Data Conservancy Update

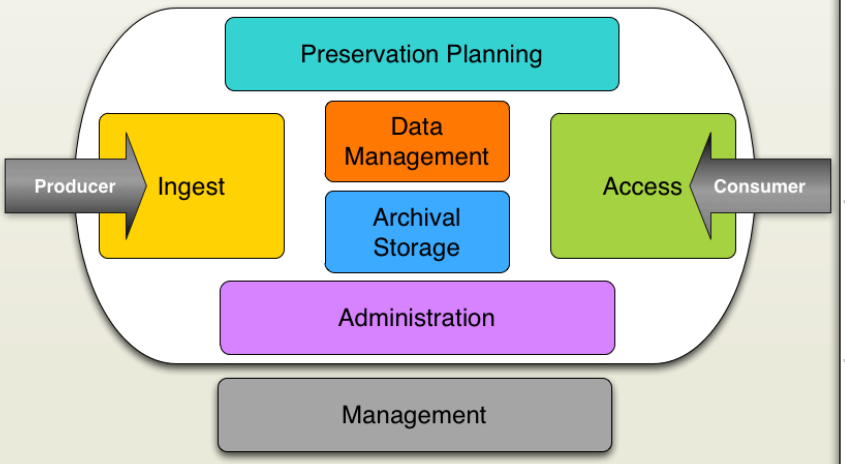
CNI 2011 Spring Membership Meeting

Elliot Metsger and Sayeed Choudhury, Johns Hopkins University

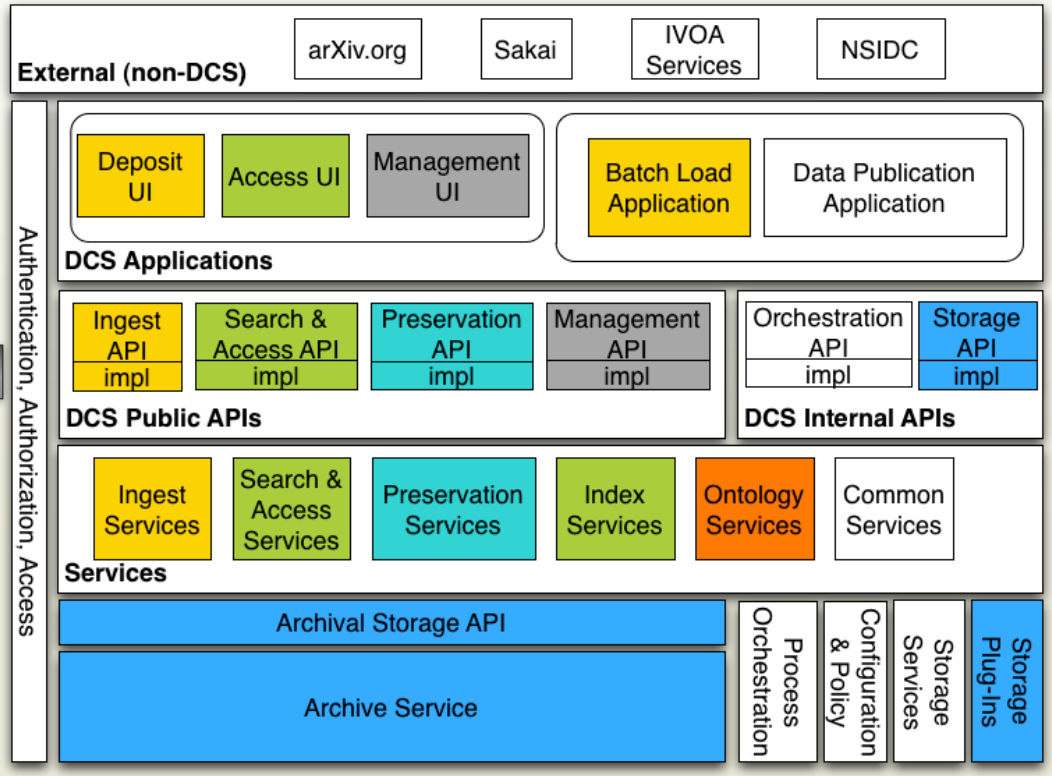
Data Conservancy Vision

- Data Conservancy embraces a shared vision: scientific 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.

Architecture Mapped to OAIS



Open Archival Information System
Functional Entities

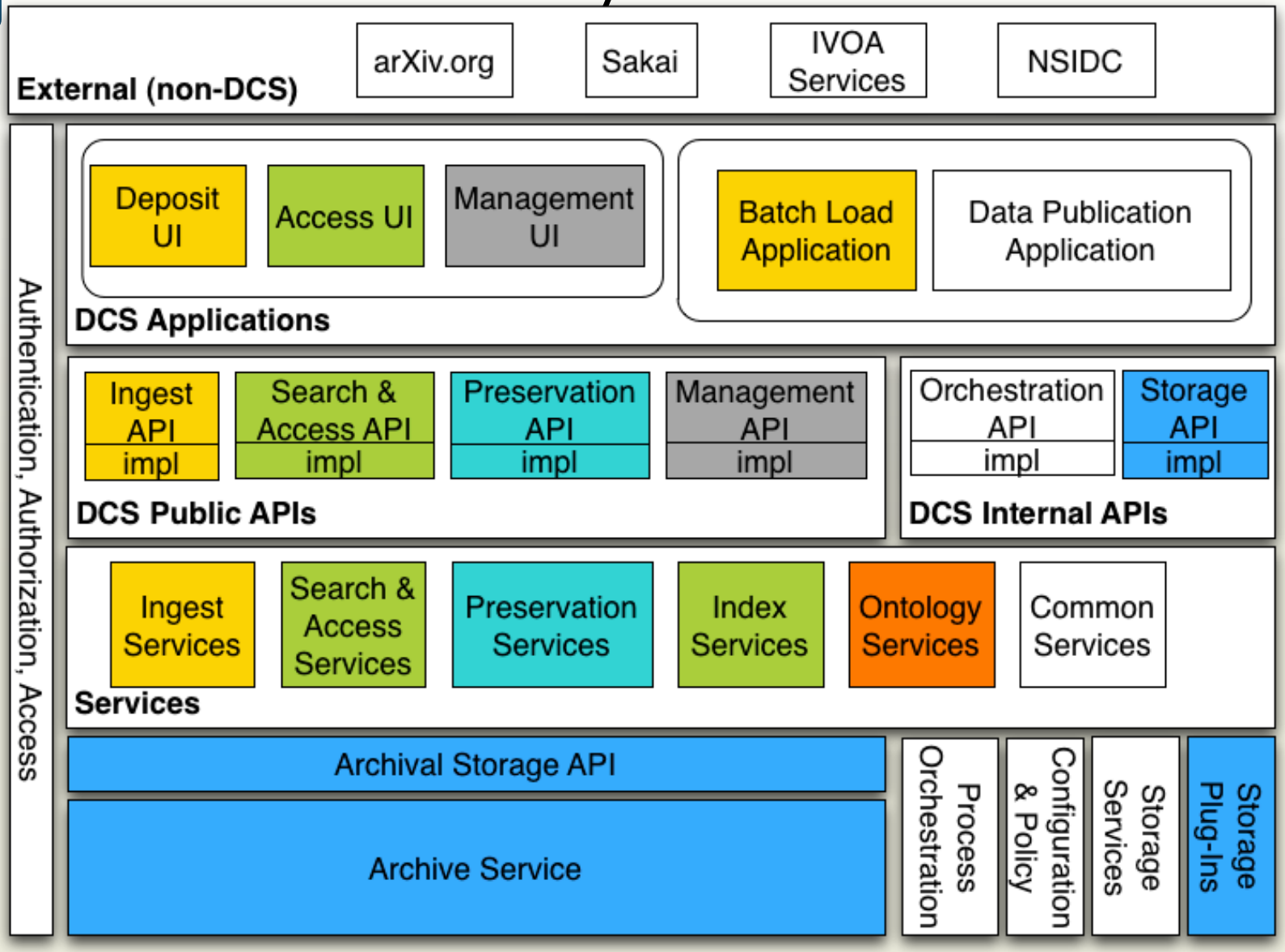


Data Conservancy Service
Architecture Block Diagram

Technical Infrastructure

- Repository initially located at JHU Libraries for data deposit
- Preservation framework
- Loosely coupled services for data access, use, discovery and integration
- Interoperability with existing and future nodes on network (including potential next-round DataNet partners)

Data Conservancy Block Architecture



Proofs-of-Concept to Pilots

- Decompose large problems into smaller, more tractable problems
- Examine the problem vis-à-vis prototyping
- Informs us regarding:
  - size of the problem space
  - feasibility of approach
  - boundaries and role of the Data Conservancy
- Expect this pattern to continue through Year Five

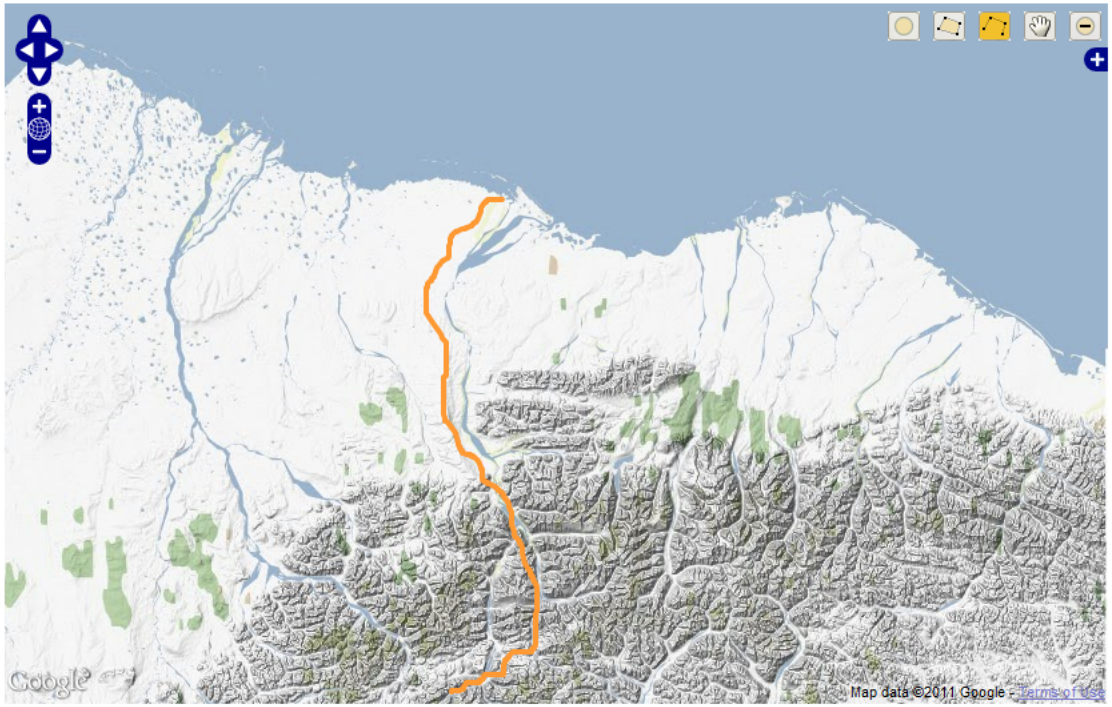
Prototyping As A Strategy

1. Cross-Dataset Queries – Ice Roads POC
2. arXiv.org – Connecting data with publications
3. IVOA – Integration with scientific frameworks
4. Sakai – Integration with teaching and learning environments
5. NSIDC Glacier Photograph Collection – Integration with existing cyberinfrastructure
6. Dry Valleys Visualization: POC to Pilot
7. Coastal Bays Visualization: POC to Pilot

Alaskan North Slope Ice Roads

Cross Dataset Discovery

Web Services



Results | Zoom

FILTER BY: VU (SHOW ALL)

NT = near threatended, VU = vulnerable,
EN = endangered, CR = critically endangered

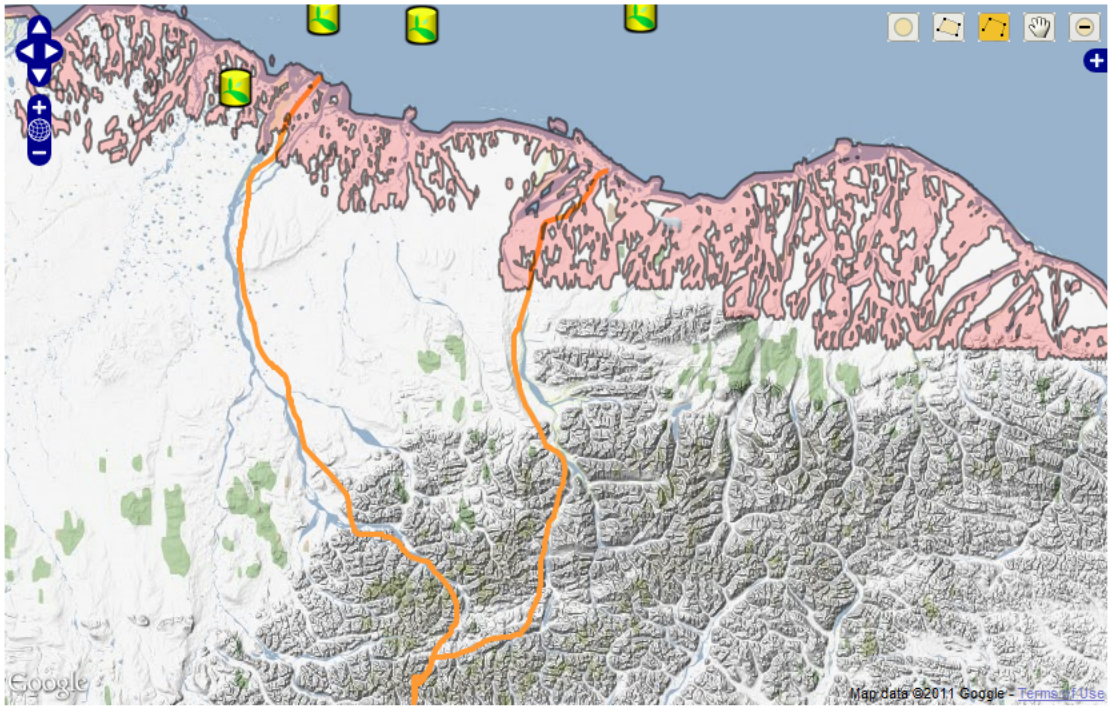
frogs
- *Lithobates sylvaticus*

mammals (terrestrial)
- *Alces americanus*
- *Alopex lagopus*
- *Canis latrans*
- *Canis lupus*
- *Castor canadensis*
- *Dicrostonyx groenlandicus*
- *Erethizon dorsatum*
- *Glaucomys sabrinus*
- *Gulo gulo*
- *Lemmus trimucronatus*
- *Lontra canadensis*
- *Lynx canadensis*
- *Marmota broweri*

Locations and Dens of Polar Bears

Cross Dataset Discovery

Web Services



Results | Zoom

FILTER BY: VU (SHOW ALL)

NT = near threatened, VU = vulnerable,
EN = endangered, CR = critically endangered

