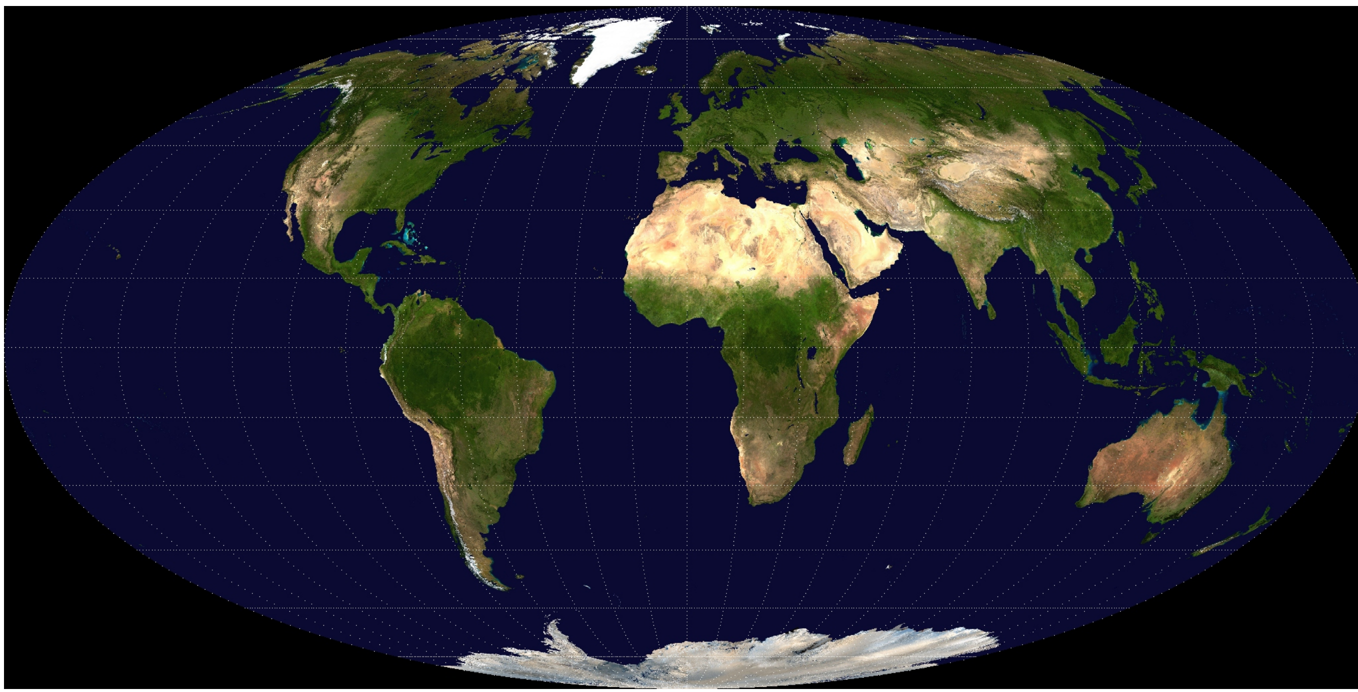
Envisioning a New Distributed Organization and Cyberinfrastructure to Enable Science



