圖書館員如何開拓研究數據管理服務新領域

Developing Data Services to Support Research Data Management

Jian Qin 秦健
School of Information Studies
Syracuse University

ADLS, Qingdao, China 2014-06-25
Types of data services
Characteristics of data services
A quick introduction of scientific data management

WHAT ARE DATA SERVICES?
Types of data services (1)

For whom?

Infrastructure type of services:
- National
- Institutional
What data services?

ARL survey report 2010

Finding relevant data 83%
Developing data management plans 79%
Finding and using available technology infrastructure and tools 76%
Developing tools to assist researchers 76%
Archiving and curating relevant data and curating it for long-term preservation and integration across datasets
Providing curatorial and data Stewardship services
Raising awareness and user training

ARL Survey report 2013 (N=72)

<table>
<thead>
<tr>
<th>Service</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Providing an institutional repository</td>
<td>89%</td>
</tr>
<tr>
<td>Locating &amp; using existing data sources</td>
<td>94%</td>
</tr>
<tr>
<td>GIS and geospatial analysis, support</td>
<td>85%</td>
</tr>
<tr>
<td>Dataset purchase, acquisition, subscriptions</td>
<td>81%</td>
</tr>
<tr>
<td>Copyright &amp; patent advising</td>
<td>75%</td>
</tr>
<tr>
<td>General statistical software support</td>
<td>58%</td>
</tr>
<tr>
<td>Data visualization support</td>
<td>36%</td>
</tr>
<tr>
<td>Data analysis support</td>
<td>39%</td>
</tr>
<tr>
<td>Data mining</td>
<td>28%</td>
</tr>
<tr>
<td>Database design &amp; management</td>
<td>28%</td>
</tr>
<tr>
<td>Programing/software development</td>
<td>24%</td>
</tr>
<tr>
<td>Other data support services</td>
<td>29%</td>
</tr>
</tbody>
</table>

(Fearon et al., 2013)
Data Management Plan (DMP) services

- Online DMP services:
  - Explanation of DMP requirements by different funding agencies and/or NSF directorates
  - Guidelines for creating DMPs
  - Template examples of DMPs
  - A tool or resource for DMP creation
  - A data planning checklist
  - Copyright considerations, data citation guidelines, metadata examples, info about digital repository services
  - ...
What data services? (2)

• Submission of data
• Data export
• Data format conversion /transformation
• Access to data (discovering and obtaining data)
• IP protection and management
• Educational offerings
• Technical assistance including data management and manipulation services
  • Access to computing facilities
  • Curation
  • Archive and preservation tools
  • Information
• Print and publication services
• Marketing
• Publicity
• Software development services

Source: Marcial & Hemminger, 2010
Data providers, managers, and users

Data center
- Acquiring, processing
- Conversion /Transforming
- Metadata tagging
- Discovering and obtaining
- Analyzing, visualization
- Archiving, curating, preservation
- Training, outreaching
- Marketing, publicizing
- Distributing, publishing

Research center

Library

Who best suits for which services?
What is considered as good data services?

- Repeatable
- Sustainable financially and technically
- A community of practice
- Institutionalization
- Collaboration and coordination
- Conformance to regulations and laws
What are data?
What are some of the major data formats?
Why data formats?

FUNDAMENTALS OF DATA
What are data? (1)

What are data? (2)
What are data? (3)
Medical and health data

A Charter for Health Data

- **Standardization**
- **Compliance**
- **Security**

http://www.weforum.org/issues/charter-health-data

Qingdao, China, June 25, 2014
The multi-dimensions of data

- Data types
- Levels of processing
- Data formats

Research orientation
Scientific data formats

- Common data format
- Image formats
- Matrix formats
- Microarray file formats
- Communication protocols

Data model:
- Hierarchical
- Relational
- Object-oriented
- Network

Scientific data formats:
- Metaformats
- Data structures
- Physical data

- DSV
- CSV
- XML

Metaformats:
- Tuple
- Set
- List
- Array
- Tree

Physical data:
- Bits
- Bytes
- Characters
- Strings
Software specific data formats

Quantitative data analysis
- SPSS
- SAS
- R

Microsoft Office
- Open Office

Qualitative data analysis
- NVivo
- QDA
- Atlas.ti

DSV  CSV  XML
Scientific & medical data formats

- Medical and Physiological Data Formats
  - BDF — BioSemi data format (.bdf)
  - EDF — European data format (.edf)
- Molecular Biology data Formats
  - PDB — Protein Data Bank format (.pdb)
  - MMCIF — MMCIF 3D molecular model format (.cif)
- Medical Imaging
  - DICOM — DICOM annotated medical images (.dcm, .dic)
- Chemical Formats
  - XYZ — XYZ molecule geometry file (.xyz)
  - MOL — MDL MOL format (.mol)
  - MOL2 — Tripos MOL2 format (.mol2)
  - SDF — MDL SDF format (.sdf)
  - SMILES — SMILES chemical format (.smi)
- Bioinformatics Formats
  - GenBank — NCBI GenBank sequence format (.gb, .gbk)
  - FASTA — bioinformatics sequence format (.fasta, .fa, .fsa, .mpfa)
  - NEXUS — NEXUS phylogenetic data format (.nex, .ndk)
Summary

