Data Management for Collaboration, Access and Interoperability - CLIR Workshop on Research Data Principles & Services

Karl Benedict & Plato Smith II - University of New Mexico

July 28, 2015

Contents

Introduction **1**

Research and Data Lifecycle Models 2

Relationship Between the Researcher and Data Lifecycles Models - Part 2 (cira 2013 to Present) . 5

Data Management Considerations 8

Data Interoperability and Linked Open Data 12

In the Shoes of the Researcher - You ... **16**

Introduction

Outline

Data Management Planning - Foundation Principles

- Context - Data Management Requirements
- Relationship Between the Researcher and Data Lifecycles Models - Part 1
- Relationship Between the Researcher and Data Lifecycles Models - Part 2
- Data Management Considerations
- Data Interoperabilty and Linked Open Data

The Researcher's Perspective and Library Research Data Services

Context - Data Management Requirements

- Data Management Plans
- Data Sharing Requirements
- Institutional Review Board (IRB) Protocols
- Interdisciplinary Collaborative research
- Data Intensive Research

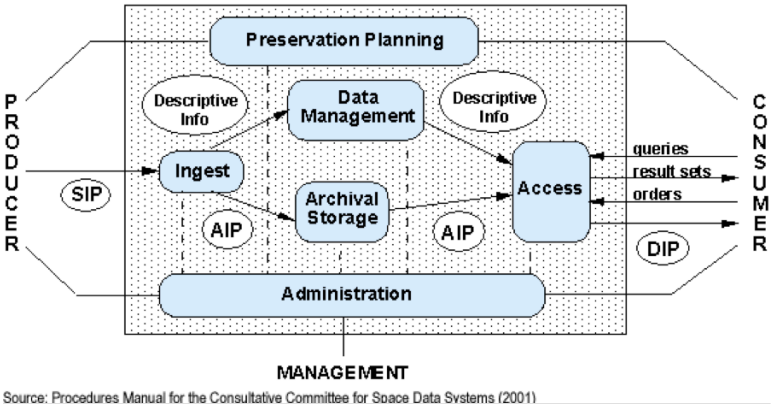
Notes

- Data Management Plan Requirements from funding agencies (e.g. DOE, NASA, NSF, NEH, USGS)
- Data Sharing requirements from funding agencies and publishers (e.g. OTSP, The Royal Society, PLOS|One)
- Institutional Review Board (IRB) Protocol requirements for explicitly defining how collected data will be managed, de-identified, shared, and/or destroyed along with expected risks
- Interdisciplinary collaborations, research, and networks require efficient sharing of data within and across research teams and domains
- Data intensive research magnifies the need for effective data management

Research and Data Lifecycle Models

Relationship Between the Researcher and Data Lifecycles Models - Part 1 (circa 2001 to 2012)

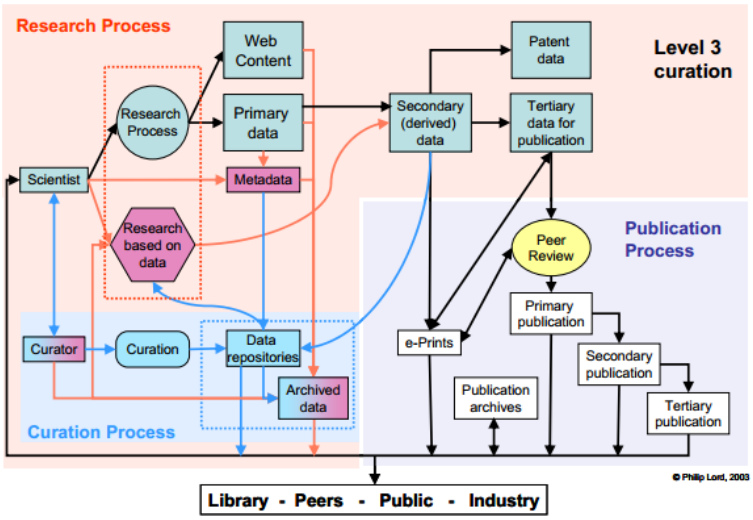
The OAIS Reference Model was developed by the Consultative Committee for Space Data Systems (CCSDS), 2001 - OAIS Functional Entities



Source: Procedures Manual for the Consultative Committee for Space Data Systems (2001)

