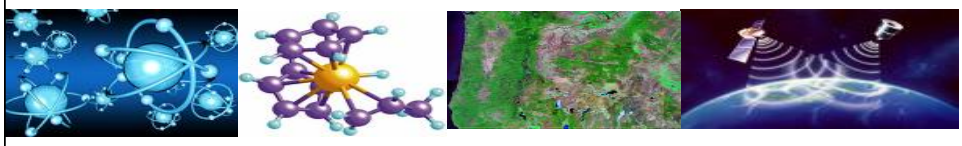
Research Lifecycle, Data Lifecycle, and Data Infrastructure Services

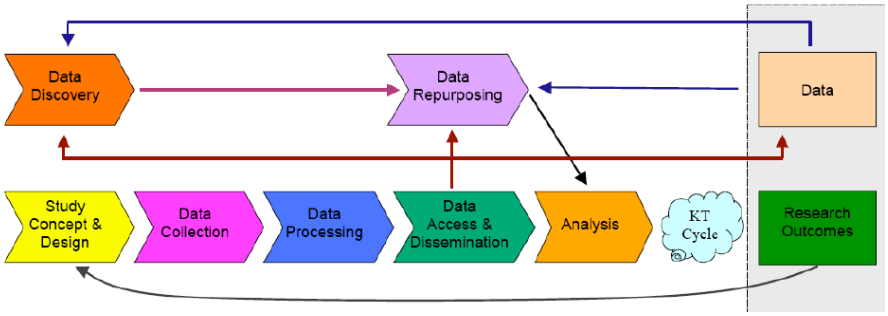
Jian Qin
School of Information Studies
Syracuse University
Syracuse, New York

National Science Library of China, Beijing, June 26, 2013

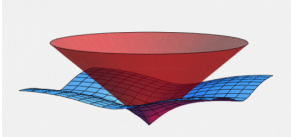


Research Lifecycle and Data Lifecycle

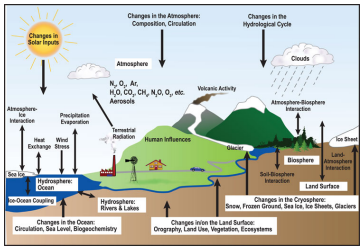
E-Science and the life cycle of research



The "KT Cycle" in the diagram represents the processes of knowledge transfer. This life cycle diagram comes from Charles Humphrey, "E-Science and the Life Cycle of Research" (2006) available online at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc



But what research?



Context of research

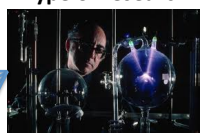
Academic

Corporate

Government

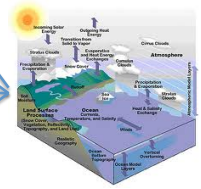
Non-Profit

Type of research

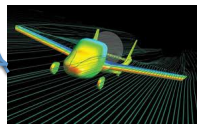


Experiment

Observation



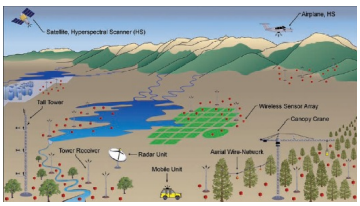
Modeling



Simulation

LCAS, Beijing, June 26, 2013 5

A scenario of data collection



NSF. (2007). *Cyberinfrastructure Vision for 21st Century Discovery.* http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Instruments that collect data

- Sensors
- Microwave towers
- Remote sensing

Level one processing:

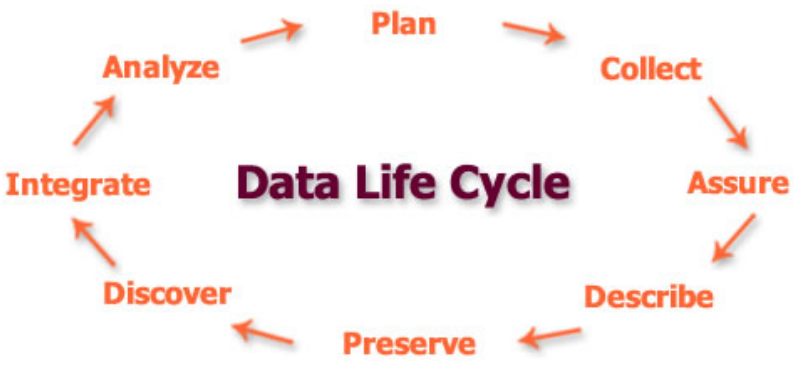
- Formatting
- Calibrating
- Documenting
- Archiving (of raw data)
- Delivering (copies to research team)

