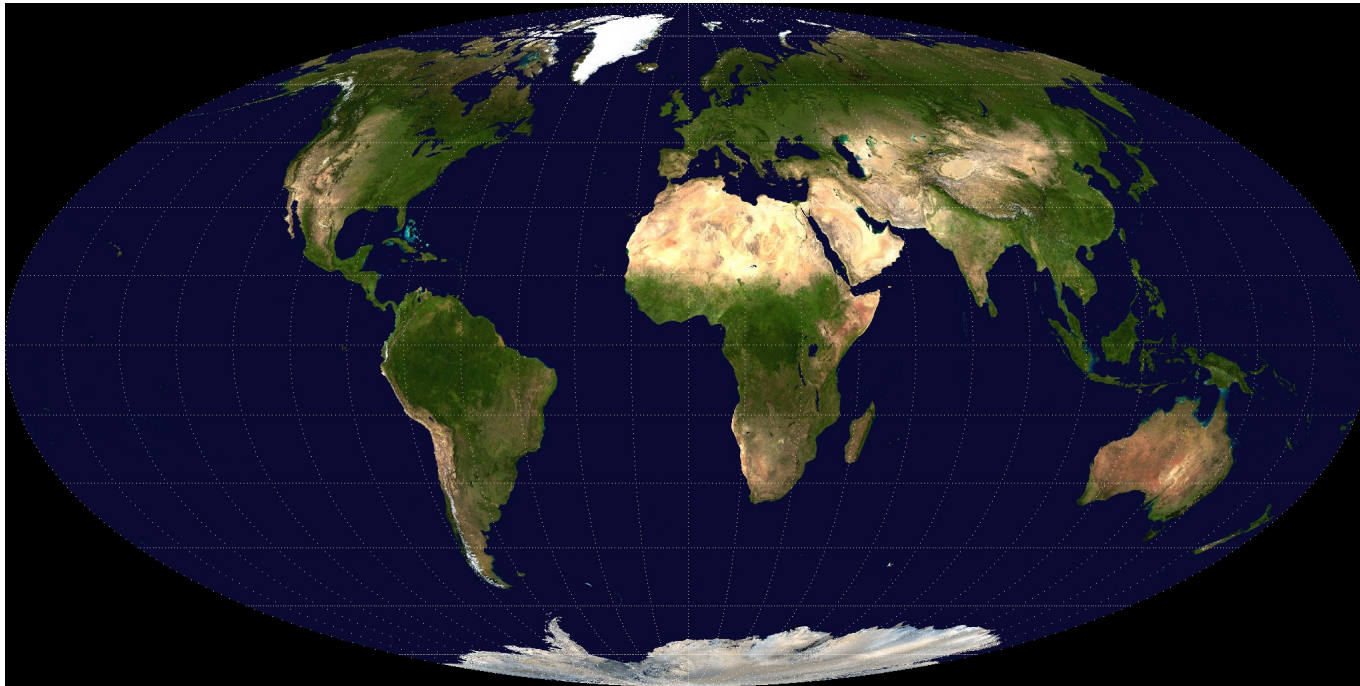




Envisioning a New Distributed Organization and Cyberinfrastructure to Enable Science



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)

Living with Cancer
The changing science

Beyond Baghdad: Where The Enemy Has Its Own Surge

The Sopranos' Last Song: What Exit Will Tony Take?

TIME

SPECIAL DOUBLE ISSUE



The Global Warming Survival Guide

51 Things You Can Do to Make a Difference

2ND-QTR SIZZLE
PROFITS AT 900 COMPANIES (P. 74)

PAYING FOR COLLEGE
BEWARE OF THOSE HIGH 529 FEES (P. 96)

TERRORISM
WHAT COMPANIES STILL NEED TO DO (P. 26)

BusinessWeek

August 16, 2004

GLOBAL WARMING

Why Business Is Taking It So Seriously

BY JOHN CAREY (P. 60)

APRIL 5, 2004

TIME

SPECIAL REPORT GLOBAL WARMING

BE WORRIED. BE VERY WORRIED.


Climate change isn't some vague future problem—it's already damaging the planet at an alarming pace. Here's how it affects you, your kids and their kids as well

EARTH AT THE TIPPING POINT
HOW IT THREATENS YOUR HEALTH

INDIA CAN HELP D—OR DESTROY IT
CRUSADERS




Adapted for A NEW GENERATION from the New York Times Bestseller



an inconvenient truth

the crisis of global warming



AL GORE

January 1999

nature

International Journal of Science

Eocene global warming

Hydrothermal vents prompt methane release

Malina parasite
Hints on extraterrestrial astigmatic variations

Photonic crystals
Perfecting the defects

Galapagos giant tortoise
Septuagennarian male seals mate



August 26, 2002

TIME

SPECIAL REPORT



HOW TO SAVE THE EARTH

The hot and wild weather is a sign of things to come. But fresh ideas and new technology can cool us down and make this a **GREEN CENTURY**

September 25, 2004

TIME

Where's the Beach?

America's Vanishing Coastline



APRIL 9, 2005 \$3.50

TIME



GLOBAL WARMING


Climbing temperatures. Melting glaciers. Rising seas. All over the earth we're feeling the heat. Why isn't Washington?

