The Data Conservancy: Building Sustainable Infrastructure for Interdisciplinary Scientific Data Curation and Preservation

PV 2009

sayeed@jhu.edu

December 3, 2009

Data Conservancy

- One of two awards through US National Science Foundation (NSF) DataNet program
- Led by Sheridan Libraries at Johns Hopkins University
- Goal of DataNet program is to build data curation infrastructure
- Next round of DataNet will add three more partners into the network

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

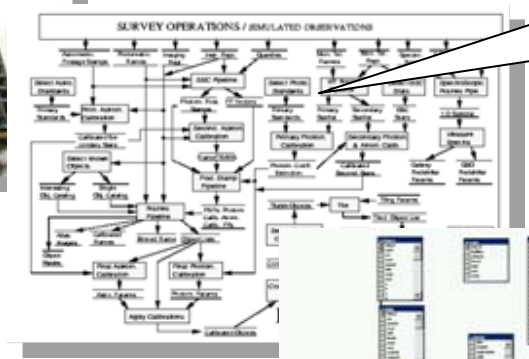
NSF Grant 0630263 • Human and Social Dynamics • Computer and Information Science and Engineering • Office of Cyberinfrastructure

…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.

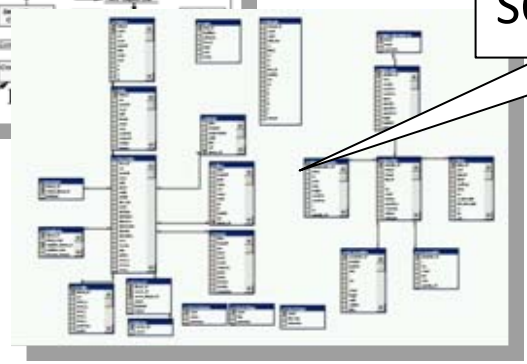
Data Flow (Levels of Data)

Pixel data collected by telescope

Sent to Fermilab for processing

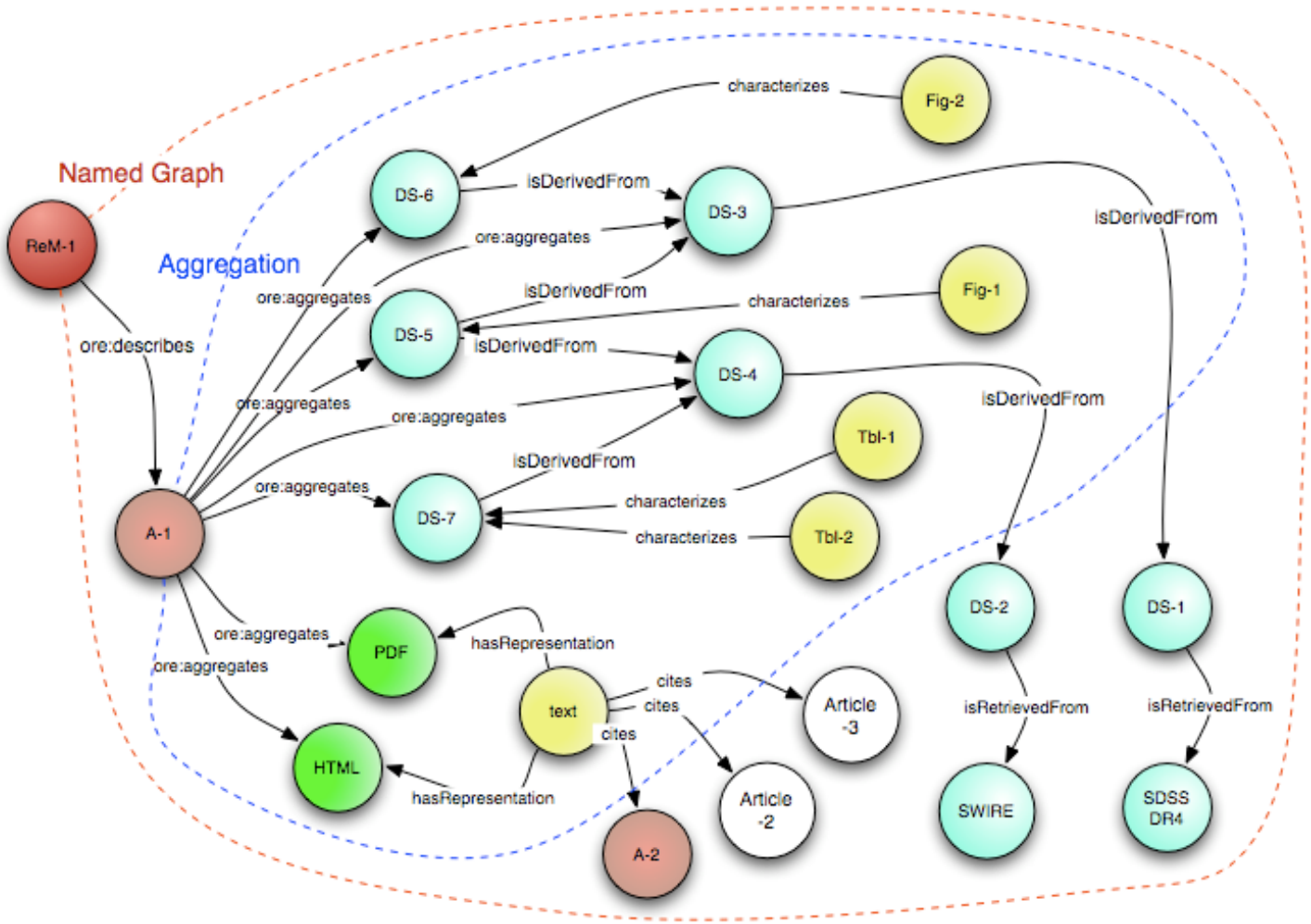


Beowulf Cluster produces catalog



Loaded in a SQL database

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>Life Sciences</th><th>Earth Sciences</th><th>Social Sciences</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>

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
- New roles for research libraries

New Roles for Research 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

Office of Cyberinfrastructure DataNet Award #0830976

NLG grant award LG0606018206

Microsoft Research

- Alex Szalay (data flow slide)
- Tim DiLauro (OAI-ORE slide)
- Carole Palmer (information science slide)