The Data Conservancy

CNI Spring Forum

sayeed@jhu.edu

April 7, 2009

Data Curation

The Data Conservancy embraces a shared vision: data curation is not an end, but rather a means to collect, organize, validate, and preserve data to address grand research challenges that face society.

Goal

The overarching goal of DC is to support new forms of inquiry and learning to meet these challenges through the creation, implementation, and sustained management of an integrated and comprehensive data curation strategy.

Understanding Infrastructure: Dynamics, Tensions, and Design

Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures"

Paul N. Edwards
Steven J. Jackson
Geoffrey C. Bowker
Cory P. Knobel

January 2007

NSF Grant 0630263 • Human and Social Dynamics • Computer and Information Science and Engineering • Office of Cyberinfrastructure

…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.

Principles

Our strategy focuses on connection of systems into infrastructure through a program informed by user-centered design and research, sustained through a portfolio of funding streams, and managed through a shared, coordinated governance structure.

Build on existing exemplar scientific projects, communities and virtual organizations that have deep engagement with citizen scientists and extensive experience with large-scale, distributed system development

Partner institutions

- Johns Hopkins University (Lead institution)
- Cornell University
- Encyclopedia of Life (Marine Biological Laboratory)
- Fedora Commons
- National Center for Atmospheric Research
- National Snow and Ice Data Center
- Portico
- Tessella, Inc.
- University of California Los Angeles
- University of Illinois at Urbana-Champaign

Unfunded partner institutions

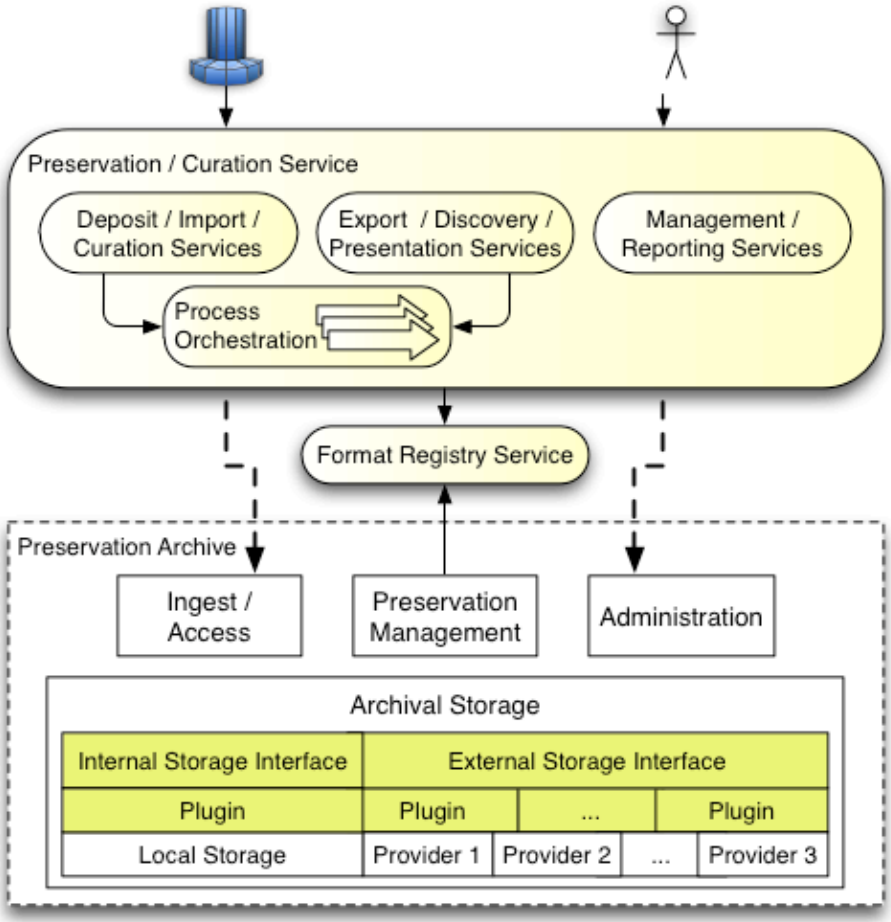
- Australian National Data Service
- Australian National University
- British Library
- Digital Curation Centre
- Microsoft Research
- Monash University
- Nature Publishing Group
- Optical Society of America
- Sakai Foundation
- Space Telescope Science Institute
- SPARC
- University of Queensland, Australia
- Zoom Intelligence

