

Content Management Report: DATA SETS

DATASETS

Domain/Scope

The Domain of this topic will address two distinct areas: a) DATASETS which are acquired from outside resources (commercial or academic) used by the faculty and students for research and coursework., and b) the DATASETS generated by the faculty and students. The uniqueness of DATASETS lies in the fact that the end product of the research requires unique access, storage and exchange issues.

The Scope of this area poses unique problems for access issues, storage and exchange of content and therefore concomitant suggested solutions.

Access issues are both external and internal: the cost of acquiring access to commercial datasets; how to make the Wesleyan community aware of the datasets to which we have access- commercial, open-access, results of internal research; how best to incorporate datasets in teaching and research; how to provide access to the Wesleyan faculty research with off-campus collaborators.

Storage of the datasets produced by research and coursework: where to house this data; how to ensure compatibility/migration of the data for future research/use; how to fund the storage.

Exchange of the data internally/externally: efficient file formats for portability and sharing of the data; how to generate awareness of locally produced datasets (residing either on Wesleyan servers or in off-campus data warehouses).

Sub group membership

Helen M. Aiello, Serials/E-resources Librarian

Manolis Kaparakis, Academic Computing Manager for Social Sciences, ITS

Henk Meij, Applications Technology Specialist, ITS

1. Existing practices, including the number of people involved in the practice, the technologies/systems involved in managing the process, budget for materials in that area, and the budget for systems/personnel in that area.

Wesleyan's sources for data sets are best described in two general areas:
Research data and data for teaching modules

Research data set sources (applies to both primary source and secondary source data sets)	Data sets for teaching modules
Vendor managed data sets, generally obtained from commercial publishers	Any of the same categories described for Research data set sources
Restricted access, e.g. confidential data sets	Data sets produced specifically with which to teach
Open access data sets (in some cases sources may be dubious)	
Researcher produced data sets (both faculty and students)	

- Acquisition/licensing of most commercial datasets acquired by the University are managed by the Library and their use is concentrated in DIV 2 (Social Studies, Philosophy, and Religion) and primarily by the Social Sciences faculty and students.

- The Library's Online Catalog currently includes information/linkage to data-sets, and this same information is used to populate the Data and Data Analysis website < <http://www.wesleyan.edu/its/acs/data.html> >
- DIV 3 (Mathematics and Natural Sciences) faculties acquire/use many types of datasets (publicly or privately created) primarily on their own.
- In both DIV 2 & 3, the faculty and students generate data sets which must be stored locally
- Primary responsibility for the selection of and instruction on the use of the commercial datasets for both faculty and students resides with ITS. More specifically, Manolis Kaparakis, Academic Computing Manager for the Social Sciences, evaluates and advises on selection of products and supports/instructs/assists faculty and students on the use of same.
- The Library's Collection Group makes the final decision on the acquisition of commercial datasets; decisions which are driven, largely, on the ability of the Library's Material Acquisitions budget to absorb the cost of the annual access license.
- The University network, data lab networks, classrooms equipped with the technology and equipment to assist in teaching about the datasets to students.
- The budget for the acquisition of commercial datasets in FY 05/06 was \$ 30,400.

2. Strengths: What is good about the existing practice?

- An excellent collection of major commercial data sources (ICPSR, CompuStat, Global Insight) that enhances Wesleyan's ability to attract quantitative candidates. Concomitantly, the faculty involve ITS to inform candidates of these resources and discover what candidates and new hires in the Social Sciences may need for their teaching/research.
- Initial collaboration between ITS and the library to describe commercial datasets in order to inform the Wesleyan University community of their existence.
- Supports one of the University's key capabilities: Quantitative analysis.
- The support the faculty receive from ITS in their selection, use and management of data sets.