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

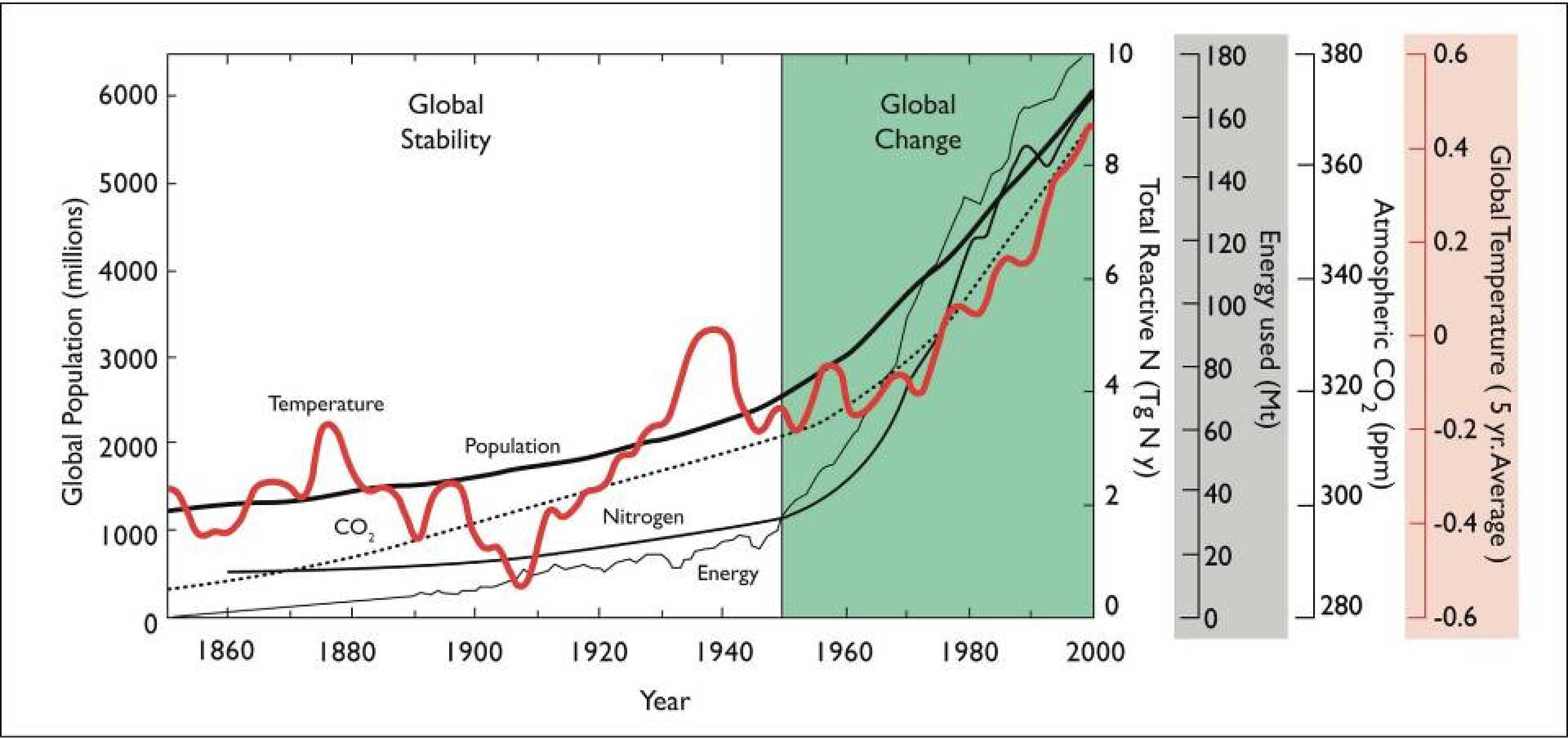
CDL

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)

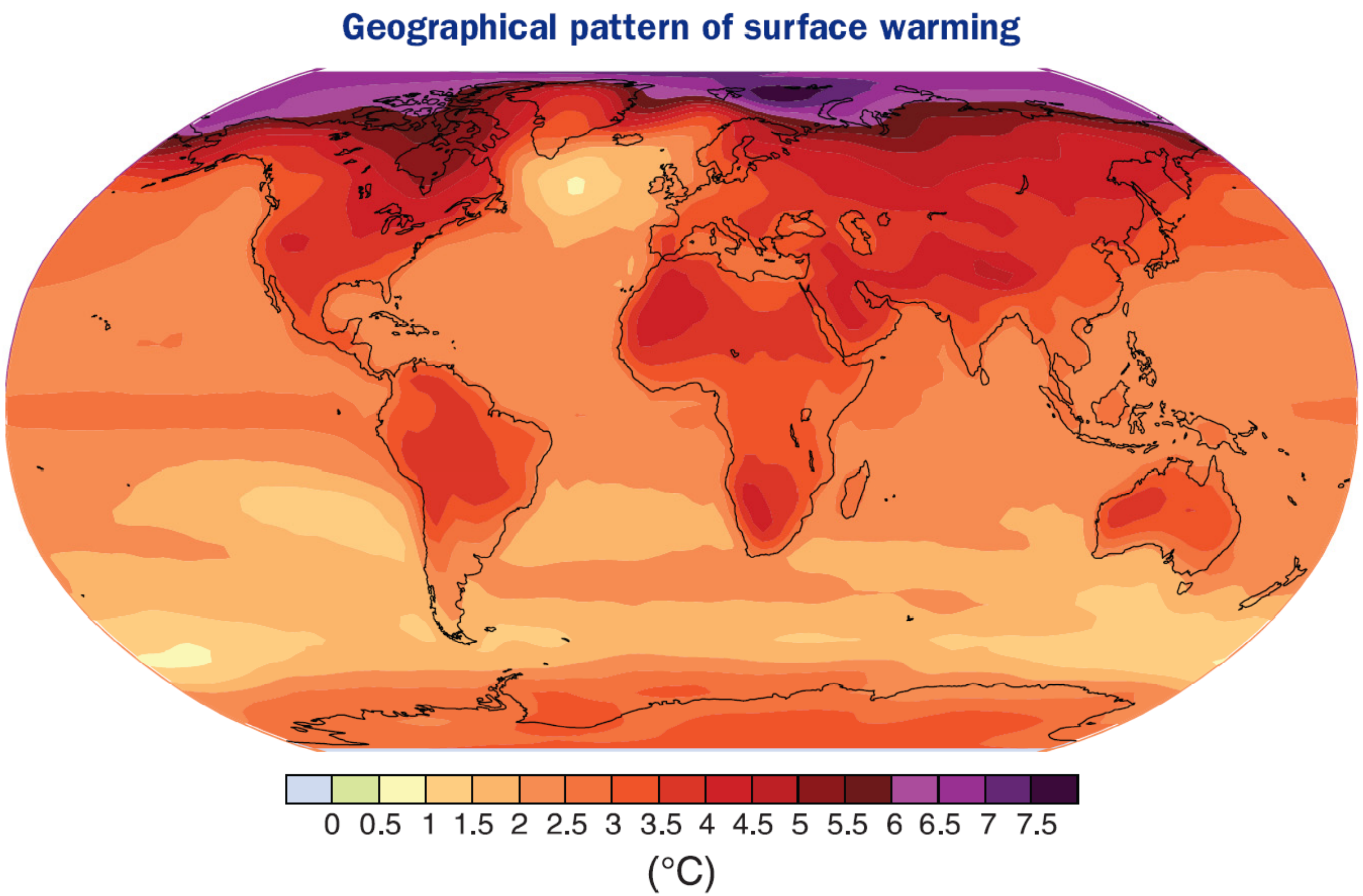
Living with Cancer
The changing science
Beyond Baghdad: Where The Enemy Has Its Own Surge
The Sopranos' Last Song: What Exit Will Tony Take?
TIME
SPECIAL DOUBLE ISSUE
The Global Warming Survival Guide
51 Things You Can Do to Make a Diffe
2ND-QTR SIZZLE PROFITS AT 900 COMPANIES (P. 74)
PAYING FOR COLLEGE BEWARE OF THOSE HIGH 529 FEES (P. 95)
TERRORISM WHAT COMPANIES STILL NEED TO DO (P. 26)
BusinessWeek
GLOBAL WARMING
Why Business Is Taking It So Seriously
BY JOHN CAREY (P. 60)
APRIL 3, 2006
SPECIAL REPORT GLOBAL WARMING
TIME
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
Adapted for A NEW GENERATION
from the New York Times Bestseller
an inconvenient truth
the crisis of global warming
AL GORE
nature
Eocene global warming
Hydrothermal vents prompt methane release
Malaria parasite
Photonic crystals
Perfecting the defects
Galapagos giant tortoise
AUGUST 26, 2002
SPECIAL REPORT
TIME
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
TIME
007 at 25
Where's the Beach?
America's Vanishing Coastline
W YORKER
LOBAL WARMING CARTOONS - 2007 - 21
HINA
JAPAN
RUSSIA
PACIFIC OCEAN
TEXAS
UTAH
LAS VEGAS
ST LOUIS
CHICAG
PENNSYLVANIA
ATLANTIC OCEAN
APRIL 9, 2001
GLOBAL WARMING
Climbing temperatures.
Melting glaciers. Rising seas.
All over the earth we're feeling the heat.
Why isn't Washington?
OCTOBER 19, 1987
TIME
SPECIAL REPORT Fighting for Global Markets
The Heat Is On
How the Earth's Climate Is Changing
Why the Ozone Hole Is Growing
Dr. Bush's Rx for Health Care
TIME
VANISHING OZONE
THE DANGER MOVES CLOSER TO HOME
THE BIG DRY
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

