DataStaR: Science Metadata Schemas Meet the Semantic Web

Brian Caruso
Brian Lowe
Gail Steinhart

Metadata Working Group Forum 12/09

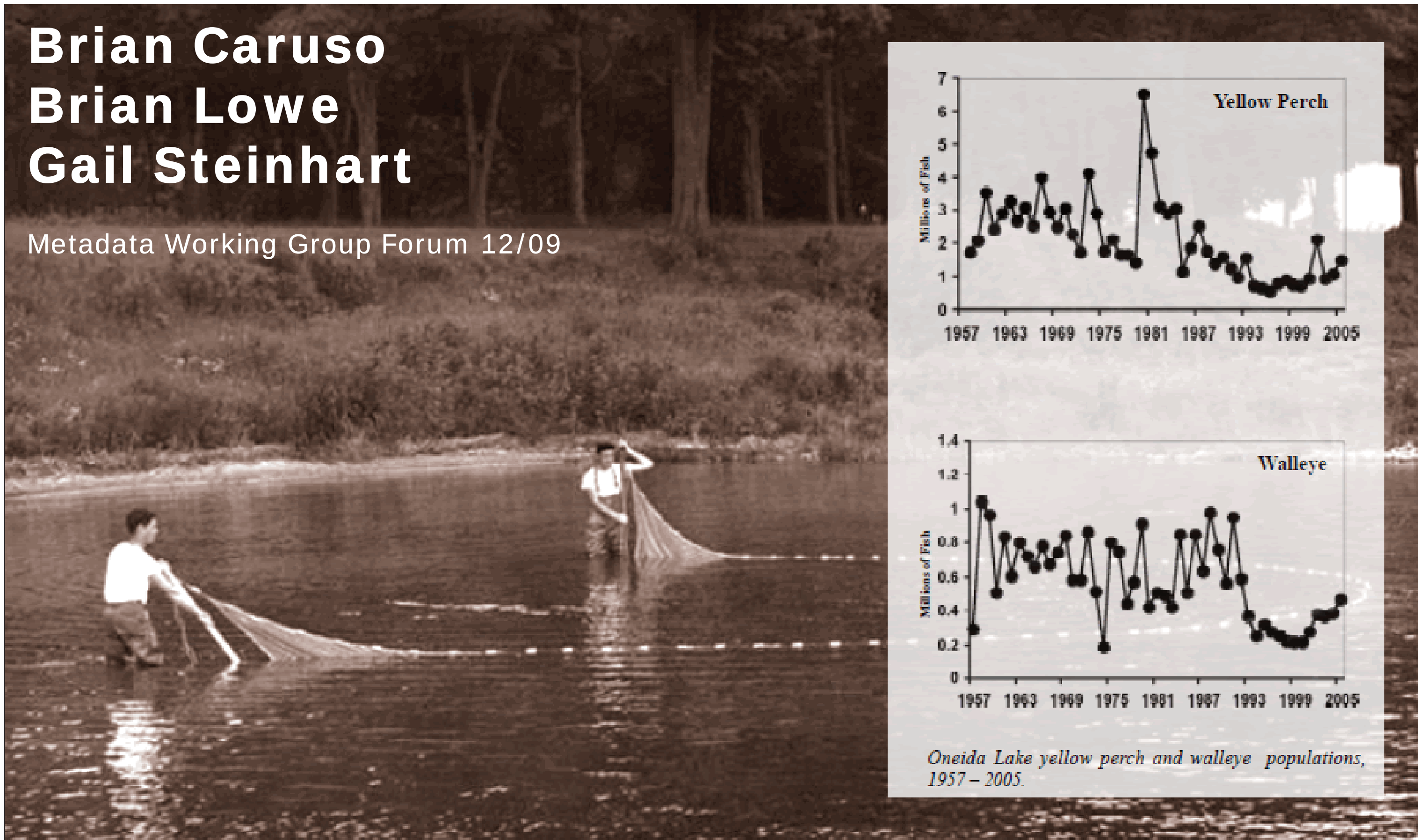
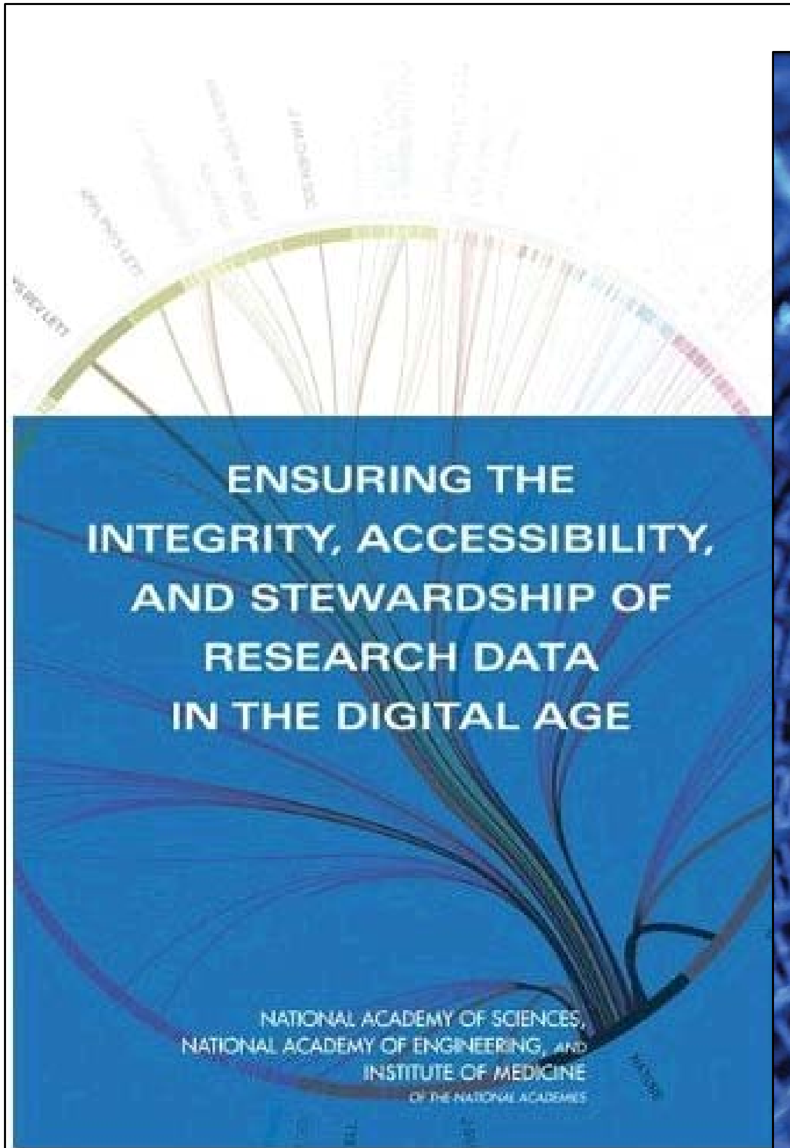


