Data Conservancy: A Web Science View of Data Curation

Sayeed Choudhury

Johns Hopkins University

sayeed@jhu.edu

Data Curation

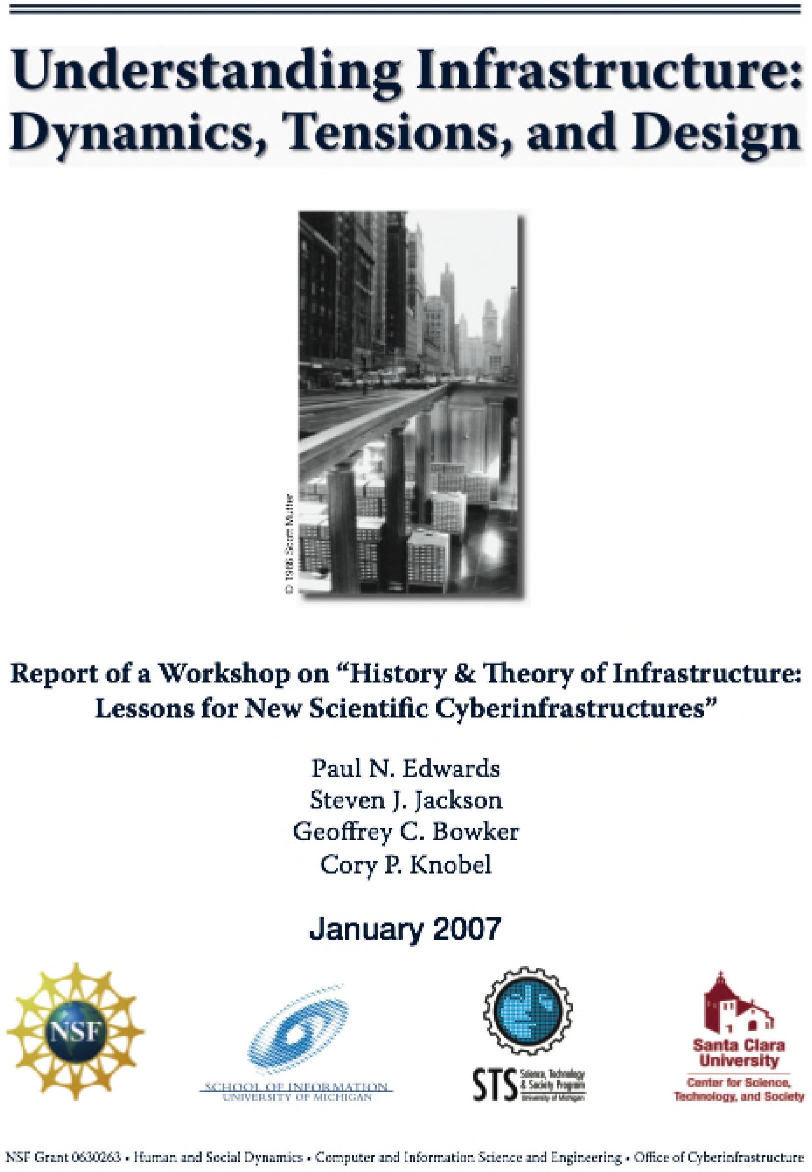
The Data Conservancy embraces a shared vision: data curation is a means to collect, organize, validate and preserve data so that scientists can find new ways to address the grand research challenges that face society.

Goal

The 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.

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



Understanding Infrastructure: Dynamics, Tensions, and Design

© 1986 Scott Mutter

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
SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN
STS Science, Technology & Society Program University of Michigan
Santa Clara University
Center for Science, Technology, and Society

