The University of Kansas

KU Libraries Digital Data Services Strategy

Scott McEathron
Rhonda Houser
Geoffrey Husic
Michele Lubbers
Julie Petr
Jennifer Roach
Mickey Waxman
Xanthippe Wedel

1/8/2013

KU Libraries Digital Data Services Strategy

January 8, 2013

Contents

Introduction

As part of the University of Kansas (KU) Libraries' strategic planning process, two teams have been tasked to implement actions supporting the strategy of enhancing KU Libraries capacity for data services, data management and e-research support. One team focused on the task "to develop, cultivate, manage and support access to data collections." The other team was tasked to "design and implement an enhanced program of consultation, interaction, and education with KU's producers and consumers of research data." The aim of this white paper is to propose a model for the provision of these services.

The University strategic planning process 'Bold Aspirations' has resulted in a number of initiatives which support campus research efforts, including the recent changes and service enhancements for storage of digital research data by KU Information Technology and Research and Graduate Studies.[1] Given these changes, what roles and services are practical for the KU Libraries to undertake to support KU researcher data needs?

If we consider the traditional or past roles of academic libraries in relation to the data lifecycle (Figure 1), researchers had focused on creating, processing and analyzing data, while librarians had been engaged in preserving, providing access and facilitating re-use of data. With the growth of digital data, these traditional roles have largely persisted; however, some of the boundaries have blurred. Library and IT staff often aid researchers in the creation, processing and analysis of their digital data at a basic level. However, the two areas in which KU Libraries are currently well-positioned to aid research on campus are: facilitating access to and understanding of digital research data. Enhancing data access (sharing, publication) and data literacy can be achieved in the following ways: 1) enhance metadata quality in systems such as KU ScholarWorks; 2) better establish and promote support for data management planning; (3) invest in new data discovery and publication tools (e.g. Data Citation Index by Thomson Reuters); 4) promote and participate in the open data movement; 5) participate in new collaborative roles on behalf of the institution (e.g. VIVO, ORCID, DOI registration); (6) improve the program of general data literacy through workshops and consultation.

[1] For a detail explanations of these services see: http://technology.ku.edu/data-storage, http://technology.ku.edu/research-file-storage, and email "Secure storage options for your research data" by Bob Lim and Steve Warren, Thursday, July 05, 2012 10:16 AM.

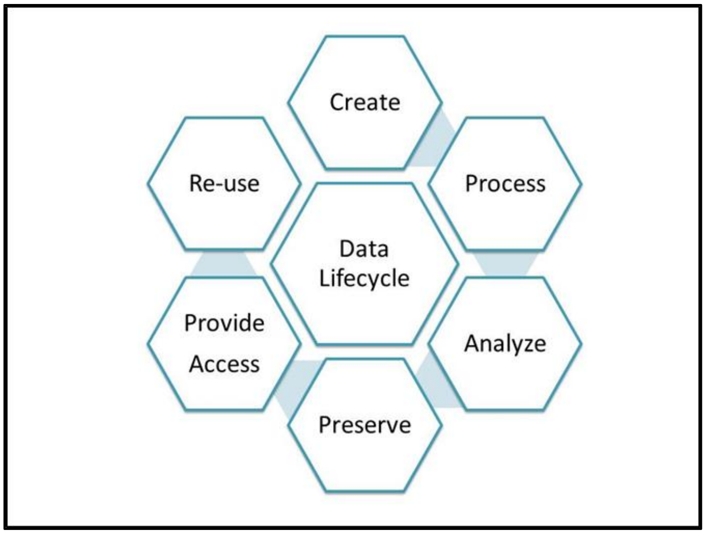

Figure 1—The Data Lifecycle.

Given uncertain financial support and rapid change in technology, long-term (indefinite) preservation of digital data is probably the area that KU Libraries, KU IT and RGS are least prepared to support. However, the Research File Storage service recently introduced to campus researchers is a step forward and adequate as a short-term[2] data preservation service. Likewise, researchers utilizing KU ScholarWorks to publish their data and make it more accessible will also benefit from its short-term preservation capabilities. As an institution, we don't yet have the instructional resources, policies, and infrastructure needed to provide long-term digital preservation. Thus, development of long-term digital preservation services is beyond the scope of our committees.