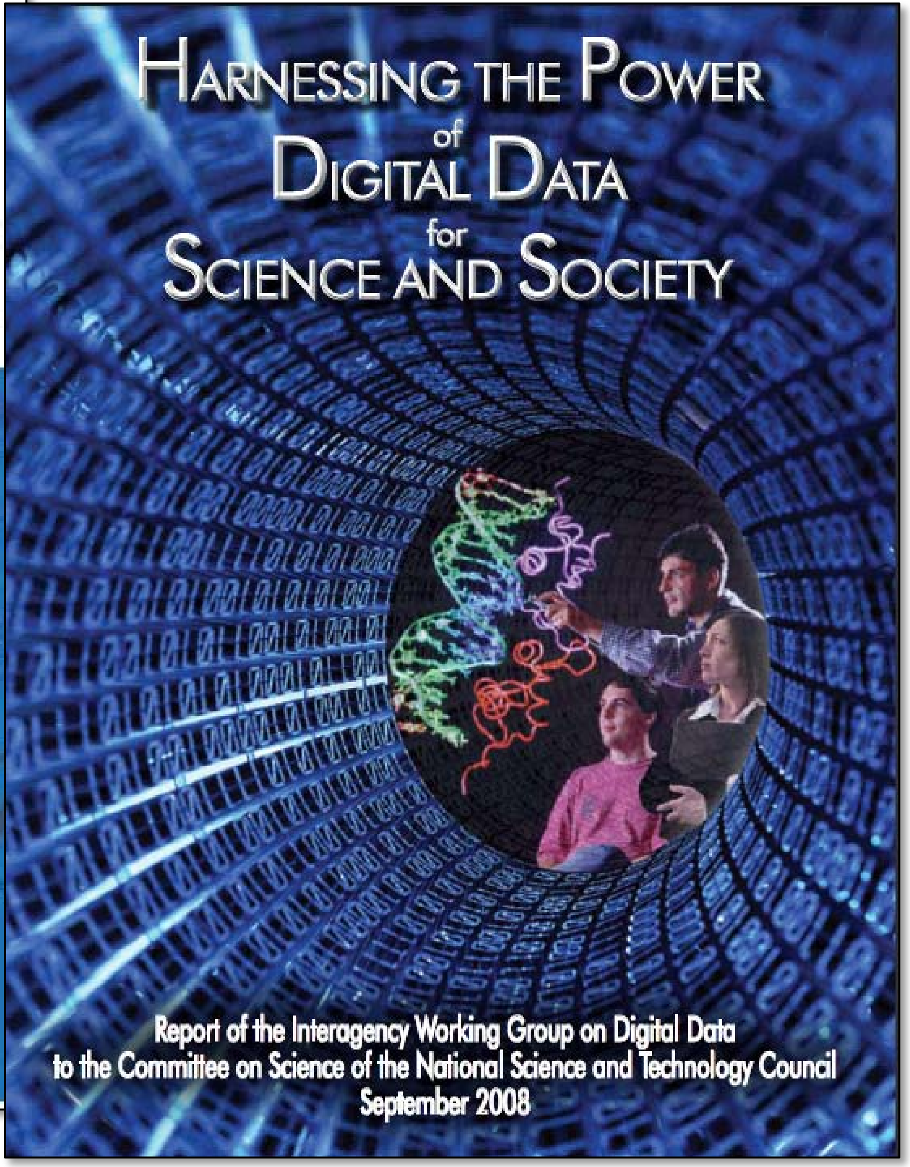
Image courtesy of the Cornell Biological Field Station

Outline

- Introduction – Gail
- System design – Brian C.
- Metadata management – Brian L.



http://www.nap.edu/catalog.php?record_id=12615



http://www.nitrd.gov/about/harnessing_power_web.pdf

Cornell University Library

Login

naturenews

nature news home | news archive | specials | opinion | features | news blog | events blog | nature journal

Specials

See all specials

Data Sharing

Sharing data is good. But sharing your own data? That can get complicated. As two research communities who held meetings in May on the issue report their proposals to promote data sharing in biology, a special issue of *Nature* examines the cultural and technical hurdles that can get in the way of good intentions.

- EDITORIAL
- FEATURE
- OPINION
- ELSEWHERE IN NATURE

Editorial

Data's shameful neglect
Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly.
9 September 2009

Feature

Data sharing: Empty archives
Most researchers agree that open access to data is the scientific ideal, so what is stopping it happening? Bryn Nelson investigates why many researchers choose not to share.
9 September 2009

Opinion

Prepublication data sharing
Rapid release of prepublication data has served the field of genomics well. Attendees at a workshop in Toronto recommend extending the practice to other biological data sets.
9 September 2009

most recent | commented

- Sex chromosomes linked to evolution of new species
  27 September 2009
- Physicists shrink X-ray source
  27 September 2009
- Mountains may be cradles of evolution
  25 September 2009
- Melting memory chips in mass production
  25 September 2009
- Oldest feathered dinosaur found
  25 September 2009

Naturejobs

Early Career Visitors
The Mathematical Biosciences Institute
Ohio, USA

Junior or Senior Faculty Position in Climate Science
Yale University
New Haven, CT

More science jobs

Post a job for free

Key:

content requires subscription or payment

http://www.nature.com/news/specials/datasharing/index.html

“The university research libraries themselves are obvious candidates to assume this role.”

http://www.nature.com/news/specials/datasharing/index.html

The Saga Of the Lost Space Tapes

NASA Is Stumped in Search For Videos of 1969 Moonwalk

By Marc Kaufman
Washington Post Staff Writer
Wednesday, January 31, 2007

As Neil Armstrong prepared to take his "one small step" onto the moon in July 1969, a specially hardened video camera tucked into the lander's door clicked on to capture that first human contact with the lunar surface. The ghostly images of the astronaut's boot touching the soil record what may be the most iconic moment in NASA history, and a major milestone for mankind.