Level two processing:

- Organizing data into appropriate data files and segments
- Converting metrics / measurements
- Delivering level two processing copies back to data archive

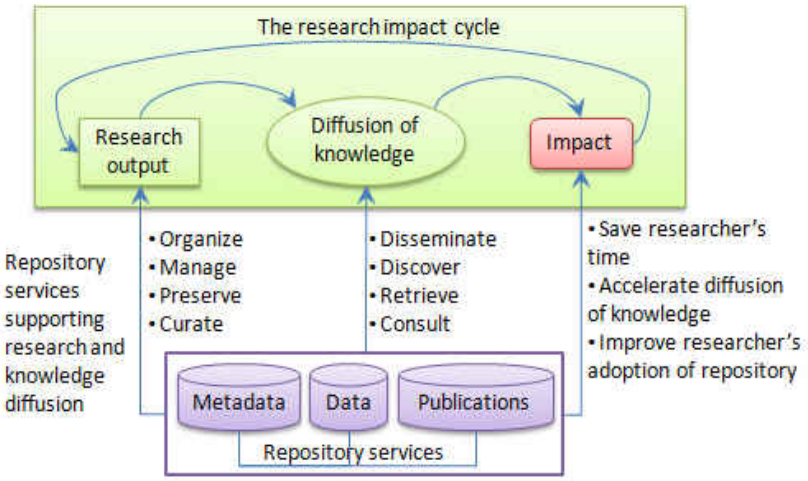
LCAS, Beijing, June 26, 2013 6

Scientific data life cycle



http://www.dataone.org/best-practices

SDM, research impact, and value



Four categories of issues

- Science-based issues
- Data management issues
- Policy issues
- Technical issues

Anderson, W. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal*, Vol. 3 (2004) P 191-201. http://dx.doi.org/10.2481/dsj.3.191

Science-based issues

- What is scientific data?
  - Observations, experiments, simulations, models, etc.
  - Active data, intermediary data, final data
  - Raw data, calibrated data, derived/calculated/ transformed data, aggregated/meshed data
- Data and metadata
- Barriers to preservation and access
  - Differences among disciplinary communities
  - Interdisciplinary work

Data management issues

- Data management is considered not part of research activity
  - How to motivate researchers to engage in data management?
- Increasing expectations on data availability and quality as well as its documentation
- Convergence of data management responsibilities and scientific work practices
- Reward system and support for data management
- Business and organizational models
  - How to sustain data management practices and services?
- Selection and appraisal of data to be archived
- Planning and requirements
- Training and education

Policy issues

- National and global perspectives:
  - Determining and insuring data privacy and confidentiality
  - Questions of cultural ownership and use of data
  - National security
  - Determinations of intellectual property protection, limits, and exceptions
  - General ethical issues, particularly when dealing with personal and public health data
- Open and fair access:
  - Legislation and defining controlling authorities
  - Supporting policies, regulations, and practices for freedom of information and access authorization
  - Financing and cost recovery policies
  - Policy enforcement mechanisms

Technical issues

- Scientific data and databases are different than literature
  - Sizes and volumes
  - Metadata is more complicated
- Diversity of data types, formats, media types, and standards
  - Data migration to keep up with technology advances
  - Specialized search options
  - Minimizing data processing (analysis-ready)

Summary

- Each stage in the science data life cycle involves some issues
  - Science-based
  - Data management
  - Policy
  - Technical
- Goal of science data management
  - Access in short and long term
  - Use for various purposes by various groups of users

Research Data Management as Infrastructure Services

What is an infrastructure?

