The Data Curation Profile as a Tool for Archivists

Jake Carlson, Data Services Specialist, Associate Professor - jrcarlso@purdue.edu

Eugenia Kim, Data Services Specialist, Visiting Assistant Professor - eugeniakim@purdue.edu

Purdue University Libraries

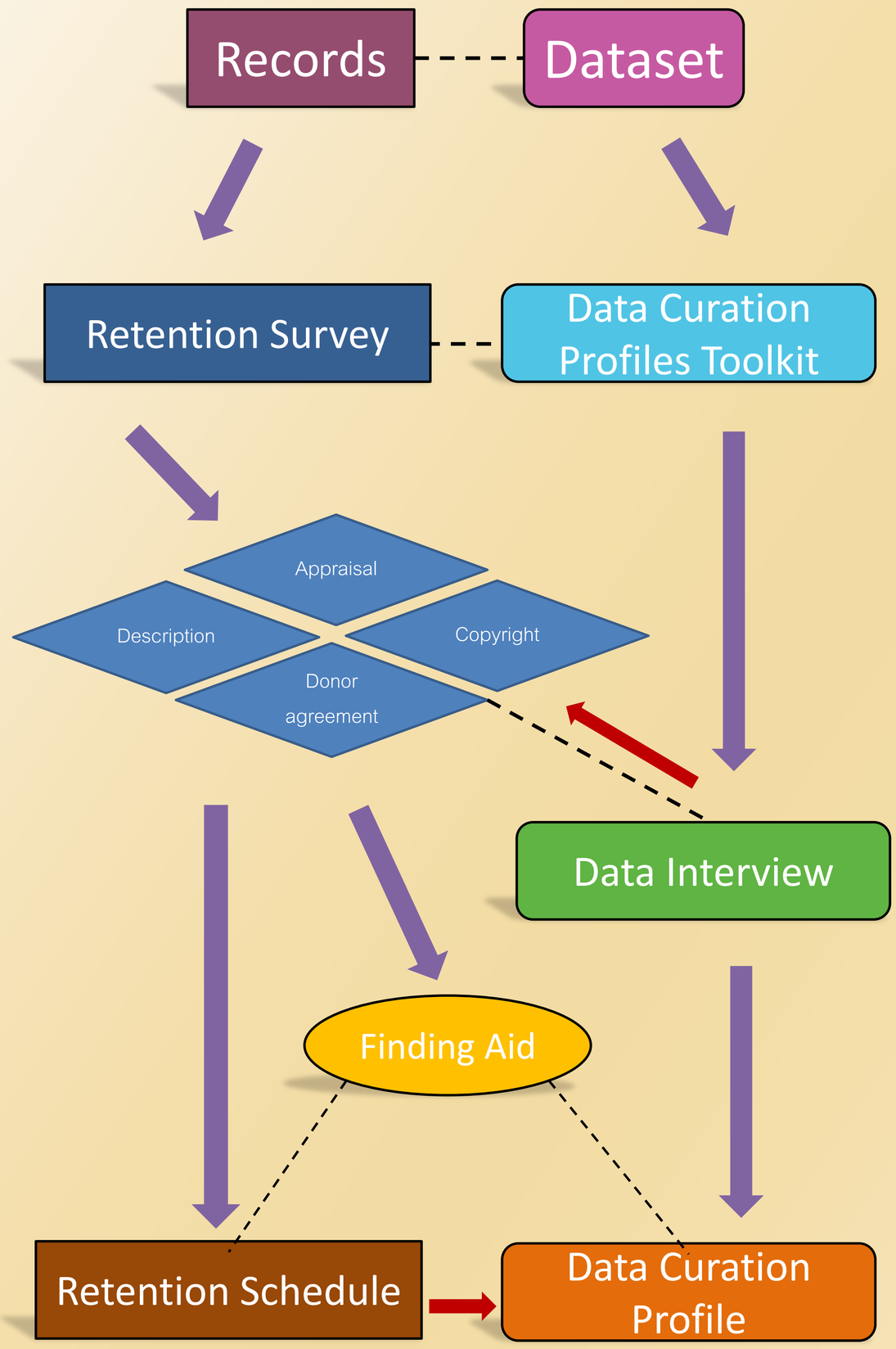
PURDUE UNIVERSITY LIBRARIES

Access. Knowledge. Success.