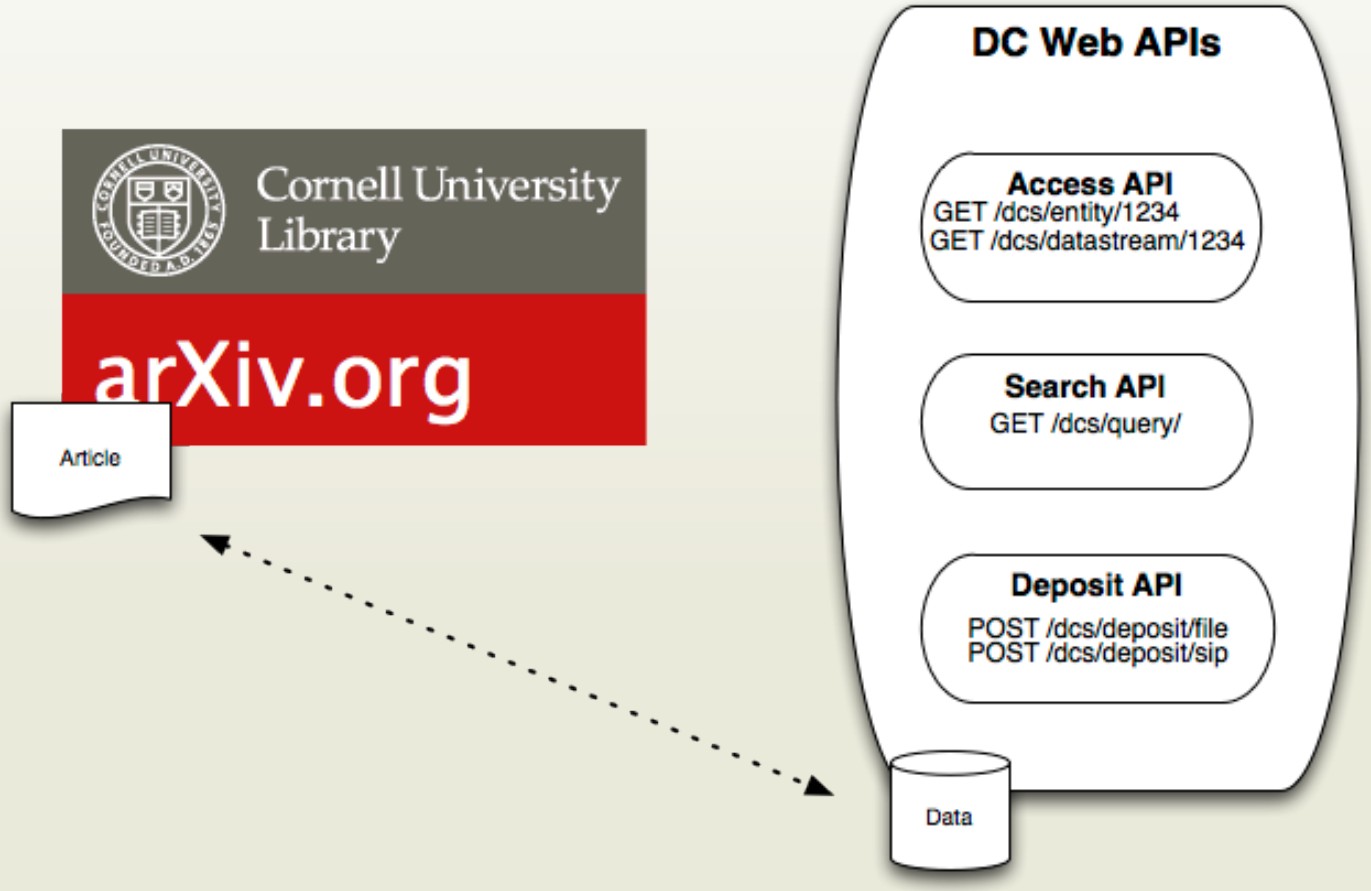
frogs

mammals (terrestrial)

Ursus maritimus VU

NSF · PBS an Atkins company · USGS science for a changing world · RED LIST · BirdLife INTERNATIONAL · GBIF www.gbif.org · eol Encyclopedia of Life · flickr

Connecting Data and Publications



Connecting Data With Publications

Start >> **Add Files** >> Process >> Metadata >> Preview

Submission files

File uploaded

Your submission to the archive must be in one of the following formats (listed in order of preference):

- (La)TeX, AMS(La)TeX, PDFLaTeX
- DOCX (Word 2007)
- PDF
- PostScript
- HTML with JPEG/PNG/GIF images

! If your submission is (La)TeX, then you must submit the source (plus necessary macros and figures), not derivative dvi, Postscript, or PDF (see Why TeX?). For more information on formats and other submission details see Submission Help.

Add files

Choose File | No file chosen

Continue: Process Files

Upload file

Files Currently Attached to Submission		Delete All
Filename	**Type**	**Actions**
datamodel.png	Image (gif/jpg etc)	Delete
example.tex	LATEX2e	Delete
fef-pipeline.png	Image (gif/jpg etc)	Delete