Objectives

- Infrastructure research and development
  - Technical requirements
- Information science and computer science research
  - Scientific or user requirements
- Broader impacts
  - Educational requirements
- Sustainability
  - Business requirements

Technical Architecture

- Well-defined interfaces
- Emphasis on modularity and interchangeability
- Reference implementation using Fedora, but other implementations desirable
- Storage layer features
  - Access to data stored / preserved elsewhere
  - Interface to third-party storage services



Data Framework

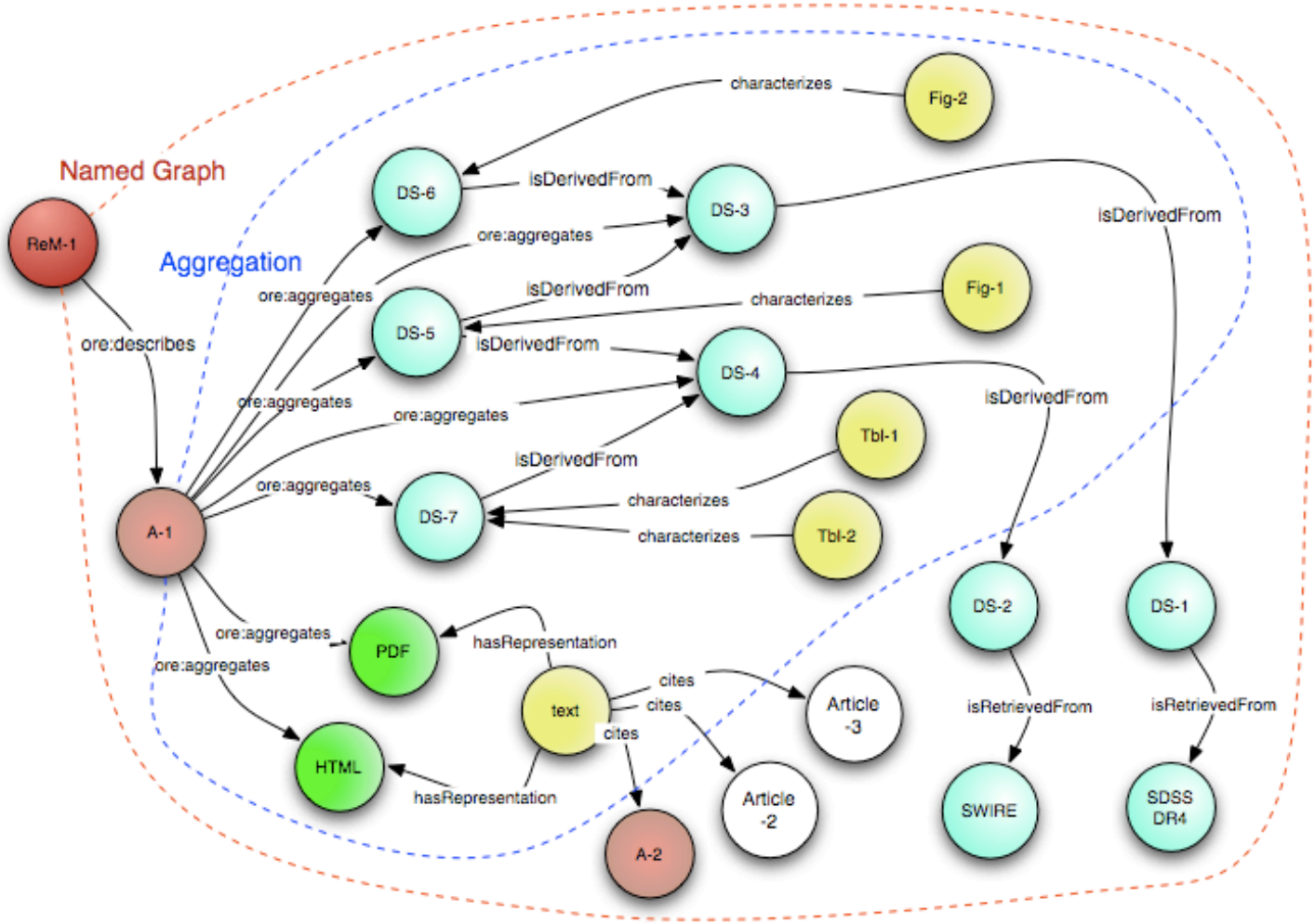
- Start with a common conceptualization that applies across scientific domains
- Exploit semantic technologies
- Leverage existing work
- Prototype the framework in target communities
  - Iteratively refine, learn from experience
  - Demonstrate success, measured in terms of new science