- Scientific data formats are closely tied to scientific computing
  - Data structure, model, and attributes
  - Self-descriptive with header/metadata
  - API for manipulating the data
  - Interoperability: conversion between different formats
- No one-format-fits-all standard
- Each standard has one or more tools for creating, editing, and annotating dataset
What is a dataset?
What are some of the metadata standards for describing datasets?
What is data management?

DATASETS, METADATA, AND DATA MANAGEMENT
Dataset classification

**Volume**
Large-volume
Small-volume

**Collection source type**

**Discipline**

**Level of interpretation**
Measurements
Observations
Derived observations
Synthetic observations
Interpretations

**Data structure**
Ecological data example: Instantaneous streamflow by watershed

http://www.hubbardbrook.org/data/dataset.php?id=1

<table>
<thead>
<tr>
<th>WS</th>
<th>MO</th>
<th>DA</th>
<th>YR</th>
<th>TIME</th>
<th>GAGE FT.</th>
<th>GAGE CM.</th>
<th>DISCH. C.F.S.</th>
<th>DISCH L/S</th>
<th>INTERV. INCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>08</td>
<td>455</td>
<td>0.240</td>
<td>7.370</td>
<td>0.0700</td>
<td>2.0100</td>
<td>0.01000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>08</td>
<td>102</td>
<td>0.240</td>
<td>7.370</td>
<td>0.0700</td>
<td>2.0100</td>
<td>0.01000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>593</td>
<td>235</td>
<td>0.240</td>
<td>7.430</td>
<td>0.0700</td>
<td>2.0500</td>
<td>0.01000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>2400</td>
<td>645</td>
<td>0.230</td>
<td>7.280</td>
<td>0.0600</td>
<td>1.9500</td>
<td>0.00643</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>1357</td>
<td>2004</td>
<td>0.230</td>
<td>7.250</td>
<td>0.0600</td>
<td>1.9300</td>
<td>0.01000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>2400</td>
<td>447</td>
<td>0.220</td>
<td>6.790</td>
<td>0.0500</td>
<td>1.6600</td>
<td>0.00922</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>1195</td>
<td>1822</td>
<td>0.220</td>
<td>6.700</td>
<td>0.0500</td>
<td>1.6300</td>
<td>0.01000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>2384</td>
<td>2400</td>
<td>0.210</td>
<td>6.670</td>
<td>0.0500</td>
<td>1.6100</td>
<td>0.01000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>2400</td>
<td>573</td>
<td>0.220</td>
<td>6.730</td>
<td>0.0500</td>
<td>1.6400</td>
<td>0.00033</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>1184</td>
<td>1725</td>
<td>0.220</td>
<td>6.880</td>
<td>0.0600</td>
<td>1.7000</td>
<td>0.01000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>08</td>
<td>2227</td>
<td>2383</td>
<td>0.220</td>
<td>6.820</td>
<td>0.0500</td>
<td>1.6700</td>
<td>0.01000</td>
</tr>
</tbody>
</table>

Qingdao, China, June 25, 2014
Diabetes data and trends—Country level estimates:

Diabetes Data & Trends home page:
http://apps.nccd.cdc.gov/ddtstrs/default.aspx
Clinical trials data management:
http://www.clinicaltrials.gov/ct2/show/NCT00006286?term=TADS+NIMH&rank=1

Treatment for Adolescents With Depression Study (TADS)

This study has been completed.

First Received on September 14, 2000. Last Updated on January 18, 2008

Purpose

TADS is designed to compare the effectiveness of established treatments for teenagers suffering from major depressive disorder (MDD). The treatments are: psychotherapy ("talking therapy"); medication; and the combination of psychotherapy and medication. Together, 32 teenagers (both males and females) ages 12 to 17, will take part in this study at 12 sites in the United States.

The TADS design will provide answers to the following questions: What is the long-term effectiveness of medication treatment of teenagers who have major depression? What is the long-term effectiveness of a specific psychotherapy ("talking therapy") in the treatment of teenagers who have major depression? How does medication treatment compare with psychotherapy in terms of effectiveness, tolerability and teenager and family acceptance? And, what is the cost-effectiveness of medication, psychotherapy and combined treatments?

