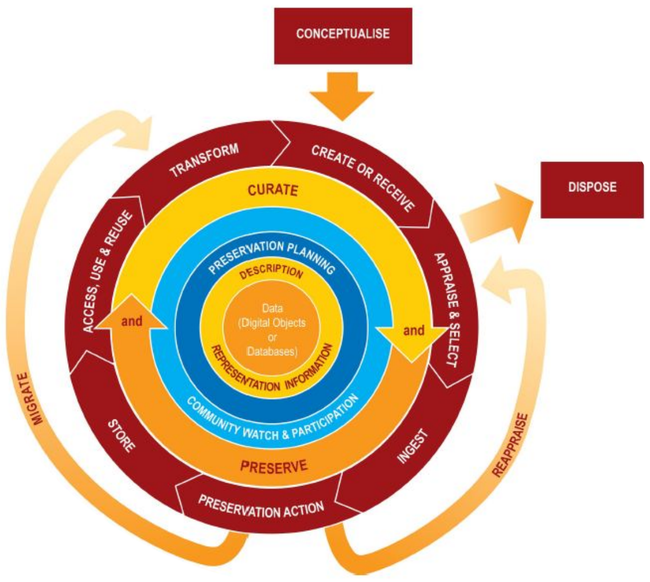
Data Curation Tasks Performed by Data Producers Throughout the Data Life Cycle

Jillian C. Wallis, *University of Michigan, School of Information*

CENTER FOR EMBEDDED NETWORKED SENSING

Introduction

- The role of the data producer in data curation appears to be to provide data to data curators (fig 1a)
- Data straight from instruments and simulation are very different from the data eventually provided for curation (fig 1b)
- ***How are data producers making their data useable?***
- Data curators could leverage the data curation work that data producers put into their personal data collections



Figures 1a & 1b: The data curation life cycle from the perspective of data curators (Higgins, 2008) and the data life cycle from the perspective of the data producers.

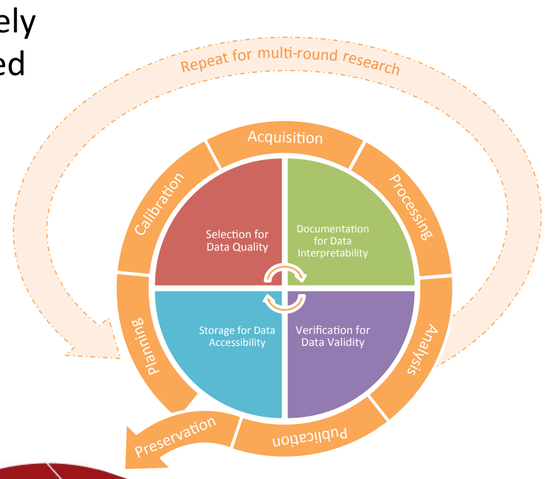
Selection	Verification	Storage	Documentation
• Selecting site/s [4] • Selecting appropriate variables, equipment, sampling frequencies [6] • Discarding outliers [5] • Selecting data to tell a story [4] • Discarding samples or data no longer of use [2] • Discarding intermediary data versions [3]	• Verification of methods [6] • Equipment calibration [4] • In-field verification [5] • Data calibration [3] • Automated tasks for removal of outliers [3] • Verification of outliers [5] • Verification of findings [5] • Verification of data presentation [4]	• Saving data to personal machines [6] • Deposit of data in lab-shared space (server, repository, cloud storage) [6] • Backing up machines [5] • Descriptive file-naming [4] • Folder structures to support access [4] • Distinguishing data versions [6] • Server/repository maintenance [6] • Deposit of data in discipline repository [1]	• Annotation in notebooks [4] • Annotation in separate, digital file [3] • Annotation in protocol sheet [2] • Ingest internal metadata [4] • Documentation of scripts/code [4] • Including research documentation in publication [6] • Including external documentation with data [5] • Including scripts/code with data [2]

Figure 3: The various data curation practices data producers reported performing during the course of their research, grouped into four categories. The number of cases in which each practice was reported is indicated in square brackets.