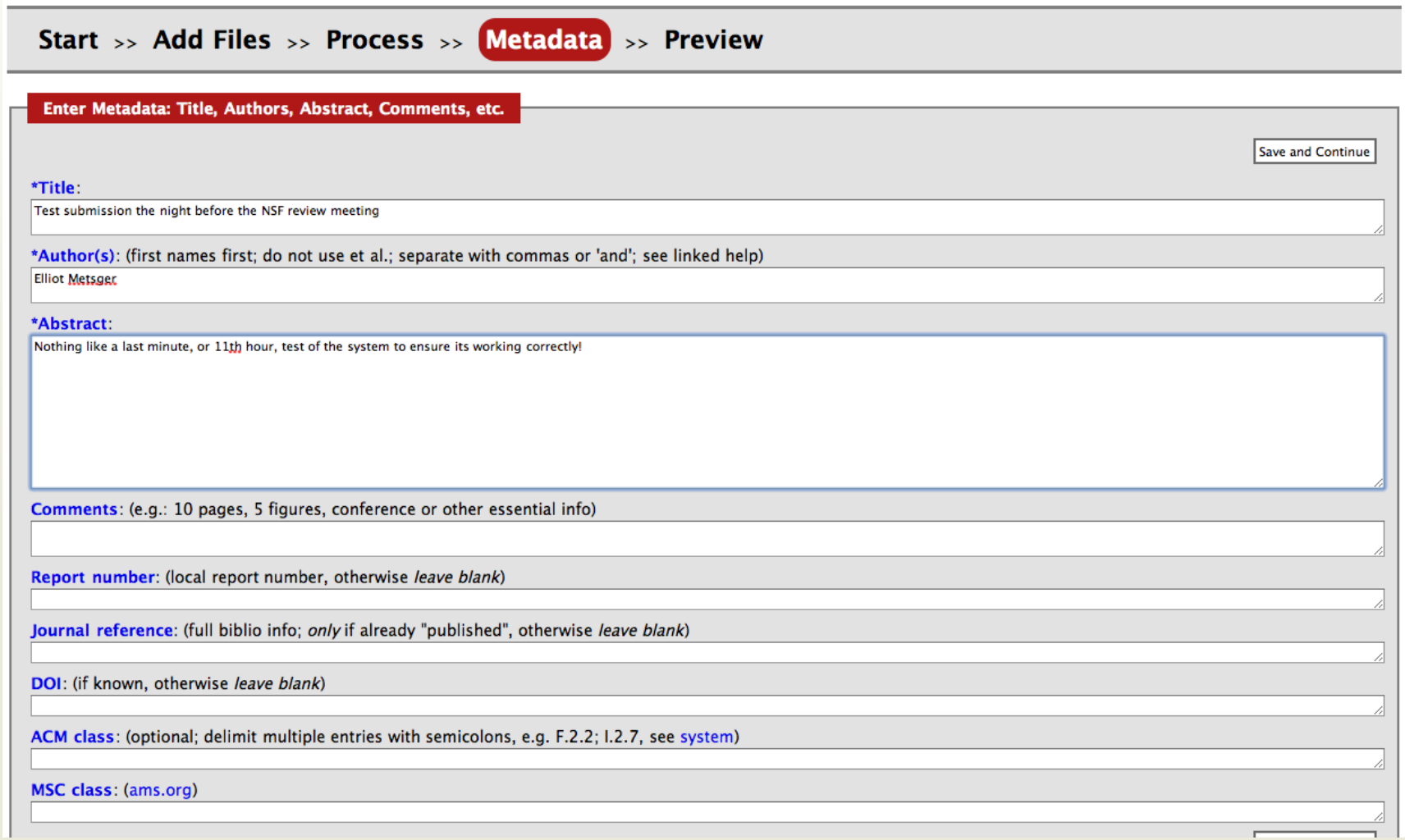
Data Conservancy Pilot Project

arXiv is currently participating in a pilot project with the Data Conservancy, which provides storage and infrastructure for scientific data. Please indicate if you are interested in storing large data sets, imagery, simulations, videos or other material associated with your submission, which exceeds the regular size limits imposed by arXiv, for storage with the Data Conservancy.

☑ Yes, please enable the extra features from the Data Conservancy pilot project.

Continue: Process Files

Connecting Data With Publications



Connecting Data With Publications

Start >> Add Files >> Process >> Metadata >> **Add Data** >> Preview

Data associated with this Submission

File uploaded

Here you may upload data sets, imagery, simulations, videos, or other material of substantial file size (10MB -- 100MB total), which is of relevance to your submission while exceeding the size restrictions imposed by arXiv. These files will be relayed to the Data Conservancy Project (DC) for archival storage and access.

There will be mutual links from arXiv to DC to connect the external material with this submission. Visit arXiv's dataset help for more information.

Data uploaded here will be distributed under the same license as the article. You may change the license on the [Start] screen.

! Material uploaded in this section will be relayed to the Data Conservancy. It will remain associated with your submission, but it will not be stored at arXiv.

This is a pilot project which explores future cooperation between arXiv and DC and the related infrastructure requirements. As such this service may undergo substantial changes and is not guaranteed to persist beyond the duration of the pilot. However reasonable efforts will be made to keep the uploaded material accessible.

Add material

Choose File | No file chosen | *Short description | Continue: Preview

Upload file

Ancillary Material Attached to this Submission			Delete All
Filename	**Type**	**SHA256 checksum**	**Actions**
dcs-model-1.1.0-SNAPSHOT-sources.jar	JAR archive	fdde6338851eef2a45cd9e26ea634f14393044e9a673004bdbd528c8c7d495d8	Delete
dcs-model-1.1.0-SNAPSHOT.jar	JAR archive	da5588c7bf3b6640856dda60427bde7eaf2984122b1598fe3032b90edf92ed12	Delete

Connecting Data With Publications

Article submitted

Your arXiv.org account: emetsger

E-mail: emetsger@jhu.edu
Name: Elliot Metsger
Default Category: astro-ph.EP
Groups: physics, nlin, cs, stat, q-fin, math, and q-bio

Affiliation: Johns Hopkins University
URL: http://www.dataconservancy.org/
Country: United States
Career Status: Staff

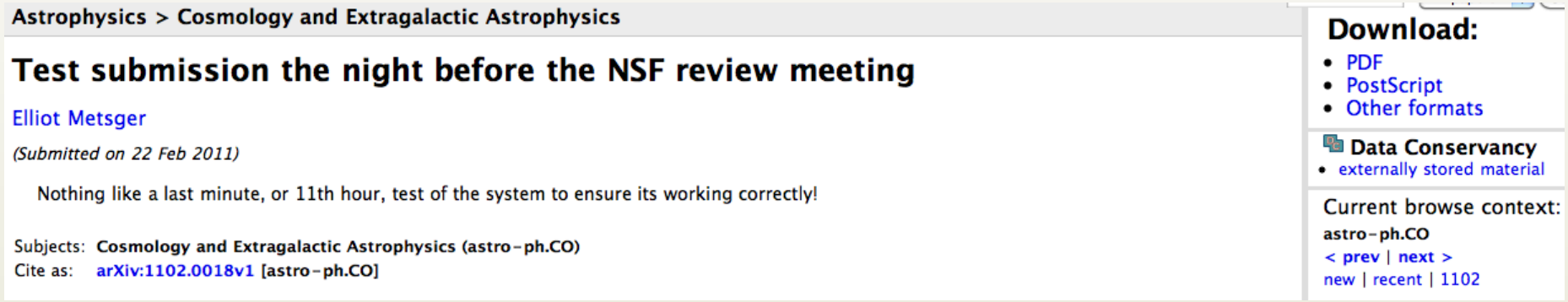
Change User Information | Change Password | Change Email

Article Submissions Update Delete Unsubmit

Identifier	Type	Title	Status	Actions	Expires
submit/0100785	New	Test submission the night before the NSF review meeting	submitted		

START NEW SUBMISSION

Connecting Data With Publications



Astrophysics > Cosmology and Extragalactic Astrophysics

Test submission the night before the NSF review meeting

Elliot Metsger

(Submitted on 22 Feb 2011)

Nothing like a last minute, or 11th hour, test of the system to ensure its working correctly!

Subjects: **Cosmology and Extragalactic Astrophysics (astro-ph.CO)**
Cite as: **arXiv:1102.0018v1 [astro-ph.CO]**

Download:

- PDF
- PostScript
- Other formats

Data Conservancy

- externally stored material

Current browse context:
astro-ph.CO
< prev | next >
new | recent | 1102

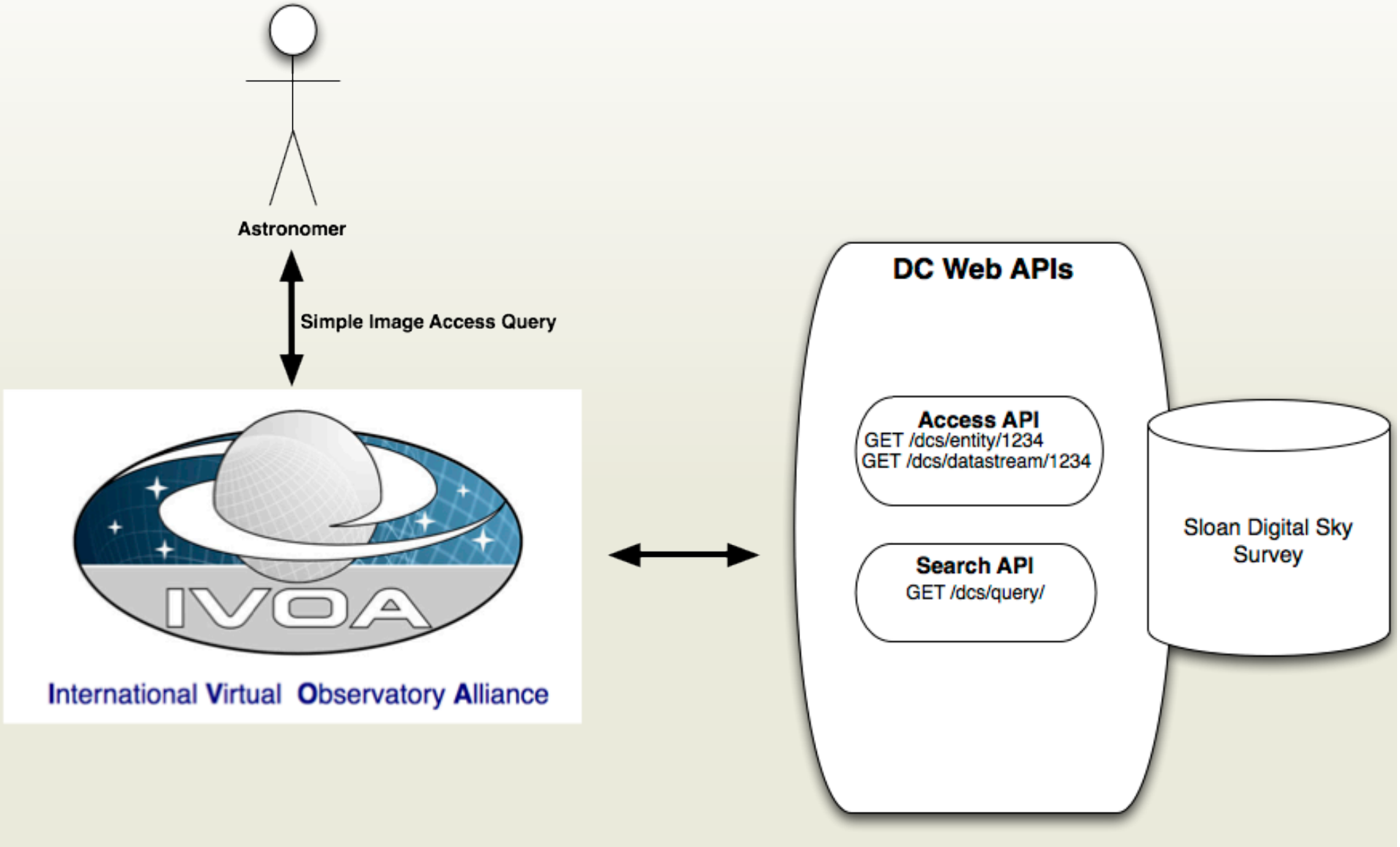
Connecting Data and Publications

- DC enables deposition of data artifacts
- DC enables preservation of data artifacts
- DC enables the linking of publications to data
- Seeing interest from other publishers

Science Frameworks

- IVOA focuses on the development of astronomy standards
- Encourages their implementation
- Working groups design community standards
- Does not provide preservation

Science Frameworks



IVOA Demo

Perform a Simple Image Access Search

- Success.
- 10 results found.

SIA Endpoint: http://dcivoa.dataconservancy.org:8080/voservice/siap/imageQuery

Right Ascension: 326.81

Declination: 13.03

Search Radius: 0.1

Execute Simple Image Access Search

Search Results

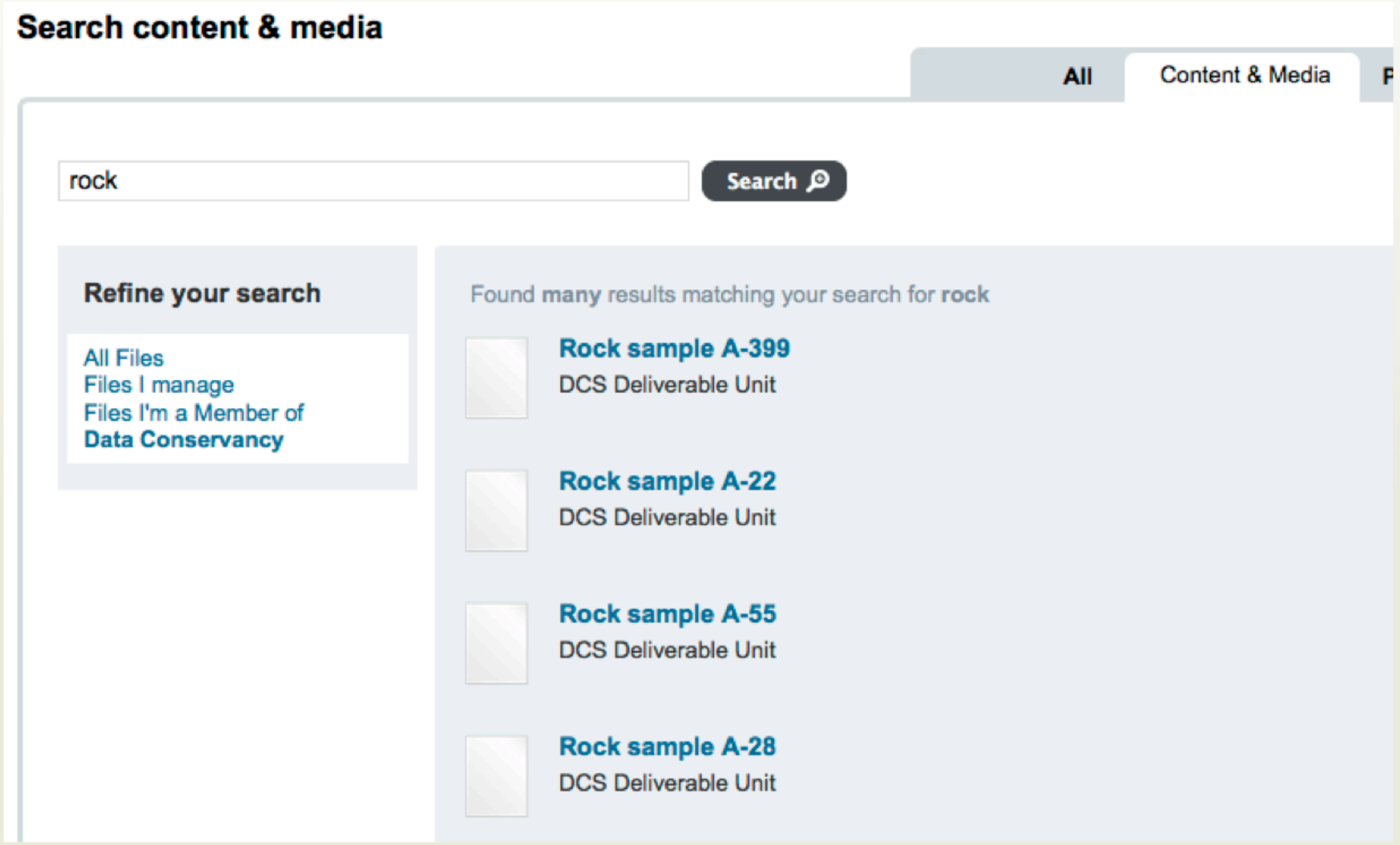
As VOTable XML

Image Title	Declination	Right Ascension	Number of axes	Axes length	Image scale	Image format	Image URL	Instrument ID	Mean date of observation	STC Coordinate Reference Frame	WCS Equinox	WCS Projection	WCS Reference Pixel	WCS Reference Value	WCS CD Matrix	Image processing flag(s)
76 N	13.03011	326.81337	2	1489,2048		image/fits	Datastream 34579			FK5	2000.0	TAN	1024.5,744.5	326.77716655,13.05558741	-1.7622398147469E-5,1.08548704315015E-4,1.08555161132813E-4,1.7619354838709E-5	C
76 N	13.02998	326.81279	2	1489,2048		image/fits	Datastream 75111			FK5	2000.0	TAN	1024.5,744.5	326.77660289,13.0554424	-1.7585792173515E-5,1.08530784181175E-4,1.08522104492187E-4,1.76099327956994E-5	C
76 N	13.03011	326.81337	2	1489,2048		image/fits	Datastream 34579			FK5	2000.0	TAN	1024.5,744.5	326.77716655,13.05558741	-1.7622398147469E-5,1.08548704315015E-4,1.08555161132813E-4,1.7619354838709E-5	C
76 N	13.02998	326.81279	2	1489,2048		image/fits	Datastream 75111			FK5	2000.0	TAN	1024.5,744.5	326.77660289,13.0554424	-1.7585792173515E-5,1.08530784181175E-4,1.08522104492187E-4,1.76099327956994E-5	C
76 N	13.03013	326.81382	2	1489,2048		image/fits	Datastream 59354			FK5	2000.0	TAN	1024.5,744.5	326.77762344,13.05562637	-1.7599173658773E-5,1.08539822946639E-4,1.0860904296875E-4,1.76087096774201E-5	C