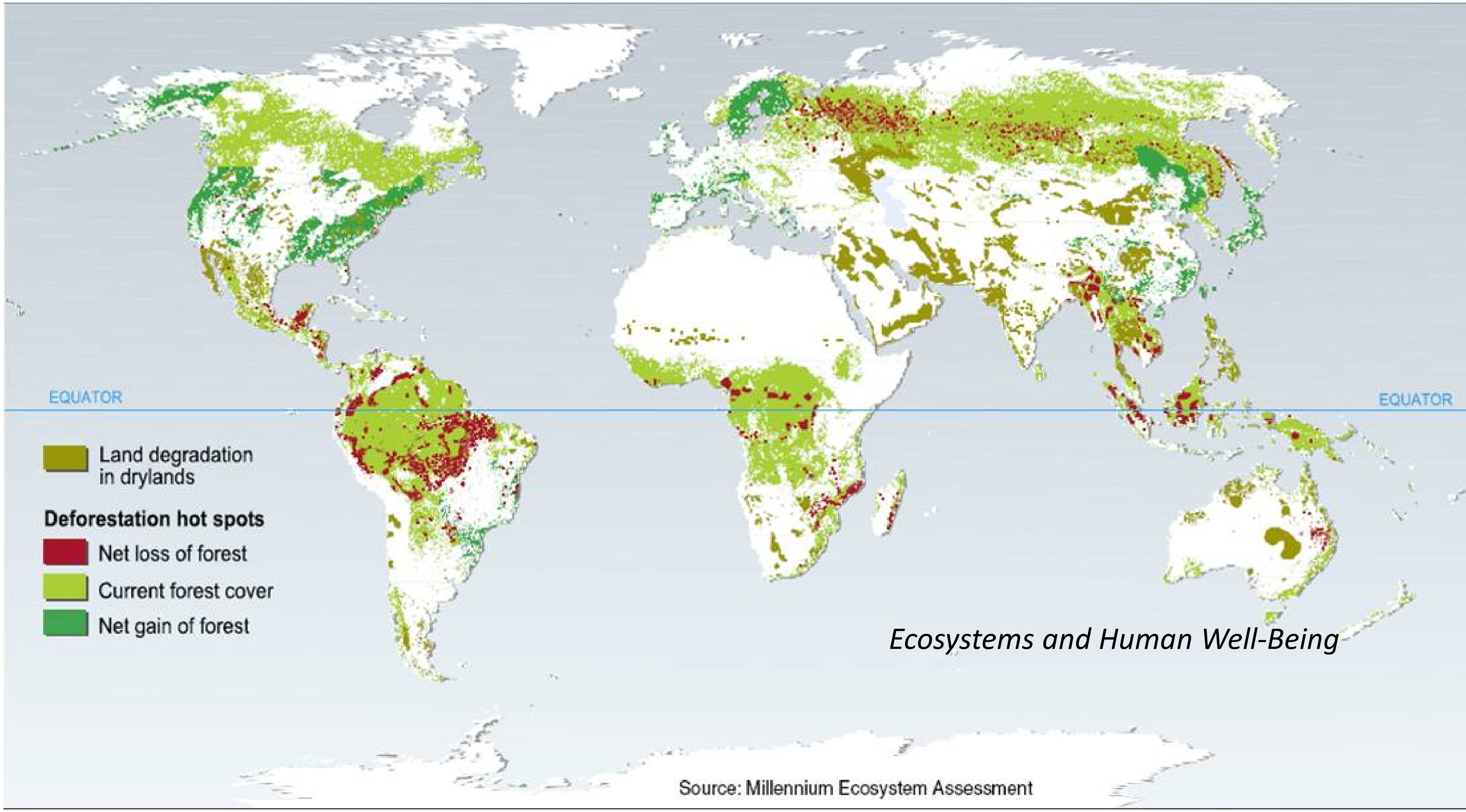


Smith, Knapp, Collins. In press.