As the need for better practice in the curation and preservation of data gains prominence in research communities, archivists, records managers, and other information professionals are being called upon to apply their skills to these burgeoning fields. In order to address the numerous challenges of providing data services within research communities, information professionals need effective tools that will enable an informed understanding of researcher needs and requirements in the acquisition, maintenance, and archiving of their data.

Basic Correlations

The diagram below identifies a set of key concepts and terminology to be explored across archives, data curation, and records management. The black dotted lines indicate possible correlations or potential equivalencies. The cluster of blue diamonds indicate a grouping of processes. Purple arrows indicate a typical process flow between concepts and red arrows indicate possible points of influence.



Conceptual Connections

In the course of conducting a Data Interview, the following archival terminology are used for the following purposes:

Provenance – Indicates how important a record of changes over time to the dataset is to a researcher.

Donor – Defines the researcher's role in ownership of the data regarding the length of use, volume of datasets, copyright, etc.

Digital preservation – An assessment of existing digital datasets for future collaboration, use and preservation.

A Sample Data Curation Profile – Traffic Flow (Excerpt)

Profile Author	J. Carlson
Institution Name	Purdue University
Contact	J. Carlson, jrcarlso@purdue.edu
Date of Creation	October 27, 2009
Purpose	Data Curation Profiles are designed to capture requirements for specific data generated by a single scientist or scholar as articulated by the scientist him or herself. They employ a standardized set of fields to enable comparison and are designed to be flexible enough for use in any domain or discipline.
Context	A profile is based on the reported needs and preferences for these data. They are derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.
Sources of Information	• An initial interview with the scientist conducted in May 2008. • A questionnaire completed by the scientist as a part of the second interview. • A White Paper/Report written by the scientist
Scope Note	The scope of individual profiles will vary, based on the author's and participating researcher's background, experiences, and knowledge, as well as the materials available for analysis.
Editorial Note	Any modifications of this document will be subject to version control, and annotations require a minimum of creator name, data, and identification of related source documents.
Author's Note	This Traffic Flow data curation profile is based on analysis of interview and document data, collected from a researcher working in this research area or sub-discipline. Some sub-sections of the profile were left blank; this occurs when there was no relevant data in the interview or available documents used to construct this profile.

Overview of the research

Research area focus - The scientist studies real-time traffic signal performance measures…

Intended audiences - Civil Engineers, Other researchers studying transportation and traffic flow issues, State Government – esp. Departments of Transportation, Industry, General Public

Data kinds and stages

Data narrative - The scientist and his research collaborators have placed sensors and video cameras to monitor traffic flow at several intersections around Indiana…

Data Stage	Output	Typical File Size	Format	Other / Notes
Primary Data				
Raw	Sensor data	100k in 1 file per day	The format is proprietary to the sensor	**FTP downloads are mostly automated.**
Processing Stage 1	Normalized, screened for outliers & errors	Roughly 6kb	.csv / .xls	**Data are formatted into .csv before bring reformatted into a mySQL database.**
Processed	Data vectors	800 records per intersection per day. Each record has about 38 fields (floating point)	SQL / .xls	**The database typically holds 3-4 months' worth of data.**
Analyzed	Pivot charts/graphs		.xls / .emf	**Data are placed into charts and graphs for interpretation.**
Published	Pivot charts/graphs		.ppt	**Data are presented to others (incl. funders) via power point.**
Augmentative Data				
Video			Several formats – primarily "Real Video" but .wmv, .mpeg as well	**Video taken are correlated with the data for verification purposes.**
Image	**Stills taken from the video**		**.gif /.jpg / .ppt**	**Images are generated as still shots from the video.**

Preservation

Duration of preservation

The scientist believes that the data should be preserved for 3-5 years. The data may not be reliable beyond this time frame as road construction and other events may impact traffic flows.

Data provenance

Documentation of any and all changes made to her data over time is a high priority for the scientist.