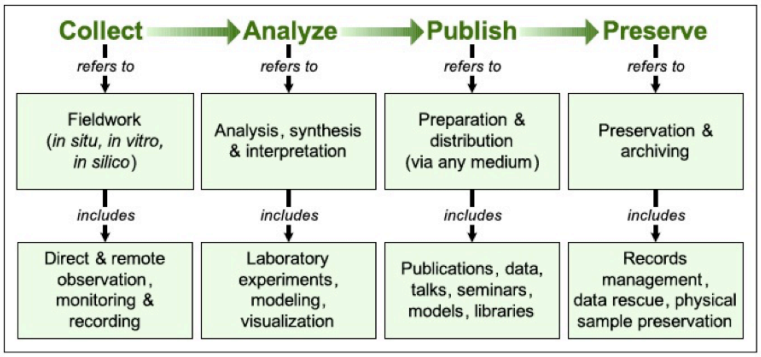
OAIS Reference Model - 2001 **"The Open Archival Information System (OAIS) Reference Model** was developed by the Consultative Committee for Space Data Systems (CCSDS) as a work item under the ISO Technical Committee 20, Sub-Committee 13. It is a framework for understanding and applying concepts needed for long-term digital information preservation (where long-term is long enough to be concerned about changing technologies). It is also a starting point for a model addressing non-digital information" (CCSDS Blue Book)/ISO 14721:2002). The OAIS Functional Entities conceptual framework describes the environment, functional components, and information objects within a long-term preservation system and is widely recognized in scientific, data management, and archival communities" (COES Data Life Cycle Models and Concepts v.8, 2011, p. 12).

Level 3 curation - information flow with data archiving (Developed by Philip Lord and Alison Macdonald, 2003) - 2003 e-SCicence Curation Report

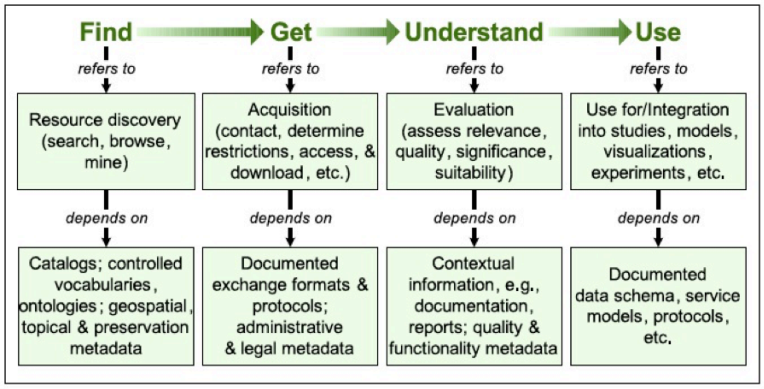


Level 3 Curation Diagram - 2003 **The Level 3 curation diagram** is the information flow with data archiving model that is comprised of (1) traditional academic flow of information **(Level 1 curation)** and (2) information flow with data curation **(Level 2 curation)**. Developed in 2003 by Philip Lord as part of the 2003 e-Science Curation Report Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, this data lifecycle has influenced the development of continually evolving data lifecycle models and concepts.

Producer Perspective (Developed by Tom Gunther and Dave Govoni, 2006 - USGS) - Contributes to future USGS Data Lifecycle Diagram



Consumer Perspective (Developed by Tom Gunther and Dave Govoni, 2006 - USGS) - Contributes to future USGS Data Lifecycle Diagram



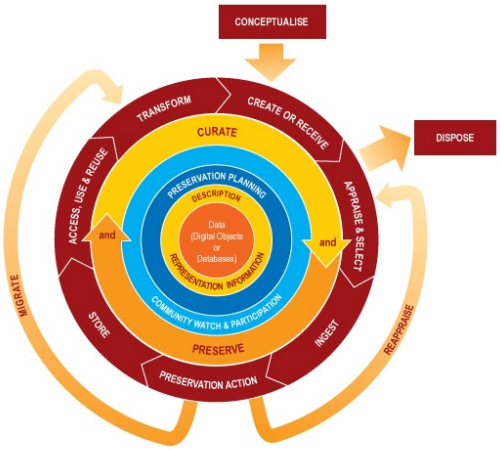
USGS Data Lifecycle Perspective Models - 2006 The USGS Producer perspective and Consumer perspective scientific information and knowledge management diagrams were developed by Tom Gunther and Dave Govoni of the US Geological Survey as part of an investigative report titled "Scientific Information Management at the U.S. Geological Survey: Issues, Challenges, and a Collaborative Approach to Identifying and Applying Solutions (Abstract) in Geoinformatics in 2006. These two perspectives are the foundation of the USGS Data Lifecycle Diagram developed in 2012 and encapsulate some of the major processes and functions of the OAIS Functional Entities and the Level 3 curation diagrams.

The **producer** of the data is concerned (implicitly or explicitly) with the processes involved in research such as (1) fieldwork, (2) analysis, (3) preparation, and (4) preservation which correlate to Level 1 curation, Level 2, and Level 3 curation.

The **consumer** of data is concerned with (1) resource discovery, (2) acquisition, (3) evaluation, and (4) integration of data that correlate to (1) dissemination information package (DIP) of the OAIS functional entities and access, use, and reuse to Level 1 curation processes in the data lifecycle model.