Science Frameworks

- DC enables preservation of SDSS data artifacts
- Reinforces idea that preservation is re-use
- DC integration into existing services framework

Integration with Sakai



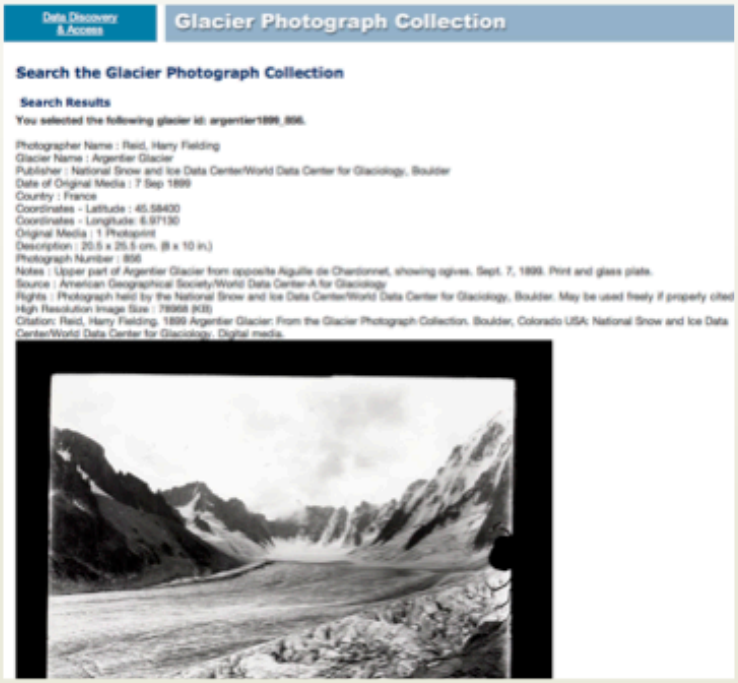
Teaching and Learning

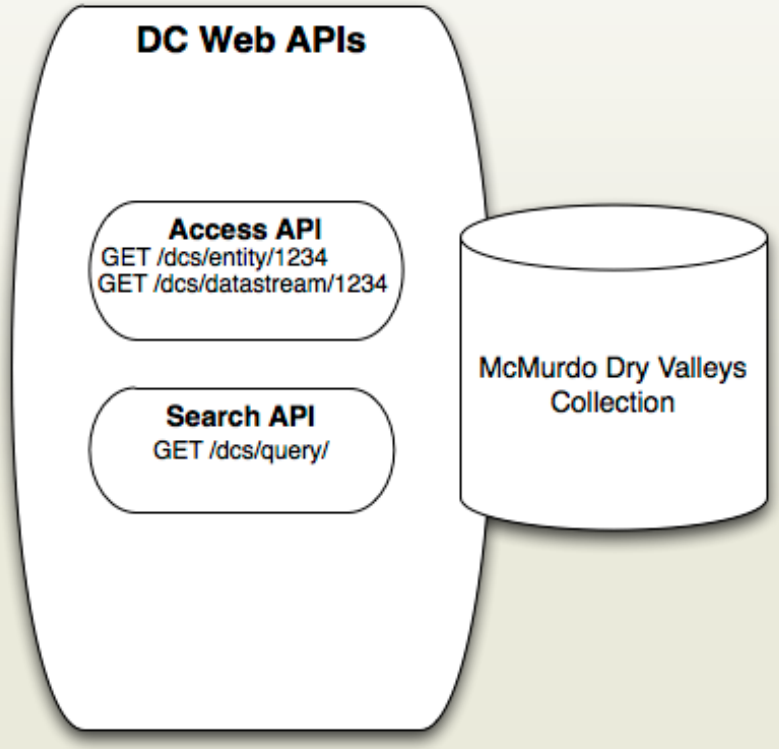
- DC integration into existing CLE software platform
- Data in repositories are now exposed to the CLE
- Reinforces idea that preservation is re-use

NSIDC Glacier Photo Collection

- Currently contains more than 13,000 geo-spatially referenced photos
- Types of users
  - scientists
  - educators and students
  - charitable groups
  - governmental institutions
  - museums
- One of the most accessed data sets at NSIDC
- Approx. 12,000 users, in 73 countries in 2010

NSIDC Glacier Photo Collection





Glacier Photograph Collection

Data Discovery & Access

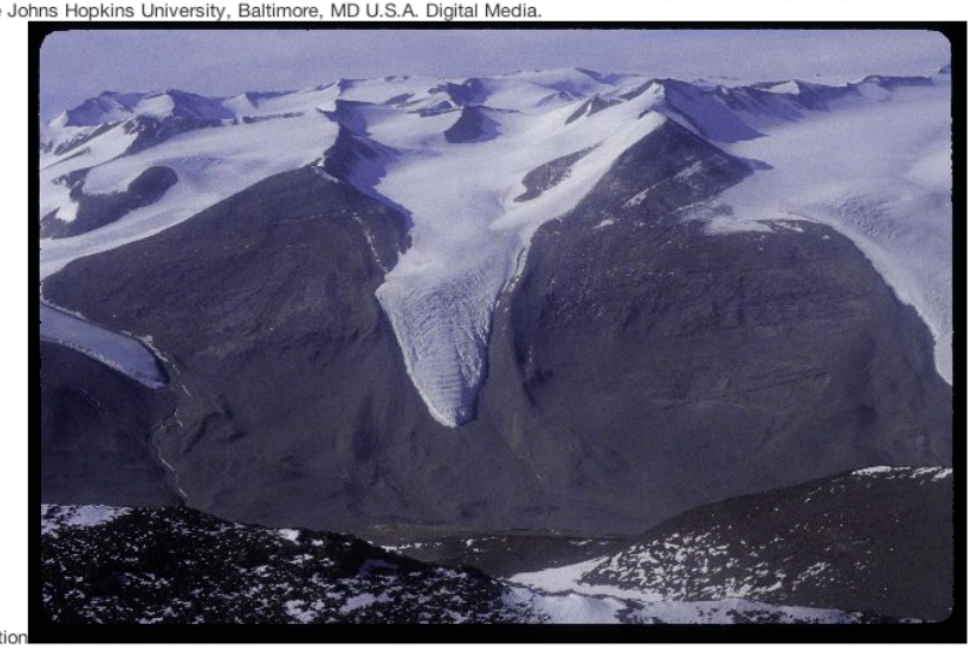
Glacier Photograph Collection

Search the Glacier Photograph Collection

Search Results

You selected the following glacier id: McDV_FieldPhoto_2000_00868.

