Repurposing Archival Theory in the Practice of Data Curation

Elizabeth Rolando| Wendy Hagenmaier |Susan Wells Parham

Georgia Institute of Technology

Introduction

- Expansion of data curation and digital archiving services at the Georgia Tech Library and Archives
- What are the points of intersection between data curation and archival science? What are the differences?
- How can comparing data curation and archival science lead to improvements in processing workflows and practices?

Methodology

- Process the same digital collection, once by data curator, once by digital archivist
- Data curation processing informed by OAIS Reference Model[1], ICPSR workflow[2], and UK Data Archive workflow[3]
- Archival processing informed by key archival science concepts, such as appraisal, *respect des fonds*, original order, and archival value[4], as well as documented practices at peer institutions
- Compare processing plans to discover areas of agreement and areas of conflict

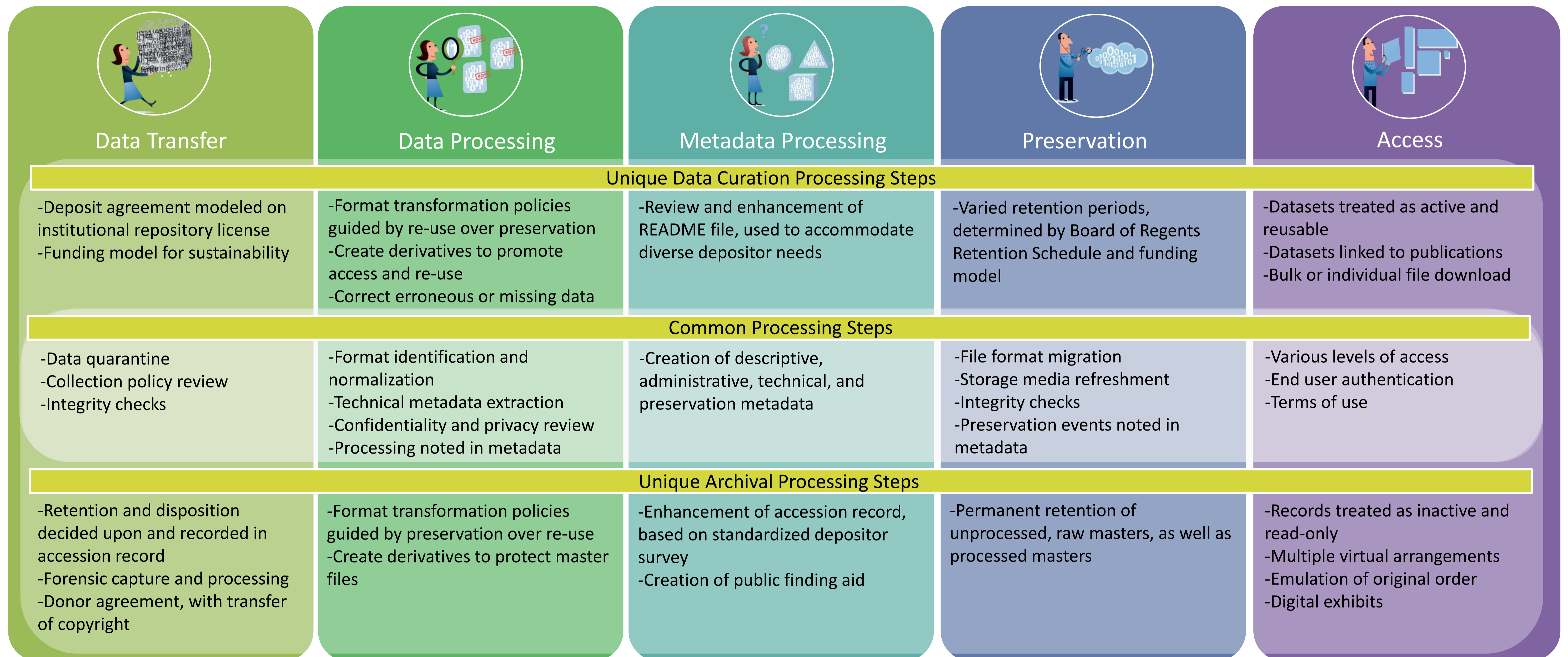| Data Transfer	Data Processing	Metadata Processing	Preservation	Access
Unique Data Curation Processing Steps				
-Deposit agreement modeled on institutional repository license -Funding model for sustainability	-Format transformation policies guided by re-use over preservation -Create derivatives to promote access and re-use -Correct erroneous or missing data	-Review and enhancement of README file, used to accommodate diverse depositor needs	-Varied retention periods, determined by Board of Regents Retention Schedule and funding model	-Datasets treated as active and reusable -Datasets linked to publications -Bulk or individual file download
Common Processing Steps				
-Data quarantine -Collection policy review -Integrity checks	-Format identification and normalization -Technical metadata extraction -Confidentiality and privacy review -Processing noted in metadata	-Creation of descriptive, administrative, technical, and preservation metadata	-File format migration -Storage media refreshment -Integrity checks -Preservation events noted in metadata	-Various levels of access -End user authentication -Terms of use
Unique Archival Processing Steps				
-Retention and disposition decided upon and recorded in accession record -Forensic capture and processing -Donor agreement, with transfer of copyright	-Format transformation policies guided by preservation over re-use -Create derivatives to protect master files	-Enhancement of accession record, based on standardized depositor survey -Creation of public finding aid	-Permanent retention of unprocessed, raw masters, as well as processed masters	-Records treated as inactive and read-only -Multiple virtual arrangements -Emulation of original order -Digital exhibits

Figure 1: Highlights from comparison between archival and data curation processing plans. The first row lists those elements of the data curation processing plan that were unique, while the bottom row lists the unique elements of the archival processing plan. The middle row identifies those elements of the processing plans that were common between the two. Images used in the diagram were created by Jørgen Stamp (www.digitalbevaring.dk) and are published under a Creative Commons Attribution 2.5 Denmark license.

Findings

What data curation might learn from archival science and processing:

- Forensic capture and processing may be valuable for certain data sets
- Existing repository license agreement models might not work for digital data sets
- Retention and disposition should be planned at the point of data transfer
- Creating virtual arrangements that emulate the data creator's original environment could be valuable
- Data curators might question how much should be done to correct data in order to facilitate re-use--how much effort is enough?

What archival science and processing might learn from data curation:

- Establish a balance between protecting the integrity of the record and facilitating future access and use—do disk images support easy access?
- Existing donor agreement and copyright transfer models might not work for digital archives acquisitions
- Funding model should be planned at the point of record transfer
- Processed records may not be "inactive"; the life of the record continues through re-use, which enriches the record and should be documented in the record itself

References

[1] Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)* (Magenta Book CCSDS 650.0-B-1). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf.
[2] Inter-university Consortium for Political and Social Research. (n.d.). *A Case Study in Repository Management.* Retrieved from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/index.html.
[3] UK Data Archive. (2014). *How We Curate Data.* Retrieved from http://www.data-archive.ac.uk/curate.
[4] Society of American Archivists, *Glossary of Archival and Records Terminology*: http://www2.archivists.org/glossary