Each of the two services (Research File Storage and KU ScholarWorks) appears to have a niche, and thus can be considered complementary. Research File Storage is appropriate for researchers needing safe, secure, scalable storage for active projects; KU

[2] Long-term preservation - Continued access to digital materials, or at least to the information contained in them, indefinitely.
Medium-term preservation - Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.
Short-term preservation - Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology. Digital Preservation Coalition (2008). "Introduction: Definitions and Concepts". Digital Preservation Handbook. York, UK. Retrieved 3 December 2012 http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

ScholarWorks for researchers seeking to publish their digital data widely and measure its impact.

Proposed model of support

We propose a tiered model of support for data services within KU Libraries. The first tier, consisting of all public services staff, will need to be knowledgeable of the basic resources and services that are offered on campus and within the Libraries, and be able to refer appropriately to the domain expert. Annual training will be required for all public services staff and liaison librarians in these areas and on the basic concepts of data literacy.

A second tier would consist of domain experts, staff with specialized knowledge in the humanities, sciences, social sciences, geospatial, or other areas (Figure 2).

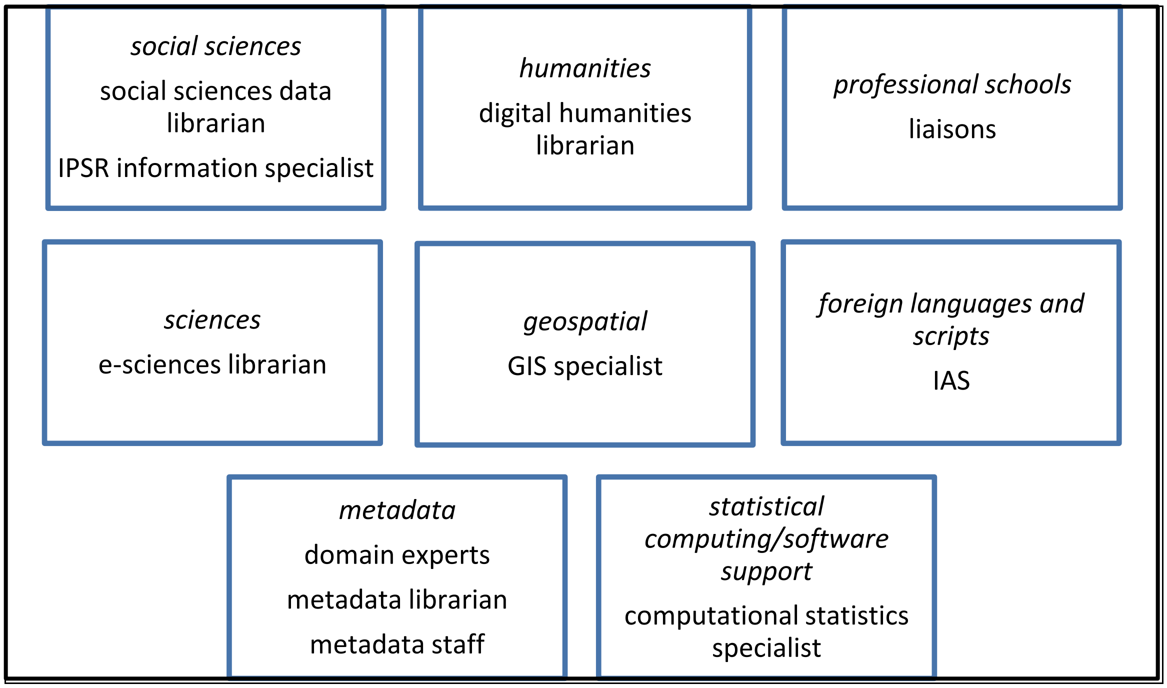
social sciences
social sciences data librarian
IPSR information specialist
humanities
digital humanities librarian
professional schools
liaisons
sciences
e-sciences librarian
geospatial
GIS specialist
foreign languages and scripts
IAS
metadata
domain experts
metadata librarian
metadata staff
statistical computing/software support
computational statistics specialist

Figure 2 –Domain Experts.

Domain experts would be responsible for data services within their discipline or specialty. These services would include instruction/training on data resources and tools, and consultation via assistance with data management planning, analysis, and visualization. Advanced training or education and ongoing cooperation among domain experts would be required.

Examples of existing services and case studies

The purpose of this section is to benchmark the breadth of current KU Library services. This is not meant to be an exhaustive list, only to provide examples.

