Data Management—It's for Libraries Too!

Annette Bailey, Tracy Gilmore, Monena Hall, Andi Ogier, and Connie Stovall
Virginia Tech, USA

Introduction

As many academic libraries work to develop data services units to address the data needs of campus stakeholders, one early, natural avenue for testing these services lies within the library itself. During the summer of 2013, a small team of librarians endeavored to test a data audit methodology on internal library data in order to determine both the viability of the methodology as an assessment tool and as a way to audit workflows for internal data processing, naming, documenting, and storage. This team included the assistant director for collections, the collections assessment librarian, the learning commons and assessment librarian, the data science and informatics librarian, and the assistant director for electronic resources and emerging technology services. The assembled team is called the Assessment Working Group, with a stated mission to gain a better understanding of available library data and its usefulness across departments.

This paper will describe how one methodology from the field of data management, the Data Asset Framework Methodology, can be applied to library data.

Need

At the Virginia Tech University Libraries, data is collected regarding every task completed, every action taken, every link clicked, and every dollar spent. This data informs many key decisions about how the organization manages resources and provides the best possible services to our patrons. The need to better manage this data was becoming increasingly clear, and through informal conversations, the need to solve several emerging problems regarding the scalability of managing the libraries' data in the future was becoming apparent. Though this data is regarded as invaluable, rigorous application of data management policies were not in place. What data the library had, how much, and where it was stored were questions that could not be easily answered. New metrics continued to emerge and an informal group formed as concerns grew that the library needed better processes in place to ensure the availability and security of data. In 2012 the director of assessment and a team of library staff conducted an environmental scan of library data being collected. This document provided a valuable foundation for deciding what first steps to take in managing the libraries' data. The group also requested ongoing consultation from the data science and informatics librarian to support this effort.

Methodology

The Data Asset Framework (DAF) Methodology developed by the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow was recommended by the data science and informatics librarian at VT University Libraries. The group used the DAF to establish a business case and an interview protocol to determine the types of data held, who managed the data, and its importance to the organization. The team at VT (which later formed into the Assessment Working Group) interviewed the director of collections and technical services and the assistant director of collections to create a robust list of data assets, according to a modified version of the DAF guidelines, from those departments. The changes to the interview instrument included a list of "Questions Currently Asked" and "Questions to Add" (see Table 1). The questions were supplemented with a description of the question and whether the question was mandatory or optional.

Table 1

Questions Currently Asked	Description of question	Mandatory/ Optional
Report Number	Identifier	M
Report Title	Unique Identifier	M
Department	Library department responsible for the intellectual content of the data asset	M
Description	A description of the information contained in the data asset	M
Who Collects?	Library employee responsible for compiling the data asset	M
System used to collect data	Call this Source—where does data originate, how is it collected?	M
Frequency	Frequency of updates to this dataset to indicate currency	M
Start Date	Date when the data asset was created started mm/dd/yyyy	M
End Date	Date when the reporting cycle ends mm/dd/yyyy	M
Purpose (how is it used?)	Description of the main reason for the data asset's creation	M
Caveats	Exceptions to rules currently in place	O
Reported To	Person who receives the report	M
Where is it stored?	Path or web address where the data asset can be found	M
Comments	Additional information relevant to the data asset	O

Questions to Ask	Description of question	Mandatory/ Optional
Size of files	Size of the data set in MB/GB	O
Expected growth of file sizes	Expected rate of growth	O
Difficulty in replacing data	Approximate—high, medium, low	M
Time invested to collect and process data	Approximate number of hours needs to complete report	O
Process for processing and organizing data	Approximate steps needed to complete report	M
Source	Where does data originate?	M
Naming convention used	FileName.extension	M
Date last modified	mm/dd/yyyy	M needs to be automatic
Original purpose	Description of what was the main reason for the data asset's creation	M
Curation to date	History of preservation and curation activities	O

Questions Currently Asked	Description of question	Mandatory/ Optional
Usage constraints	Access restrictions applied to the data asset	O
Responsibility for the asset in the long term	Description of the retention policy and management of the data asset for the longer term	O
Preservation policy	Description of any digital preservation or curation activities planned or applied to the data asset	O
File formats	File format(s) and their version(s) the data asset is using	M
Hardware and software requirements	Description of any specialized hardware or software requirements the data asset has	M
Relation	Description of relations the data asset has with other data assets	O
Level	What level is the current description being applied (e.g., an entire collection of data objects, an individual database, a coding table used in conjunction with main database)	O
Keywords	Relevant Keywords that describe the data asset	O
Scope	Number of years of data available	O
How long do you plan to collect the data?	If this data has an end date or is ongoing	
Do you and/or colleagues review the data for current relevancy?	Is it collected because it always has been? Do people look and review to see if it's still needed	
Do you have any plans for the data in the future?		

Table 1—Questions from DAF

There were two interviews conducted with the interview ID "COLTS-ER1," named for the department and data area being audited: **COL**lections and **T**echnical **S**ervices—**E**lectronic **R**esources. In Audit Form 2, each data asset was assigned a unique identifier COLTS-XX, a data asset type, and a description. The manager of each asset was listed, occasionally being the person who was interviewed, but more frequently not. The source of the data was recorded and this question of provenance is important for data around electronic resources. It includes the systems that we use to manage our electronic resources, as well as external providers of data. The location of where data was stored and a classification of whether the data was "vital," "important," or "minor" was assigned. This was an extremely useful classification to use when discussing with the director of IT about data needs going forward and is mentioned below.