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

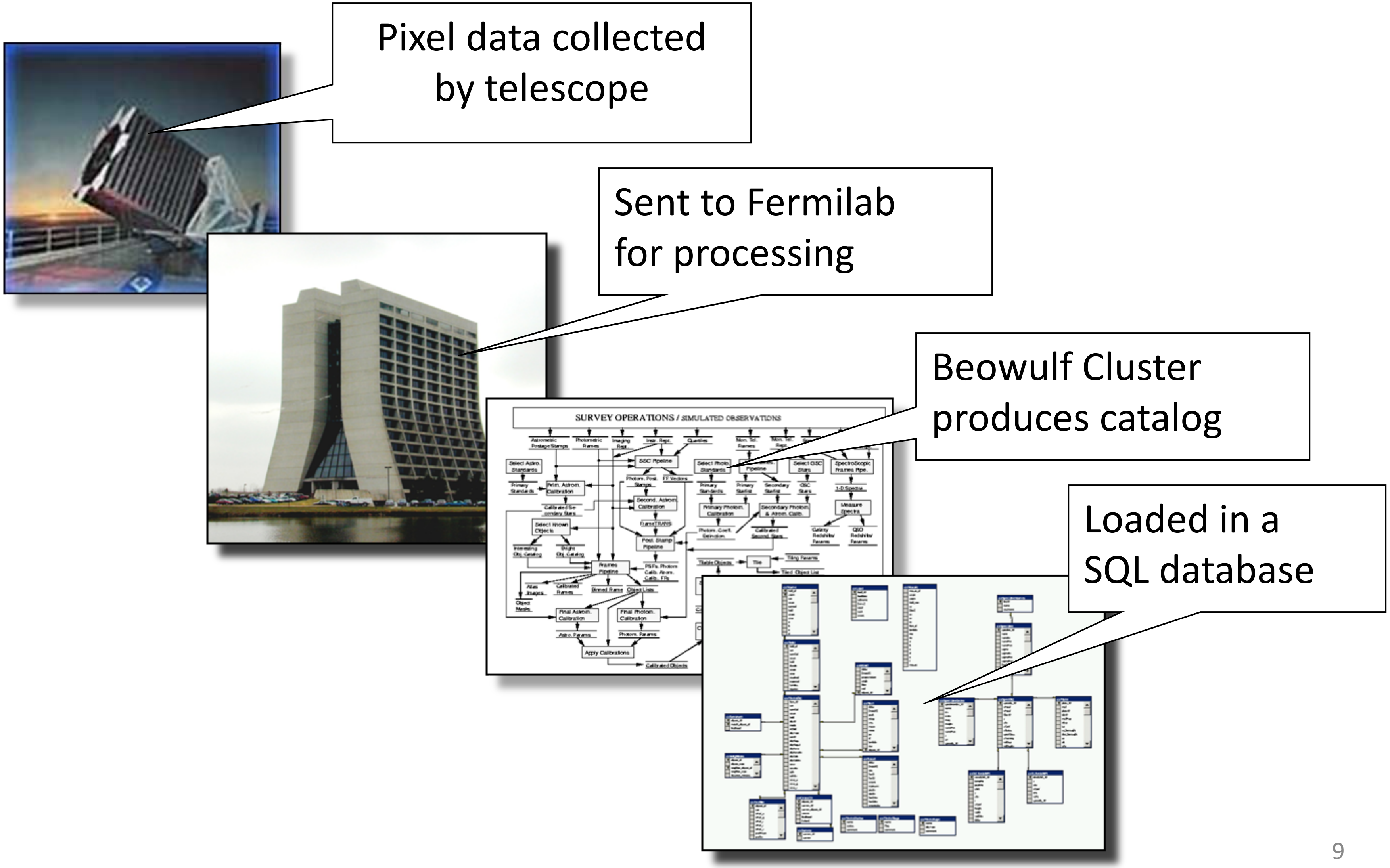
Objectives

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

Objectives

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

Data Flow (Levels of Data)