Dr. Bush's Rx for Health Care

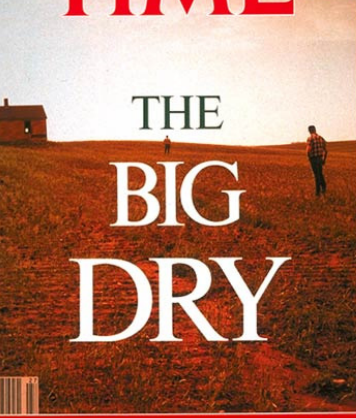
TIME

VANISHING OZONE

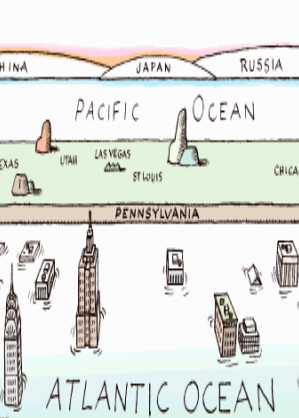
THE DANGER MOVES CLOSER TO HOME



THE BIG DRY



GLOBAL WARMING CARTOONS - 2007 - 210



OCTOBER 19, 1997 \$1.95

TIME

The Heat Is On

How the Earth's Climate Is Changing

Why the Ozone Hole Is Growing



NOVEMBER 28, 2005

JOE KLEIN ON IRAQ ■ BROKEBACK MOUNTAIN: GIDDY-YEP, I'M GAY

TIME

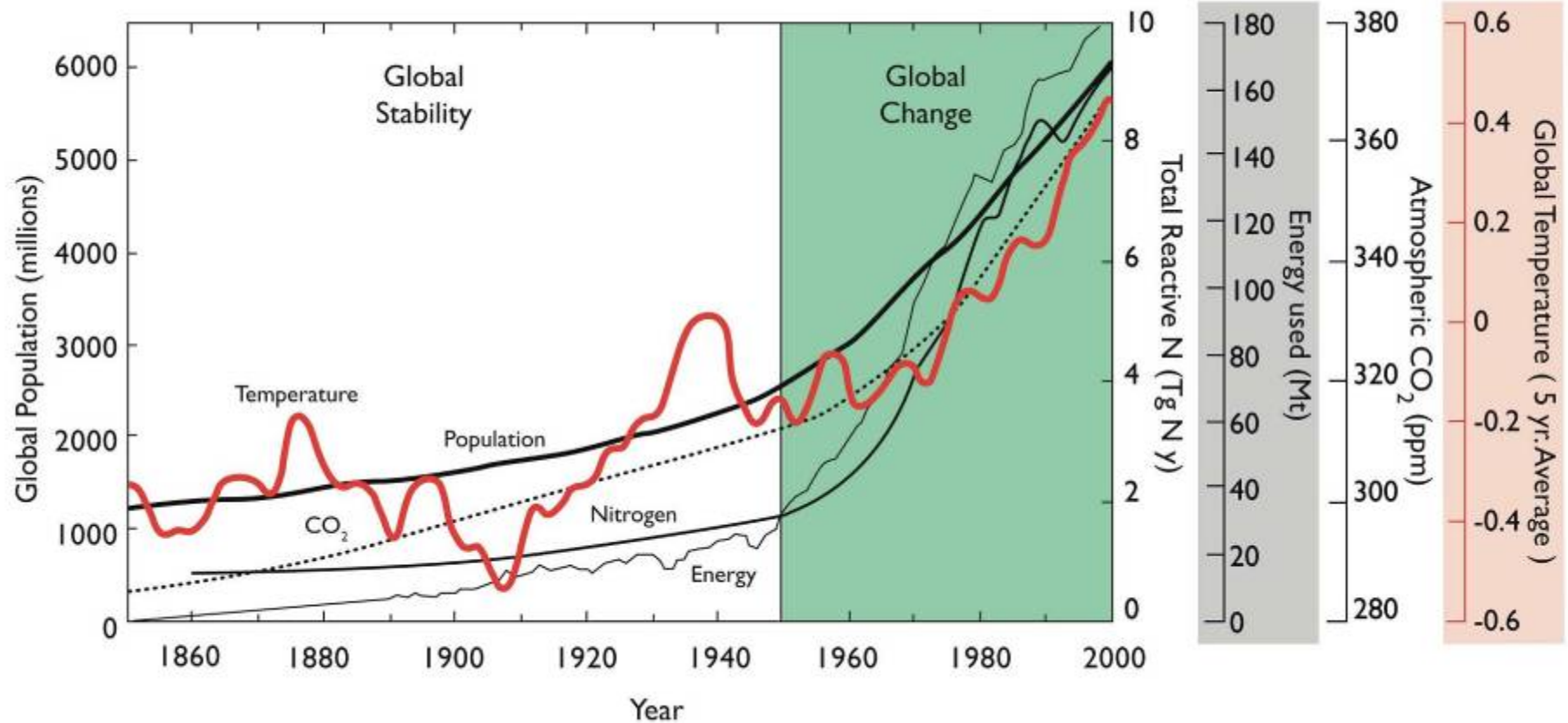


New Orleans Blues

It's worse than you think. Three months after Katrina, the city still suffers

BY CATHY BOOTH THOMAS

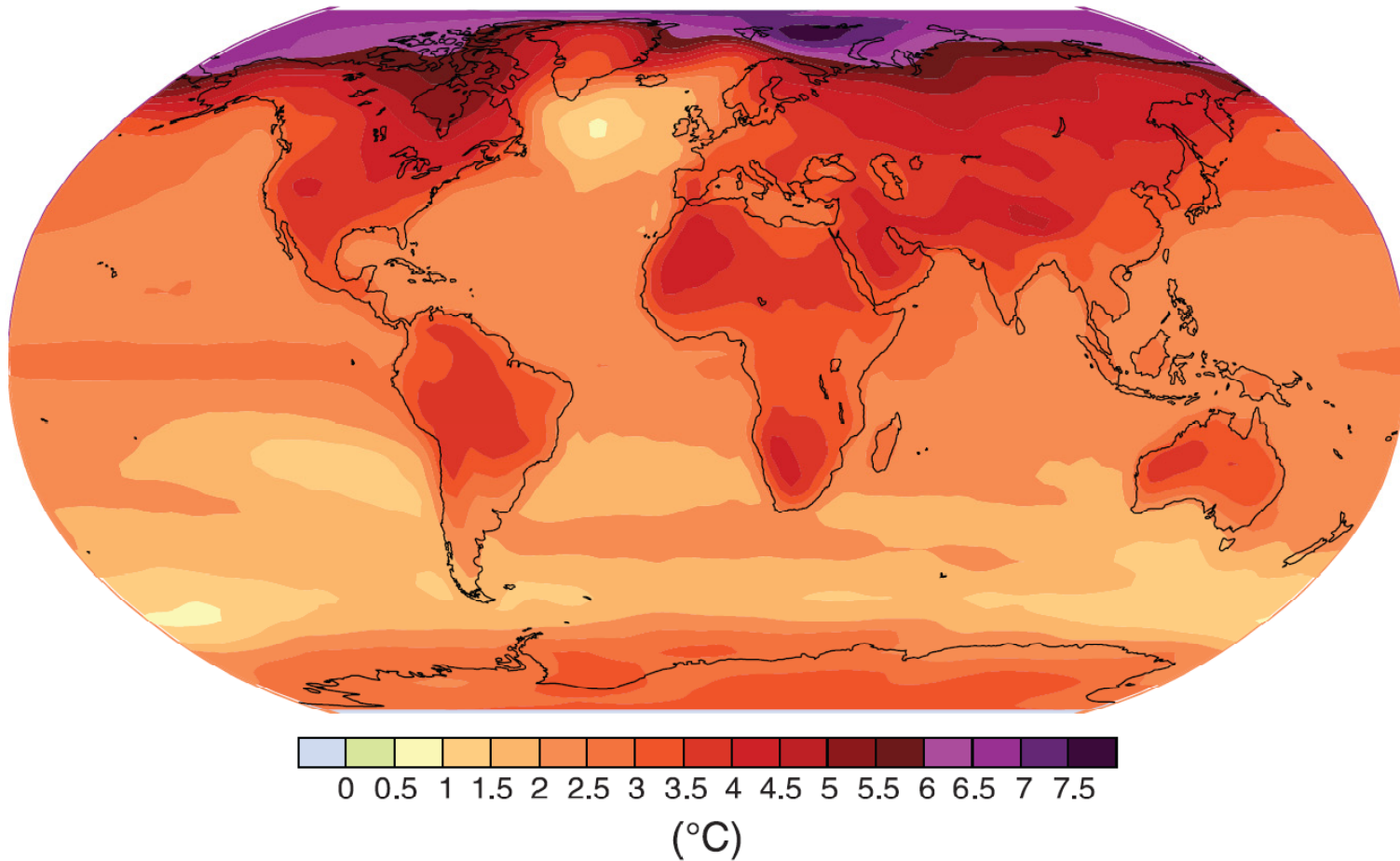
The complexities of global change



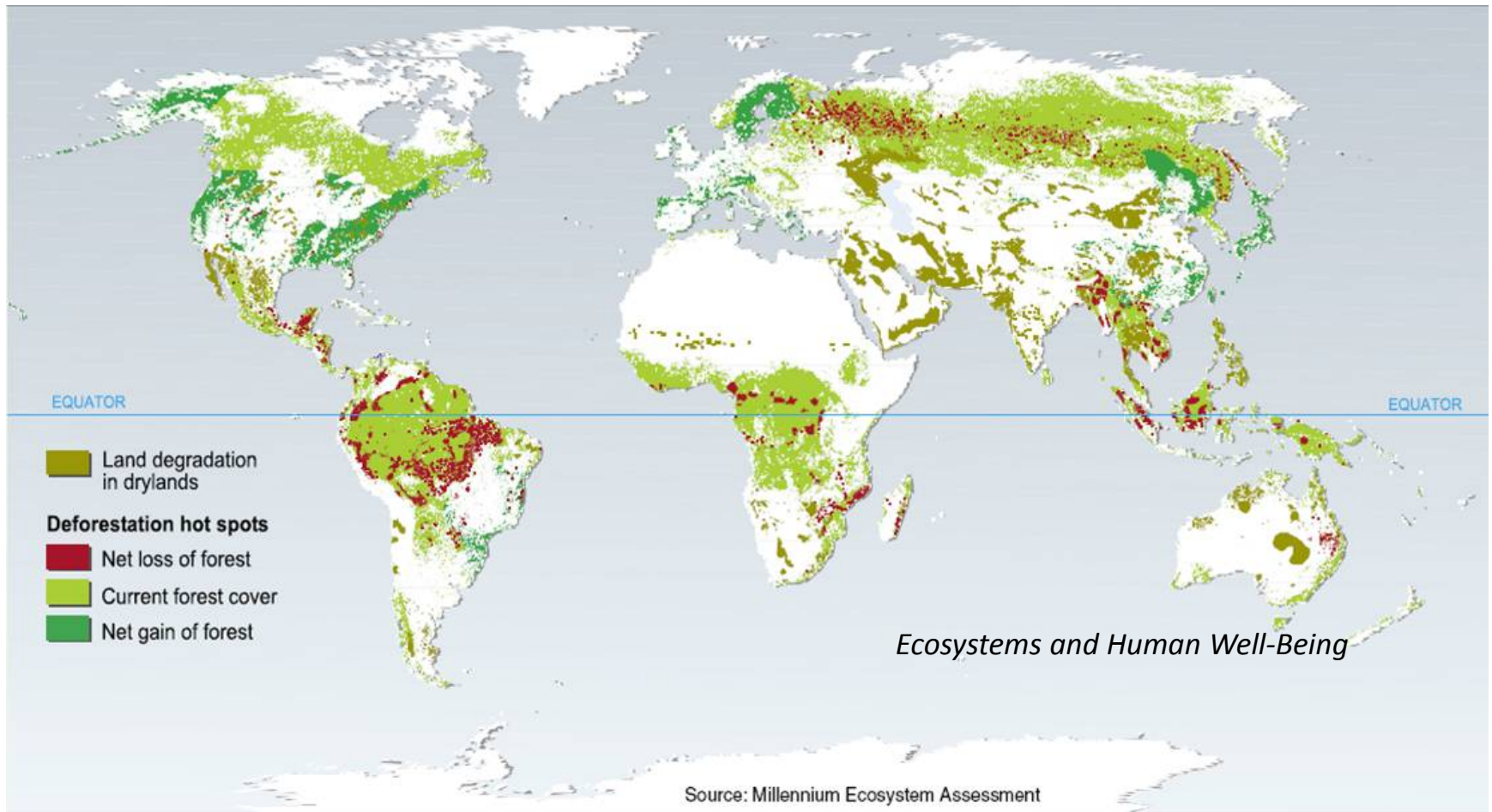
Smith, Knapp, Collins. In press.