Govoni, D.L. and T.M. Gunther, 2006. Scientific Information Management at the U.S. Geological Survey: Issues, Challenges, and a Collaborative Approach to Identifying and Applying Solutions (Abstract). *Geoinformatics 2006—Abstracts. Scientific Investigations Report 2006-5201*, p. 19-20. U.S. Geological Survey, Reston, Virginia

The Digital Curation Centre's (DCC) Data Curation Lifecycle (2007) - 2007/2015 DCC Curation Lifecycle Model

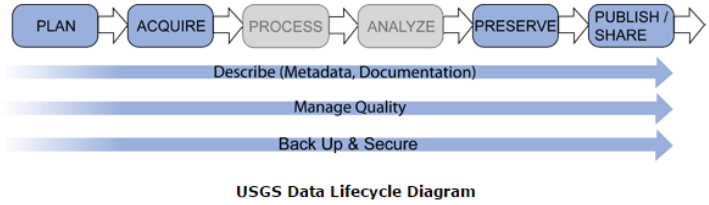


DCC Data Curation Lifecycle

http://www.dcc.ac.uk/resources/curation-lifecycle-model

DCC Curation Lifecycle Model - 2007 The **DCC Curation Lifecycle Model** describes major processes of the curation and preservation processes of data throughout its usefulness to research, teaching, and learning. The DCC Curation Lifecycle Model was introduced to the research and learning communities at the 3rd International Digital Curation Conference (IDCC) in December 2007 in Washington, DC.

USGS Data Lifecycle Diagram, 2012 - Contributes to future USGS CDI Science Framework

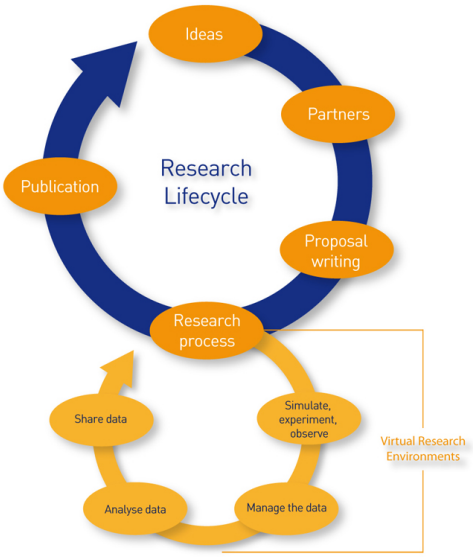


USGS Data Lifecycle Diagram

USGS Data Lifecycle Diagram - 2012 The **USGS Lifecycle Diagram** builds on the previous work of Tom Gunther and Dave Govoni (USGS) and describes key processes in the research lifecycle from research plan (must include data management planning) to data acquisition (data capture) to description (metadata, documentation) to analysis to preservation to publication. This model includes aspects of the the Level 3 curation model and the DCC Curation Lifecycle Model in a condensed version (infographic).

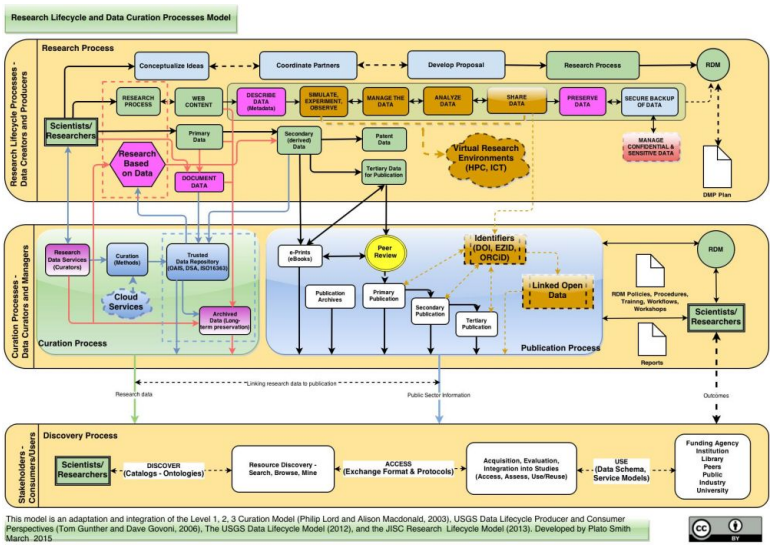
Relationship Between the Researcher and Data Lifecycles Models - Part 2 (cira 2013 to Present)

2013 JISC Research Lifecycle Diagram - link



JISC Research Lifecycle - 2013 JISC developed their Research Lifecycle diagram as a means to map a standard sequence of steps in the research process into the suite of services and capabilities that they provide to researchers. This diagram provides a simple and understandable representation of the research process in terms that are familiar and understood by researchers, a critical step in linking the processes that researchers already follow into the services and associated data curation activities that are central to the effective management, documentation, preservation, discovery and access to research data.