- Individual or group consultation on where to find data for coursework, graduate study and teaching
    - Group of architecture students looking for 2-foot contour (elevation) data for city of Lawrence, KS
    - Student looking for weather data for Ecuador and Panama for a date in 1981
- Help in analyzing and manipulating data
    - Graduate student locates data for research project but data are in GIS format; student is affiliated with IPSR and meets with their information specialist; during meeting data is extracted from GIS and student is referred to Library GIS specialist for additional
    - consultation, as project would be greatly enhanced by the use of spatial analysis
    - Graduate student in Ecol. and Evol. Bio. needs to reclassify raster (pixel) data as suitable or unsuitable (binary values) warbler habitat
    - Graduate student in American Studies needs help assessing whether GIS is appropriate tool for dissertation research in food history and habits.
    - Graduate student needing help analyzing statistical data for Dissertation
- Course-integrated instruction and open workshops
    - Urban Planning: Land Use and Site Planning, Biology and Environmental Studies: Field and Lab Methods in Ecology, etc.
    - Finding GIS Data on the Web, Introduction to ArcGIS
    - SPSS I: Getting Started
    - Geography and Public Adminstration: Geostatistical Analysis (IPSR with help from Libraries)
- Collect, maintain, and make accessible digital data sets
    - City of Lawrence geospatial data in GIS and Data Lab via ArcGIS

    - Hundreds of layers that accompany ArcGIS
    - IPSR Kansas Data Archive
    - Survey datasets of The International City/County Management Association (ICMA)
    - KU ScholarWorks as digital data repository
        - Jackson and Feddema, http://hdl.handle.net/1808/8320
        - Peterson and Holder, http://hdl.handle.net/1808/9875
        - Rosenbloom and Ash, http://hdl.handle.net/1808/5491
- Data management planning (several librarians have assisted researchers in documenting their DMP)

Examples of new or enhanced services and case studies

The purpose of this section is to provide specific examples of potential new or enhanced services. A full analysis of the specific roles and resources required to implement each point will need to be developed. Again, this is not meant to be an exhaustive list, only to provide examples.

- Improve the amount of digital data deposited into KU ScholarWorks
    - Graduate students deposit the research data with electronic thesis & dissertation (ETD) into KU ScholarWorks upon completion of their degree
    - When appropriate, researchers publish their data along with copies of articles in KU ScholarWorks
- Better establish and promote support for data management planning (DMP)
    - Publish DMP website and LIBGuide
    - Offer new workshops
        - University of Virginia workshop "Data Management Bootcamp for Graduate Students"
        - University of North Carolina workshop "Understanding Data Management Plans & Funder Requirements"
    - Customize DMP tool [https://dmp.cdlib.org/]
    - Publish new KU ScholarWorks data deposition guidelines
    - Improve professional development programming
    - Develop a communications and marketing plan

- Invest in new data discovery and publication tools when appropriate and enhance metadata quality for institutional digital data assets including KU ScholarWorks
    - VIVO participation (an open source semantic web application--when populated with researcher interests, activities, and accomplishments, it enables the discovery of research and scholarship across disciplines at that institution and beyond)
    - Increase the quantity and quality of metadata in KU ScholarWorks
- Improve the program of general data literacy through workshops, undergraduate/graduate level research based learning, and consultation
    - Basic training for public services staff and subject liaisons
    - New software/programing tools and workshops, e.g. "R for statistical computing"--Center for Research Methods and Data Analysis/ LAS 792
- Improve campus digital data awareness; contribute to national data policy solutions
    - Support and participate in 1st North American Data Documentation Initiative User Conference April 1-3, 2013 (KU Libraries and the Institute for Political and Social Research (IPSR) are joint sponsors)

Models at peer institutions

Most of KU Libraries' peer organizations are at similar places in changing their organizational structures from primarily supporting traditional scholarly communications toward emerging areas such as digital data services and e-research support. A few notable organizations have led the way including: Purdue, the University of New Mexico, and the University of Virginia. Many of the services and approaches at these institutions have been utilized by others and have assisted us in our service enhancements suggestions. The purpose of this section is to explore the staffing approach at each of these Libraries.

