Spring February, 2013

Data Management in the Digital Humanities

Thea P Atwood, *University of Massachusetts Amherst*

Data Management for the Digital Humanities

Some terms (for our term)

- Data
  - What *is* data?
  - What *is not* data?
  - Who (or what) generates data?

Data per the US Government

- Research data is defined as the recorded factual material commonly **accepted in the scientific community as necessary to validate research findings**, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples). Research data also do not include:
  - Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
  - Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.

Office of Management and Budget, Circular A-110, September 30, 1999. White House Website, OMB Home. http://www.whitehouse.gov/omb/circulars/a110/a110.html

Data Management Defined

- Lots going on in various fields
  - MANY publications:
    - Google Scholar Search on “Data Management”: ~1.25 M hits
    - Digital Curation Bibliography: Preservation & Stewardship of Scholarly Works: http://digital-scholarship.org/dcbw/dcb.htm
    - Research Data Curation Bibliography: http://digital-scholarship.org/rdcb/rdcb.htm
  - Much that is published covers very specific topics or projects

Data Management Defined

- A sample of training websites/materials on data management:
  - MANTRA – Research Data:
    - http://datalib.edina.ac.uk/mantra/
  - SoDaMaT – Audio-Visual Data Management:
    - http://rdm.c4dm.eecs.qmul.ac.uk/category/project/sodamat
  - Managing Research Data, by Graham Pryor
  - Specifically for DH – NEH points to ICPSR

Data Management Defined

- Lots happening at Conferences & Workshops-
  - Digital Humanities Data Curation Institutes Workshop, from DH Curation
  - Digital Humanities Summer Institute
    - @Oxford, @Leipzig, @Brown
  - THATCamp (The Humanities And Technology Camp)
  - IASSIST (International Association for Social Science Information Services & Technology)
  - Librarians are (clearly) talking about it too:
    - ALA, SLA, ACRL

Challenges of Data Management

- Takes time & effort
- Easy to procrastinate
- Might not know where to start or what to do

Importance of Data Management

- **“Data Sharing and Management Snafu in 3 Short Acts, or Why Data Management is Important”**
  - http://youtu.be/N2zK3sAtr-4

Importance of Data Management

- **Meet grant and funding requirements**
  - Many funders require a description of a data management plan as part of a grant proposal
  - Some funders, such as NIH & Wellcome Trust, are withholding or discontinuing funding if a grantee doesn't follow certain guidelines (this is serious business!)
    - Right now, limited to OA pubs, but that's just a start...

Importance of Data Management

- **Increase the visibility of your research**
  - With additional information, like metadata ('data about data'), you help make your research searchable by search engines, and increase its findability.
  - Sharing research data also increases citation rate
    - See: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308
      - Findings: Publicly available data was significantly associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.

Importance of Data Management

- **Save time and money**
  - Managing something as you're working with it is much easier than trying to go back and fill in the blanks
    - Also reduce human errors! Tricky brains....

Importance of Data Management

- **Save time and money (ii)**
  - From two reports on Keeping Research Data Safe – Acquisition and Ingest of data are the most expensive components of a project
  - Sharing & preservation – much lower costs
    - See: Most WC. Keeping Research Data Safe: Cost issues in digital preservation of research data. 2:5–6. Available from: http://www.beagrie.com/KRDS_Factsheet_0910.pdf

Importance of Data Management

- **Save time and money (iii)**
  - Save money by only collecting data *once* – no needless duplication!
  - Findings from a 2012 audit state that the US loses an estimated $69M to duplicate studies.
    - See: Oransky, I. (2013, January 30). Has “double-dipping” cost U.S. science funding agencies tens of millions of dollars? « Retraction Watch. Retrieved February 8, 2013, from http://retractionwatch.wordpress.com/2013/01/30/has-double-dipping-cost-u-s-science-funding-agencies-tens-of-millions-of-dollars/

Importance of Data Management

- **Maintain data integrity and reliability**
  - "Bit rot"
  - Systematically taking care of data
  - Ensure that data is in a stable, re-usable format
    - PDF instead of .docx, CSV vs. .xls, etc.
    - In general – use non-proprietary, well-documented, commonly used & unencrypted file formats

Importance of Data Management

- **Preserve your data**
  - Know what goes where
  - Know who does what

Importance of Data Management

- **Increase research efficiency**
  - Spend less time figuring out where that data went!
  - Less time duplicating another study
  - When asked to share – *no lag time*!
    - Disseminate your findings more quickly!
    - Get your hands on relevant data in a timely manner!
    - Science is better, faster, stronger!
      - (I can't emphasize that enough – research is better and more impactful)

Importance of Data Management

- **Support Open Access (#OA)**
  - Main tenant of OA – Make research more available, at a faster rate, with as few impediments to access and re-use as possible
    - “Trapping” data or publications behind a paywall helps only the elite, only those already in the Ivory Tower
  - Opening up data facilitates this movement

Importance of Data Management

- **Facilitate new discoveries**
  - Greatest advancements occur when data from one discipline becomes useful and accessible to researchers in another
    - e.g., Galileo & the moon
      - From Ione, A. (1999). Multiple Discovery. In M.A. Runco & S.R. Pritzer (Eds). *Encyclopedia of Creativity, Vol. 2, I-Z* (pp. 261-272). San Diego, California: Academic Press.
    - Bayer
    - Other examples?