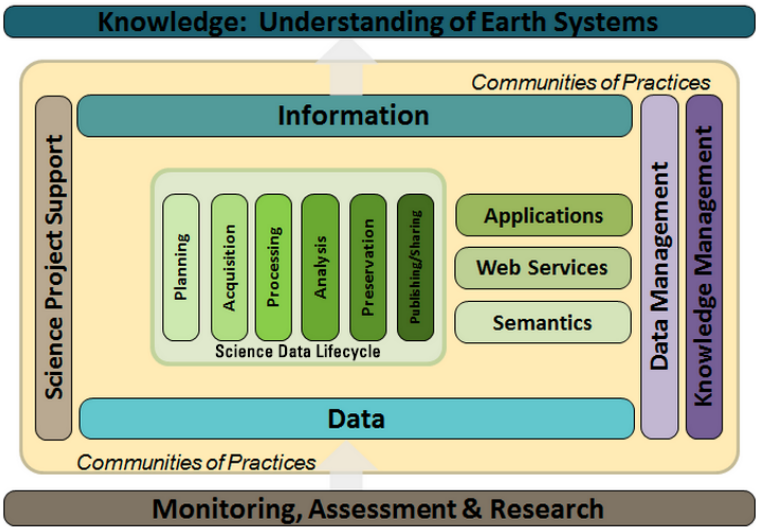
Integrated Research Lifecycle and Data Curation Processes Model (2015)



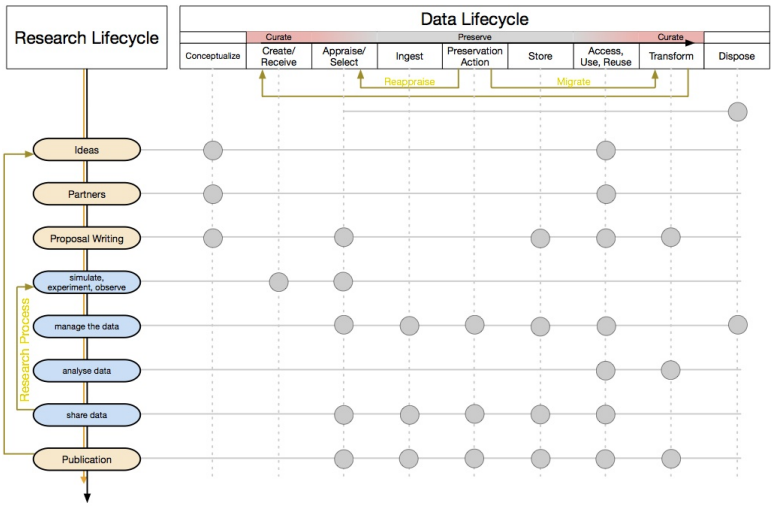
Research Lifecycle and Data Curation Processes Model - 2015 The Research Lifecycle and Data Curation Processes Model integrates Level 3 curation (Lord & Macdonald, 2003), USGS Data Lifecycle Producer and Consumer Perspectives (Gunther & Govoni, 2006), USGS Data Lifecycle Model (2012), and the JISC Research Lifecycle Model (2013) for a comprehensive view of major research data management

processes. This general integrated data management model describes a part of the complex architecture involved in data management and prepares for development of domain-specific data management as developed by the USGS and illustrated by the Community for Data Integration Science Framework Model (the next and last data management model in this series).

USGS Community for Data Integration (CDI) Science Support Framework (2015) - link



USGS CDI Science Support Framework - 2015 "The U.S. Geological Survey (USGS) Community for Data Integration (CDI) represents a dynamic aggregation of multiple communities of practice, focused on the advancement of scientific data and information management and integration capabilities across the USGS and external organizations to enhance earth science research" (USGS, 2015).



Mapping Between Models This diagram provides a preliminary mapping of the JISC research lifecycle steps into corresponding elements in the DCC Data Curation Lifecycle. Each of the circles in the central area of the diagram represents a point in the research lifecycle where there are likely data curation activities and potential service opportunities. This conceptual model is helping the RDS team at UNM identify the elements in our service catalog that are potentially relevant in our work with researchers throughout their research process.

Data Management Considerations

Some Definitions

- Data
- Data Curation
- Documentation (Metadata)
- Open Access
  - Consent to Share
- Embargo
  - Consent to Restrict
- License
- Data Repositories
- Long-term preservation
  - Standards

Notes **Data** - Within the scope of this presentation/document, data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret data. This includes digitized and "borne digital" (no analog surrogates) data.

Data Curation - The integration of descriptive and representative information (metadata) for data to facilitate the efficient management, effective preservation, and usefulness of data over its lifecycle.

Documentation (Metadata) - Metadata is the description and representation information about data, datasets, and/or databases (analog and/or digital). Metadata provide administrative and technical content, context, structure, interrelationships, and provenance information about data.

Open Access (OA) - Open access is freely-available access to data with limited to no copyright restrictions.

- **Consent to Share** - If data collected from human subjects if to be published, then Informed Consent, Institutional Review board (IRB), University Policies, and any other relevant policies must be invoked for compliance (e.g. FERPA, HIPPA, etc.). Research data sharing involving human subjects must protect the confidentiality and rights of participants while upholding ethical behavior in all facets of research from data inception to publication to data destruction.