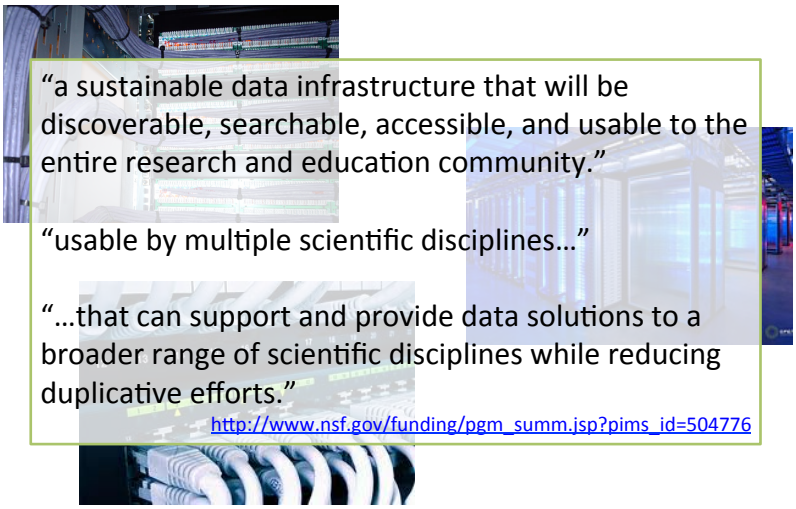
The underlying foundation or basic framework (as of a system or organization).

The system of public works of a country, state, or region.

The resources (as personnel, buildings, or equipment) required for an activity.

http://www.merriam-webster.com/dictionary/infrastructure

What is data infrastructure?



"a sustainable data infrastructure that will be discoverable, searchable, accessible, and usable to the entire research and education community."

"usable by multiple scientific disciplines…"

"…that can support and provide data solutions to a broader range of scientific disciplines while reducing duplicative efforts."

http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776

Nature of an infrastructure

- **Embeddedness**. Infrastructure is sunk into, inside of, other structures, social arrangements, and technologies.
- **Transparency**. Infrastructure does not have to be reinvented each time of assembled for each task, but invisibly supports those tasks.
- **Reach or scope beyond a single event or a local practice**.
- **Learned as part of membership**.
- **Links with conventions of practice**.
- **Embodiment of standards**.
- **Built on an installed base**.
- **Becomes visible upon breakdown.**
- **Is fixed in modular increments**, not all at once or globally.

Star, S.L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information space. Information Systems Research, 7(1): 111-134.

Relevant concepts (1)

E-Infrastructure: the integration of networks, grids, data centers and collaborative environments, and are intended to include supporting operation centers, service registries, credential delegation services, certificate authorities, training and help-desk services.

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

Relevant concepts (2)

Cyberinfrastructure: consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked by high speed networks to make possible scholarly innovation and discoveries not otherwise possible.

Definition source: http://kb.iu.edu/data/auhf.html

Relevant concepts (3)

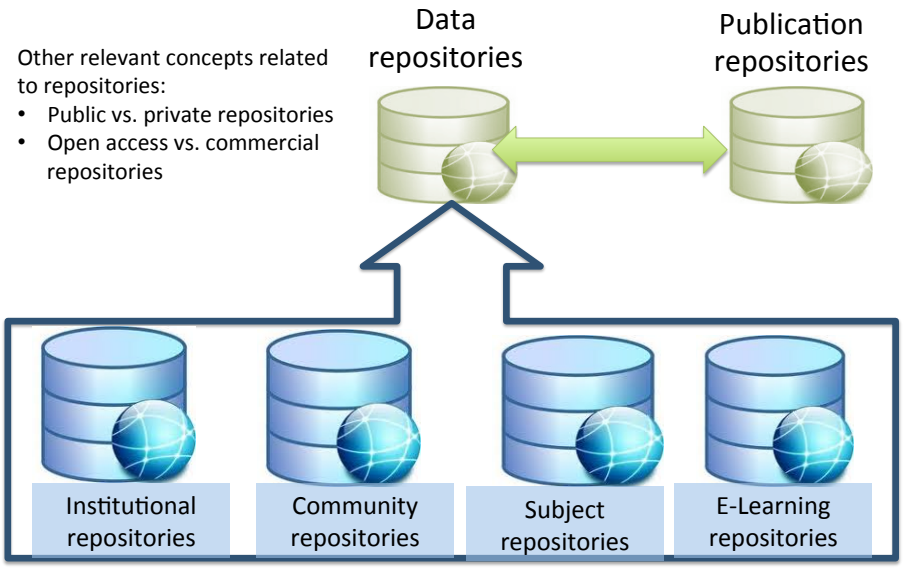
Collections: information items brought together for some specific purpose or with at least one feature in common. They may be
- generated by an institution or project
- gathered for a discipline or by an individual