The medication being used in this study is called fluoxetine. Fluoxetine is also known as Prozac. Research has shown that medications like Prozac help depression in young persons. Fluoxetine has been approved by the FDA for use in the treatment of child and adolescent (ages 7 to 17 years) depression.

The psychotherapy or "talking therapy" being used in this study is called Cognitive Behavioral Therapy (CBT). CBT is a talking therapy that will teach both the teenager and his or her family member (e.g., parent) new skills to cope better with depression. Specific topics include education about depression and the causes of depression, setting goals, monitoring mood, increasing pleasant activities, social problem-solving, correcting negative thinking, negotiation, compromise and assertiveness. CBT sessions may also help with resolving disagreements as they affect families.
Common in the examples

• Attributes of a dataset tell users/managers:
  – What the dataset is about
  – How data was collected
  – To which project the data is related
  – Who were responsible for data collection
  – Who you may contact to obtain the data
  – What publications the data have generated
  – ??
Metadata standards in medical & health sciences

**Structure**
- Healthcare
- Medical images
- Bioinfomatics

**Semantics**
- NCBI Taxonomy
- NCBO Bioportal
- UMLS
- MeSH (Medical Subject Headings)
- SNOMED CT (Systematized Nomenclature of Medicine--Clinical Terms)
Datasets are the objects in data collections...
<table>
<thead>
<tr>
<th>Reference collection</th>
<th>Size</th>
<th>Metadata Standards</th>
<th>Management</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Larger, discipline-based</td>
<td>Multiple, comprehensive</td>
<td>Organized Institutionalized, Heroic individual inside the team</td>
</tr>
<tr>
<td>Resource collection</td>
<td>Smaller, team-based</td>
<td>None or random</td>
<td></td>
</tr>
<tr>
<td>Research collection</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Datasets, data collections, and data repositories

- **Data collections** are built for larger segments of science and engineering.

- **Datasets**
  - typically centered around an event or a study
  - contain a single file or multiple files in various formats
  - coupled with documentation about the background of data collection and processing

System for storing, managing, preserving, and providing access to datasets.

- A repository may contain one or more data collections.

- A data collection may contain one or more datasets.

- A dataset may contain one or more data files.
Exciting opportunities,

Practical challenges
Planning tool: SWOT analysis
Example of SWOT analysis:
ARL/DLF’s E-Science Institute

• Baseline module: develop a basic understanding of eResearch
• Context module: environment scan
• Building blocks module: SWOT analysis of organizational change, collaboration, sustainability, policy

This example is based on Gail Steinhart’s guest lecture to the Data Services course at Syracuse iSchool, February, 2012.
ARL/DLF’s E-Science Institute: Baseline module

- Self-assessment:
  - Organization of research, IT, cyberinfrastructure at your institution
  - Major sources of funding, research areas
  - Key research centers, individuals, partnerships

- Identifying interview candidates: university administration, IT leaders, key faculty/researchers
ARL/DLF’s E-Science Institute: Context module

- Self-assessment:
  - Institutional culture & priorities
  - Current infrastructure and staffing
  - Library & change
  - Sustainability and assessment of initiatives
  - Other activities
ARL/DLF’s E-Science Institute: Building blocks module

- Turn interview transcripts and self-assessments into “statements”
- Classify statements as **Strength**, **Weakness**, **Opportunity**, **Threat**
Starting the Conversation: University-wide Research Data Management Policy

Ricky Erway
Senior Program Officer
OCLC Research

http://oclc.org/content/dam/research/publications/library/2013/2013-08.pdf
Key issue 1: Staffing

- ARL survey report (2013):
  - Subject librarian or liaison (50%)
  - Digital (38%)
  - Data librarian (18%)
  - Metadata (17%)
  - Data services (13%)
  - GIS or Geospatial (12%)
  - Research data (11%)
  - Curation (11%)
  - Repository (10%)
  - Software or systems (9%)
  - Data management (9%)
Key issue 2: funding

- Internal library regular budget 98%
- Direct administrative funding (separate from library funds) 11%
- External grant funding 11%
- Internal library temporary or special project budget 9%
- Department or research institute/project group funds 6%
- Endowment fund 6%
- Fee to researcher or researcher’s grant 4%
- Facilities and administrative (F&A) funding 2%
- Other source of funding 9%

(N=57, from Fearon et al., 2013)
Key issue 3: training

- Training / Experience most important to RDM:
  - Subject domain expertise
  - Digital/data curation training
  - IT technology or services experience
  - Library MLS/MLIS training

- Particularly important:
  - Research methods and data analysis
  - Research data management
  - Data curation
  - Scholarly communication

- Other skills and training:
  - Identifying and applying metadata standards
  - Digital preservation
  - Data ownership policies
  - Ethical and legal issues
  - Data security
  - Data sharing and access
  - Data storage and backup planning
  - Data retention policy
  - Data citation
Key issue 4: data policies

- Institutional data policies
- Project data policies

Data policies are a major component of the institutionalization of research data management
Cornell Research Data Management Services

Services and expertise related to the management of research data are distributed throughout multiple units at Cornell University. This is meant to serve as a directory to research data services at Cornell.