Pixel data collected by telescope

Sent to Fermilab for processing

Beowulf Cluster produces catalog

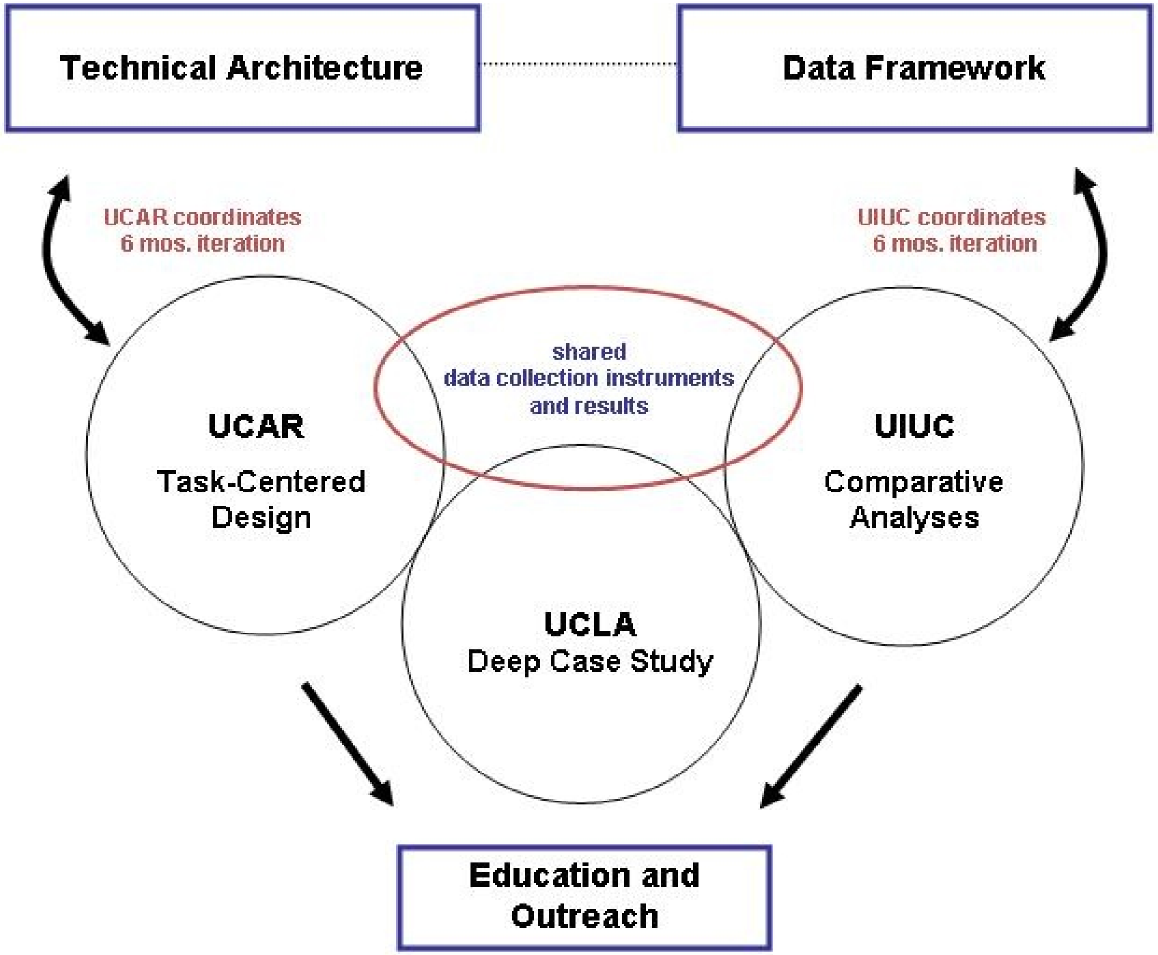
Loaded in a SQL database

Domain coverage/methods

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

<table>
<tr><th></th><th>Astronomy</th><th>Earth Sciences</th><th>Life Sciences</th><th>Social Sciences</th><th></th></tr>
<tr><td><b>UCAR</b></td><td colspan="4">Task-based design and usability testing ⇒ Use cases, data requirements, system recommendations</td><td><b>UCAR</b></td></tr>
<tr><td><b>UCLA</b></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><b>UIUC</b></td></tr>
</table>

Information science research



Technical Architecture

Data Framework

UCAR coordinates
6 mos. iteration

UIUC coordinates
6 mos. iteration

shared
data collection instruments
and results

UCAR
Task-Centered
Design

UCLA
Deep Case Study

UIUC
Comparative
Analyses

Education and
Outreach

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.