Critical areas in the Earth's system

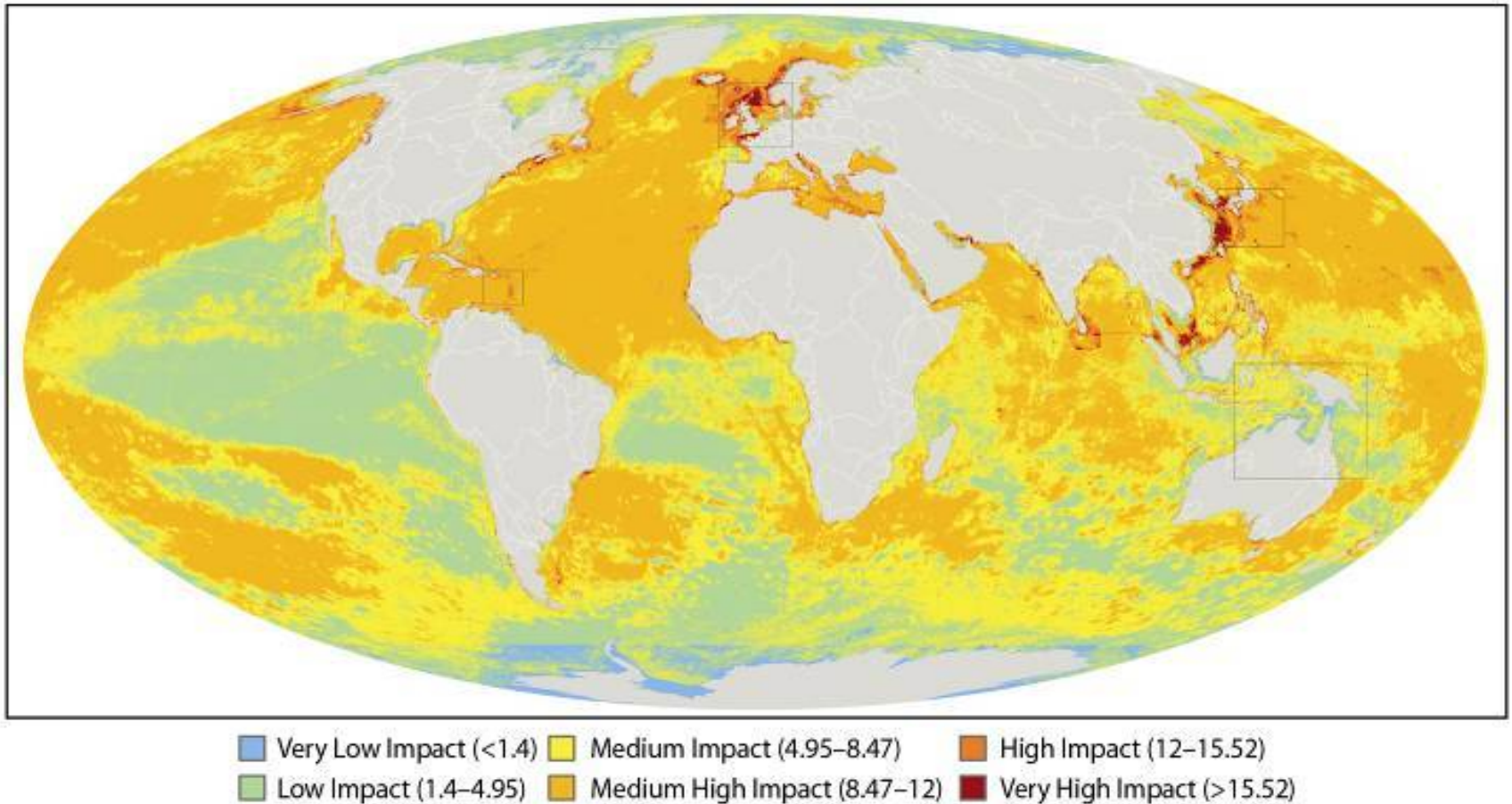
Geographical pattern of surface warming



Human impacts on land-based ecosystems



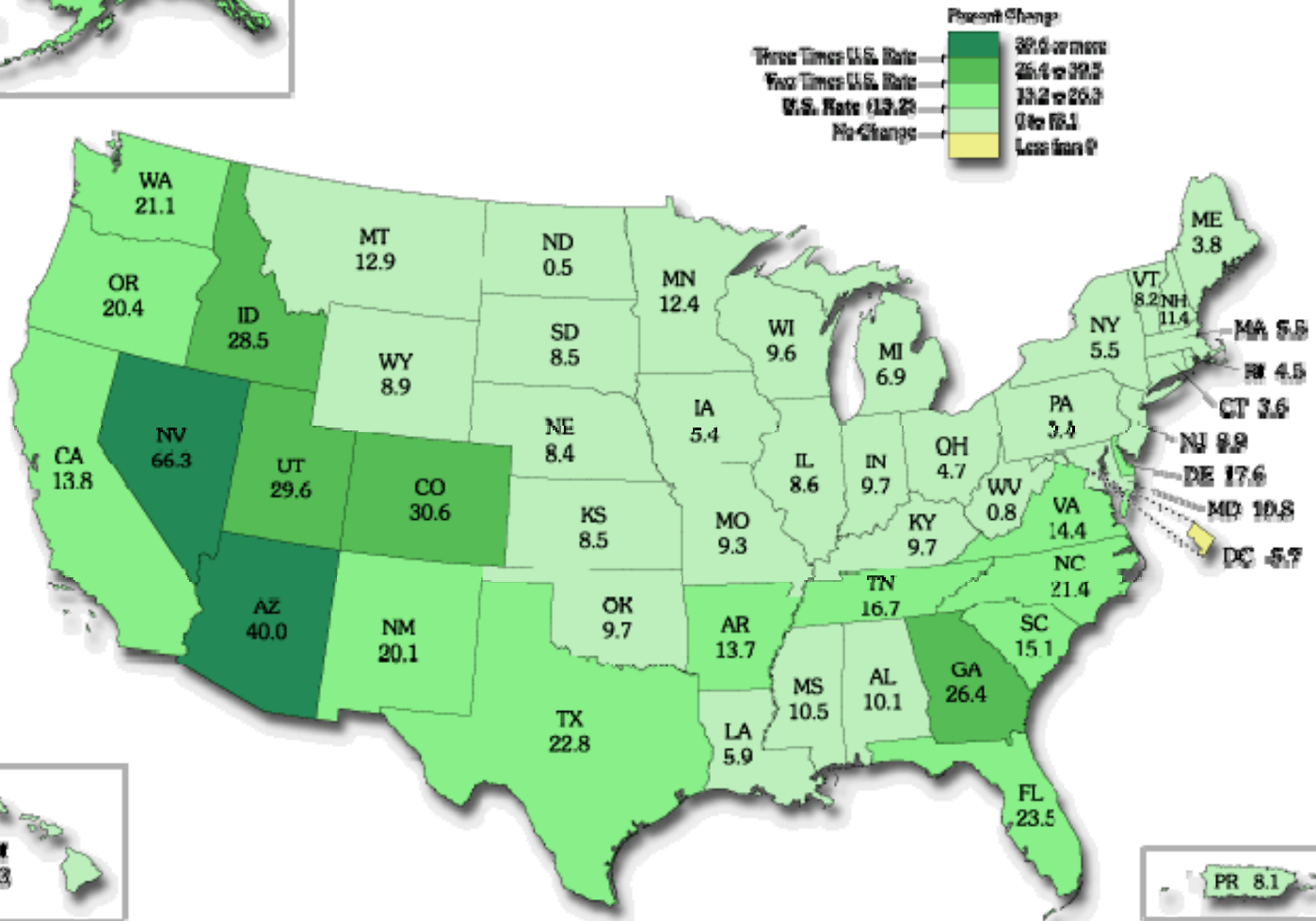
Human impacts on the world's oceans



Human population change



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



Figures by Geographic Edition



U.S. CENSUS BUREAU
Making the Nation Informed



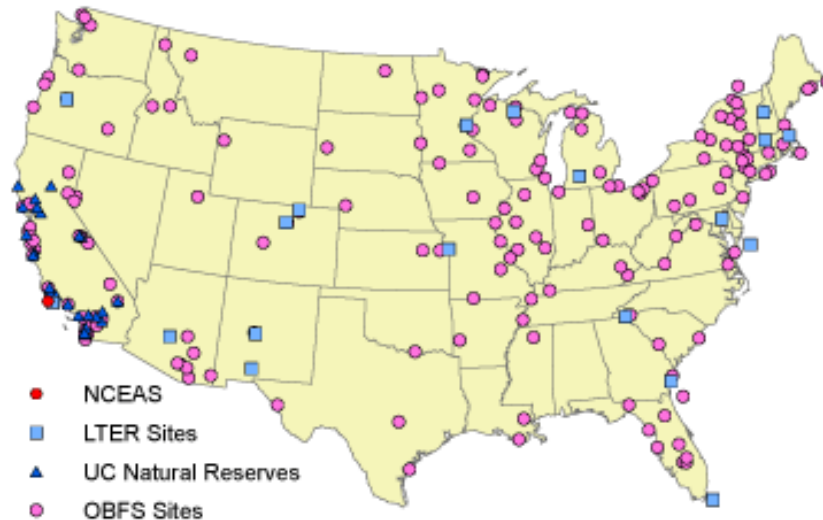
Outline of today's talk

- 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)

Study A

METADATA (from EML)	Study A: White Mountains				
	Area col. units: sq. meter				
	PIRU = <i>Picea rubens</i>				
	BEPA = <i>Betula papyifera</i>				
DATA	date	site	species	area	count
	10/1/1993	N654	PIRU	2	26
	10/3/1994	N654	PIRU	2	29
	10/1/1993	N654	BEPA	1	3

Study B

METADATA (from EML)	Study B: Green Mountains			
	Area sampled: 1 sq. meter			
	picrub = <i>Picea rubens</i>			
	betpap = <i>Betula papyifera</i>			
DATA	date	site	picrub	betpap
	31 Oct 1993	1	13.5	1.6
	14 Nov 1994	1	8.4	1.8