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

Repositories: the constructs that hold collections and facilitate their use.
- Narrowly, they mean storage equipment and supporting computer programs
- Wider definition: they include the management framework, services, and tools associated with a repository as well as the storage machinery itself.

Relevant concepts (3, cont'd)

Other relevant concepts related to repositories:
- Public vs. private repositories
- Open access vs. commercial repositories

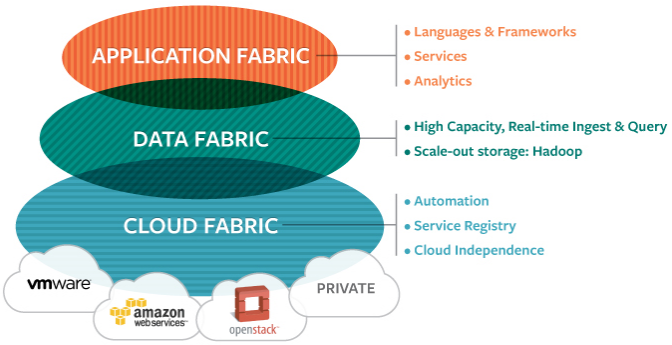


Data infrastructure services

Two cases of commercial data infrastructure services

Amazon Web Services offers a complete set of infrastructure and application services that enable you to run virtually everything in the cloud.

PIVOTAL ONE (previously "Greenplum")



Repositories as an important part of data infrastructure services

- Content is deposited in the repository, whether by content creator, owner, or third party
- The repository architecture manages content as well as metadata
- The repository offers a minimum set of basic services such as put, get, search, and access control
- The repository must be sustainable and trusted, well-supported and well-managed

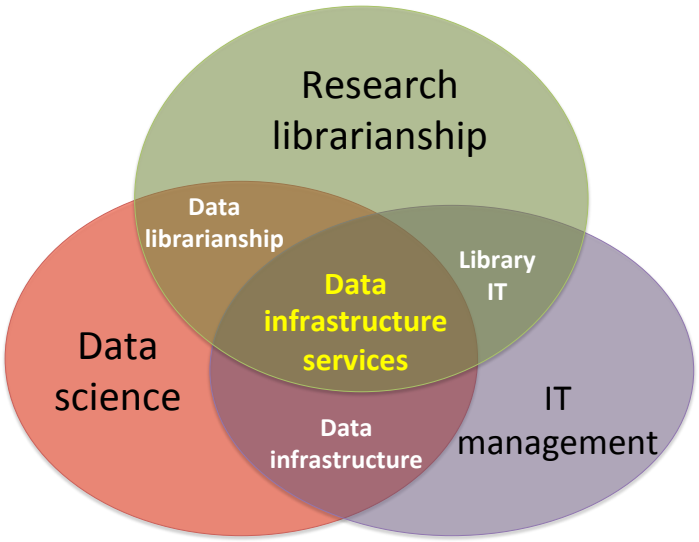
http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

What is data science?

"An emerging area of work concerned with the collection, presentation, analysis, visualization, management, and preservation of large collections of information."

Stanton, J. (2012). Introduction to Data Science.
http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

Data infrastructure services and research libraries



Top three issues in the long-term preservation, management, and curation of scientific data in digital form:

Linkage of data sets:
"If data libraries are to be truly useful for social sciences, they must provide users with the software tools to link these very-large and unwieldy data-sets easily, reproducibly, and reliably."

http://www.arl.org/storage/documents/publications/nsfworkshop2006-brady.pdf

Confidentiality:
"Although the availability and the linkage of social science data provide tremendous opportunities for answering important social science questions, they also exponentially increase the dangers of disclosing personal information through the possibility of identifying individuals…"

Institutional models:
"At the moment, we have nothing comparable to the 'University Library' which has historically made rational acquisitions through 'collection specialists' working with researchers, developed meta-data through classification and indexing…"

The benefit of data is the same in all fields, but the problems of data management are different in these fields.