Embargo - A period during which access to research data is not allowed to certain types of users. This is either to protect the revenue of the publisher or (more generally) to protect the interests of other parties (for example, partner research organizations). [Source: University of Bristol: Data Management Glossary]

- **Consent to Restrict** - Copyright/intellectual property rights owners access restrictions invoked by (1) Embargoes, (2) Internet Protocol (IP) restrictions, or (3) No access (complete restriction) must be respected at all times and any protocols to circumvent access restrictions should be prohibited.

License/Copyrights - Within the scope of this presentation/document, a license is a legal instrument for a rights holder to permit how and to what extent a second party may use copyrighted material. It is imperative that the intellectual property rights (IPR) pertaining to the data are well-established and articulated before any licensing takes place (e.g. Creative Commons, Science Commons, SHERPA/RoMEO, SPARC)

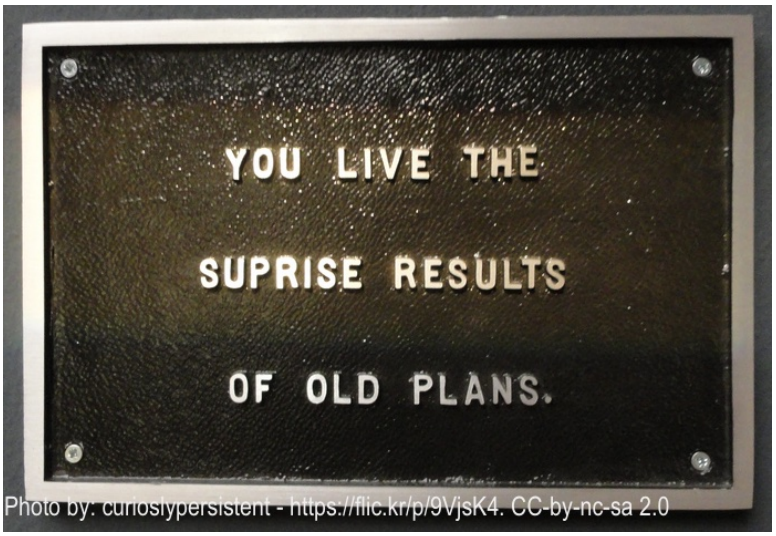
Data Repositories - technology and platform infrastructure used in the aggregation, dissemination, and preservation of data. Some data repositories included **(1) Dryad** (multiple disciplines), **(2) arXiv** (STEM disciplines), **(3) Figshare** (multiple disciplines), **(4) Morhbank** (Biological Sciences), and **(5) XSEDE** (Engineering) to name a few.

- **Long-term preservation** - the long-term management, storage, and archival preservation of data for current and future use that includes but not limited to (1) authentication, (2) integrity, and (3) security checks of data throughout its lifecycle.
  - **Standards** - generally a set of best practices and guidelines governing processes and/or activities involved in data management and curation
    * Data Seal of Approval - 16 assessement guidelines for developing a trusted repository
    * Open Archival Information System (OAIS) - CCSDS 650.0-M-2 Recommended Practice (Magenta) for open archival informaiton system
    * ISO 16363/TDR - A standard for trusted repositories developed from the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) and CCSDS 652.1-R-1 (Red) Draft Recommended Practice

Some Recommendations

- What do you need to know?
  - Four Kinds of Expertise
    * Domain (Subject)
    * Analytical
    * Data Management
    * Project Management
  - Professional Development and Training

- Data Assessment
  - Organization
  - Structure/Content
  - Formats
  - Documentation



What do you need to know?

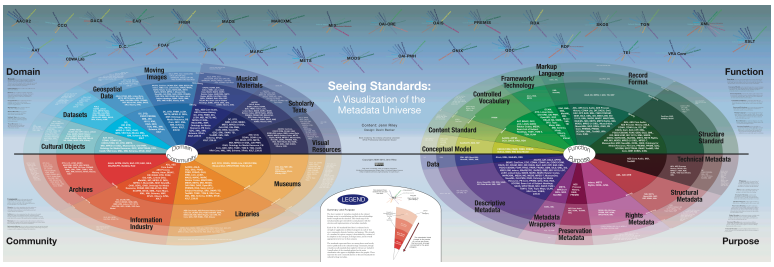
- Four Kinds of Expertise [One Culture]
  - Domain (Subject) - an understanding of the concepts, methods, models, and practices within a discipline and/or chosen profession
  - Analytically - the capability to explore, identify, and leverage data, information, and knowledge pragmatically and technically
  - Data Management - knowledge of domain-specific best practices, guidelines, and standards for data life-cycle management
  - Project Management - the ability to initiate, develop, and lead projects, teams, and workshops from start to finish

Data Assessment Questions

- Who is responsible for data management? (e.g. research lab, researchers, sponsors, postdocs)
- Who owns the data?
- What data will be collected?
- Where will the data be collected?
- Who will collect the data?
- When will the data be collected?
- How will the data be collected?
- What is the format of the data?
- How will access be provided to the data?
- What privacy and/or security issues exist for the data?