Integrated Data

study	date	site	species	density
A	0/1/1993	N654	<i>Picea Rubens</i>	13.0
A	0/3/1994	N654	<i>Picea Rubens</i>	14.5
A	0/1/1993	N654	<i>Betula papyifera</i>	3.0
B	10/31/1993	1	<i>Picea Rubens</i>	13.5
B	10/31/1993	1	<i>Betula papyifera</i>	1.6
B	11/14/1994	1	<i>Picea Rubens</i>	8.4
B	11/14/1994	1	<i>Betula papyifera</i>	1.8

metadata
'promoted'
to become
data

format
normalized
using
metadata

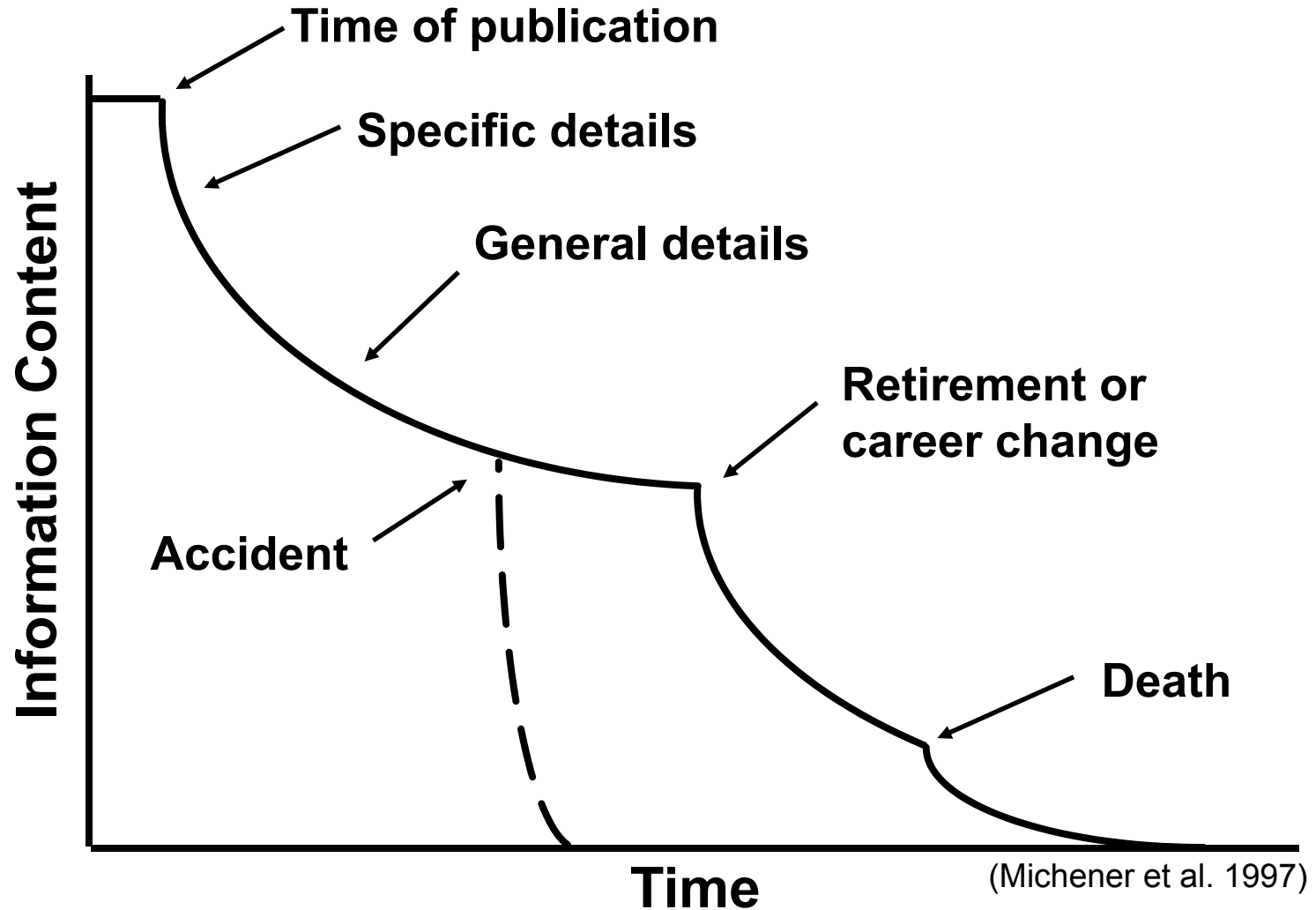
species metadata
from study B
is now data
(picrub/betpap
column headings)

density
calculated
using
metadata

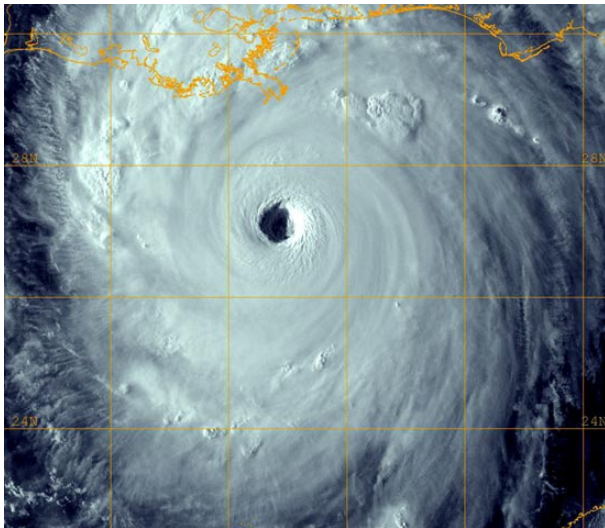
Jones et al. 2007

Data challenge 3: poor practice

“data entropy”



Data challenge 4: loss

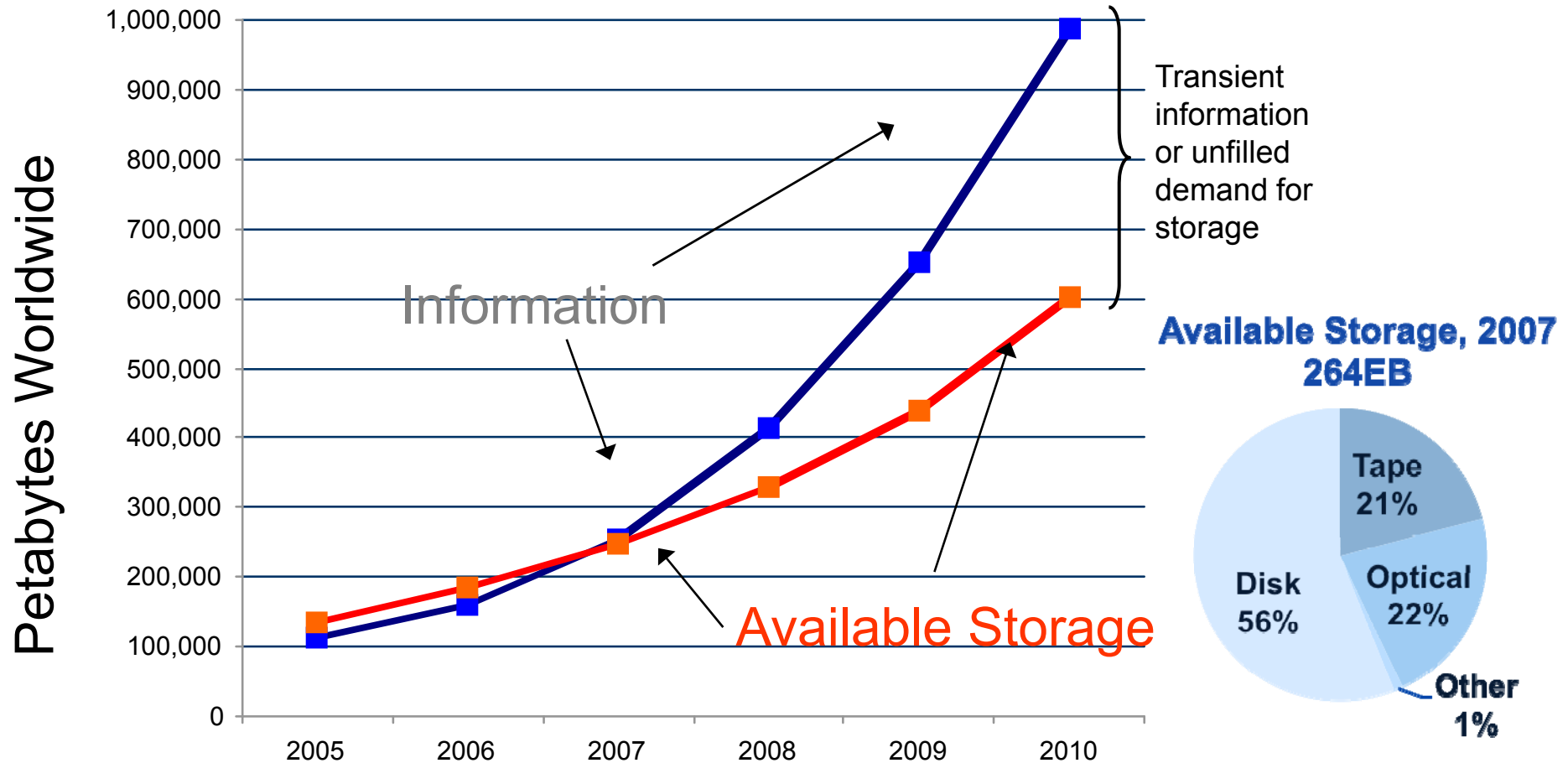


- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing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

Source: S. Abrams, CDL



Data challenge 4: more loss



Source: John Gantz, IDC Corporation: The Expanding Digital Universe

Cumulative impact: data longevity

Study	Resource Type	Resource Half-life
Rumsey (2002)	Legal Citations	1.4 years
Harter and Kim (1996)	Scholarly Article Citations	1.5 years
Koehler (1999 and 2002)	Random Web Pages	2.0 years
Spinellis (2003)	Computer Science Citations	4.0 years
Markwell and Brooks (2002)	Biological Science Education Resources	4.6 years
Nelson and Allen (2002)	Digital Library Objects	24.5 years

Koehler, W. (2004) *Information Research* 9(2): 174.

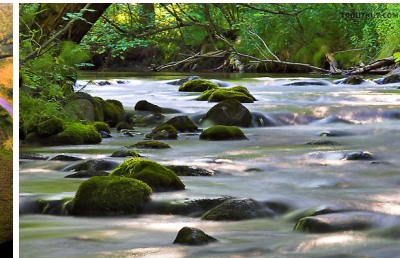
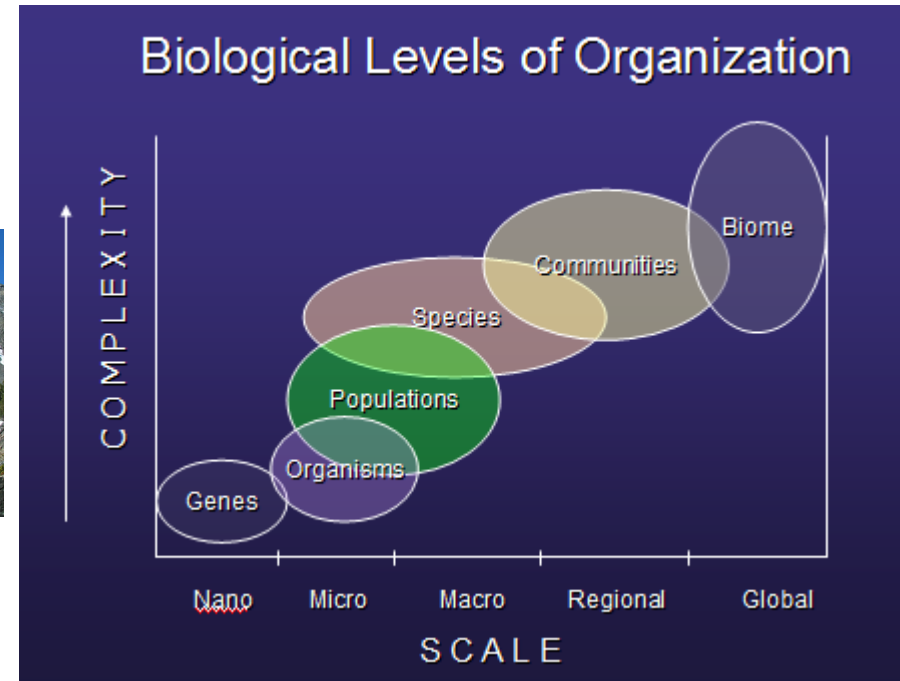
Outline of today's talk

- 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

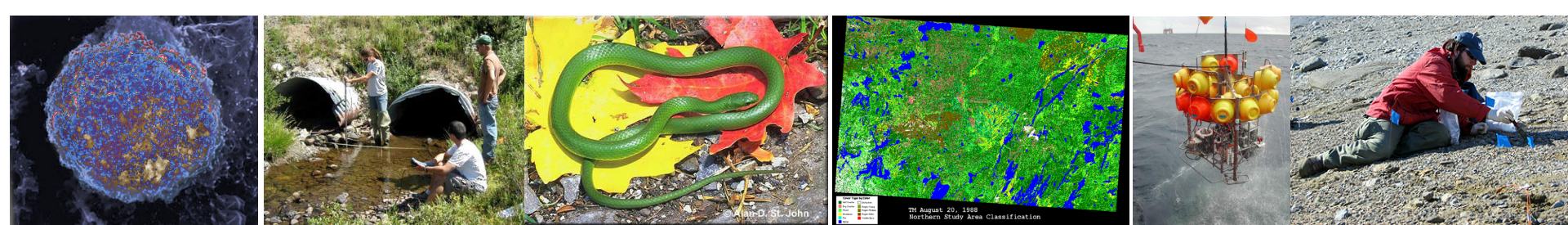
- The goal of DataONE is to enable new science through universal access to data about life on earth by:
 - engaging the scientist in the data preservation process
 - supporting the full data life cycle,
 - encouraging data stewardship and sharing
 - promoting best practices
 - engaging citizens
- One of two DataNet awardees recommended for funding by NSF

Data types

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



- 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

Data Archive	Types of Data Managed	Metadata Standard(s)
	Biodiversity, taxonomic, ecological	BDP, DwC, DC, OGIS
	Biogeochemical dynamics, terrestrial ecological Earth observation imagery	DIF, BDP, ECHO
	Ecological, biodiversity, biophysical, social, genomics, and taxonomic	EML
	Avian populations and molecular biology	DC
	Biological and taxonomic	DC subset
	Biophysical, biodiversity, disturbance, and Earth observation imagery	EML
	Biodiversity, biotic structure, function/process, biogeochemical, climate, and hydrologic	EML
EML=Ecological Metadata Language		

BDP=Biological Data Profile

DC subset=Dublin Core subset

DwC=Darwin Core

OGIS=OpenGIS

DC=Dublin Core

DIF=Directory Interchange Format

ECHO=EOS ClearingHouse

DataONE Providing one-stop shopping for data

Simple Pilot Catalog Interface

(searches entire metadata record)

40,000 Data Set Records

NBII 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)

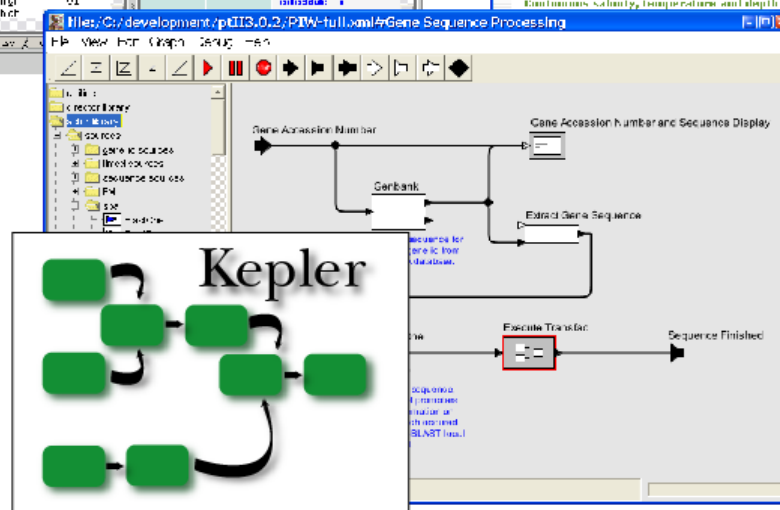
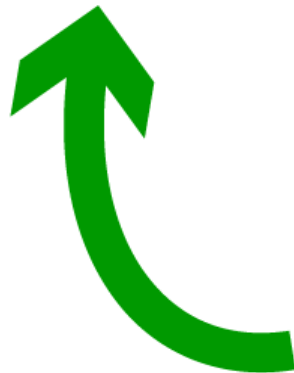
Organization of Biological Field Stations (124)

Inter-American Institute for Global Change Research (IAI) (79)

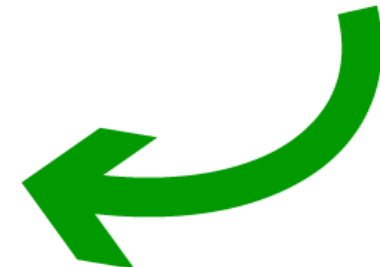
MODIS and ASTER Products (LPDAAC) (38)

National Phenology Network (USANPN) (29)

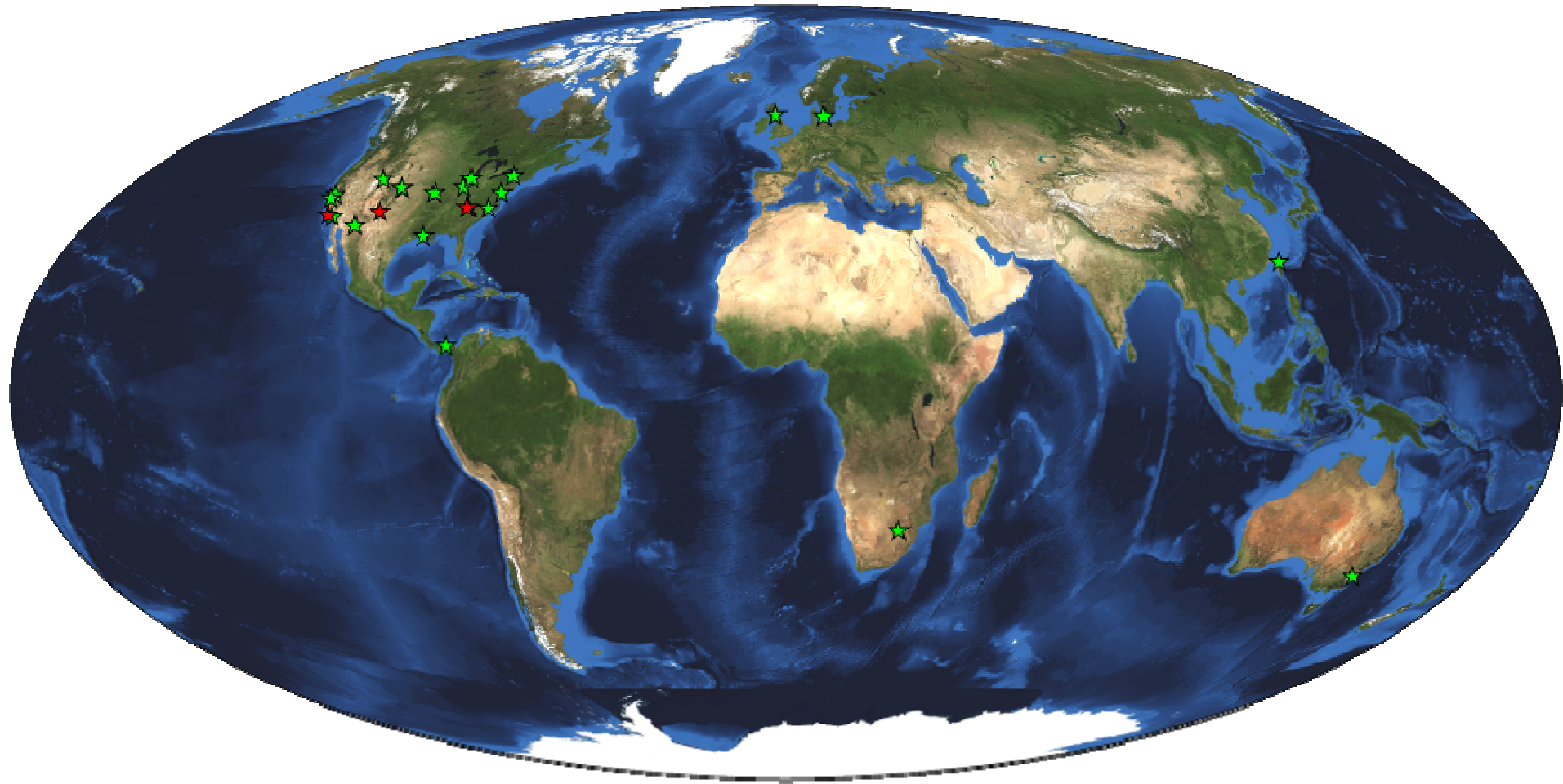
The screenshot shows the DataNetONE Pilot Catalog Interface. At the top, the logo "DataNetONE" is on the left and "A Pilot Catalog For Earth Observations" is on the right. Below the logo is a black bar with "DataNetONE Metadata Clearinghouse" and a "HELP" button. The main interface has two tabs: "Simple Search" (selected) and "Advanced Search". Below the tabs is a section titled "Search All Records For" with a search input field and a "SEARCH" button. A hint below the input field reads: "Hint: boolean operators, wildcards and phrases are allowed. ex: precipitation or (rain* and 'moisture content')". To the right of the input field is a "Results/Page" dropdown menu set to "10". Below the search field is a section titled "Query being built:" with a text area that says "Not Editable". At the bottom left is a "CLEAR QUERY" link. At the bottom center is a "Mercury" logo. At the bottom right are links for "Disclaimer and Privacy Statement", "Feedback Form", "Mercury", "Email Us", and "Security Warning".



The screenshot shows a web browser window titled "Mozilla" displaying the "Biocomplexity Data Search" page. The page has a blue header with the KNN logo and the text "Home". Below the header, there is a search bar with the text "database" and a "Search KNN" button. To the right of the search bar, there is a link to "Advanced search tool". Below the search bar, there is a section titled "You ARE logged in (Logout)" and a message about the search results. The search results are displayed in a table with columns "Title", "Contacts", and "Organization". The first row shows "Biology and Ecology" with contact "Marianne" and organization "Marianne". The second row shows "Productivity, Diversity and Soil Data from two North American Grasslands" with contact "Doe" and organization "Doe". The third row shows "ED: Bioclassified" with contact "Doe" and organization "Doe". A large "MetaCat" logo is overlaid on the right side of the page. At the bottom, there is a circular "EcoGrid" diagram with the text "EcoGrid" in the center. The diagram is divided into segments labeled "Query", "Read", "Write", "Access", "Replicate", "Compute", and "Store".



DataONE Building new global cyberinfrastructure



Coordinating Nodes

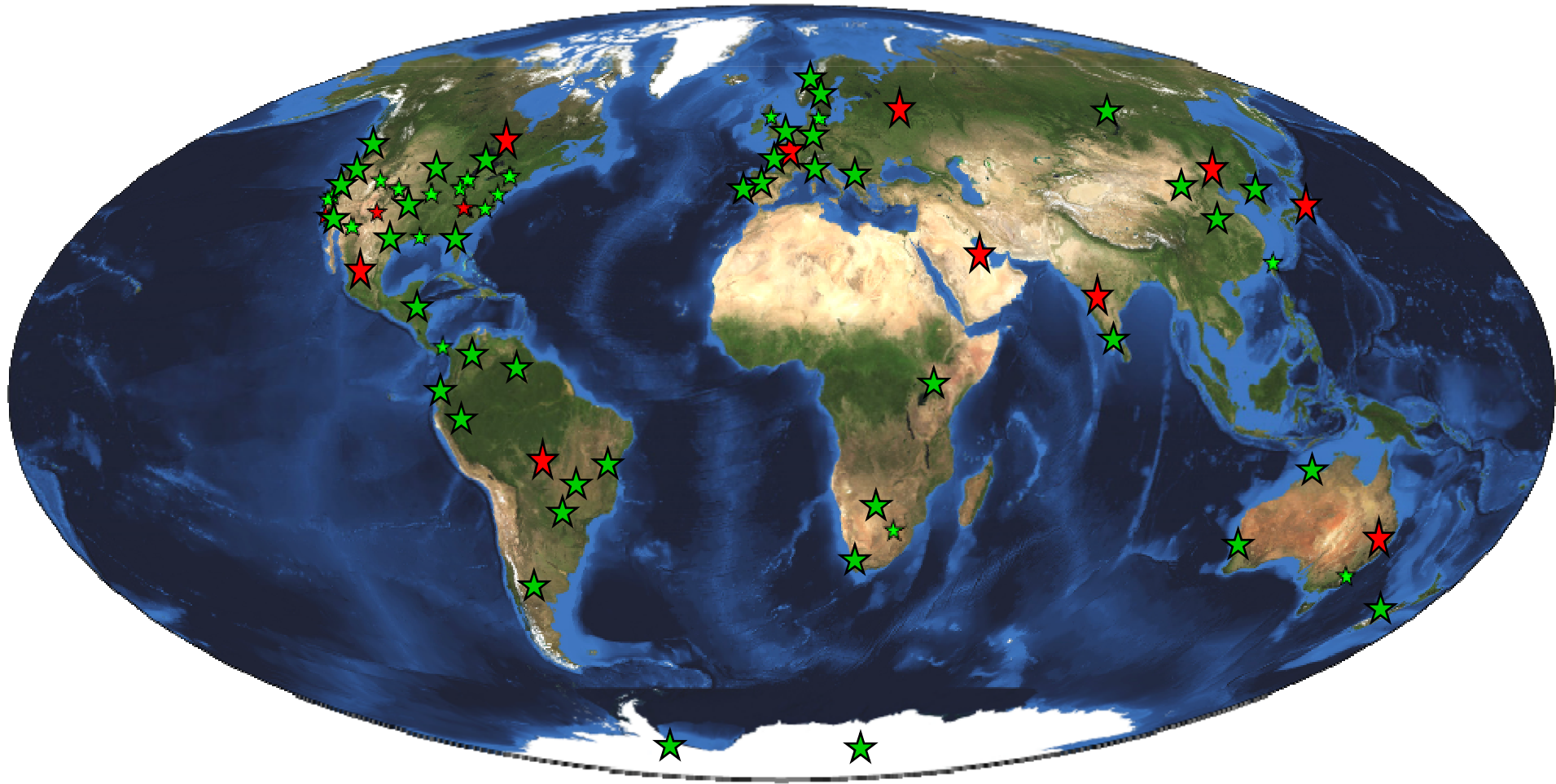
- retain complete metadata catalog
- subset of all data
- perform basic indexing
- provide network-wide services
- ensure data availability (preservation)
- provide replication services

Flexible, scalable,
sustainable network

Investigator 1..N Toolkit



DataONE Building new global cyberinfrastructure



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

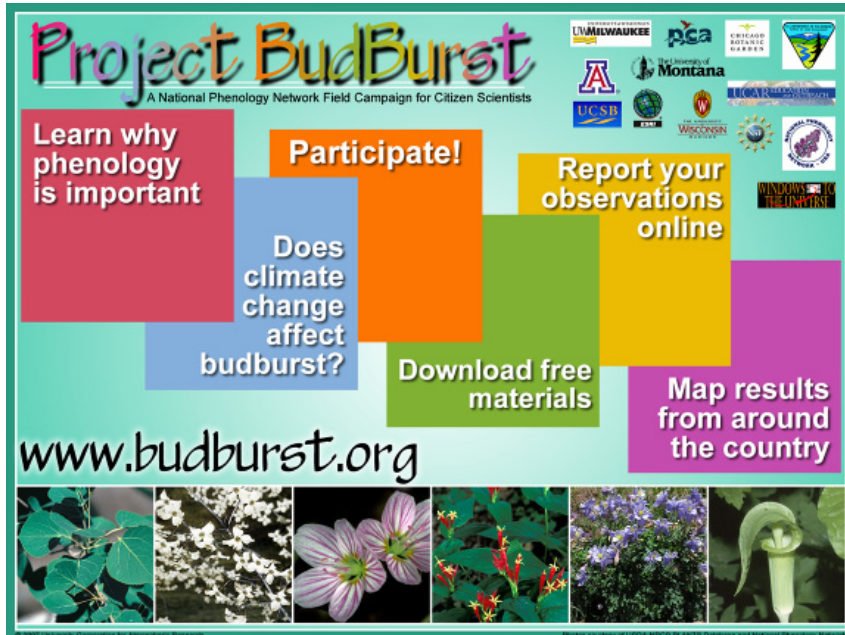
DataONE Partners plus Kepler-CORE and SEEK/KNB Teams

We welcome your involvement!



THE UNIVERSITY of
NEW MEXICO





Project BudBurst
A National Phenology Network Field Campaign for Citizen Scientists

Learn why phenology is important

Participate!

Report your observations online

Does climate change affect budburst?

Download free materials

Map results from around the country

www.budburst.org

Logos for participating institutions: UW Milwaukee, PCA, University of Montana, UCSB, etc.