The Data Curation Profiles (DCP) Toolkit is a semi-structured interview built to assist information professionals in identifying the data needs of faculty researchers. Developed by the Purdue University Libraries and Graduate School of Library and Information Science at the University of Illinois, the DCP Toolkit serves as a means of exploring issues surrounding the sharing, curation and preservation of research data from the researcher's perspective. Although developed by librarians, the DCP Toolkit was informed in part by archival theory and practice. Using the toolkit, it is then possible to create a DCP (see above) that could serve as a finding aid and retention schedule for a dataset. We envision the toolkit as a means to help archivists and other information professional engage with researchers to address a growing area of need.

Questions for Consideration

The Data Curation Profiles document a researchers' data needs through an interactive assessment process. Issues to address through collaboration between archivists, records managers, and other information professionals include:

- How does one handle datasets of major value?
- How does one distinguish between the life-span of use and the long-term value of a dataset?
- How does the use of common terminology across disciplines shape definitions over time?

Data Curation Profile – Traffic Flow

http://www.datacurationprofiles.org

Profile Author	J. Carlson
Institution Name	Purdue University
Contact	J. Carlson, jrcarlso@purdue.edu
Date of Creation	October 27, 2009
Purpose	Data Curation Profiles are designed to capture requirements for specific data generated by a single scientist or scholar as articulated by the scientist him or herself. They employ a standardized set of fields to enable comparison and are designed to be flexible enough for use in any domain or discipline.
Context	A profile is based on the reported needs and preferences for these data. They are derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.
Sources of Information	• An initial interview with the scientist conducted in May 2008. • A second interview with the scientist conducted in December 2008. • A questionnaire completed by the scientist as a part of the second interview. • A White Paper/Report written by the scientist
Scope Note	The scope of individual profiles will vary, based on the author's and participating researcher's background, experiences, and knowledge, as well as the materials available for analysis.
Editorial Note	Any modifications of this document will be subject to version control, and annotations require a minimum of creator name, data, and identification of related source documents.
Author's Note	This Traffic Flow data curation profile is based on analysis of interview and document data, collected from a researcher working in this research area or sub-discipline. Some sub-sections of the profile were left blank; this occurs when there was no relevant data in the interview or available documents used to construct this profile.

Overview of the research

Research area focus - The scientist studies real-time traffic signal performance measures…

Intended audiences - Civil Engineers, Other researchers studying transportation and traffic flow issues, State Government – esp. Departments of Transportation, Industry, General Public

Data kinds and stages

Data narrative - The scientist and his research collaborators have placed sensors and video cameras to monitor traffic flow at several intersections around Indiana…

Data Stage	Output	Typical File Size	Format	Other / Notes
Primary Data				
Raw	Sensor data	100k in 1 file per day	The format is proprietary to the sensor	FTP downloads are mostly automated.
Processing Stage 1	Normalized, screened for outliers & errors	Roughly 6kb	.csv / .xls	Data are formatted into .csv before bring reformatted into a mySQL database.
Processed	Data vectors	800 records per intersection per day. Each record has about 38 fields (floating point)	SQL / .xls	The database typically holds 3-4 months' worth of data.
Analyzed	Pivot charts/graphs		.xls / .emf	Data are placed into charts and graphs for interpretation.
Published	Pivot charts/graphs		.ppt	Data are presented to others (incl. funders) via power point.
Augmentative Data				
Video			Several formats – primarily "Real Video" but .wmv, .mpeg as well	Video taken are correlated with the data for verification purposes.
Image	Stills taken from the video		.gif /.jpg / .ppt	Images are generated as still shots from the video.

Target data for sharing - The ingest package would consist of the spreadsheets containing the processed vector data from the sensor and the corresponding images of vehicles that pass through the site...

Use/re-use value of the data - The data could be used for longitudinal studies on traffic flow and congestion issues...

Contextual narrative - The data are considered dynamic as they are still being generated and the data set continues to grow. The scientist was unable to provide an estimation of the eventual size of the data sets.

Preservation

Duration of preservation - The scientist believes that the data should be preserved for 3-5 years. The data may not be reliable beyond this time frame as road construction and other events may impact traffic flows...

Data provenance - Documentation of any and all changes made to her data over time is a high priority for the scientist.

Data audits - The ability to audit the data to ensure its structural integrity is a high priority for the scientist.

Format migration - Given the short term duration of preservation actions needed for this data, format migration is a low priority for the scientist.