The Data Conservancy

April 5, 2012
Data Conservancy Objectives

The Data Conservancy is a community that develops solutions for data preservation and sharing to promote cross-disciplinary re-use:

- Preserve – collect and take care of research data
- Share – reveal data’s potential and possibilities
- Discover – promote re-use and new combinations
Architecture mapped to OAIS
Service Oriented Architecture

- Well-defined APIs
  - Public HTTP-based APIs
  - Internal Java APIs
- Loosely coupled
  - Minimal dependencies between system components
- Principle adhered to throughout the Data Conservancy, not just at the Service layer
- Facilitates interoperability
- Promotes sustainability
  - (e.g. update the archival storage module to leverage more cost-effective storage)
- Allows independent evolution and extension of Data Conservancy modules
Some Definitions

**Archiving** – actions to support long-term data protection (e.g., storage, backup, media migration, fixity)

**Preservation** – policy and actions to ensure long-term (perhaps as short as 5 years) access and sharing (e.g., metadata, format migration)

**Curation** – adding value to foster re-use and unanticipated use (e.g., feature extraction, query framework)
Data Model

- Multiple Data Models
- Content models for describing the contents of a Manifestation
- General Model used to correlate model entities across heterogeneous datasets
  - geo-reference, time of observation, etc…
Feature Extraction Framework

- Must accommodate a variety of data formats
- No assumption made regarding the form of data input or output
- Not coupled to a specific execution model
Feature Extraction Framework

- **Subsetting**
  - Returning a portion of a dataset
- **Indexing**
  - Output suitable for indexing by the Query Framework
- **Workflows**
  - Process Orchestration, Meandre, Taverna, Kepler
- **Execution environment for analysis**
  - Stateless Mappings basis for MapReduce
Status

- First version of the software to be publicly released in August
- Probably quarterly releases there after
- arXiv.org and NSIDC pilots operational for more than 1 year
- JHU Data Management Services
- SEAD collaboration
- NSIDC/CU instance
Powered by the Data Conservancy

• JHU Data Management Service (DMS) represents the culmination of two years of research, design, development and implementation of Data Conservancy
• Service launched in July 2011
• DC instance launched in October 2011
• Important, essential foundations in place
• There remains work to be done so join the community!
Citation in the Data Conservancy

• Citable locator technology agnostic design
  • However DOI's through EZID will be the first locators supported
• All collections will have a citation, but two use cases have been identified
  • User enters a citation
  • Citation auto-generated based on entered metadata
Acknowledgements and Resources

- NSF Award OCI-0830976
- Sheridan Libraries financial support
- Johns Hopkins University financial support
- http://dataconservancy.org
- http://dmp.data.jhu.edu