Millions of television viewers around the world saw those fuzzy, moving images and were amazed, even mesmerized. What they didn't know was that the Apollo 11 camera had actually sent back video far crisper and more dramatic -- spectacular images that, remarkably, only a handful of people have ever seen.

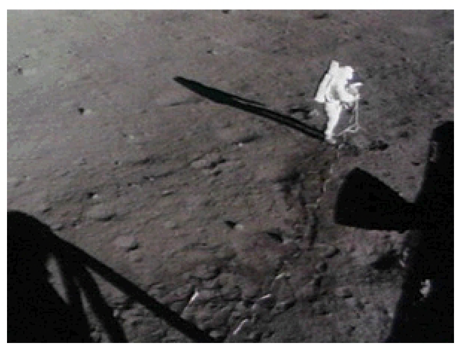
NASA engineers who did view them knew what the public was missing, but the relatively poor picture quality of the broadcast images never became an issue because the landing was such a triumph. The original, high-quality lunar tapes were soon stored and forgotten.

http://www.washingtonpost.com/wp-dyn/content/article/2007/01/30/AR2007013002065.html

Houston, We Erased The Apollo 11 Tapes

by NELL GREENFIELDBOYCE

Listen to the Story
Morning Edition



Enlarge NASA

After the camera recorded the astronauts' descent onto the moon's surface, they placed it on the moon to record their other activities.

slipped throug

http://www.npr.org/templates/story/story.php?storyId=106637066

http://www.wired.com/wired/archive/15.01/nasa.html

One Giant Screwup for Mankind

NASA put a man on the moon - then lost the videotape. A grizzled crew of ex-rocket jockeys are on a star-crossed mission to find it.

By David Kushner Page 1 of 3 next »

PHOTO GALLERY

Click above thumbnails for online extras; including additional images and video of the first lunar landing.

APOLLO MISSION PANORAMAS

***Apollo 11* Fullscreen QuickTime VR**

Launched: July 16, 1969 9:32 a.m. EDT
Landed on moon: July 20, 1969 4:17 p.m. EDT
Neil A. Armstrong, commander
Michael Collins, command module pilot
Edwin E. Aldrin Jr., lunar module pilot
Apollo 11 NASA Journal

WHEN THE EAGLE LUNAR MODULE TOUCHED DOWN ON JULY 20, 1969, all eyes were on astronaut Neil Armstrong. But Stan Lebar's ass was on the line.

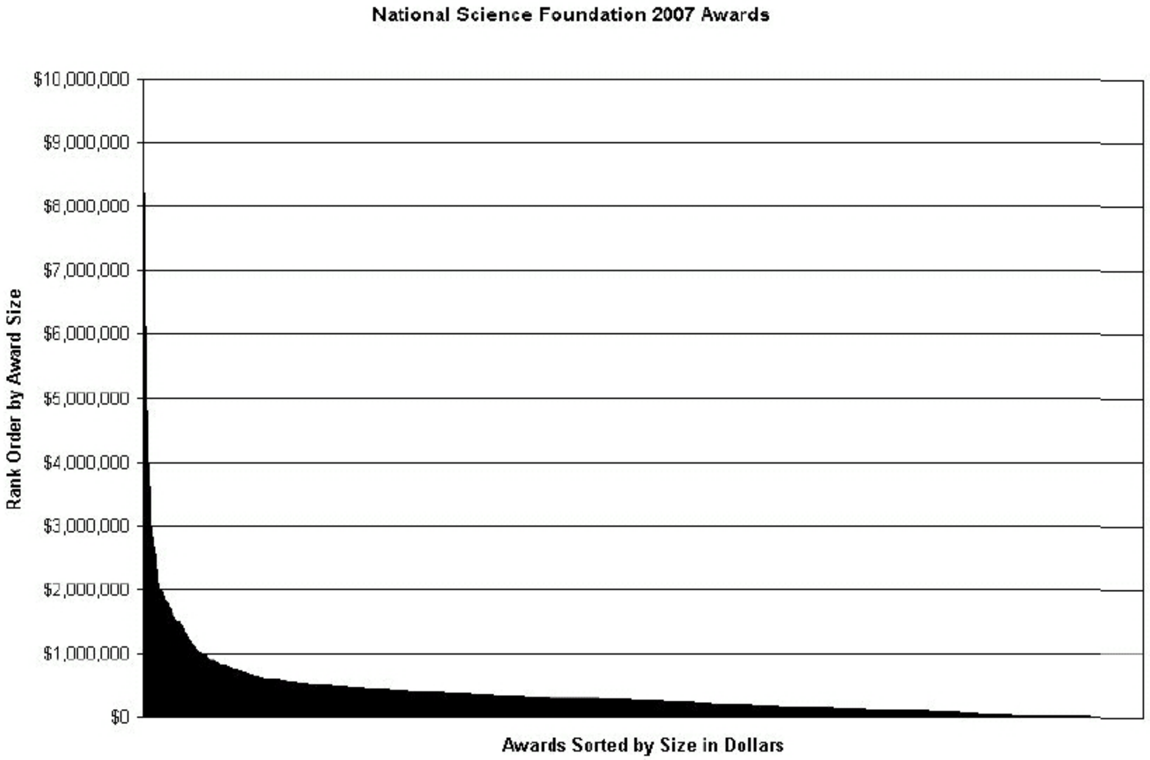
A young electrical engineer at Westinghouse, Lebar had been tasked with developing a camera that could capture the most memorable moment of the 20th century – the *Apollo 11* moon landing. The goal of the mission wasn't merely to get a man on the moon. It was to send back a live television feed so that everyone could see it – particularly the Soviets, who had initiated the space race in 1957 by launching Sputnik. If the feed failed, Lebar, the designated spokesperson for the video setup, would turn the camera on himself at Mission Control in Houston and apologize to more than half a billion TV viewers. "It was my responsibility," he says. "I'd have to stand up and take the hit."

"The long tail of dark data"

- **In 2007, NSF awarded ~12000 grants >$500, worth a total of $2,865,388,605**

- **80% between $579-$300,000**

- **That 80% was worth $1,117,431,154, or about 40% of the funds NSF awarded**



Heidorn, P.B. 2009. Shedding light on the dark data in the long tail of science. Library Trends 57(2): 280-299.