Table 2

Interview ID	Asset Unique Identifier	Data Asset Type	Description of the Data Asset	Asset Manager	Source	Location	Classification (Vital, Important, Minor)	Classification Comments	General Comments	Raw Data Report or Derived Report

Table 2—Column Titles from Audit Form 2

A space for "General Comments" was used to record comments from the interviewees regarding the data that did not fit into the other categories. General comments included observations indicating that certain types of data "will become more important" and other data, such as journal citation reports that "we haven't downloaded in a while." The Collections and Technical Services management can use these reflections to prioritize use of resources in strategic planning for the near future. Some data collection is a result of newer systems such as data from the discovery layer, both data from the discovery layer's usage reports and harvesting of raw live "click data." The live "click data" has not yet been used for collections analysis, but it is seen as a potentially valuable source of data for understanding collections usage. The final column in the interview report is an indication of whether the data is "raw" or is "derived." We classify title lists from vendors as "raw" and cost per use reports generated in-house as "derived."

Audit Form 3B was completed after the interviews were completed, given the depth of knowledge needed to answer the questions. For each unique identifier from Audit Form 2, COLTS-XX, a list of detailed data was compiled. We have included our Audit Form 3B (see Table 2). Note the level of detail needed in understanding every data asset. Once again, this instrument was invaluable in talking with key stakeholders outside of the project about technological and other resource needs regarding the library's data. From these interviews, AWG considered long-term storage and archival needs of institutional data resources.

Once the interviews were conducted, the metadata recorded, and the document discussed by AWG, the next step taken was to invite the director of collections and technical services to review the group's findings. This has led to the restructuring of directories and files on library shared server space. It also led to a conversation with the director of IT regarding the IT data storage needs of the Collections and Technical Services Department. This conversation was fruitful because of the findings from application of the DAF.

Table 3

Unique Identifier	From Audit Form 2
Title	Official name of the data asset
Variant Title	Alternative or commonly used name, if available
Structural Type	Structural type of the data asset (e.g., database, photo collection, text corpus, tabular)
Content Type	Contents of the data asset (e.g., numerical, text, mixed, photos, code, etc.)
Owner	Formal owner of the data asset in terms of intellectual rights
Rights	Indication of the user's rights to view, copy, redistribute or republish all or part of the information in the data asset
Usage Constraints	Access restrictions applied to the data asset
Source/Creator	Source of the data asset—may be a third party vendor
Asset Manager	Name of the person responsible for the current management of the asset
Original Purpose	Description of the main reason for the data asset's creation
Description	Description of the information contained in the data asset
Creation Date	Date in which the data asset was first created or received (mmddyy)
Cycle Start Date	Date when the reporting cycle begins (mmddyy)
Cycle End Date	Date when the reporting cycle ends (mmddyy)
Updating Frequency	Frequency of updates to the data asset
Completion Date	Data when the data asset was/will be completed
Date Last Modified	Latest date of data asset modification
Management History	History of maintenance of the data asset
Curation History	History of preservation and curation activities
Former Asset Managers	Chain of custody for the data asset
Usage Frequency	Estimated frequency of use and use cycles
Protection/Ethical Issues	Description of any protection or ethical issues related to the content of the data asset
Potential Reuses	Description of any potential re-uses that the asset managers can envision
Current Location	Path, location, or internet address of the asset
Source Location	Path, location, or internet address of the asset source (if third party sourced)
Relationships with Other Assets	Description of any relationships asset may have with other assets or departments
Version	Current version of the dataset
Long-Term Responsibility	Description of retention policy and management plans for the asset in the long-term
Long-Term Value	Description of the value of the data asset in the long term

Unique Identifier	From Audit Form 2
Backup/Archiving Policy	Number of copies of the data asset that are currently stored, frequency and location of backups and archiving procedures
Disaster Recovery Measures	Description of recovery processes in case of damage
Retention Period	Planned end date or retention period for the data asset (if applicable)
Preservation Policy	Description of any planned preservation or curation activities
File Format	File formats and versions the data asset is using
Software Source	Software that created the data asset
Fixity	Description of measures for ensuring the authenticity of the data asset
Current Maintenance Cost	Current maintenance costs of the data asset
Funding	Source of funding available for the data asset now and likelihood of continued devoted resources
Size	Size of the data asset in GB
Hardware/Software Requirements	Description of any specialized hardware or software requirements

Table 3—Audit Form 3B

The same small team, comprised of librarians from public services and assessment, data services, and collection management and technical services, is currently using the same methodology to audit library data globally.

Findings

The formation of the Assessment Working Group and its focus on multiple spheres of library data revealed library department silos. Such was the case with the circulation and collection management departments, where both have use for data generated by the other department. Circulation data stored by user type and subject classification proved to be incredibly useful to our collections management team who can apply this usage data to decision making when purchasing new materials or when weeding collections. Consistent and purposeful documentation, management, storage, and unencumbered internal access to these data will improve reporting, assessing, and creating vital returns on investment metrics.

References

S. Jones, S. Ross, and R. Ruusalepp, “The Data Audit Framework Methodology,” (2009), http://www.data-audit.eu/DAF_Methodology.pdf.

A. Ogier, M. Hall, A.Bailey, and C. Stovall, “Data Management Inside the Library: Assessing Electronic Resources Data Using the Data Asset Framework Methodology,” *Journal of Electronic Resources Librarianship* 26, no. 2 (2014): 101–113, doi:10.1080/1941126X.2014.910406.