3. Weaknesses: What could be improved?

- Inadequate data conversion utilities.
- Significant time is spent on having to know/work with a wide variety of proprietary formats.
- Handling of historical data; e.g. retrieval of data sets from a SAS version that is no longer running.
- Standardized metadata to enable permanent access to existing data sets and better enable collaboration across institutions through a 'common language'.
- Ease of pulling down massive data sets results in no incentive to economize by providing integrated access and extraction
- Not having in place, nor the staffing, a strategy to be aware of vendor conversion tools we have currently or may need in near future.
- Studied and continuous analysis of data quality.
- Balancing ease of use with the need to maintain/migrate/convert existing data sets.
- Need for better documentation for available data and reference services

4. Opportunities: What are things that other schools or institutions do that we might emulate? What trends are emerging from other sectors that we could use to our advantage?

- Sophisticated and organized presentation of what is available for use. E.g. see these sites for examples of good organization/access: <http://odwin.ucsd.edu/idata/>
<http://ssrs.yale.edu/statcat/Welcome.do;jsessionid=7A2BC133147FB2343AF352F523830B31>

- Tracking system of faculty usage of specific data to inform them of corrections and updates to the software used to manipulate the data sets.
- Sustained development of data use:
 - Technical: upgraded software, improvements/advances in data storage and migration,
 - Institutional: faculty, students,
- Collaborate with other institutions in 1) cataloging publicly available data, b) development of federated data archives.

5. Threats: What bad things could happen if we keep doing what we are doing? What bad things could happen if we change the way we are doing? What is happening in the environment that we are unprepared for now and would suggest that we need to change our ways?

What bad things could happen if we keep doing what we are doing?

- A lack of an institutional policy on data management could result in loss of data sets generated/used by faculty and students.
- Inadequate metadata can 'lose' data sets since there is no consistent/reliable access method for retrieval.
- Our ad hoc approach to data sets and their management is dependent on staff members that are stretched thin and, if they should leave, would create an immediate vacuum.

What bad things could happen if we change the way we are doing?

- The University could invest heavily (monetarily) in data storage/conversion utilities but have insignificant faculty involvement.

What is happening in the environment that we are unprepared for now and would suggest that we need to change our ways?

- Increased use of data in teaching and student projects may be hindered by cumbersome access issues.
- Granting agencies require grant recipients to make data produced under the grant available; universities can facilitate this by providing appropriate data publishing tools.

6. Short-term recommendations: What are some short-term (next six months a year) things we could do to address issues and opportunities identified in this report? What would be some concrete next steps towards making this happen? What would it cost?

- Develop data referencing services for undergraduate research.
- Quantitative Analysis Center (QAC) to provide data analysis support and explore best data publishing practices. .
- Develop a plan to involve faculty in the formation and need of an institutional program to support data set management, preservation/migration/storage, and research.
- Always include existing – and accommodate for emerging - national/international standards.

5. Long-term recommendations: What are some longer-term (two-five year) things we could do?

Implement the most critical 'best practices' identified by the QAC.

Identify systems and methods for data set storage, preservation and migration that can be integrated into a University repository (assuming we have one in place). Otherwise, identify a best stand alone model. Concomitantly, include the development of a data library to implement and manage the system.

Develop a plan to have the above stated recommendations embedded in the University budget.

8. Readings and links: What are the key readings in this area that would help others become informed about this area? What are some examples of technologies and applications of those technologies that would help us think through how to approach content management in this particular domain?

A. Chervenak, I. Foster, C. Kesselmann, C. Salisbury, and S. Tuecje. The Data Grid: towards an architecture for the distributed management and analysis of large scientific datasets.

<http://www.cs.umd.edu/class/fall2002/cmsc818s/Readings/JNCA-datagrid.pdf>

For a general, introductory article on the issues surrounding data sets and how ITS and libraries are addressing these issues:

S. Carlson (2006). Lost in a Sea of Science Data [Electronic version]. *Chronicle of Higher*

Education: INFORMATION TECHNOLOGY Section; 52 (June 23) Pg. 35

(Access restricted. Ask Helen M. Aiello, Olin, for e-mail copy of article).

C. Lynch, J. Lippincott (2005). Institutional Repository Deployment in the United States as of Early 2005. *D-Lib Magazine*; 11 (Sept.) <http://www.dlib.org/dlib/september05/lynch/09lynch.html>