DataStaR: A Data Staging Repository

The purpose of DataStaR is to support collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centers, and/or to Cornell's own digital repository.

datastar.mannlib.cornell.edu

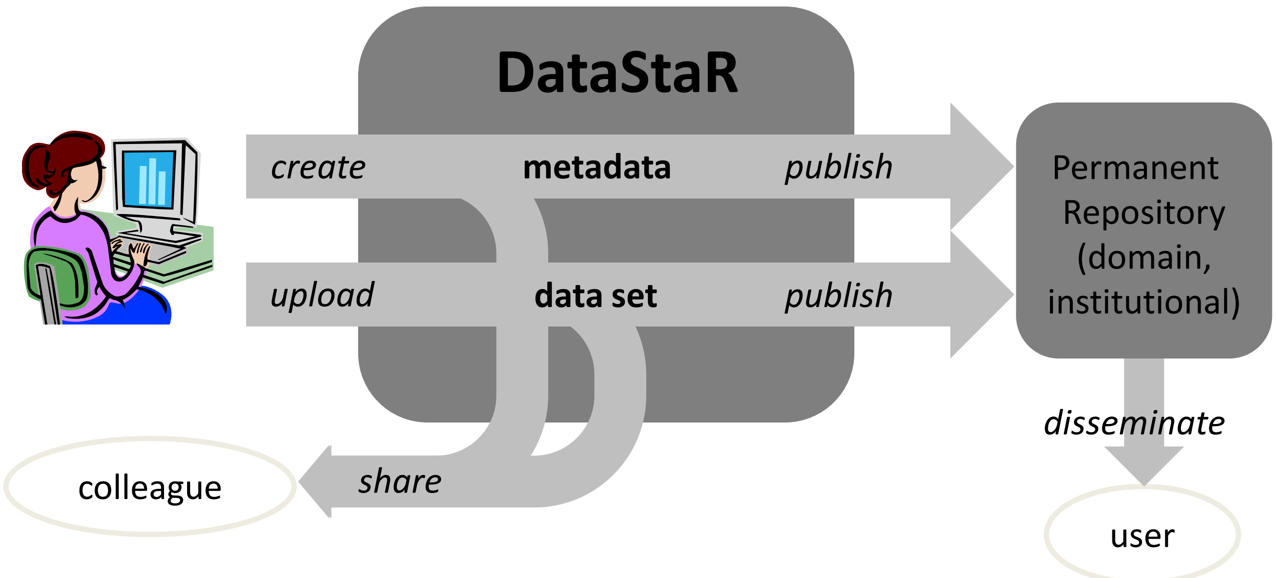
Common needs:

- I need a place to share (large) data files with colleagues.
- I want to make a data set related to a publication available online.

Common questions:

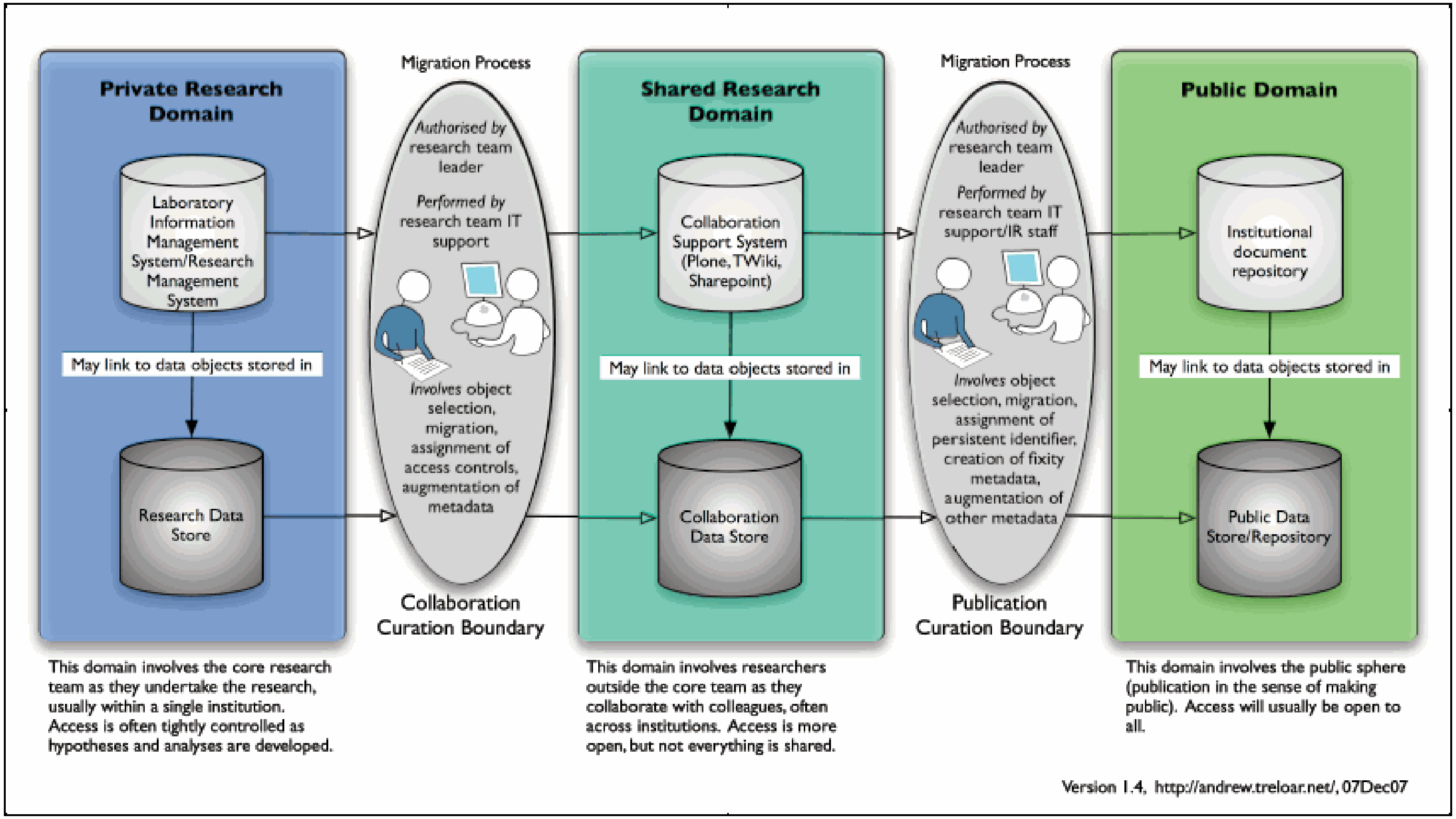
- Which data should I archive?
- How should data be formatted?
- Can I get people to ask permission to use my data?

