Envisioning a New Distributed Organization and Cyberinfrastructure to Enable Science

Presenter Name
Stephen Abrams
Patricia Cruse
John Kunze

California Digital Library
Outline of today’s talk

- Complexities of global change
- Challenges for cyberinfrastructure and data intensive research
- A solution: **DataONE**
- An approach: curation micro-services
Scientific challenges and data needs

- Global change is a complex scientific and societal challenge
- Community needs good data
- Good data...
  - builds good science
  - makes possible wise management
  - enables sound decisions
- Good data needs...
  - solid technical infrastructure
  - sound organization
  - community engagement (you)
The complexities of global change

Critical areas in the Earth’s system

Geographical pattern of surface warming

(°C)
Human impacts on land-based ecosystems

Ecosystems and Human Well-Being

Source: Millennium Ecosystem Assessment
Human impacts on the world’s oceans

Halpern et al. 2008 Science 319.
Human population change

Figure 1. Percent Change in Resident Population for the 50 States, the District of Columbia, and Puerto Rico: 1990 to 2000

Legend:
- Madison, 26.5 to 29.5
- High Rate, 24.6 to 26.9
- Very High Rate, 21.1 to 24.5
- High Rate, 18.2 to 21.0
- Low Rate, 14.0 to 18.1
- Lower Rate, 10.2 to 13.9
- Lowest Rate, 6.1 to 10.1
- Decline, Less than 6.1

Map showing the percent change in resident population for the 50 states, the District of Columbia, and Puerto Rico between 1990 and 2000. The states are color-coded to indicate the rate of change.
Outline of today’s talk

- Complexities of global change
- Challenges for cyberinfrastructure and data intensive research
- DataONE: A solution
- An approach: curation micro-services
Data challenge 1: dispersed sources
(“finding the needle in the haystack”)

- Data are massively dispersed
  - Ecological field stations and research centers (100’s)
  - Natural history museums and biocollection facilities (100’s)
  - Agency data collections (100’s to 1000’s)
  - Individual scientists (1000’s to 10,000s to 100,000s)
Data challenge 2: diversity
“the flood of increasingly heterogeneous data”

- Data are heterogeneous
  - Syntax (format)
  - Schema (model)
  - Semantics (meaning)

Jones et al. 2007
Data challenge 3: poor practice
“data entropy”

Information Content

Time of publication
Specific details
General details
Retirement or career change
Accident
Death

Time

(Michener et al. 1997)
Data challenge 4: loss

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements

Source: S. Abrams, CDL
Data challenge 4: more loss

Petabytes Worldwide

Available Storage

Available Storage, 2007 264EB

Source: John Gantz, IDC Corporation: The Expanding Digital Universe
Cumulative impact: data longevity

<table>
<thead>
<tr>
<th>Study</th>
<th>Resource Type</th>
<th>Resource Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rumsey (2002)</td>
<td>Legal Citations</td>
<td>1.4 years</td>
</tr>
<tr>
<td>Harter and Kim (1996)</td>
<td>Scholarly Article Citations</td>
<td>1.5 years</td>
</tr>
<tr>
<td>Koehler (1999 and 2002)</td>
<td>Random Web Pages</td>
<td>2.0 years</td>
</tr>
<tr>
<td>Spinellis (2003)</td>
<td>Computer Science Citations</td>
<td>4.0 years</td>
</tr>
<tr>
<td>Markwell and Brooks (2002)</td>
<td>Biological Science Education Resources</td>
<td>4.6 years</td>
</tr>
</tbody>
</table>

Outline of today’s talk

- Complexities of global change
- Challenges for cyberinfrastructure and data intensive research
- DataONE: a solution
  - Building on existing cyberinfrastructure
  - Creating new cyberinfrastructure
  - Changing science culture and institutions
- An approach: curation micro-services
The goal of DataONE is to enable new science through universal access to data about life on earth by:

- engaging the scientist in the data preservation process
- supporting the full data life cycle,
- encouraging data stewardship and sharing
- promoting best practices
- engaging citizens

One of two DataNet awardees recommended for funding by NSF
Data types

- Biological (genes to biomes)
- Environmental
  - Atmospheric
  - Ecological
  - Hydrological
  - Oceanographic
Data sources

- Research networks and environmental observatories
- Biological specimens
- Individual Scientists
- Citizen scientists’ data
- Natural resources and conservation data
- Observational data
- Global and continental land cover/land change and biogeochemical data
Existing biological data archives

- ESA’s Ecological Archive
- Distributed Active Archive Center
- National Biological Information Infrastructure
- Fire Research & Management Exchange System
- Long Term Ecological Research Network
- Knowledge Network for Biocomplexity
Examples of data holdings

Metadata Interoperability Across Data Holdings

<table>
<thead>
<tr>
<th>Data Archive</th>
<th>Types of Data Managed</th>
<th>Metadata Standard(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Biological Information Infrastructure</td>
<td>Biodiversity, taxonomic, ecological</td>
<td>BDP, DwC, DC, OGIS</td>
</tr>
<tr>
<td>DAAC</td>
<td>Biogeochemical dynamics, terrestrial ecological Earth observation imagery</td>
<td>DIF, BDP, ECHO</td>
</tr>
<tr>
<td>The US Long Term Ecological Research Network</td>
<td>Ecological, biodiversity, biophysical, social, genomics, and taxonomic</td>
<td>EML</td>
</tr>
<tr>
<td>Avian Knowledge Network</td>
<td>Avian populations and molecular biology</td>
<td>DC</td>
</tr>
<tr>
<td>ALA Atlas of Living Australia</td>
<td>Biological and taxonomic</td>
<td>DC subset</td>
</tr>
<tr>
<td>South African Environmental Observation Network</td>
<td>Biophysical, biodiversity, disturbance, and Earth observation imagery</td>
<td>EML</td>
</tr>
<tr>
<td>TAIWAN ECOLOGICAL RESEARCH NETWORK</td>
<td>Biodiversity, biotic structure, function/process, biogeochemical, climate, and hydrologic</td>
<td>EML</td>
</tr>
</tbody>
</table>

BDP=Biological Data Profile
DC subset=Dublin Core subset
DwC=Darwin Core
OGIS=OpenGIS
DC=Dublin Core
DIF=Directory Interchange Format
ECHO=EOS ClearingHOuse
Providing one-stop shopping for data

Simple Pilot Catalog Interface
(searches entire metadata record)

A Pilot Catalog For Earth Observations

40,000 Data Set Records

DataNetONE Metadata Clearinghouse (31,864)
Long Term Ecological Research (LTER) Network (6,897)
ORNL Distributed Active Archive Center for Biogeochemical Data (810)
Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) (783)
Organization of Biological Field Stations (124)
Inter-American Institute for Global Change Research (IAI) (79)
MODIS and ASTER Products (LPDAAC) (38)
National Phenology Network (USANPN) (29)
Existing cyberinfrastructure: tools
Building new global cyberinfrastructure
New distributed framework

Coordinating Nodes
- retain complete metadata catalog
- subset of all data
- perform basic indexing
- provide network-wide services
- ensure data availability (preservation)
- provide replication services

Flexible, scalable, sustainable network

Investigator 1..N Toolkit
- MERCURY
- Kepler
- MetaCat
- Sp

Member Nodes
- diverse institutions
- serve local community
- provide resources for managing their data
DataONE Building new global cyberinfrastructure
DataONE management and partners

William Michener, University of New Mexico
Suzie Allard – University of Tennessee
Bob Cook – Oak Ridge National Laboratory DAAC
Patricia Cruse – California Digital Library
Mike Frame – USGS, National Biological Info. Infrastructure
Matt Jones – University of California Santa Barbara
Steve Kelling – Cornell Lab of Ornithology

