Apr 7th, 9:30 AM

Data Conservancy: A Life Sciences Perspective

G. Sayeed Choudhury  
*Johns Hopkins University, sayeed@jhu.edu*

Follow this and additional works at: [http://escholarship.umassmed.edu/escience_symposium](http://escholarship.umassmed.edu/escience_symposium)  
Part of the [Library and Information Science Commons](http://escholarship.umassmed.edu/escience_symposium) and the [Life Sciences Commons](http://escholarship.umassmed.edu/escience_symposium)

[http://escholarship.umassmed.edu/escience_symposium/2010/program/2](http://escholarship.umassmed.edu/escience_symposium/2010/program/2)

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in University of Massachusetts and New England Area Librarian e-Science Symposium by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Data Conservancy: A Life Sciences Perspective

Sayeed Choudhury
Johns Hopkins University
sayeed@jhu.edu
Data Conservancy

• One of two current awards through the National Science Foundation DataNet program
• Other award is DataONE led by William Michener at University of New Mexico
• Each award is $20 million, 5 year award with multiple partners
Data Curation

The Data Conservancy embraces a shared vision: data curation is a means to collect, organize, validate and preserve data so that scientists can find new ways to address the grand research challenges that face society.
...not a rigid road map but principles of navigation. There is no one way to design cyberinfrastructure, but there are tools we can teach the designers to help them appreciate the true size of the solution space – which is often much larger than they may think, if they are tied into technical fixes for all problems.
Objectives

• Infrastructure research and development
  – Technical requirements
• Information science and computer science research
  – Scientific or user requirements
• Broader impacts
  – Educational requirements
• Sustainability
  – Business requirements
What are Life Sciences?
Long Tail of Biology

Small number of providers with lots of data. High-throughput biology, monitoring, simulation.

Large number of providers with small amounts of data. Observational, experimental
21st Century Biology

- Molecular biology drivers and promise of informatics
- Fundamental unity of biology
- Data generated within one domain can also serve another
- System that captures most possible value
- Add higher level thinking within the discipline
Data Driven Discovery

• Discoveries made from aggregating data and querying in new ways
• Need data management tools
• Need data analysis tools
• Need data visualization tools
The Problem

• How do we make data sharing part of the normal work flow across the Life Sciences?
  – Social
  – Technical

• Address barriers

• Accommodate needs

• Do this for all Life Science sub-disciplines!
For each Life Science Sub-discipline

• Data culture
• Data policy
• Metadata standards
• Ontologies

• How to address each sub-discipline in four years?
Data Flow (Levels of Data)

- Pixel data collected by telescope
- Sent to Fermilab for processing
- Beowulf Cluster produces catalog
- Loaded in a SQL database
Domain coverage/methods

- Multi-site user research methods are a blend of:
  - Case study & domain comparisons
  - Depth & breadth
  - Local & global

<table>
<thead>
<tr>
<th></th>
<th>Astronomy</th>
<th>Earth Sciences</th>
<th>Life Sciences</th>
<th>Social Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCAR</td>
<td>Task-based design and usability testing ⇒ Use cases, data requirements, system recommendations</td>
<td></td>
<td></td>
<td>UCAR</td>
</tr>
<tr>
<td>UCLA</td>
<td>Ethnography, virtual ethnography, oral histories ⇒ Use cases, data requirements</td>
<td>Interviews, Surveys, Worksheets, Content analysis ⇒ Curation requirements, taxonomy, metadata/provenance framework</td>
<td></td>
<td>UIUC</td>
</tr>
</tbody>
</table>
Data Framework

• Start with a common conceptualization that applies across scientific domains
• Exploit semantic technologies
• Leverage existing work
• Prototype the framework in target communities
  – Iteratively refine, learn from experience
  – Demonstrate success, measured in terms of new science
Common Conceptualization

Observations are the foundation of all scientific studies, and are the closest approximation to facts.

Emergence

• Emergence: The Connected Lives of Ants, Brains, Cities, and Software by Steven Johnson

• The movement from low-level rules to higher-level sophistication is what we call emergence.
Data Model using OAI-ORE
Acknowledgements

Office of Cyberinfrastructure DataNet Award #0830976

NLG grant award LG0606018206

• Anne Thessen and David Patterson (Life sciences slides)
• Alex Szalay (Data Flow)
• Carole Palmer (Domain coverage/methods slides)
• Carl Lagoze (Data Framework slides)
• Tim DiLauro (OAI-ORE)