Data infrastructure: User community requirements

- Long-term data preservation including authenticity and other checks that guarantee the data quality
- Data access and data curation services of various types
- Increasing request for usage of computational capacities
- Data distribution, data federation and data grid solutions not only for data preservation but also for access optimization

These translates into the functionality of future data infrastructure

Functionality required for future data infrastructure

- **Open deposit**, allowing community centers to easily store data
- **Bit-stream preservation**, i.e., ensuring that data authenticity will be guaranteed for specified number of years
- **Format and content migration**, executing CPU intensive transformations on large data sets on the command of the communities
- **PID servi**ce, allowing centers to register a huge amount of persistent identifiers with the required granularity and data extensions and resolve them
- **Metadata support** to allow effective management
- **Maintaining proper access rights** as the basis of all trust
- A variety of **access and curation services** that will vary between disciplines and over time
- **Execution services** that allow a large group of researchers to operate on the stored date
- **Security** to ensure that only partners of the trust federation will have access
- **High availability of service** so that researchers can rely on them
- **Regular quality assessment** to ensure adherence to all agreements
- **Authentication** should be supported by distributed Authentication and Authorization Infrastructure (AAI) mechanisms enabling single identity and single sign-on
- The principles of **Services Oriented Architecture** should be followed
- Achieving **a higher degree of interoperability at format and semantic level** is one of the major goals.

Building data infrastructure services

"It's all about transformation"

- To change in composition or structure (what we are/what we do)
- To change the outward form or appearance (how we are viewed/understood)
- To change in character or condition (how we do it)

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

Predicted involvement of research librarians in the next decade

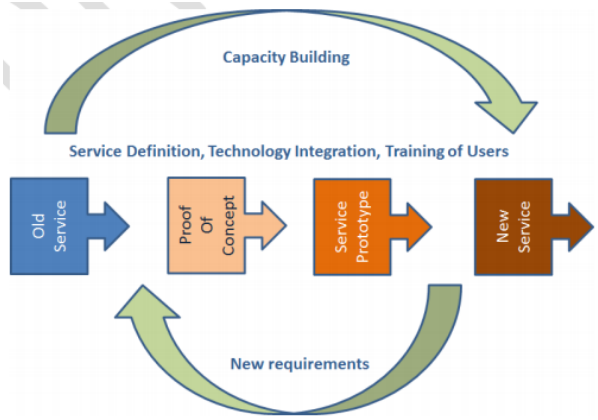
- Mobility of user
- Cloud computing
- Semantic Web
- Open content
- Globalization
- Collective intelligence
- Online education/ MOOC
- Assessment
- Customization/personal web
- Radical collaboration
- E-Research
- Information policy
- Mutability
- …

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

The keyword for data infrastructure services is:

Capacity Building

Capacity building lifecycle



capacity building lifecycle

Source: https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/data_einfra_h2020_fiches_on-line_consult.pdf , p. 9

Types of data services

Characteristics of data services

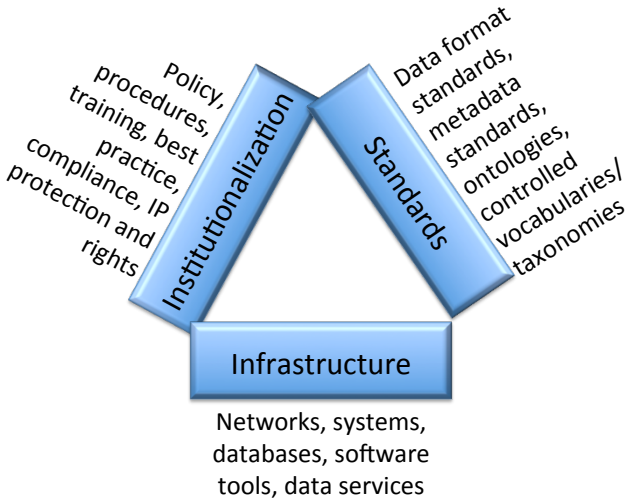
A quick introduction of scientific data management

WHAT ARE DATA SERVICES?