Common Conceptualization

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

Wiens, J. A. (1992). Cambridge studies in ecology: The ecology of bird communities. *Foundations and Patterns*, 1; *Processes and Variations*, 2

Data Model using OAI-ORE

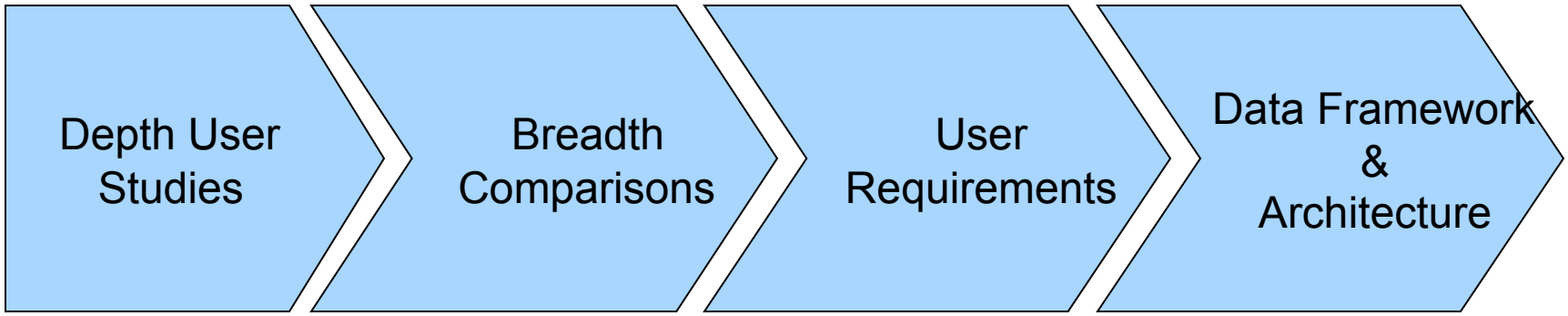


Domain coverage/methods

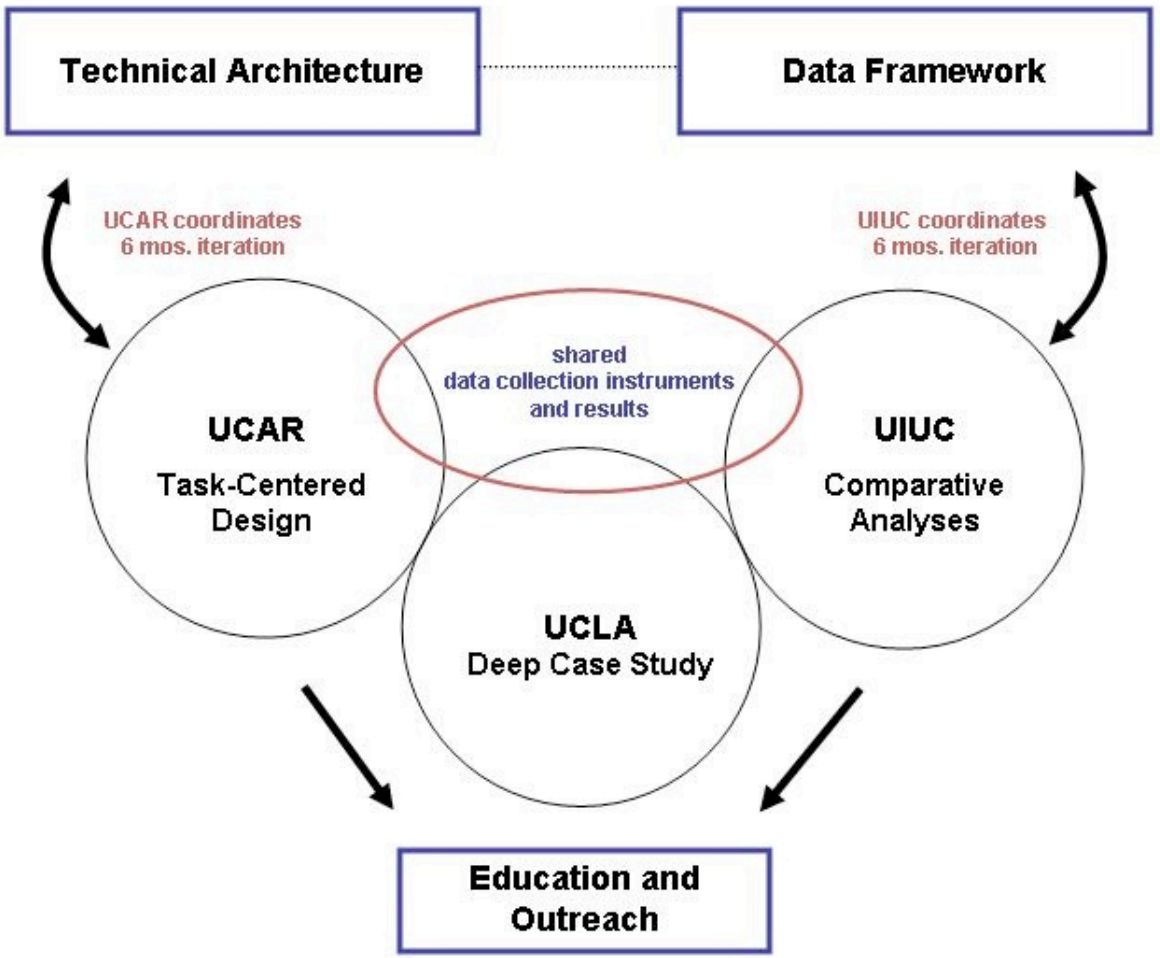
- Multi-site user research methods are a blend of:
  - Case study & domain comparisons
  - Depth & breadth
  - Local & global

<table>
<tr><th></th><th>Astronomy</th><th>Biodiversity</th><th>Earth Science</th><th>Social Science</th><th></th></tr>
<tr><td>UCAR</td><td colspan="4">Task-based design and usability testing ⇒ Use cases, data requirements, system recommendations</td><td>UCAR</td></tr>
<tr><td>UCLA</td><td>Ethnography, virtual ethnography, oral histories ⇒ Use cases, data requirements</td><td colspan="3">Interviews, Surveys, Worksheets, Content analysis ⇒ Curation requirements, taxonomy, metadata/ provenance framework</td><td>UIUC</td></tr>
</table>

Design Flow


Depth User Studies
Breadth Comparisons
User Requirements
Data Framework & Architecture

Information science research



Sustainability

- Diversified portfolio of funding and perspectives
- Align with existing institutional priorities
- Leverage partners' sustainability mechanisms
- Focus on economies of scale and economies of scope
- Consider business requirements as equal to other requirements from inception of activity
- Incorporate findings from Blue Ribbon Task Force on Sustainable Digital Preservation and Access

Implications for Libraries

- Libraries as part of a distributed network
- Data as collections
- Data as services
- Librarians as data scientists
- “Data centers are the new library stacks”
  -- Winston Tabb (JHU Dean of Libraries)

Acknowledgements

DataNet award for “The Data Conservancy”

NLG grant award LG0606018206

Microsoft Research

- Tim DiLauro (technical architecture slide)
- Carl Lagoze (data framework slides)
- Carole Palmer (information science slides)