Organization

- **Define folder, file names, and structure** (e.g electronic notebooks)
  - Use meaningful names that include basic information (e.g. date, measurement, collection, PI, etc.)
  - Unique
  - Avoid spaces
  - ASCII Characters only
  - Security of Files & Backups
- **Structure/Content**
  - Consistent content
  - Separate data from analysis
  - Focus on tabular structure for tabular data
  - Explicitly encode missing data and document that encoding
  - Use meaningful column headings - while keeping short without spaces
  - Include units
  - Data dictionary
- **Formats**
  - Plan for data & metadata integration into an archive (e.g. Metadata Interoperability and Standardization - A Study of Methodology Part I, D-Lib June 2006, v.12(6))
  - Open Standards
  - Proprietary ASCII
  - Proprietary Binary - Documented
  - Proprietary Binary
- **Documentation**
  - Many documentation standards (e.g. See Standards: A Visualization of the Metadata Universe, Riley (2009/2010))
  - Machine and human readable
  - Commonly based on Extensible Markup Language (XML)
  - Wide variety of strategies, methods, and tools for creating documentation
  - Enables Discovery, Use, and Understanding
  - Work with experts in documentation for your discipline to identify relevant standards for your data



Jenn Riley (2009-2010). Seeing Standards. A Visualization of the Metadata Universe. http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards.pdf

Overall Recommendations

- Procure assistance - consult early and often (e.g. collaborate, network)
- Maintain documentation from the project planning stage and throughout your work
- Adopt a systematic model for organizing your data: naming, file structure, formats, storage, backups
- Adopt consistent and documented data structures
- Always have the entire data and research life-cycle models in mind when you are managing your data

Data Interoperability and Linked Open Data

> The Semantic Web isn't inherently complex. The Semantic Web language, at its heart, is very, very simple. It's just about the **relationships between things**.

Tim Berners-Lee (2007). "Q&A with Tim Berners-Lee". Bloomberg Business, April 9, 2007. http://www.bloomberg.com/bw/stories/2007-04-09/q-and-a-with-tim-berners-leebusinessweek-business-news- stock-market-and-financial-advice

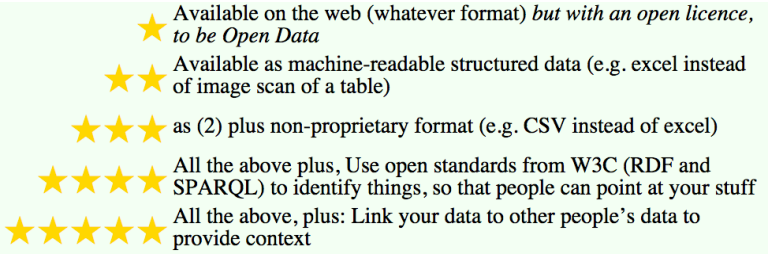
> … the most important thing that was new was the idea of URI – or URL [it was UDI back then, universal document identifier]. The idea that any piece of information anywhere should **have an identifier**, which will not only identify it, but allow you to **get hold of it**. That idea was the basic clue to the universality of the Web. That was the only thing I insisted upon.

Tim Berners-Lee (1999). "Interview with the Web's Creator" by Chris Oakes. Wired, October 23, 1999. http://archive.wired.com/science/discoveries/news/1999/10/31830?currentPage=all

Definitions

- Interoperability
- Linked Open Data Models

- Internet Standards
  - Web Services (REST, SOAP)
- Domain Specific Standards & Protocols
  - Open Geospatial Consortium (OGC) Web Map, Web Feature and Web Coverage Services (WMS, WFS, WCS)
  - DataONE, CUAHSI



Tim Berners-Lee (2006). Linked Data. http://www.w3.org/DesignIssues/LinkedData.html

Notes

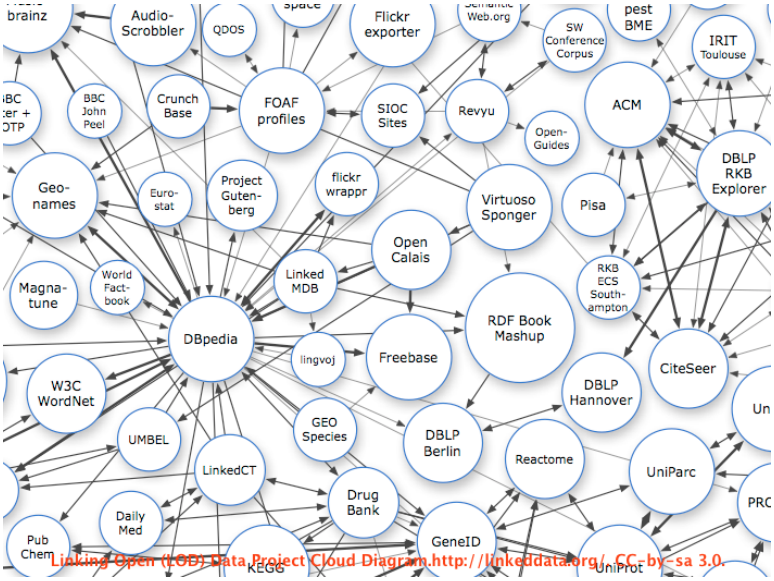
Interoperability "interoperability is the ability of different information technology systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged." - National Alliance for Health Information Technology. (2005) "What Is Interoperability?" 2005. Available online at www.nahit.org

