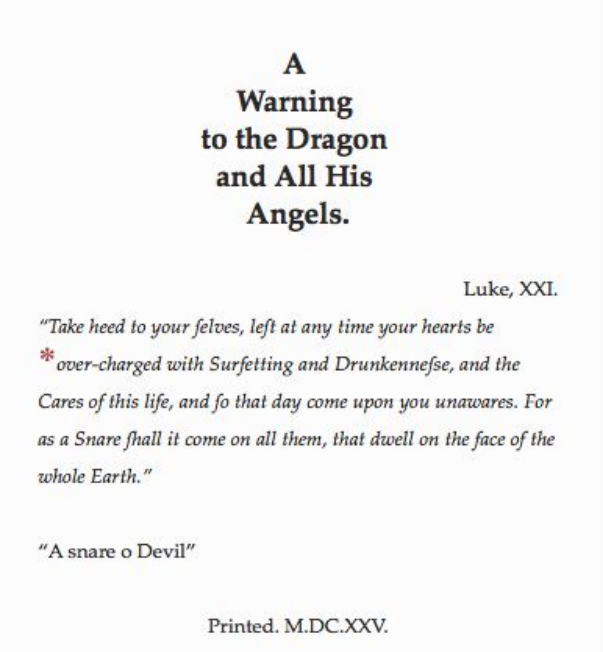
Princeſſe,

The

Ducheſs of Newcaſtle.

London,

Printed by A. Maxwell, in the Year M.DC.LX.VIII.



A
Warning
to the Dragon
and All His
Angels.

Luke, XXI.

"Take heed to your ſelves, leſt at any time your hearts be
** over-charged with Surfetting and Drunkenneſſe, and the*
Cares of this life, and ſo that day come upon you unawares. For
as a Snare ſhall it come on all them, that dwell on the face of the
whole Earth."

"A snare o Devil"

Printed. M.DC.XXV.

http://wwo.wwp.northeastern.edu/WWO/search?browse-all=yes#!/view/behn.emperor.xml
http://wwo.wwp.northeastern.edu/WWO/search?browse-all=yes#!/view/cavendish.blazing.xml
http://wwo.wwp.northeastern.edu/WWO/search?browse-all=yes#!/view/davies.warning.xml

File Versioning

Allows for the tracking of progress and reversion of earlier instances of the dataset

Useful for collaborative analysis and/or working on multiples machines

Predetermine key points for versioning (x times per day, after each type of analysis)

Differentiate versions by v_x or v-x and v_final when completed. (20170111Group1Regression_v_4.csv)

Data Backup and Security

Save files frequently, even between versions

Store copies in multiples places

- At least two physical copies (hard drive/flash drive)
- One cloud-based backup (DropBox, GoogleDrive)

LOCKSS: Lots of copies keeps stuff safe!

Open Science Framework

Open source project administered by the Center for Open Science

Makes data documentation, organization, and dissemination easier

Integrates with existing tools like:

- GoogleDrive
- DropBox
- GitHub
- FigShare

How can I apply what I learned today to my research?

- Use some type of versioning platform (like github)
- Document data collection and cleaning decisions extensively
- Open up as much of your code and data as possible
- Schedule a consultation with us!

Works Cited

Acknowledgement: The content related to data organization and documentation was based on Dr. Kristin Briney's 2015 book, *Data Management for Researchers: Organize, maintain, and share your data for research success*.

- Borgman, C.L. (2007) Scholarship in the Digital Age : Information, infrastructure, and the Internet. Cambridge, MA. : MIT Press
- Gold, A. (2007) Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. D-Lib Magazine, Volume 13 Number 9/10 http://www.dlib.org/dlib/september07/gold/09gold-pt1.html
- Office of Management and Budget Requirements, Uniform Administrative Requirements, Cost Principles, and Audit Requirements for Federal Awards, 2 CFR 200.315 (2016)
- Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities." *Journal of Digital Humanities,* vol. 2, no. 3, 2013.
- Whyte, A., Tedds, J. (2011). 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/briefing-papers