Purdue Libraries has become best known for its D2C2 (Distributed Data Curation Center) which advances understanding of issues in curating research data sets in distributed environments. However, this is only one component of a much broader suite of services. The D2C2 program is part of the Research Department of the Libraries, led by an Associate Dean (Scott Brandt). The Research Department "supports three areas: library and archival science research; development of innovative services that respond to changes across the scholarly communication spectrum; and its research administration.” Besides D2C2, they “also engage in and support research in other areas of library and archival science research. It is also the central driving force behind the Libraries' Data Services, a distributed service across the Libraries that works with researchers to address practical problems related to research data.” Staffing in the department includes: two Digital Library Software Developers, Data Services Specialist, Digital Data Repository Specialist, and Interdisciplinary Research Librarian.[3] Moreover, subject librarians and other domain specialists such as the GIS Specialist support the digital data services strategy.

The University of New Mexico Libraries has a similar framework to what we propose. Several domain experts (Data Librarian for Science & Engineering, Data Librarian for Business & Economics, and Data Librarian for Social Science/Humanities) report through the Libraries Outreach & Research Services department. Further, two high-level administrative positions are held within the Libraries’ (e-Research Center Director and Director of e-Science Initiatives). The University of New Mexico is a leader in the *DataONE* (Data Observation Network for Earth) –an NSF funded consortium of distributed data centers, science networks or organizations for the preservation, access, use and reuse of multi-scale, multi-discipline, and multi-national environmental science data.

The University of Virginia Libraries has digital data domain experts spread across its organization. Several positions are with the Digital Research and Scholarship unit as well as the Science, Engineering, and Education units. This structure suggests a more robust level of staffing (Social Sciences Data Librarian, two GIS Specialists, and two Data Specialists).

Glossary of terms and concepts

One challenge of forging a path forward is the abstract nature of many of the terms and concepts associated with the use and management of digital data. Because various professions and subject domains also have different understandings and usage of these same terms, it is important to state clear definitions to support shared

[3] http://www.lib.purdue.edu/research/

understanding. A shared understanding is essential to forge a cooperative approach to support research and establish a path forward. It is also important to improve the overall institutional confidence in using these terms and understanding the concepts.

Data: an abstract term that forms the lowest level of abstraction from which information and then knowledge are derived; may be structured or unstructured; digital or analog; factual numbers, words, images, etc., accepted as they stand that are often used as a basis for reasoning, discussion, or calculation.[4]

> **Research data**: facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analyzed, experimental or observational. Data includes: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artifacts; specimens; samples. Provenance information about the data might also be included: the how, when, where it was collected and with what (for example, instrument). The software code used to generate, annotate or analyze the data may also be included.[5]

Data archiving or digital archiving: The library and archiving communities often use it interchangeably with *digital preservation* (see below). Computing professionals tend to use digital archiving to mean the process of backup and ongoing maintenance as opposed to strategies for long-term digital preservation.

Data management plans (DMP): the documentation of research data management practices and any responsibilities such as university policies, ethics, intellectual property, attribution etc. Its purpose is to ensure the quality of your research data and outputs, integrity and repeatability, appropriate access to data, and appropriate reuse of data for subsequent research. A DMP may be created for a department, a project or collaboration. The responsibility of implementing and following the DMP lies with the involved researchers, IT managers and data managers.[6]

[4] Wikipedia; http://www.merriam-webster.com/dictionary

[5] Courtesy of The University of Melbourne, https://policy.unimelb.edu.au/MPF1242

[6] Courtesy of The University of Melbourne; https://policy.unimelb.edu.au/MPF1242

Digital data curation: *"digital curation"* or *"data curation"* often used interchangeably: the process of establishing and developing long term repositories of digital assets for current and future research. Steps include selection, description (via metadata), maintenance, preservation, and providing access.

Digital data management: or *"data management,"* the processes of creating, organizing, and making accessible and preserving digital research data (may include conventions for naming and structuring files and folders, version control, backing up of data; and metadata documentation of provenance).

Digital preservation: refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation is defined very broadly for the purposes of this study and refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change.[7]

E-Research: a broader term than *"e-Science,"* research which utilizes digital technology within the research process, including sciences, social sciences and humanities.

E-Science: an area of scientific research characterized by intensive use of computing infrastructure, highly networked environments and vast amounts of digital data.

Metadata: loosely define as *data about data*; data or information about one or more aspects of the data content. For example, card catalogs of libraries are a form of metadata.

[7] Digital Preservation Coalition (2008). "Introduction: Definitions and Concepts". Digital Preservation Handbook. York, UK. Retrieved 3 December 2012
http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts