Johns Hopkins University Data Management Services

Sayeed Choudhury

ARL eScience Institute – December 16, 2011

Johns Hopkins University Sheridan Libraries

Powered by Data Conservancy

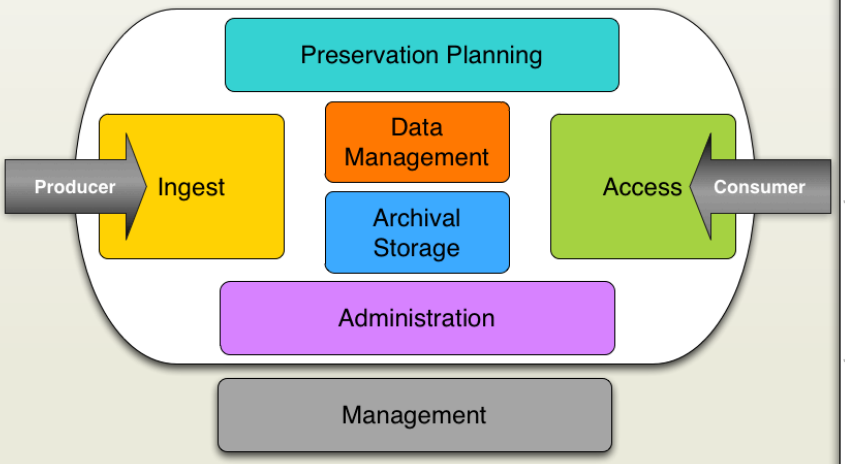
- JHU Data Management Service (DMS) represents the culmination of two years of research, design, development and implementation of Data Conservancy
- Service launched in July 2011
- DC instance launched in October 2011
- Important, essential foundations in place
- There remains work to be done

Data Conservancy Objectives

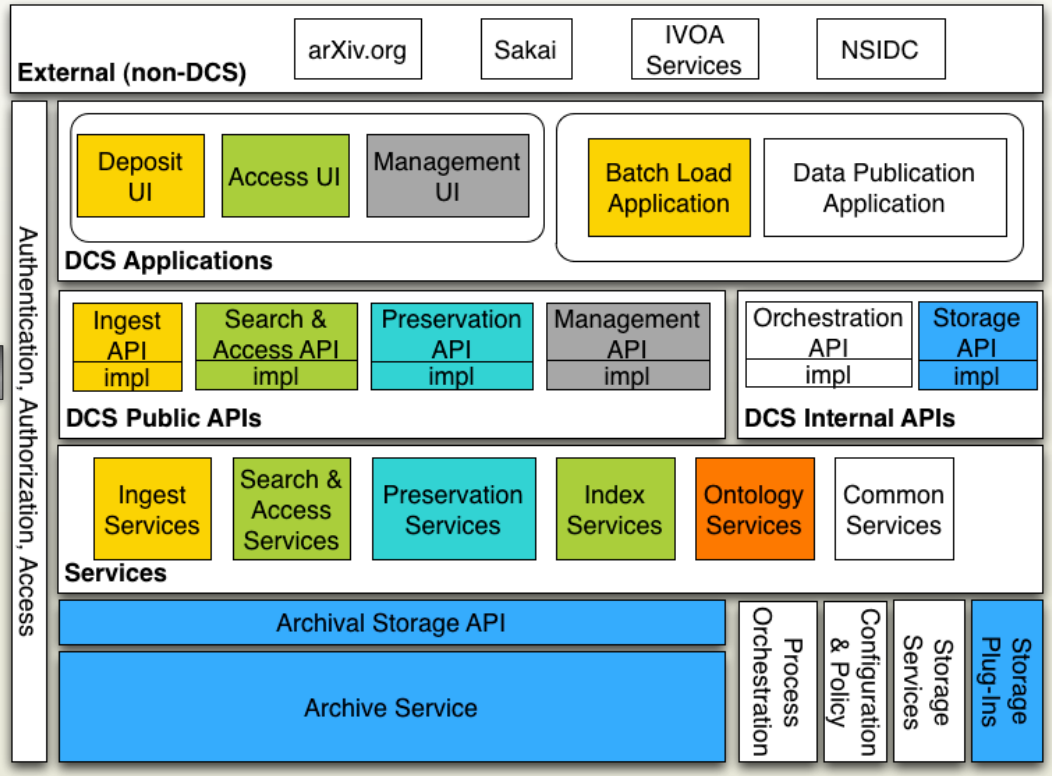
- Data Conservancy is a community that develops solutions for data preservation and sharing to promote cross-disciplinary re-use.

- 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



Open Archival Information System
Functional Entities



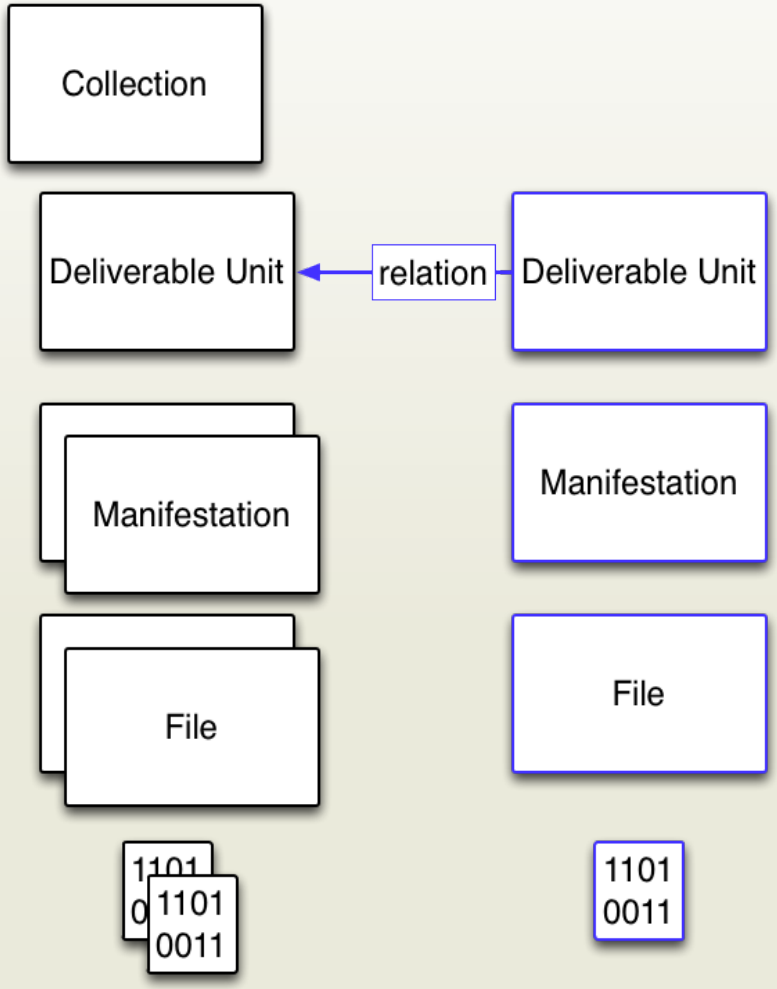
Data Conservancy Service
Architecture Block Diagram

Definition of Data Preservation

- "Data preservation involves providing enough representation information, context, metadata, fixity, etc. such that someone other than the original data producer can use and interpret the data."
  - Ruth Duerr, National Snow and Ice Data Center

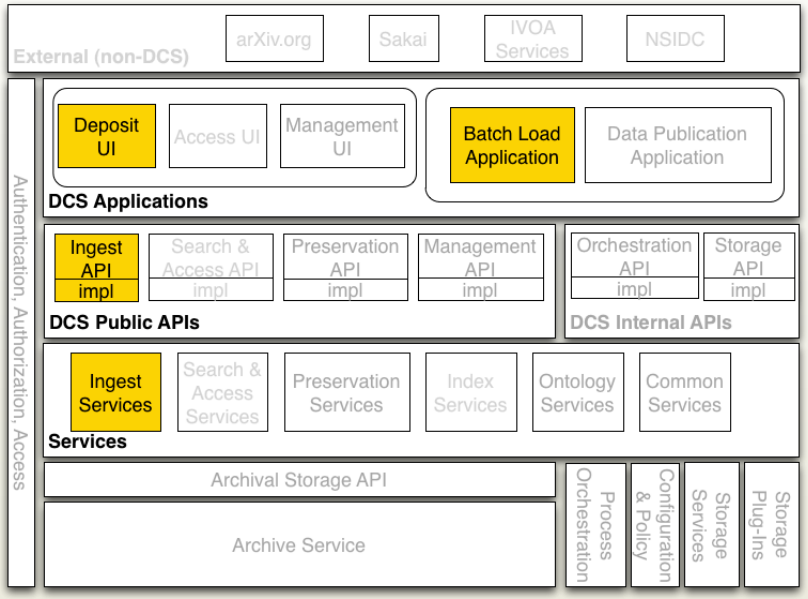
Data Model: Application

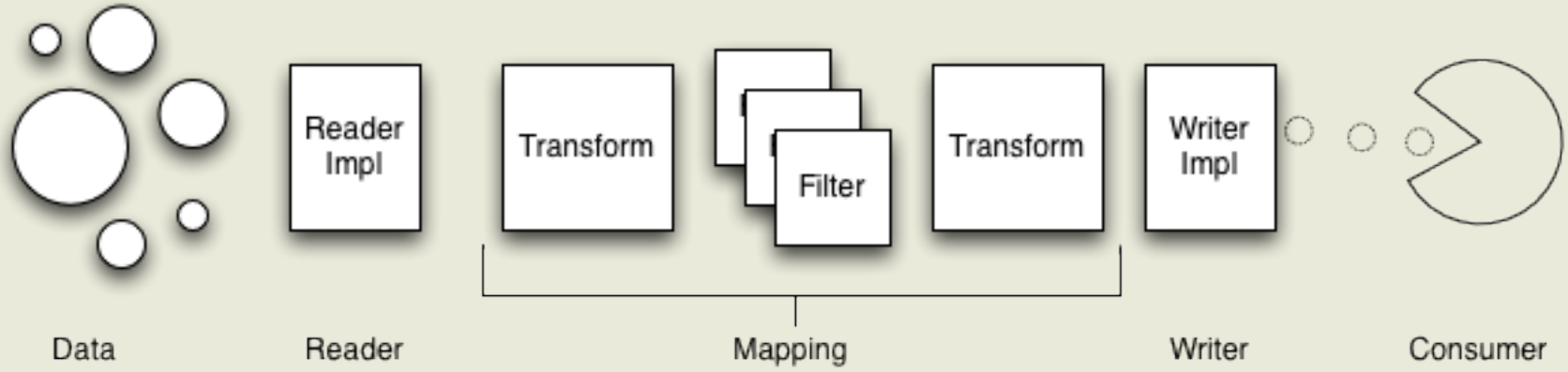
- Multiple Data Models
- Content models for describing the contents of a Manifestation
- General Model used to correlate model entities across heterogeneous datasets
  - geo-reference, time of observation, etc…



Feature Extraction Framework: Design

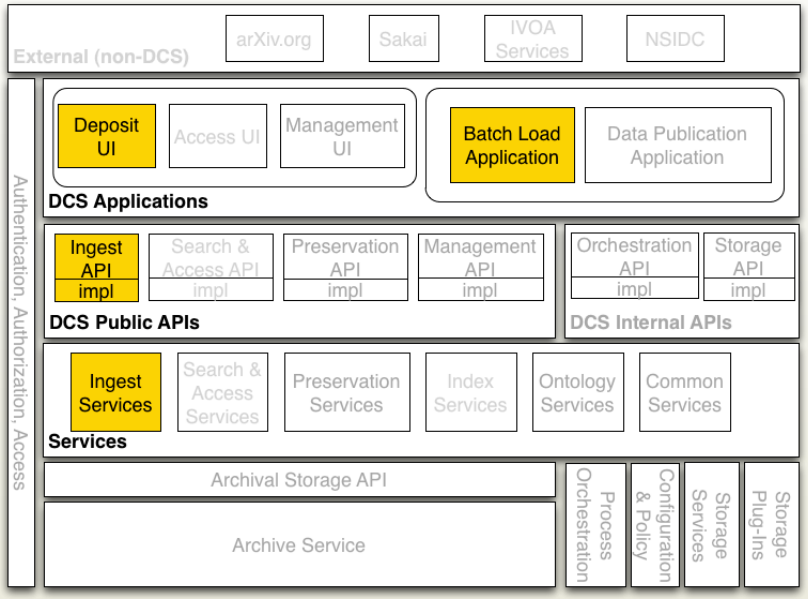
- Must accommodate a variety of data formats
- No assumption made regarding the form of data input or output
- Not coupled to a specific execution model





Feature Extraction Framework: Application

- 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



Data Management Layers

Layers	Examples	Implication for PI	Implication relative to NSF
Curation	Future JHU Data Archive and other DCS instances	• Feature Extraction • New query capabilities • Cross-disciplinary	• Competitive advantage • New opportunities
Preservation	JHU Data Archive Portico ICPSR	• Ability to use own data in the future (e.g. 5 yrs) • Data sharing	• Satisfies NSF needs across directorates
Archiving	CUAHSI NEES Dataverse	• Provides identifiers for sharing, references, etc.	• Could satisfy most NSF requirements

Storage	Server in Lab
	Website
	Amazon S3
		• Responsible for:
		  • Restore
		  • Sharing
		  • Staffing
			• Could be enough for now but not near-term future

Defining Sustainability

- "Ensuring that valuable digital assets will be available for future use is not simply a matter of finding sufficient funds. It is about mobilizing resources—human, technical, and financial—across a spectrum of stakeholders diffuse over both space and time."

Questions?

- Before we move onto the JHU Data Management Service (DMS), are there questions about the Data Conservancy?

Establishing the JHU DMS

- May 2010 NSF announces DMP expectations
- Services incubated and scoped summer/fall 2010
  - Build on Data Conservancy expertise
- Proposed in January and launched in July 2011
  - Consultative data management planning services to support NSF proposals
  - Post award data management services
- Assessment of service in March 2012

Background work to scope services

- Review of data management plan best practices and development of questionnaire
- Piloted data management consultations as cases
- Short data survey with over 70 JHU researchers
- Analysis of JHU NSF proposal and award activity
- Business school capstone project on storage options and costs
- Review of past data archiving projects and work

Proposing data management services

- Services scoped to support anticipated NSF requirements and to reflect system capabilities
  - Defined time limits, volume of data deposited per project, unencumbered data only for now
- Prepared budget for services
  - Five year timeframe for costs
  - All costs included: staffing, hardware, overhead, etc.
  - Cost assumptions included: total data archived, complexity of data prep for ingest

Developing financial model

Support secured and financial model established

- Data management planning for NSF proposals
  - Service directly funded by schools
  - Each school pays percentage according to 3 year average of total NSF proposals submitted
- Post award data management
  - Fee based service billed through a service center
  - First year fee a percent of total direct costs on grant

JHU Data Management Services team

Dedicated group (that collaborates with DC and Digital Research and Curation Center)

- Two data management consultants
- Senior technical consultant *(Part-time)*
- Software developer
- System administrator *(to be hired)*
- Interim manager *(Part-time)*

Service marketing

- Reach out through all stakeholders
  - Announcements through Deans
  - Work with research projects administration
  - Outreach to department administrators
  - Briefings with library colleagues/departments
  - Presentations to researchers, graduate students
- More to do….and then repeat!

Experiences and lessons so far…

- Initial NSF DMP guidelines are less clear than anticipated
- Researchers don't distinguish between storage, archiving and preservation they just want to meet requirement
- There is no such thing as a good boiler plate plan
- DMP requirement creates opportunity to discuss overall data management

Opportunities

- Grow researcher/graduate student understanding of data management
- Establish an archive specifically designed for data, enabling future discovery and use
- Expand services to support:
  - Other granting agency DMP requirements
  - Research community data management needs
- Build collective expertise across communities

Acknowledgements and Resources

- NSF Award OCI-0830976
- Sheridan Libraries financial support
- Johns Hopkins University financial support
- Elliot Metsger for infrastructure slides
- Tim DiLauro for inspiration about layers
- Data Conservancy colleagues for their exceptional work and patience
- http://dataconservancy.org
- http://dmp.data.jhu.edu