What exactly is a data staging repository?



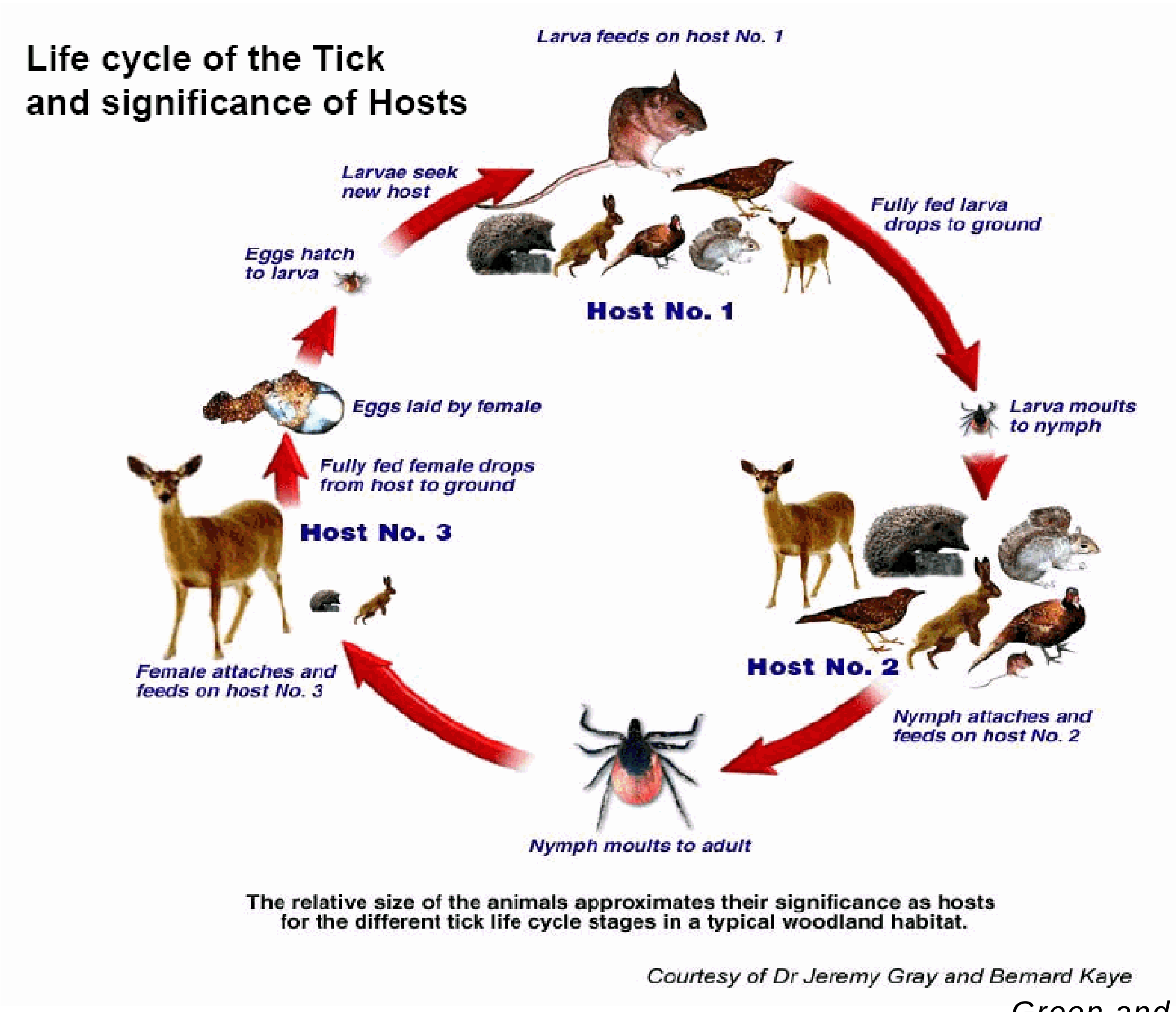
Private > shared > public domains

Figure 1: Domains, Data Stores and Curation Boundaries



Treloar and Harboe-Ree, 2008

Moving data between repositories



Green and Gutmann, 2006

Partners

- Upper Susquehanna River Basin Agricultural Ecology Program
- Cornell Biological Field Station
- Cornell Plantations Natural Areas Program
- Cayuga Lake Watershed Network
- Submission mechanism for CUGIR
- Virtual Center for Language Acquisition
- Individual researchers

Repositories and metadata

Repository	Metadata requirements
Knowledge Network for Biocomplexity (KNB)	Ecological Metadata Language (EML)
Cornell University Geospatial Information Repository (CUGIR)	Content Standard for Digital Geospatial Metadata (FGDC-CSDGM)
eCommons (Cornell's IR)	DSpace metadata
Virtual Center for Language Acquisition (VCLA)	Open Language Archives Community (OLAC)