Photographer Name : Marsh, Bruce D.
Glacier Name : Hart Glacier
Publisher : The Data Conservancy, Sheridan Libraries, The Johns Hopkins University, Baltimore, MD U.S.A.
Date of Original Media : Jan 2000
Country : Antarctica
Coordinates - Latitude : 162.24460
Coordinates - Longitude: -77.46860
Other glaciers in image: Goodspeed Glacier; Meserve Glacier
Notes : Looking southeast across Wright Valley from Olympus Range; Goodspeed Glacier (left), Hart Glacier (center), and Meserve Glacier (right).
Source : The Data Conservancy
Rights : Copyright Bruce D. Marsh, licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license.
http://creativecommons.org/licenses/by-nc-sa/3.0/
High Resolution Image Size : 120723 (KB)
Citation: Marsh, Bruce D.. 2000 McDV_FieldPhoto_2000_00868: From the Data Conservancy McMurdo Dry Valleys Field Photo Collection, Sheridan Libraries, The Johns Hopkins University, Baltimore, MD U.S.A. Digital Media.



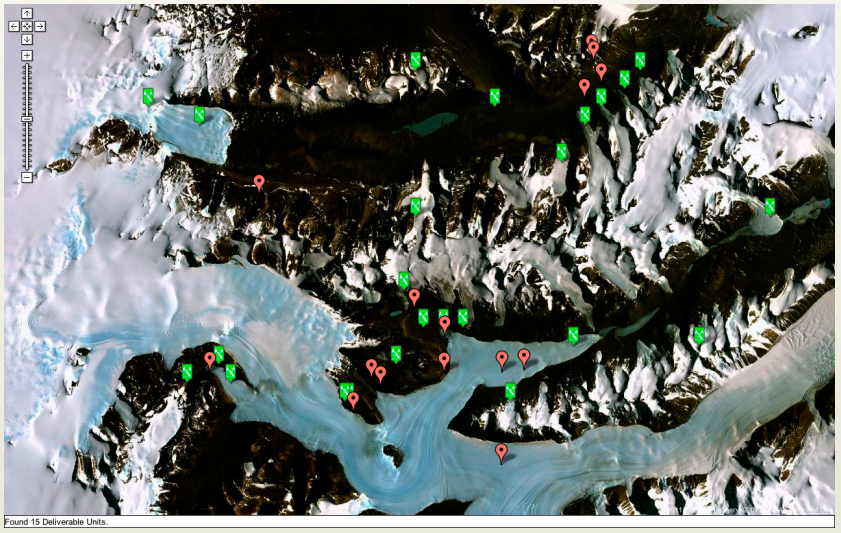
Finished Citation

NSIDC Glacier Photo Collection

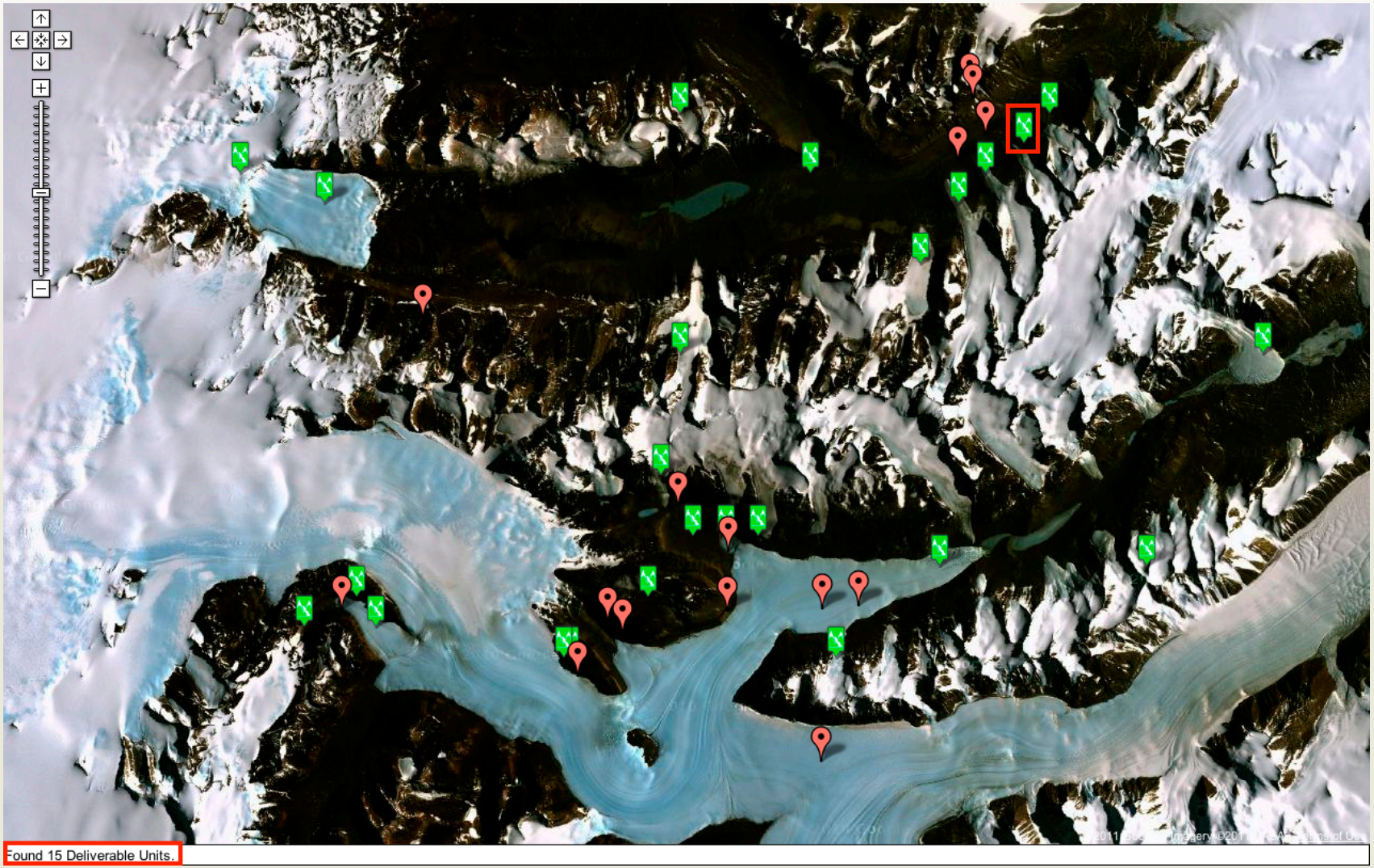
- DC enables simple, cross-discipline re-use of data
  - Glaciologist and Earth Scientist
- Enabled by
  - FDGC metadata profiles
  - DC Search and Access API