"Geospatial Interoperability is the ability for two different software systems to interact with geospatial information. Interoperability between heterogeneous computer systems is essential to providing geospatial data, maps, cartographic and decision support services, and analytical functions." - National Aeronautics and Space Administration, Geospatial Interoperability Office (2005) Geospatial Interoperability Return on Investment Study Report. p. iii. http://lasp.colorado.edu/media/projects/egy/files/ROI_Study.pdf

Linked Open Data "The Semantic Web is a Web of Data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. The collection of Semantic Web technologies (RDF, OWL, SKOS, SPARQL, etc.) provides an environment where application can query that data, draw inferences using vocabularies, etc."

"To achieve and create Linked Data, technologies should be available for a common format (RDF), to make either conversion or on-the-fly access to existing databases (relational, XML, HTML, etc). It is also important to be able to setup query endpoints to access that data more conveniently. W3C provides a palette of technologies (RDF, GRDDL, POWDER, RDFa, the upcoming R2RML, RIF, SPARQL) to get access to the data."

World Wide Web Consortium - Linked Data. *http://www.w3.org/standards/semanticweb/data*



A snapshot of a subset of the Linking Open Data Cloud Diagram

Linked Open Data Rules (Tim Berners-Lee (2006). Linked Data. http://www.w3.org/DesignIssues/LinkedData.html)

Use URIs as names for things

Use HTTP URIs so that people can look up those names.

When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)

Include links to other URIs. so that they can discover more things.

Internet Standards and Web Service Protocols *Hypertext Transfer Protocol* - HTTP is a core protocol that is used for machine to machine communication on the Internet. It defines a number of request types and the corresponding responses to those requests. HTTP is an open standard that managed by the Internet Engineering Task Force (IETF) through a set of Request for Comment documents (7230, 7231, 7232, 7233, 7234, 7235, 7236, 7237). HTTP provides the common foundation upon which broadly used web service protocols and Application Programming Interfaces (API) are based.

Simple Object Access Protocol - SOAP is a W3C Recommendation that defines the messaging model between computer systems for exchanging structured data over a network. It is based on an Extensible Markup Language (XML) and is commonly exchanged over HTTP, but not required to be HTTP supported.

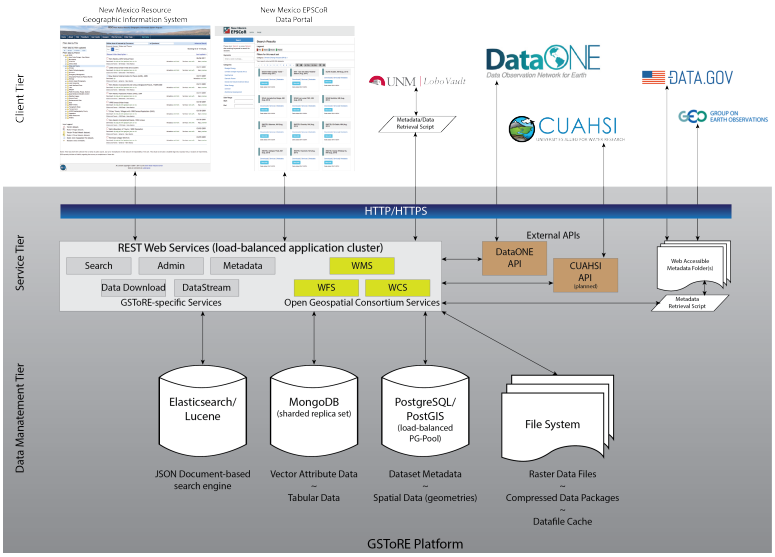
Representational State Transfer - REST web services are built upon a Resource Oriented web service architectural model in which service endpoints represent resources and the various actions that can be taken relative to those resources are defined through the standard "verbs" defined in the HTTP protocol - GET (list or get a resource), PUT (replace), POST (create), DELETE (delete).