DataONE Partners plus Kepler-CORE and SEEK/KNB Teams

We welcome your involvement!
Engaging citizens in science
Building global communities of practice and long-lived cyberinfrastructure

- Community engagement
  - Involve library and science educators
  - Engage new generations of students in best practices
  - Build on existing programs
- Involvement of cultural memory organizations brings centuries of preservation experience to datasets
Outline of today’s talk

- Complexities of global change
- Challenges for cyberinfrastructure and data intensive research
- DataONE: A solution
- An approach: curation micro-services
Data curation is hard

- Data sets encompass everything, including “regular” object types
  - Documents, images, audio, video, etc.
- Data is like software, but even more specialized
- Tension between establishing standards and fostering innovation
- Heavy processing requires a tricky long-term migration/emulation of custom data/software
- Heavy provenance and snapshot coherence requirements
- Instability: value of some preserved data depends on ongoing change, in particular, on researcher annotation
Imagining the Non-Repository

What are micro-services?

- Unbundled alternative to monolithic systems with single archival “culture”; avoiding the deadly embrace
- Low barrier, low commitment tools
  - Leverage native operating system file handling tools
- Decoupled in design
- Recoupled in deployment
  - Late binding, e.g., Unix pipes
- Creates flexible systems, mix-and-match depending on need
The wisdom of the web

- Resist urge to design user and programming interfaces without using the web’s interfaces
  - The web is the *de facto* distributed filesystem (M. Nelson)
  - Make interactions web-browser-friendly
  - … and RESTful to make them program-friendly
- “Wget” is the basic automated client, e.g., for known-item ingest and outgest
  - Very high speed obtained by multiple wget’s in parallel
Upload Comparison

Source: Rasan Rasch
The wisdom of files

- After 30 years, we’re *really* good at modern filesystems
  - Files and directories (folders) are fast, plentiful, stable, highly interoperable across platforms
  - They form an implicit standard for holding generic content
  - You can use native OS tools to create, list, change, & backup
- What’s the least work to make an “objectsystem”?  
  - Object system = File system plus minimal naming conventions
**Pairtree: hierarchy-based collection**

- *Pairtree* to hold a collection of object containers (directories)
  - Pairs of id/en/ti/fi/er characters create paths to objects
  - End of path is start of object
  - Early adopter: Hathi Trust for scanned books

You can import a pairtree and, knowing *nothing* about object purpose or structure, can reliably
-Enumerate all objects and their ids
- Produce any object by requested id
- Maintain and back it up with ordinary OS tools
- Rebuild the collection simply by walking the filesystem
Directory-based objects and object parts

- *Dflat* (digital flat) as residence for a generic digital object, with common amenities, if present, under reserved file names
- *ReDD* (reverse directory deltas) for simple file-level diffs
- *CAN* (content access node) for a repository instance
  - A Pairtree with Dflats for leaves and
  - ReDD-tinged versions
Directory typing for humans and machines

- We have lots of directory types to declare
  - ReDD versions
  - Dflat object residences
  - Pairtree roots
  - CAN instances
  - and of course Bagit bags for import/export
- *Namaste* (NAMe AS TExt) tags are *filenames* for humans
  - Example filename: “0=dflat_1.1”
  - File content has the non-lossy version for machines
A Name Value Language (ANVL) – back to basics
- An ANVL record is a sequence of elements in email header format:
  \[ \Rightarrow \text{label, colon, value} \]
- Long values are continued on indented lines
- A blank line ends a record

Based on cross-domain kernel distilled from Dublin Core
- **who** – a responsible person or party
- **what** – a name or other human-oriented identifier
- **when** – a date important in the object’s lifecycle
- **where** – a location or a machine-oriented identifier
Extended Namaste “greeting files”

- Other Namaste tags hold Dublin Core Kernel metadata, and greet a visitor who requests a directory listing with
  - 0 = one of \{bagit, redd, dflat, pairtree, can, etc.\}

```
$ ls 12/34/5
0=dflat_1.8 admin/ splash.txt
1=Twain,_Mark annotations/ v001/
2=Huckleberry.. data/ v002/
3=1898 log/ v003/
4=12345 manifest.txt
```

- (1, 2, 3, 4) = Kernel elements (who, what, when, where)
Other micro-service tools

- *BagIt* for opaque content import and export
- *Checkm* manifest format to support:
  - import, export, fixity, replication, harvesting
- *NOID* for opaque identifier minting, resolving
- *JHOVE2* for object characterization
- *XTF* for index and search
A possible data protocol: THUMP

The HTTP URL Mapping Protocol (THUMP)

- A set of URL-based conventions for retrieving information and conducting searches
- Can be used for focused retrievals or for broad database searches
- Based on commands put in the query string after ‘?’

http://example.com/?db(books)find(war and peace)show(full)
THUMP requests

The HTTP URL Mapping Protocol (THUMP)
Shortest request is a URL ending in `?', as in

http://example.foo.com/object321?

Which is shorthand for the common request:

http://example.foo.com/object321?show(brief)as(anvl/erc)

Naked `?' and `??' are designed to support the known-item query convention from the ARK persistent id scheme
THUMP responses

Responses consist of HTTP response headers, and one or more ANVL records

1  C:  [opens session]
   C:  GET http://ark.cdlib.org/ark:/13030/ft167nb0vq? HTTP/1.1
   C:  
   S:  HTTP/1.1 200 OK

5  S:  Content-Type: text/plain
   S:  THUMP-Status: 0.5 200 OK
   S:  
   S:  erc:
   S:  who:  Stanton A. Glantz and Edith D. Balbach
10 S:  what:  Tobacco War: Inside the California Battles
   S:  when:  20000510
   S:  where:  http://ark.cdlib.org/ark:/13030/ft167nb0vq
   S:  [closes session]
Broad searching in THUMP

General form of broad query

Key ? in(DB) find(QUERY) list(RANGE) show(ELEMS) as(FORMAT)

Many details to be worked out; watch for


“DataLab” project extending THUMP for tabular data integration and visualization (Nassib Nassar, RENCI)
A sample Flex application plotting data retrieved from DataLab (Kevin Gamiel)
Sample Java visualization by ZIP code

A sample Java application visualizing data retrieved from DataLab by ZIP code (Ketan Mane)
Integrating THUMP DataLab extensions

Submit a request via URL

- Web browser
- java.net.URL
- wget

Parse tab-delimited results

- Web browser
- java.lang.String.split("\t")
- R
- awk

Source: Nassib Nassar

Integrating Sarcomere and THUMP-DL with other tools and programming languages
Representing the quadratic formula

\[ x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

In LaTeX:
\[
x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}
\]

In troff/eqn:
\[
x = \{-b \pm \sqrt{b \sup 2 - 4ac}\} \over 2a
\]

In OpenOffice:
\[
x = \{-b \pm \sqrt{b^2 - 4ac}\} \over 2a
\]

Why not use XML?
\begin{align*}
\frac{x}{b \pm \sqrt{b^2 - 2ac}}
\end{align*}
Micro-services and curation in DataONE

- We will keep working to apply our micro-services approach to the problems presented by DataONE
- Much depends on community uptake of best-practices via education about early intervention as close to data producers as possible
- Our micro-services are all works-in-progress, the specifications, and some software, are summarized at
  
  http://www.cdlib.org/inside/diglib/

Micro-services eventual roster:

1. Ingest
2. Identity
3. Storage
4. Catalog
5. Fixity
6. Replication
7. Characterization
8. Description
9. Index
10. Search
11. Annotation
12. Publication

- More details in Stephen Abrams’ talk on 1pm Tuesday!
Come join us!
San Francisco
October 5-6, 2009
http://www.cdlib.org/iPres/

Contact Perry Willett for more info:
perry.willett@ucop.edu