Images of various plants and flowers.



eBird



Welcome to Citizen Science Central! - CITIZEN Science Project Toolkit Mozilla Firefox

<http://www.birds.cornell.edu/csc>

CORNELL LAB of ORNITHOLOGY

Citizen Science Central

Welcome to Citizen Science Central!

A clearinghouse for ideas, news, and resources in support of citizen science—partnerships between volunteers and scientists that answer real world questions.

Home
About
Project Gateway
References
Toolbox
Conference Proceedings
Discussion Forum

IDEAS
About this initiative
Discussion Forum

NEWS
News
Events

RESOURCES
Toolkit
References
Projects
Proceedings

Citizen Science at the Cornell Lab of Ornithology

© 2007 Cornell Lab of Ornithology
Home | Log In

CORNELL LAB of ORNITHOLOGY

100 Supercenter Way, Ithaca, NY 14850
607-255-5890 | cornellbirds@cornell.edu

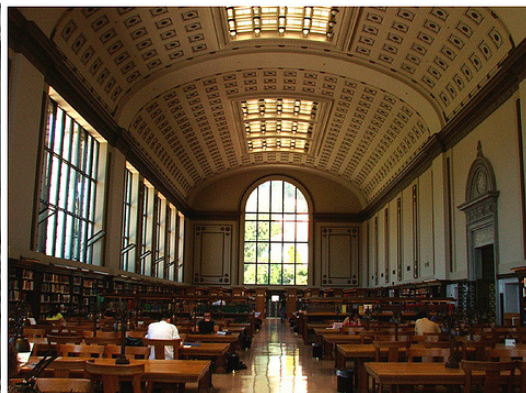
Open Notebook

www.CitizenScience.org



DataONE 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

Transfer Rate (KB/s) - upload

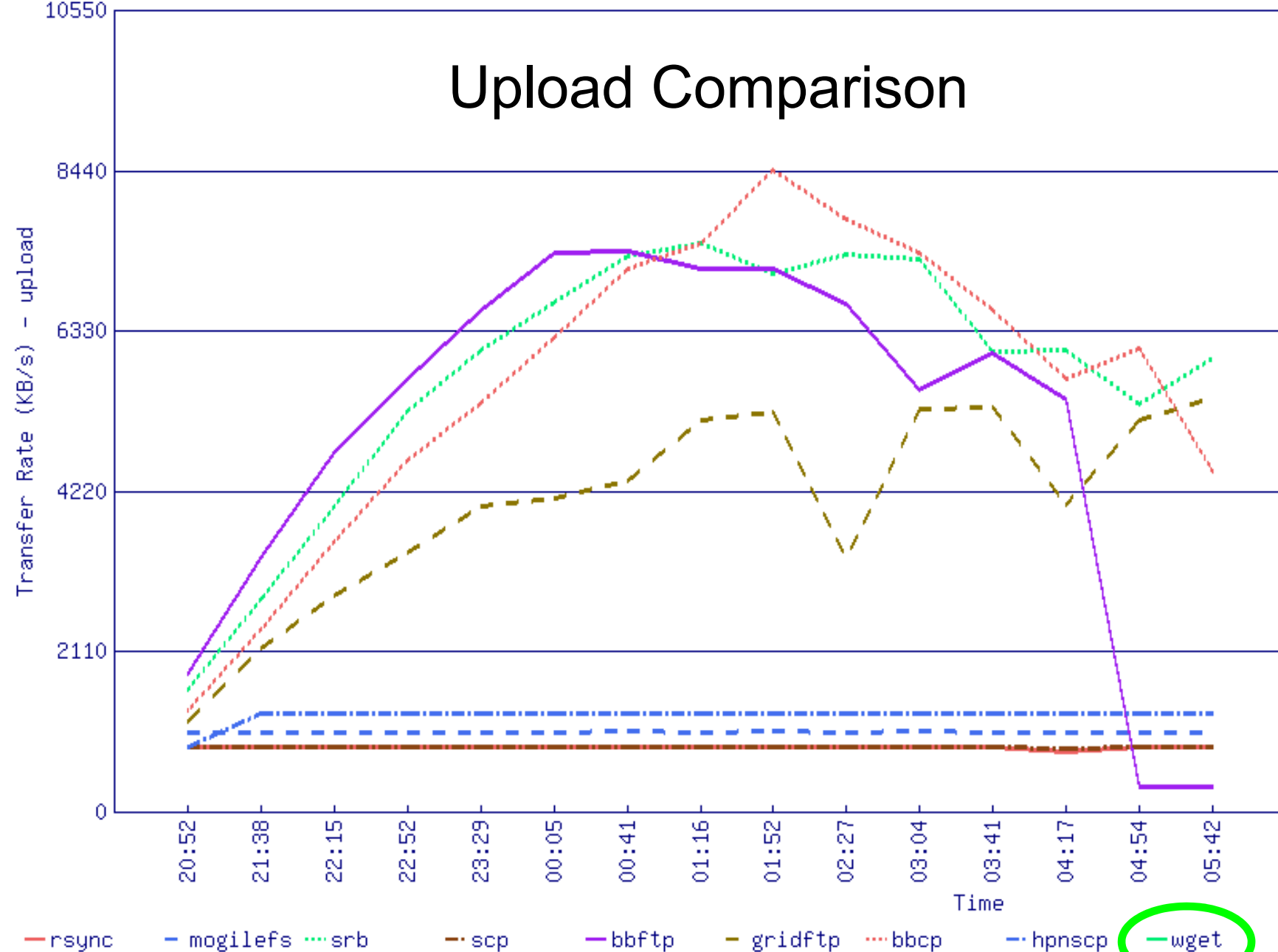
10550
8440
6330
4220
2110
0

20:52 21:38 22:15 22:52 23:29 00:05 00:41 01:16 01:52 02:27 03:04 03:41 04:17 04:54 05:42

Time

rsync mogilefs srb scp bbftp gridftp bscp hpnsdp wget

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



cyocum

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 amentities, 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

Minimalism: ANVL and Dublin Core Kernel

A Name Value Language (ANVL) – back to basics

- An ANVL record is a sequence of elements in email header format:
 ⇒ 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(anv1/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

<http://www.cdlib.org/inside/diglib/ark/thumpspec.pdf>

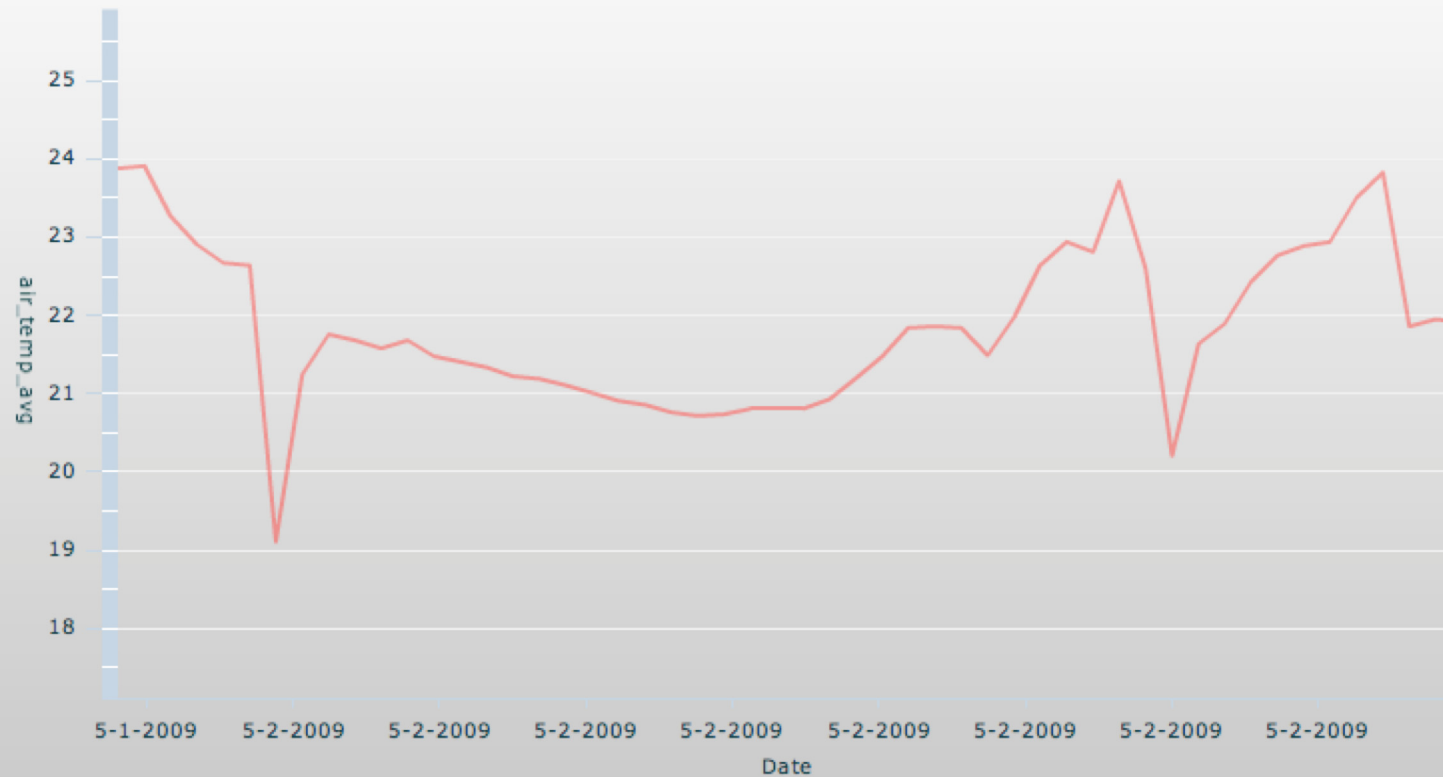
“DataLab” project extending THUMP for tabular data integration and visualization (Nassib Nassar, RENCI)