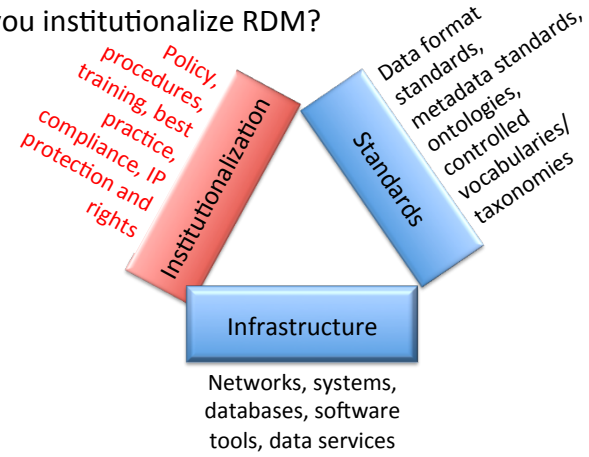
What is research data management (RDM)?

Research data management is essentially a service that an organization develops and implements through institutionalized data policies, technological infrastructures, and information standards.

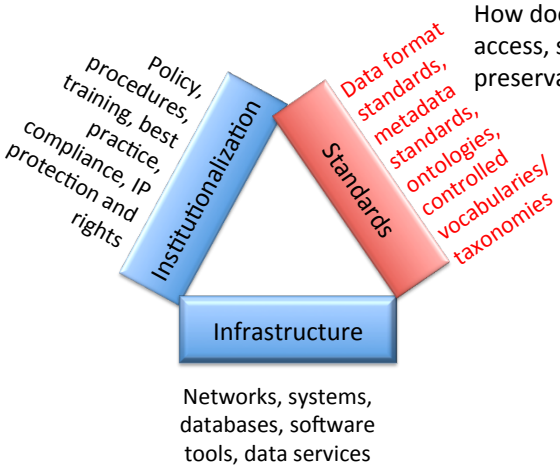
Research Data Management

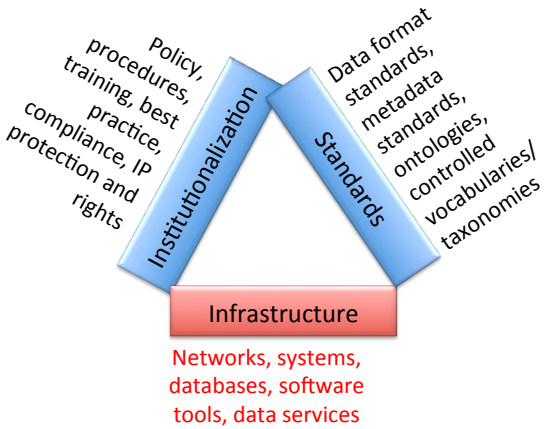


What is institutionalization?
Why do you need institutionalize research data management?
How can you institutionalize RDM?



How much do you know about data and metadata?
How does the nature of data affect metadata?
How does metadata affect data access, sharing, reuse, and long-term preservation?





What is data infrastructure and Data infrastructure services?
Why do you need to build a data infrastructure?
What is the key in building a data infrastructure?

What lies before research libraries?

Challenges

- Extremely rapid growth and sheer volume of data and metadata, in and out of repositories
- Developing and maintaining standards, database schemas, ontologies, etc. to improve interoperability
- Initiatives started to create global information infrastructures at semantic level
- Difficulties relating to data deposit, both technically and behaviorally
- Repositories are under-staffed

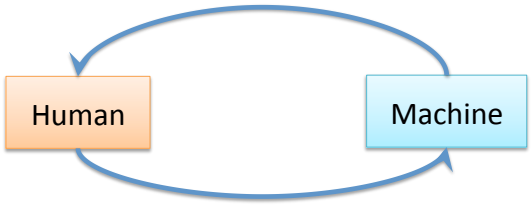
Where are the gaps?

