Data Conservancy and the US NSF DataNet Initiative

Fourth Workshop on Data Preservation and Long-Term Analysis in HEP

Sayeed Choudhury

Johns Hopkins University

NSF DataNet

- Science and engineering research and education are increasingly digital and data-intensive
- New methods, management structures and technologies necessary
- NSF DataNet solicitation addresses challenge by creating exemplar data infrastructure organizations

NSF recent actions

- Five DataNet partners funded at $20 million each for 5 years – seed funding
- Data Conservancy and DataONE are first two awards – up to three more awards in next round
- Part of broader initiatives at NSF including requirement for data management plans and (separate) Johns Hopkins grant for feasibility study of open access repository

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.

Goal

The goal of Data Conservancy is to support new forms of inquiry and learning that address grand research challenges. The Data Conservancy will accomplish this goal 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
- DuraSpace
- Marine Biological Laboratory
- 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

Objectives

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

Domain coverage/methods

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

	Astronomy	**Earth Sciences**	**Life Sciences**	**Social Sciences**	
UCAR	Task-based design and usability testing ⇒ Use cases, data requirements, system recommendations				**UCAR**
UCLA	Ethnography, virtual ethnography, oral histories ⇒ Use cases, data requirements	Interviews, Surveys, Worksheets, Content analysis ⇒ Curation requirements, taxonomy, metadata/provenance framework			**UIUC**

Data Framework

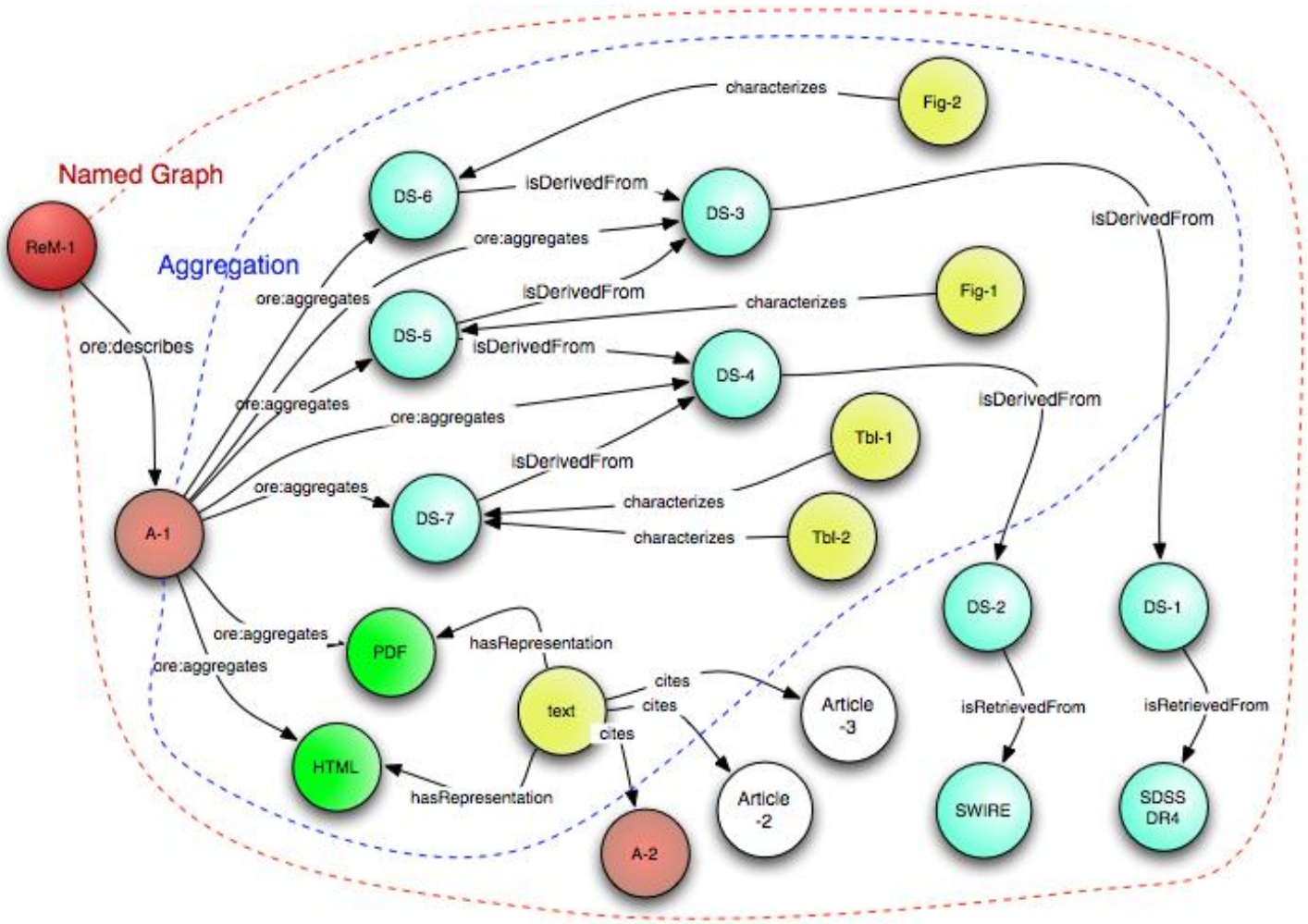
- Start with a common conceptualization that applies across scientific domains
- Exploit semantic technologies
- Leverage existing work
- Prototype the framework in target communities
  - Iteratively refine, learn from experience
  - Demonstrate success, measured in terms of new science

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



Acknowledgements

Office of Cyberinfrastructure DataNet Award #0830976

Office of Cyberinfrastructure EAGER Award #0948134

- Carole Palmer (information science slides)
- Carl Lagoze (Data Framework slides)
- Tim DiLauro (OAI-ORE)