DataLab URL:

[http://datalab.renci.org/?in\(csd/ocean_met_last24hrs\)](http://datalab.renci.org/?in(csd/ocean_met_last24hrs))

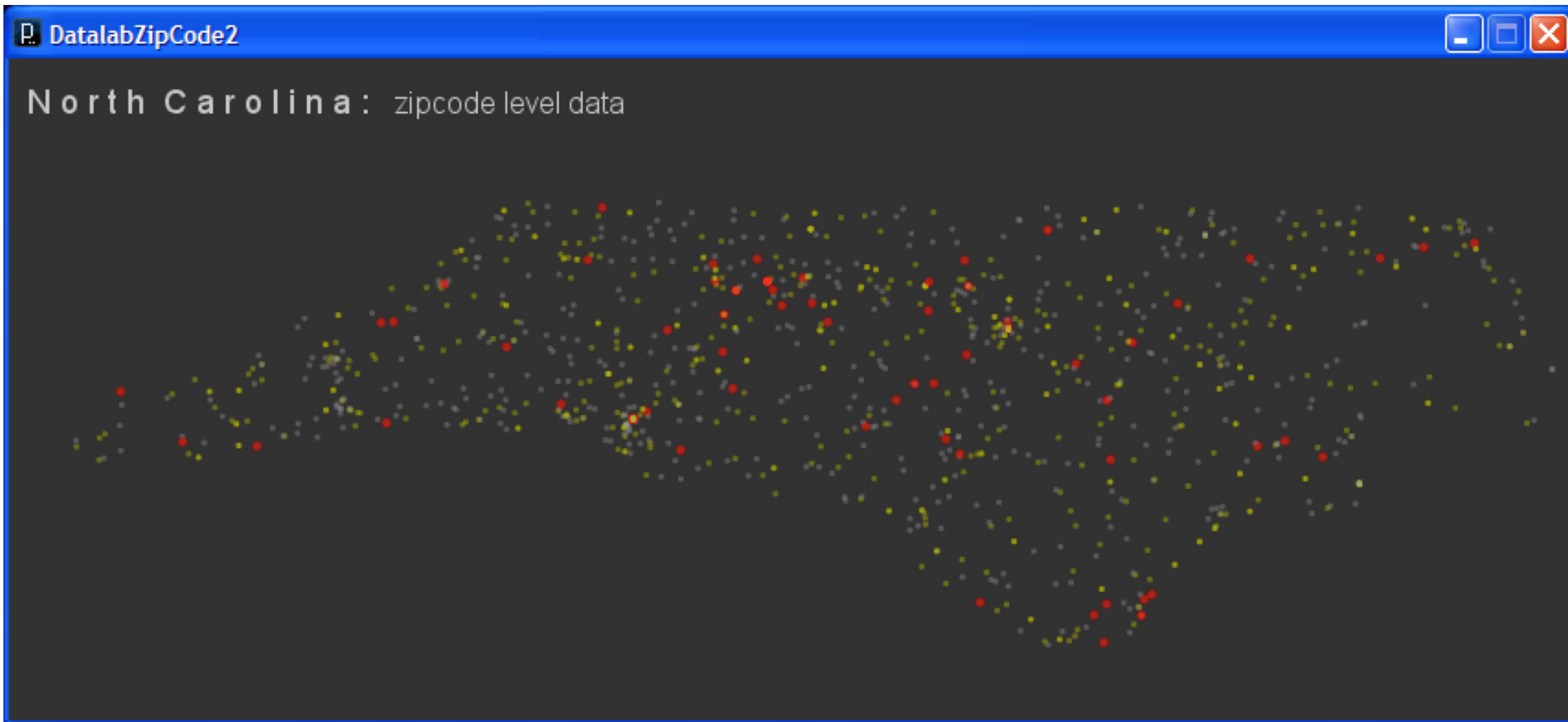
Search

air_temp_avg



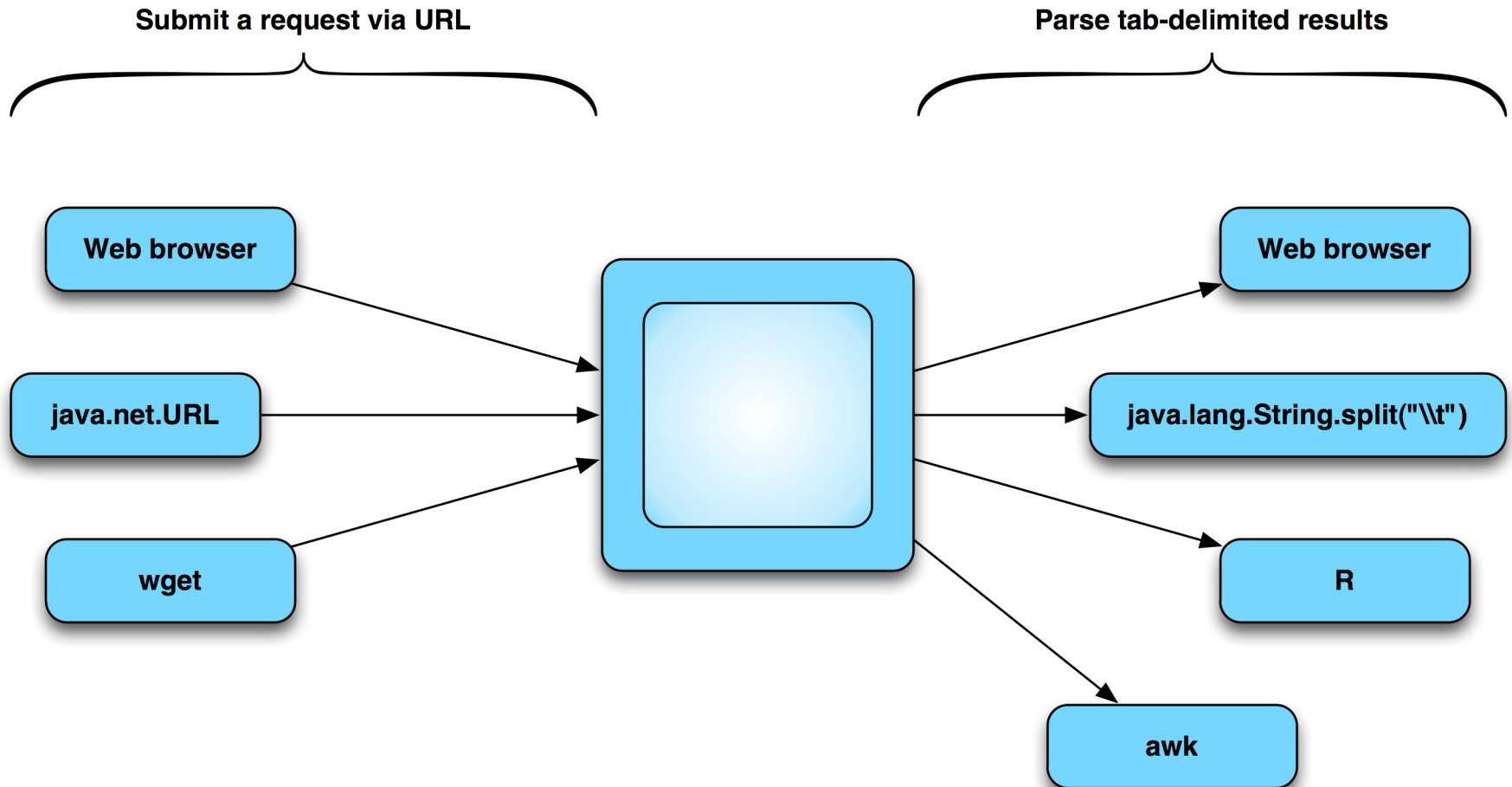
Status:

Sample Java visualization by ZIP code



QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Integrating THUMP DataLab extensions



Source: Nassib Nassar

Integrating Sarcomere and THUMP-DL with other tools and programming languages

Question:
Why use THUMP?
Why not XML?

Answer:
XML is not the
solution to
every problem

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 +- sqrt{b sup 2 - 4ac}} over 2a
```

In OpenOffice:

```
x={-b plusminus sqrt {b^2 - 4 ac}} over {2 a}  
x={-b +- sqrt {b^2 - 4ac}} over 2a
```

Why not use XML?

```

<math mode="display" xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <mi>x</mi>
    <mo>=</mo>
    <mfrac>
      <mrow>
        <mo form="prefix">&minus;</mo>
        <mi>b</mi>
        <mo>&PlusMinus;</mo>
        <msqrt>
          <msup>
            <mi>b</mi>
            <mn>2</mn>
          </msup>
          <mo>&minus;</mo>
          <mn>4</mn>
          <mo>&InvisibleTimes;</mo>
          <mi>a</mi>
          <mo>&InvisibleTimes;</mo>
          <mi>c</mi>
        </msqrt>
      </mrow>
      <mrow>
        <mn>2</mn>
        <mo>&InvisibleTimes;</mo>
        <mi>a</mi>
      </mrow>
    </mfrac>
  </mrow>
</math>

```


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	7. Characterization
2. Identity	8. Description
3. Storage	9. Index
4. Catalog	10. Search
5. Fixity	11. Annotation
6. Replication	12. Publication
- More details in Stephen Abrams' talk on 1pm Tuesday!



Moving into the mainstream. Enabling our digital future.

iPRES 2009

THE SIXTH INTERNATIONAL CONFERENCE ON PRESERVATION OF DIGITAL OBJECTS

CDL
↑

Come join us!
San Francisco
October 5-6, 2009
<http://www.cdlib.org/iPres/>

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

CDL
↑