Dry Valleys Visualization

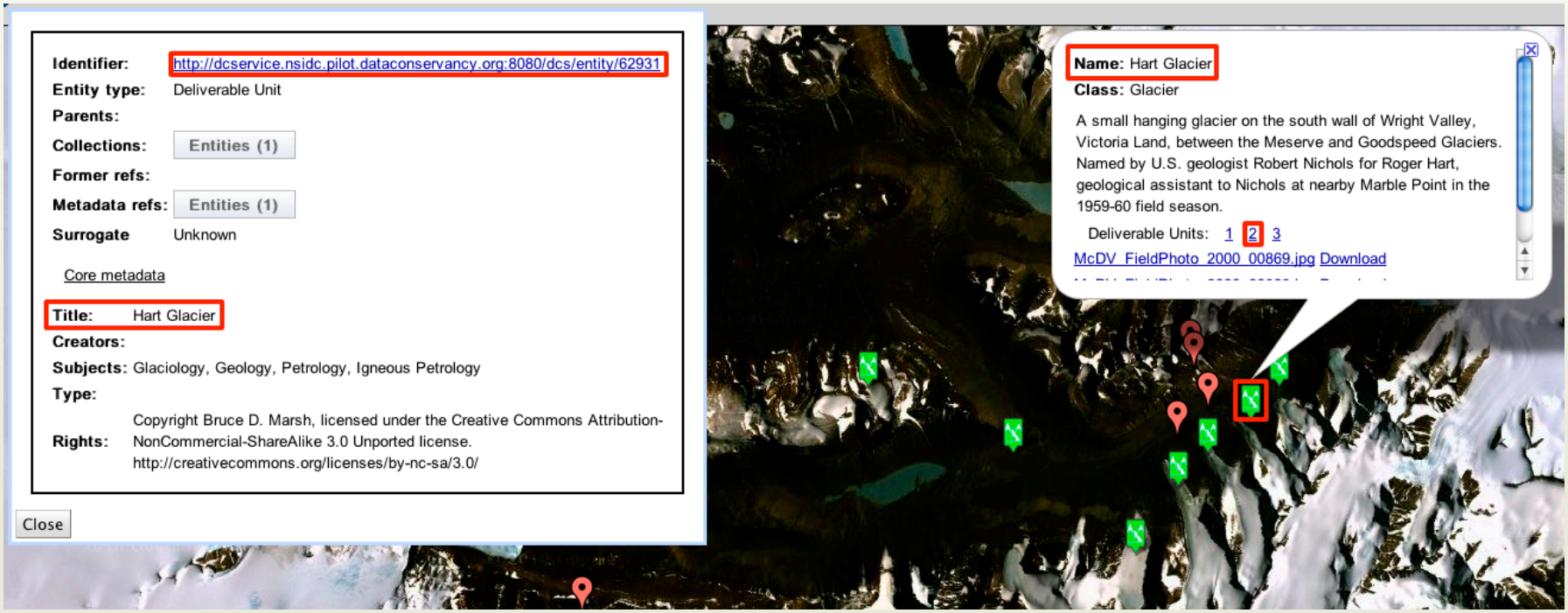
- Inspired by proof of concept
- Uses NSIDC pilot DC instance
- Exercises DC Search and Access API



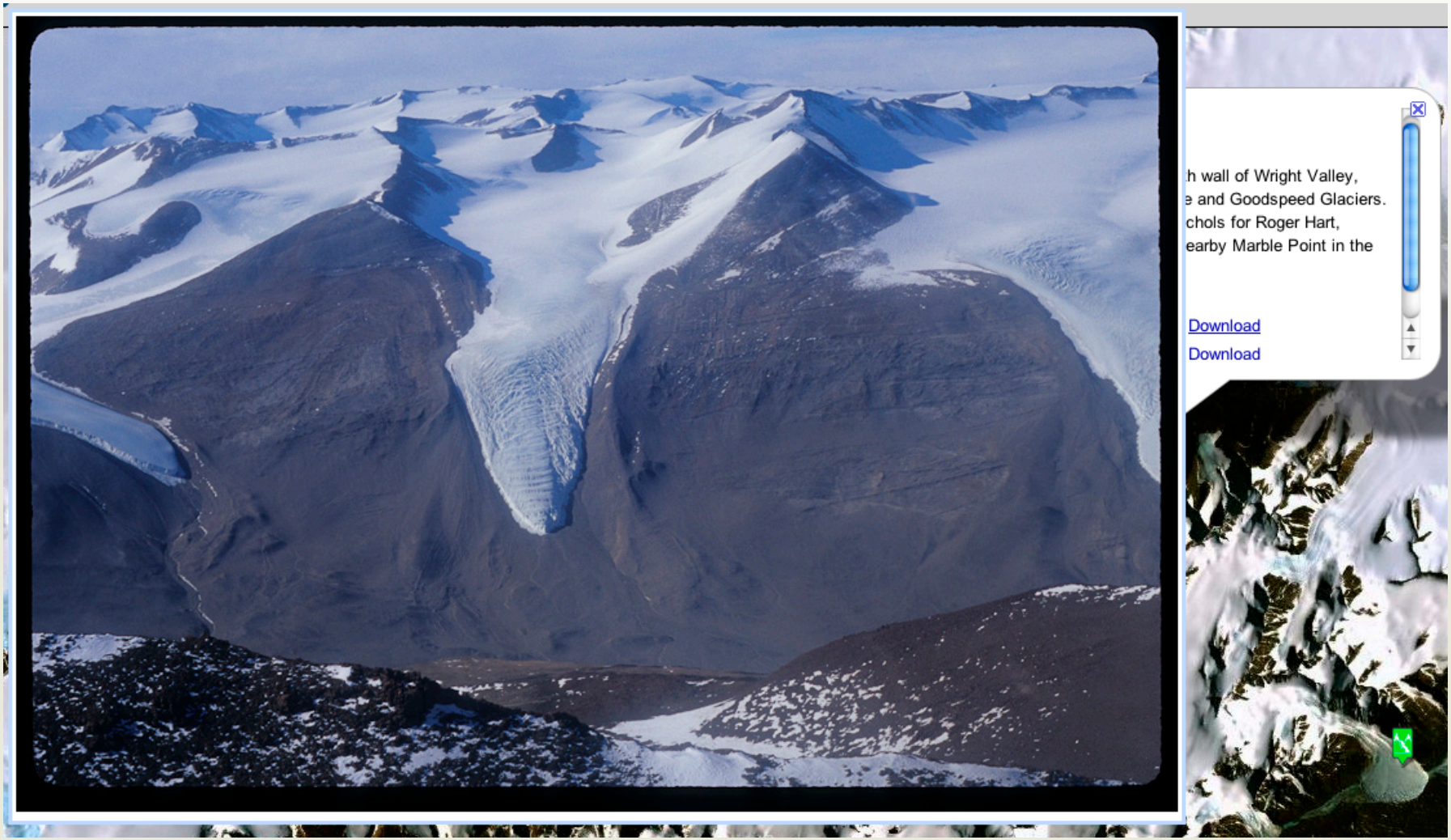
Dry Valleys Visualization (1)



Dry Valleys Visualization (2)



Dry Valleys Visualization (3)



Coastal Bays Visualization

- Illustrates tool re-use
- Rapid prototyping
  ~3 days of work
- Uses development DC instance
- Exercises DC Search and Access API

Coastal Bays Visualization (1)

Coastal Bays Visualization (2)

Coastal Bays Visualization (3)

Summary

- Prototyping of cross-dataset queries
- Proves the utility of prototyping processes
- Demonstration access and re-use of data in a preservation environment
- Integration with
  - existing scientific frameworks
  - teaching and learning environments
  - existing cyberinfrastructure
  - publishing systems
- Simple interoperability of data between disciplines
- Proves utility of DCS APIs and services

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."

Financial Sustainability (1)

- Four capstone projects with Johns Hopkins Carey Business School focused on:
  - Analysis of current and potential future storage systems
  - Total cost of ownership for storage systems
  - Business continuity and disaster recovery
  - Initial market and competitive analysis
- Formal agreements such as SDSS MOU and NSIDC and arXiv.org pilots

Financial Sustainability (2)

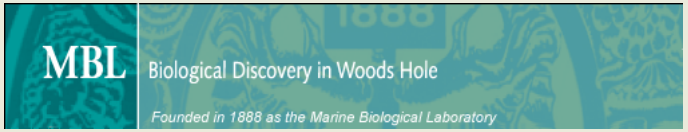
- Development of first version of DC services stack

- First version of sustainability plan including:
  - Explore potential models for providing sustainable services
  - Developing DCS services and models
  - Document findings from first two elements for wider review and distribution

Financial Sustainability (3)

- JHU as "laboratory" for testing and implementing results from financial sustainability analysis and planning
- Worked with JHU central administration to gather relevant information, engage NSF PIs and offer initial support
- Resulted in major budget request to JHU central administration for fiscal year 2012

Data Conservancy Award Partners



Acknowledgements and Resources

- National Science Foundation Award OCI-0830976

- http://dataconservancy.org
- http://twitter.com/dataconservancy
- http://www.facebook.com/pages/Data-Conservancy/128786170522661