Current status

Cornell University Library

DataStaR System Design

Brian Caruso
bdc34@cornell.edu

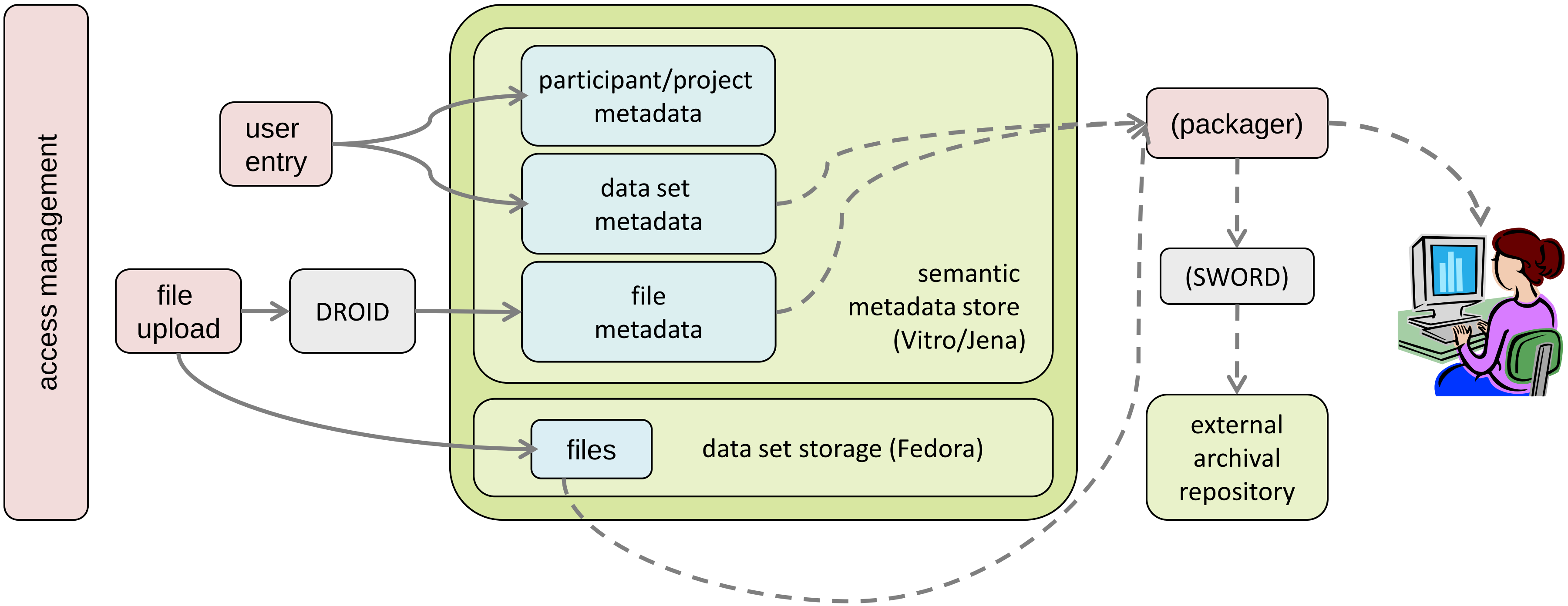
Presentation for
The Cornell University Library
Metadata Working Group
December 2009

DataStaR System Design Overview

Based on Vitro http://vitro.mannlib.cornell.edu

customizations for file management
customizations for handling metadata
customizations for access controls

Architecture



Based on Vitro

Pros

- Flexible data model
- Flexible access controls
- Already existed

Cons

- In house software
- Not designed with hooks for extension

Vitro

Vitro is a web application with a flexible data model based on java, JENA RDF library, MySQL, JSPs and Tomcat.

How Vitro was extended to create DataStaR

Minor changes

Static custom forms
Access policy

Major changes

Generate XML from RDF
Generate ontology from XML schema
Dynamic custom forms from ontology
File upload and download
Fedora integration
Modifications to support privacy

DROID and PRONOM

When uploading a file the client browser sets a CONTENT-TYPE header as part of the POST.

There is no reason to trust the that CONTENT-TYPE is set correctly. It is usually based on the file extension.

DROID examines a file's content to provide a good guess at the format of the file. It provides MIME type and PRONOM PUID.

PRONOM is the database of file formats used by the DROID software.

Use of Fedora by DataStaR

DataStaR uses Fedora as a file repository.

- Not using Fedora for searching
- Not using Fedora's RELS-EXT
- Not mirroring RDF from DataStaR in Fedora
- Not using Fedora to index RDF

This is not an exemplary use of Fedora which is unfortunate since we have experience with RDF.

DataStaR and Fedora objects

One file in DataStaR is a digital object in Fedora with

a DC XML data stream for basic file metadata and
a data stream for the file data.

DataStaR, Fedora and Identifiers

The Fedora PID is stored in the DataStaR RDF.

The DataStaR URI is stored in the Fedora object's DC.

Reason: DataStaR is intended to use dereferenceable URIs. Fedora uses the “info:fedora/” namespace.

DataStaR, Fedora and changes to files

When a file is updated in DataStaR a new digital object is created in Fedora.

Which file is a previous version of another is stored in the DataStaR RDF model, not using Fedora data stream revisions.

Reason: File name and PRONOM type may change and are stored in the DC XML of a Fedora object and that is one per an object.

Better Integration of Fedora and DataStaR

Why not mirror the RDF in DataStaR in Fedora?

Fedora places restrictions on what RDF statements can go in an object's RELS-EXT. We did not have the resources to explore this.

Learning to Use Fedora

FedoraClient and FedoraAPIM classes from FEDORA client JARs.

Unit test are an excellent resource
example: in Fedora 3.0 see the file file
/src/test/junit/test/api/testAPIM.java

We have more to learn.

Downloading datasets

Datasets are comprised of multiple files
They must be download as a group
DataStaR provides a zip of a dataset for download

Used ZipOutputStream

Access Control

Access levels are associated with a data set

- no public access
- public access to metadata only
- public access to metadata and files

Additional group based access control with similar levels.

Lesions from Building DataStaR on top of Vitro

Vitro was not designed to be as extensible as a project like DataStaR requires.

Familiarity with Vitro allowed us to overcome this.

Difficult to asses what work was avoided by reusing Vitro compared to other approaches.

How to Compare Approaches?

Time for ground up build of new flexible data model platform with DataStaR features.

Time to integrate DataStaR features into Vitro.

Time to integrate DataStaR features into other existing platform.

We have no information about how the other approaches would have gone.

Questions?

Cornell University
Albert R. Mann Library

DataStar

Metadata Management

Brian Lowe
Semantic Applications Programmer
Albert R. Mann Library
Cornell University

DataStaR: A Data Staging Repository

- Data sets themselves do not remain in DataStaR for long-term storage

- Metadata remains in DataStaR so that it can serve as a discovery tool and pointer to data repositories

Metadata Management in DataStaR

Goals

•Support multiple metadata schemas for various disciplines

•Enter basic metadata in a consistent way

•Avoid unnecessary repetition when editing. Describe a observation method once; refer to it from related dataset descriptions

•Promote thinking of metadata less as a “record” or a big form full of field, and more as part of a larger network of relationships.

Metadata Management in DataStaR

With Semantic Web technologies, we can:

•Use RDF (Resource Description Framework) as a convenient way of representing different types of metadata

•Use referenced resources named with URIs as a standard way of reusing metadata

•Use standard Semantic Web languages and tools for reasoning to make logic portable to other systems.

•Build a metadata repository accessible through standard query protocol (SPARQL) or Linked Data

Metadata Management in DataStaR

Some assumptions

•Metadata will increasingly be expressed using Semantic Web technologies, with a greater emphasis on ontology semantics.

•Metadata records conforming to syntactic schemas (especially in XML Schema) will continue to be important and widely used.

Metadata Management in DataStaR

The vision

- Where a scientific domain has established ontologies defining semantic metadata standards, they should be readily incorporated in DataStaR.

- Where a desired metadata standard is available only as a syntactic XML schema, a librarian or metadata specialist should be able to convert it to ontology form and use it with minimal effort (low upfront investment).

- Add additional rules for reasoning or mappings to additional ontologies where desirable

DataStaR

Metadata Management in DataStaR

Core metadata ontology

Extends SWRC (Semantic Web for Research Communities) and a DL-ified version of Dublin Core

Includes data set properties such as:

- title
- abstract
- owner
- contact
- metadata provider
- relationship to research group
- temporal and geographic coverage
- file properties (type, checksum, size, etc.)

First Application: Ecological Metadata

- Mann Library has partnered with researchers to describe and share ecological observation data
- EML (Ecological Metadata Language) metadata
- Knowledge Network for Biocomplexity (KNB) primary destination repository
- Cornell's DSpace installation (eCommons@Cornell) is a destination institutional repository (Dublin Core)

DataStar

Limnological summary and depth profile for six standard sampling sites on Oneida Lake, New York, 1975 - 2006 *data set*

download from eCommons | download from KNB

title

Limnological summary and depth profile for six standard sampling sites on Oneida Lake, New York, 1975 - 2006

abstract

The Cornell Biological Field Station (CBFS) serves as a primary field site for aquatic research at Cornell University (more information can be found at http://www.dnr.cornell.edu/fieldst/cbfs.htm) and is part of the Department of Natural Resources, College of Agriculture and Life Sciences. The centerpiece of the station's research program is a 50-year database on the food web of Oneida Lake, New York, that has been collected with support from the Cornell University Brown Endowment and from the New York State Department of Environmental Conservation. The data are collected by personnel from the Cornell Biological Field Station and include limnology, benthos, zooplankton, phytoplankton, and fish survey data, primarily from Oneida Lake and spanning 1957 to the present. This data package includes three tables. The first is a summary of limnological data gathered during standard sampling of Oneida Lake from 1975 - 2006. The second provides profile data by depth for temperature, dissolved oxygen, pH, and conductivity. A supplemental table provides the coordinates of the six sampling sites.

owner

Mills, Edward
Rudstam, Lars

primary contact person

Holeck, Kristen

New requirements: “Lifting” and “Lowering”

- We want to “lift” existing XML metadata documents into DataStaR
- More important, need to generate schema-compliant XML documents for submission to destination repositories
- We *don’t* want a lot of manual mapping just to lift and lower.

Ontology Axioms vs. Constraints

- OWL isn't a schema constraint language
- Open World Assumption (OWA), lack of Unique Names Assumption (UNA)
- It's attractive, however, to be able to use certain axioms as constraints in certain circumstances

OWL Restrictions vs. Schema Constraints

In an XML schema we might “require” a name element or attribute for a Person.

```
<person>
    <name>Brâncuşi, Constantin</name>
    <type>sculptor</type>
</person>
```

If the name value is missing, the document does not validate against the schema.

```
<person>
    <type>sculptor</type>
</person>
```

ERROR

OWL Restrictions vs. Schema Constraints

In an ontology, we might say something like:

`All persons have names.`

- No guarantee that we know what the name actually is.
- Maybe someone else has a document with the name.
- Maybe no one does.
- Maybe we don't care what the name is.

```
:person2234567
    a ex:Sculptor .
```

`OK – no error here`

Background: Lifting XML Schemas into OWL Ontologies

Several tools are available to do this, often employing XSLT

General approach:

- Complex types produce OWL classes.
- elements and attributes turn into object or datatype properties.
- Required types generate **`someValuesFrom`** or **`allValuesFrom`** axioms.
- constraints such as **`minOccurs`** or **`maxOccurs`** turn into cardinality axioms such as **`owl:minCardinality`** or **`owl:maxCardinality`**.

XML Schemas and OWL

- We discovered that Gloze, a tool for Jena created by Steve Battle, was a close match to DataStaR's needs.
- Available at http://sourceforge.net/projects/jena/files/
- Gloze is explicitly designed for "round tripping" between XML and RDF
- For the most part, works quite well in practice
  - Need to massage some OWL Full constructs

Lifting issues

- **Individuals as purely syntactic devices:**

```
:dataSet eml:Coverage :coverage .
:coverage eml:geographicCoverage :geoCoverage
…
:coverage eml:temporalCoverage :temporalCoverage
…
:coverage eml:taxonomicCoverage :taxonomicCoverage
```

We add direct properties and fill in the extra node later for lowering.

- **Classes that do not necessarily align well with other ontologies**

Making this all work in practice

To incorporate a metadata standard into DataStaR, we need to:

- Tweak Gloze output to keep things in OWL-DL
- Make mappings to DataStaR's core ontology
- Make editing forms
- Add extra validation queries
- Hide extra things to keep the user from being overwhelmed

Editing workflow

- Users edit properties only from core DataStaR ontology until they signal desire to submit to a repository requiring a particular metadata schema
- This triggers a type assertion using a class in another ontology, e.g. **`eml:DataSetType`**
- Additional properties/inferences are then available

Transforming simple to complex: SPARQL CONSTRUCT "rules"

DL-safe rules do not allow us to create "new" individuals

But we can CONSTRUCT blank nodes using SPARQL
(and then given them URIs)

```
CONSTRUCT {
  ?dataset eml:geographicCoverage  _:geoCoverage .
  _:geoCoverage eml:geographicDescription ?coverageTextStr .
} WHERE {
    datastar:geographicCoverage ?coverageTextStr .
}
```

DataStar

Transforming complex to simple: DL-safe SWRL rules

```
:dataset1212347  eml:geographicCoverage    :individual216 .
:individual216   eml:geographicDescription "Gobi desert" .
:individual216   eml:boundingCoordinates   :individual99341 .
```

versus

```
:dataset1212347 datastar:geographicCoverage "Gobi Desert" .
```

Generating editing forms

Editing forms automatically generated from ontology axioms as much as possible

E.g., **`owl:someValuesFrom`** prompts for a “required value”