- Data Management Planning Overview - resources for meeting funding requirements, and general planning guidance
- Storage and backup
- Metadata
- Data analysis
- Collaboration tools
- High performance computing
- Privacy and confidentiality
- Intellectual property and copyright
- Data publication

See a list of all service providers listed in these pages.

https://confluence.cornell.edu/display/rdmsgweb/services
CMM for RDM: Guidelines for research data management

http://rdm.ischool.syr.edu/xwiki/bin/view/Main/

Welcome to CMM for RDM

The Model

Go straight to the model:
Capability Maturity Model for Research Data Management

About

The broad goals of this project are to document, foster and promulgate best practices in research data management (RDM), practices that support research transparency and the replication of scientific results. We do so in order to cultivate a new generation of researchers and data managers who are both the best practice beneficiaries and contributors.

Furthermore, as more organizations invest in RDM, it has become increasingly important for administrators, researchers, and managers to be able to evaluate RDM processes for sustainability, efficiency, and effectiveness, which requires a baseline for comparison.
Case Discussions
Case Study #1: To build or not to build a data repository?

A university library has developed an institutional repository for preserving and providing access to the scholarly output by the researchers in this institution. Now the new challenge arises from e-science research demanding data management plan by the funding agency and the linking between publications and data by the authors and users. You already know that some faculty use their disciplinary data repository for submitting their datasets (e.g., GenBank for microbiology research data). The problem you face now is whether an institutional data repository should be built for those who do “small science” and don’t have funding nor expertise to manage their data.

Questions to be addressed:
• What are the strategies you will use to approach the problem?
• What are the possible solutions for the problem?
• What are some of the tradeoffs for the solutions you will adopt?
Case study #2: Developing a data taxonomy

The concept of research data management is a stranger to many faculty as well as your library staff. What is data? What is a data set? These seemingly simple terms can be very confusing and have different interpretations in different contexts and disciplines. As part of the data management strategies, you decide to develop an authoritative data taxonomy for the campus research community. This data taxonomy will benefit the creation and use of institutional data policies, data repository or repositories, and data management plans required of funding agencies.

Questions to be addressed:
• What should the data taxonomy include?
• What form should it take, a database-driven website or a static HTML page?
• Who should be the constituencies in this process?
• Who will be the maintainer once the taxonomy is released?
Case study #3: Developing a data policy

Data policies play an important role in governing how the data will be managed, shared, and accessed. It is also an instrument that will fend off potential legal problems. Data policies have several types: data access and use, data publishing, and data management. Your university’s Office of Sponsored Research has some existing policy on data, but it is neither systematic nor complete. Many of the terms were defined years ago and did not cover the new areas such as the embargo period of data. As the university has decided to build a data repository for managing and preserving datasets, a data policy has become one of the top priorities for both the institution and the data repository.

Questions to be addressed:
• What should the data policy include?
• Who should be the constituencies in this process?
• Who will be the interpretation authority for the data policy?
Case study #4: Cataloging datasets

Describing datasets is the process of creating metadata for datasets. In scientific disciplines, several metadata standards have been developed, e.g., the Content Standard for Digital Geospatial Metadata (CSDGM), Darwin Core, and Ecological Metadata Language (EML). Each of these metadata standards contains hundreds of elements and requires both metadata and subject knowledge training in order to use them. Besides, creating one record using any of these standards will require a tremendous time investment. But your library does not have such specialized personnel nor have the fund to hire new persons for the job. The existing staff has some general metadata skills such as Dublin Core. In deciding the metadata schema for your data repository, you need to address these questions:

- Should I adopt a metadata standard or develop one tailored to our need?
- How can I learn what metadata elements are critical to dataset submitters and searchers?
- What are some of the benefits and disadvantages for adopting a standard or developing a local schema?
Case study #5: Evaluating data repository tools

Research data as a driving force for e-science is inherently a tool-intensive field. Tools related to data management can be divided into two broad categories: those for creating metadata records and those for data repository management. An academic institution decided to build their own data repository as part of the supporting service for researchers to meet the data management plan requirement of funding agencies. This data repository development task was handed down to the library. You the library director have to decide whether to develop an in-house system or use an off-the-shelf software system. As usual, you put together a taskforce to find a solution to this challenge. The questions to be addressed by the taskforce include:

- What are the options available to us?
- What evaluation criteria are the most important to our goal?
- What are the limitations for us to adopt one option or the other?
- How will this option be interoperable with existing institutional repository system? Or, can the existing repository system used for data repository purposes?
References