Vision

- A reliable, sustainable data infrastructure and services for data repositories
- A high quality information space
- Information that is readily available
- A well-managed, accountable repository infrastructure

Understanding data services

For whom?



Infrastructure type of services:

- National
- Institutional

What data services? (1)

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

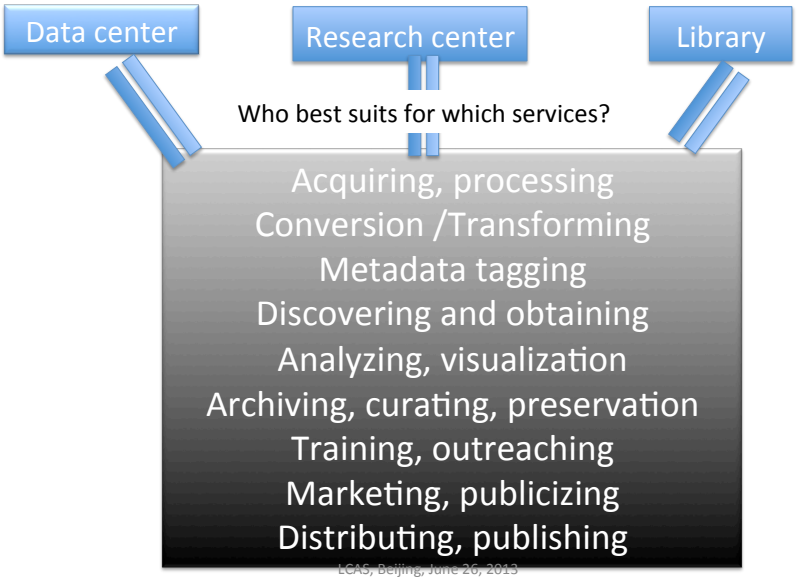
Source: ARL survey report, 2010

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



Characteristics of data services

- Repeatable
- Sustainable financially and technically
- A community of practice
- Institutionalization
- Collaboration and coordination
- Conformance to regulations and laws

Data infrastructure services that research libraries can provide

- Data management consultation
- Assistance and support for data processing and data product development
- Build and promote data repositories as an infrastructure service
  - Linkage between research project data and community repositories
  - Discovery, access, and curation
  - Metadata, semantic Web, and ontologies
- Training in data management: librarians and researchers

Are you ready for

打铁先要自身硬

How can you make a contribution? (1)

- Capabilities of the 21st century academic library information professional
  - Deep Subject, Process, or Technical Expertise
  - Deep Service Commitment
  - Commitment to Research and Development
  - Commitment to Assessment and Evaluation
  - Communication and Marketing Skills
  - Project Development and Management Skills
  - Political Engagement
  - Resource Development Skills
  - Commitment to Rigor
  - Entrepreneurial Spirit
  - Commitment to Collaboration
  - Leadership/Inspirational Capacity

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

How can you make a contribution? (2)

"Academic library staff must integrate and mainstream digital library services, digital archiving and preservation, repository development, digital publishing, and instructional technologies into the core of library budgeting, staffing and organization."

"Academic library staff must be embedded in the e-research cyberinfrastructure and scholarly communication processes, and be integral to the systems of research information management."

- Support the needs of data-driven research
  - Federal/funding agency
  - Massive data sets
  - Unstructured data/ curation
  - Extraction
  - Distribution
  - Collaboration
  - Visualization
  - Simulation
  - Preservation

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

Concluding remarks

- Data management and services are a new territory and require new thinking of research library's roles
- There are ample opportunities for research librarians to make a contribution to capacity building
- Learn the lessons from earlier initiatives
- Transforming the library image needs to transform the library tradition first

Examples of (open) data infrastructure services

- The Institute for Quantitative Social Science repository: http://www.iq.harvard.edu/
- Inter-University Consortium for Political and Social Research (ICPSR): http://www.icpsr.umich.edu/icpsrweb/landing.jsp
- The Dryad Digital Repository: http://datadryad.org/
- Data Observation Network for Earth: http://www.dataone.org/
- Datalib: http://databib.org/ (a registry/directory/catalog of research data repositories)
- Registry of Research Data Repositories: http://www.re3data.org/