Temporal Situation of Data Curation Activities

- Data curation tasks performed by data producers were temporally situated, ***depending on the stage of the data life cycle*** from the perspective of the data producer (fig 4)
- Selection tasks happen largely at the beginning and end of the data life cycle
- Storage tasks only start when there is something to store and, once initiated, these tasks, such as backing-up machines, persist throughout the life cycle
- Verification and documentation tasks happen more regularly throughout the life cycle, with a different curation-oriented tasks for each stage
- Verification tasks only need to happen occasionally to assure the research is progressing appropriately
- Documentation tasks are performed consistently throughout the life cycle.

Figure 5: Combining the data production life cycle and Higgins data curation life cycle (2008) to highlight the data producer contribution to overall curation. Data curation core functions are used to stand in for all of the processes reported by data producers to achieve these ends.



Method

- Exploratory research to capture a ***rich description of data curation tasks*** performed within the context of a single, published study by each of six research groups (fig 2)
- Groups were selected from the Center for Embedded Networked Sensing (CENS), an NSF-funded Science and Technology Research Center
- Document analysis, semi-structured interviews, and field observations were coded and analyzed for emergent themes and used to ***construct models of data curation practices***.

Case	Data Selected to Tell a Story
Fungi Case	Reduced spatial scales to construct model
Hypoxia Case	Only those locations where the hypoxia event occurred
Power Case	Not mentioned
Glider Case	Only the location data collected during experiments and in simulation
Stream Case	Not mentioned
Webcam Case	Subset of webcam streams to demonstrate proof-of-concept

Figure 2: An example of researchers across cases mentioning that a type of data curation task was performed. This task was mentioned by members of four of the six groups. Cases are grouped by their domain, either science, or technology, or some blend of the two.

Typology of Data Curation Activities

- Data curation is ***"the active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities"*** (GSLIS DCEP)
- The data producers studied here performed many different tasks to ***make the data fit for their own use and support data reuse***, essentially curating the data
- The tasks performed fall within a set of four broad categories: selection, verification, storage, and documentation (fig 3)
- ***Selection tasks are similar to archival appraisal***, in which a given item is assessed for its inherent value and whether it fits some criteria, and then either selected or discarded depending on the value and fit
- ***Data producers actively verify the data***, how the data are presented in the publication, and the conclusions made from the data
- ***Storage tasks contribute to long-term integrity of the data***, which is affected by where the data are kept, who has access to the data, and whether the data are discoverable
- ***Data producers capture documentation*** in a variety of formats and at different points of the acquisition, processing, and analysis phases to support their own access and interpretation of data, ***similar to the archival task of description***

Life Cycle Stage	Selection Tasks	Verification Tasks	Storage Tasks	Documentation Tasks
Planning	Selecting site/s [4] Selecting appropriate variables, equipment, sampling frequencies [6]	Verification of methods [6]		
Calibration		Equipment calibration [4]		Annotation in lab notebooks [2]
Collection		In-field verification [5]	Saving data to personal machines [6] Deposit of data in lab-shared space (server, repository, cloud storage) [6] Backing up machines [5]	Annotation in separate, digital file [1] Annotation in field notebooks [3] Collection protocol sheet [1] Ingest internal metadata [4]
Processing	Discarding outliers [5]	Data calibration [3] Automated tasks for removal of outliers [3] Verification of outliers [5]	Descriptive file-naming [4] Folder structures to support access [4] Distinguishing data versions[6]	Annotation in lab notebooks [2] Processing protocol sheet [1] Annotation in separate, digital file [2] Documentation of scripts/code [4]
Analysis		Verification of findings [5]		
Publication	Selecting data to tell a story [4]	Verification of data presentation [4]		Including research documentation in publication [6]
Preservation	Discarding samples or data no longer of use [2] Discarding intermediary data versions [3]		Server/repository maintenance [6] Deposit of data in discipline repository [1]	Including external documentation with data [5] Including scripts/code with data [2]

Figure 4: The various data curation practices data producers reported performing during the course of their research, grouped into four categories and then mapped to the data life cycle stage in which they were reported to occur. The number of cases in which each practice was reported is indicated in square brackets.

Conclusion

- ***Data producers are contributing to the long-term curation of their own data*** through careful practices to make data useable for themselves (fig 5)
- Data management plan requirements should encourage data producers to ***follow core functions rather than best practices***
- There are ***many opportunities for data curators*** to get involved in the data production lifecycle to improve and facilitate practices