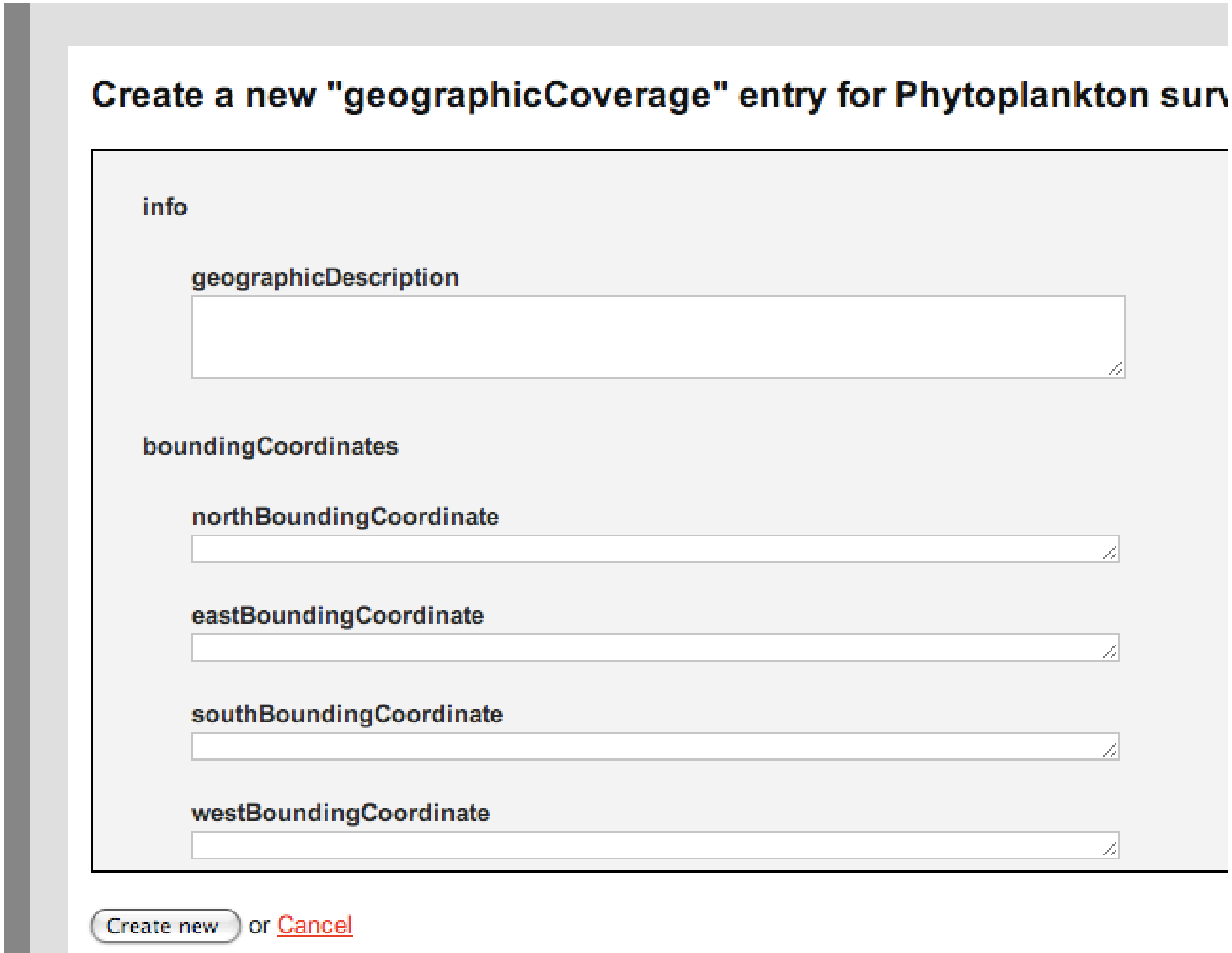
Individuals with human-readable label properties are offered as options on picklists

Additional annotation properties control ordering, hiding, and labeling

Editing system can create and edit complex subgraphs via a single HTML form

DataStar

A automatic start to a form



DataStar

What the form produces

```
:PhytoplanktonsurveyofOneidaLake
  eml:geographicCoverage :individual281180169.
 :individual281180169
     rdf:type eml-coverage:GeographicCoverage ;
     eml:geographicDescription “Standard
  phytoplankton
         sampling sites, Oneida Lake, New York,
         1975 - 2006"^^xsd:string ;
     eml:boundingCoordinates :individual762138544.
 :individual762138544
     rdf:type eml-coverage:BoundingCoordinates ;
     eml:southBoundingCoordinate "43.18083” ;
     eml:northBoundingCoordinate ” 43.22111” ;
     eml:westBoundingCoordinate ” -76.04444” ;
     eml:eastBoundingCoordinate "-75.77083” .
```

Challenges

Important consideration:

Avoiding playing games of “Where’s the assertion?”

- the problem of wanting to edit a property value that’s been inferred, when the original assertion was using a different ontology

Hiding things

Annotation properties control hiding.

- Simplify interface

- Configure how certain properties should cause others to be hidden so user can't edit the same thing in two ontologies at once,

e.g.:

```
eml:geographicCoverage
    vitro:masksProperty
        datastar:geographicCoverage .
```

Challenges: Ordering of axioms

- Gloze uses RDF reification. Can't have that in DataStaR.

- List structures for OWL have been proposed (e.g. Drummond et al., OWLED 2006) but we're not interested in reasoning on the sequence

- We create simple OWL-DL compatible sequences using intermediate reification individual (semantics understandable only by Vitro code)

- Vitro converts this to RDF reification for handoff to Gloze

DataStar

Challenges: Text Markup

- Text markup (paragraphs, emphasis, super/subscripts, etc.) is difficult to deal with and not very useful represented as an RDF graph
- EML uses a portion of the DocBook standard for text
- Currently populate only simple paragraph structures in RDF graph
- Would be preferable to store use XSLT transformations

Summary & Conclusion

- DataStaR incorporates OWL/RDF versions of metadata schemas into a web application for end-user metadata production and discovery.
- Automated lift; automated forms; hide/refine where necessary
- May not be appropriate for highly complex metadata requiring heavily customized interfaces.
- For other types of metadata, it is an effective way of bridging the syntactic and semantic worlds.
- Interoperate with established infrastructure while generating data for the Semantic Web.

Cornell University Library

Thank you.

Image courtesy of the Cornell Biological Field Station

DataStaR team:

Brian Caruso
Kathy Chiang
Jon Corson-Rikert
Dianne Dietrich
Ann Green
Janet McCue
Gail Steinhart

This material is based upon work supported by the National Science Foundation under Grant No. III-0712989. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.