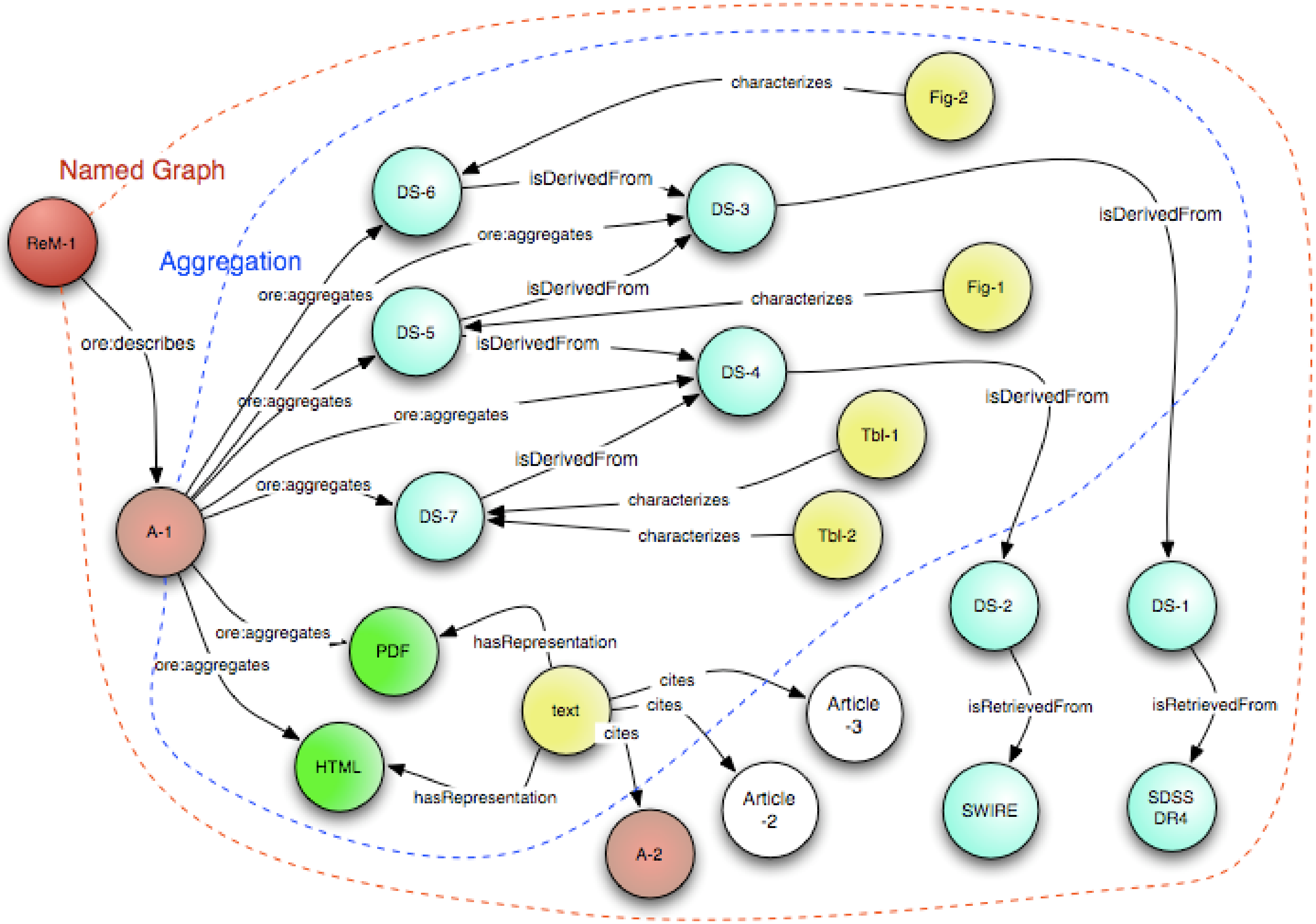
Wiens, J. A. (1992). Cambridge studies in ecology: The ecology of bird communities. *Foundations and Patterns*, 1; *Processes and Variations*, 2

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

NSF

Office of Cyberinfrastructure DataNet Award #0830976

INSTITUTE of Museum and Library SERVICES

NLG grant award LG0606018206

Microsoft Research

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