[Tangent on Data Sharing]

- Data Management also helps make data sharing easier, & data sharing has its own benefits:
  - Increase citations
  - Increases public understanding
  - Maintaining credibility of research
  - Mine data for new and interesting findings
  - Combating fraud
  - Respond to demands for evidence
  - Take on *big* issues not addressable by one lab or country, e.g. climate change, migration

[Tangent on Data Sharing]

- Rights Management
  - Important for those who want to publish their work to argue to NOT sign over copyright to publishers
  - Important for data generators to place clear use & reuse guidelines on data
    - Some, but not all, data repositories require this declaration
- How? Creative Commons Licensing!
  - http://creativecommons.org/

Dissecting the NEH Requirements

- June 2011 – NEH announced that their new program, the Digital Humanities Implementation Grant, will require a data management plan (DMP)
  - Comes on the heels of the NSF's announcement in January 2011
  - Not the only funders requiring/suggesting data management or sharing
    - Department of Energy, Environmental Protection Agency, Institute of Museum & Library Services, NASA, National Oceanographic & Atmospheric Administration, & NIH

Dissecting the NEH Requirements

- Like NSF – NEH DMP is no more than two pages in length
- Address two main topics:
  - What data are generated by your research?
  - What is your plan for managing the data?

Dissecting the NEH Requirements

- “Proposals must include enough information to enable peer reviewers to assess an applicant's data management plan.”
- “The plan should reflect best practices in the applicant's area of research, and it should be appropriate to the data that the project will generate.”

Dissecting the NEH Requirements

- “The plan should describe how the project team will manage and disseminate data generated by the project.”

Dissecting the NEH Requirements

- “The NEH Office of Digital Humanities is aware of the need for flexibility in the assessment of data management plans.”

Contents of the DMP

- **Roles and Responsibilities**
  - Who is responsible for what role?
    - e.g., Lab assistant prepares metadata; PI finalizes and deposits data, etc. etc.
  - What is the back-up plan for when an individual moves from the university?

Contents of the DMP

- **Expected Data – generated data (i)**
  - Describe what types of data will be generated or collected
    - Again, data can be many things: samples, physical collections, software, curriculum materials, code, etc.
  - Describe how that data will be generated or collected
  - Describe any issues that might restrict you from managing the data
    - e.g., legal, ethical restrictions

Contents of the DMP

- **Expected data – generated data (ii)**
  - Describe the data!
    - Data format
    - Data about the data (aka metadata, documentation, README file)
    - Remember – open-sourced, well-documented, widely-accepted formats are ideal
    - How much data you expect to generate (in whatever measurement is appropriate – data points, bytes, etc.)

Contents of the DMP

- **Period of data retention**
  - “NEH is committed to timely and rapid data distribution. However, it recognizes that types of data can vary widely and that acceptable norms also vary by discipline. It is strongly committed, however, to the underlying principle of timely access.”
  - So: outline how long you think it will take to share data if you will have an embargo period.

Contents of the DMP

- **Data formats and dissemination**
  - Sort of discussed in the generate data portion of the DMP, but -
  - Important here to flesh out what data you will make accessible, and how you will make it accessible.
    - Use this section to elaborate on your future 'home' for the data – if you have identified a repository, include that information here, a well as the repository's standards for sharing/dissemination
  - Very important to talk about how you will protect privacy, confidentiality, security, intellectual property, or other rights or requirements.

Contents of the DMP

- **Data storage and preservation of access**
  - Use this section to describe the physical and e-resources and facilities that will allow long-term access to and storage of the research data.
    - Some best practices for this too -
      - “Here, near, and far away”
        - Eg – local HD/departmental server; at sister campus; at repository in Europe
        - LOCKSS principle
      - Discipline specific repository vs. Institutional Repository (IR)
        - Is there a backup plan for these if they go out of business?
        - Is this the standard for your field?

Post-Award Monitoring

- “After an award is made, data management will be monitored primarily through the interim and final performance reports”
  - This isn't a 'one-and-done' deal – data management is an ongoing process
  - Good documentation means that this write-up will be much easier! Means you don't have to try to remember *'what did I do with that data last year/month/week?'*

Discussion

- Let's look at the two files added for today's class:
  - NEH-ODH offers suggested guidelines for what to include in an NEH DMP
  - NSF-GEN sample is a sample DMP from the DMPTool website

Data Management Planning

- DMPTool: https://dmp.cdlib.org/

Some General Resources for Data Management

- DataONE: http://www.dataone.org/best-practices
- U Oregon: http://www.nceas.ucsb.edu/news/2009/borer
- ICPSR: http://www.icpsr.umich.edu/icpsrweb/datamanagement/
- Metadata standards from DCC: http://www.dcc.ac.uk/resources/metadata-standards
- And really just anything from DCC: http://www.dcc.ac.uk/digital-curation
- Coming soon: Hampshire LibGuide on Data Management! Keep checking out the LibGuide list - http://libguides.hampshire.edu/browse.php or Thea's Twitter account: @librarianThea

- Questions?
- Email! taLO@hampshire.edu

Thanks!