Domain Specific Standards and Protocols *Open Geospatial Consortium* (OGC) - The geospatial and location standards of the OGC define data format, representation, visualization, and web service standards for geospatial data. While there are numerous OGC standards, three are relevant in the context of this presentation: Web Map Services (WMS) for data visualization, Web Feature Services (WFS) for access to feature data (i.e. geometries and their associated attributes), and Web Coverage Services (WCS) for access to coverage data (i.e. gridded data).

Data Observation Network for Earth (DataONE) - DataONE has defined an API that defines how *member nodes* within their network share information (in the form of specifically structured RDF documents) about data and metadata holdings with *coordinating nodes* that retrieve published metadata and provide network-wide data discovery and brokered access to data.

Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) - CUAHSI has developed the Hydrologic Information System (HIS) as a distributed system that enables the publication and access to water resource data. The HIS implements a set of standards that define a database schema model specifically designed to represent point-time-series data values, an XML schema (WaterML) that provides a data and metadata transfer specification, and a set of web services (WaterOneFlow) that define the methods for computer systems to exchange water data.

An Illustration



Data Management Resources/Tools

1. Open Science Framework (OSF) - OSF is part network of research materials, part version control system, and collaborative software - https://osf.io/4znzp/
2. Digital Research Tools (DIRT) - Registry of digital research tools for scholarly use - http://dirtdirectory.org/
3. OPENRefine - A free, open source, powerful tool for working with messy data - http://openrefine.org/
4. Tabula - A tool for liberating data tables locked inside PDF files - http://tabula.technology/
5. import io - Instantly turn web pages into data - https://import.io/
6. Australian National Data Service (ANDS) - http://www.ands.org.au/
7. DCC Tools and applications - http://www.dcc.ac.uk/resources/tools-and-applications
8. DCC Tools & Services - http://www.dcc.ac.uk/resources/external/tools-services
9. Digital Curation Centre: Disciplinary Metadata - http://www.dcc.ac.uk/resources/metadata-standards
10. Library of Congress Sustainability of Digital Formats - http://www.digitalpreservation.gov/formats/
11. PLOSONE Data Sharing Requirements - http://www.plosone.org/static/policies
12. UC3 University of California Curation Center - http://www.cdlib.org/uc3/

Professional Development and Training

- Your Campus Institutional Review Board (IRB)
- Your Campus Office of the Vice President for Research (OVPR)
- Your Campus High-Performance Computing (HPC) Center
- CITI - Collaborative Institutional Training Initiative at the University of Miami (e.g. Human Subjects Research - Responsible Conduct for Research (RCR)

- ICPSR - Data Management and Curation
- MANTRA - Research Data Management Training
- DigCurV - A Curriculum Framework for Digital Curation

In the Shoes of the Researcher - You ...

Your Data

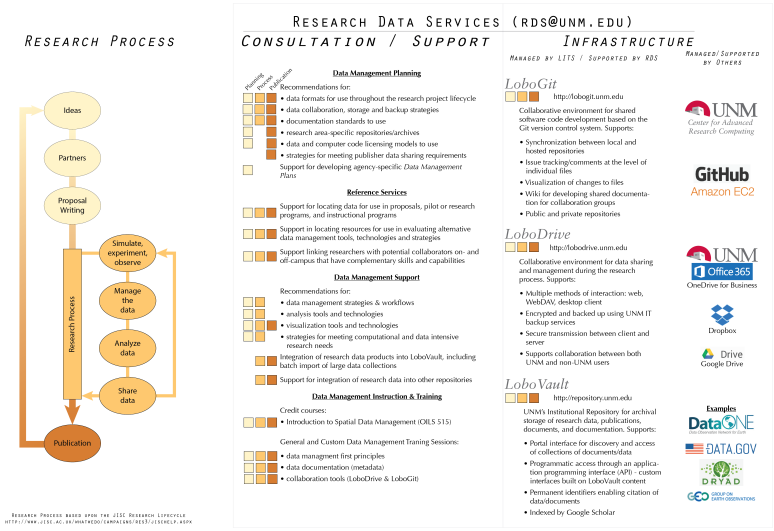
How does your personal experience with data management match these goals?

What have we learned from our data management experiences that can inform how we communicate with and support the researchers we work with?

Activity Outline

1. A Data Interview (in pairs)
    i. Types of Data
    ii. Documentation Needed for Discovery, Understanding & Use
    iii. Formats - both for analysis/visualization & preservation
    iv. Privacy and confidentiality issues/concerns
    v. Sharing & Licensing
    vi. Preservation Needs
2. Grouping Exercise
3. Discussion and Identification of Disciplinary Gaps
4. Map into Library Research Data Services

UNM's RDS Service Catalog - a Point of Reference



Acknowledgments

- CLIR Postdoctoral Fellowship Program in Data Curation at the University of New Mexico
- NSF EPSCoR Program (Track 1 [Awards: 0447691, 0814449, 1301346] and Track 2 awards [0918635, 1329470])
- New Mexico Resource Geographic Information System
- NASA ACCESS Program
- UNM's College of University Libraries and Learning Sciences