Critical areas in the Earth’s system

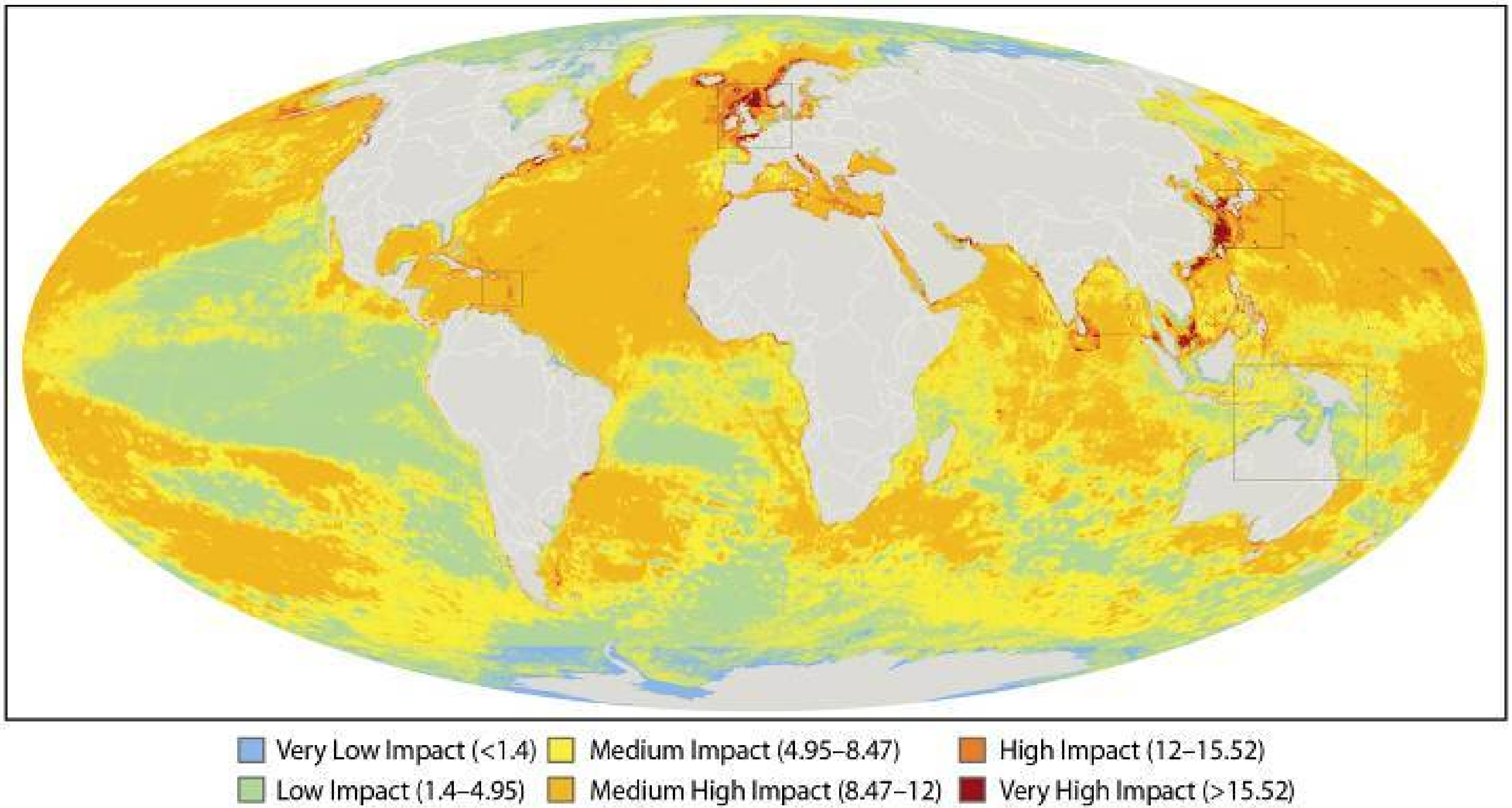


Human impacts on land-based ecosystems



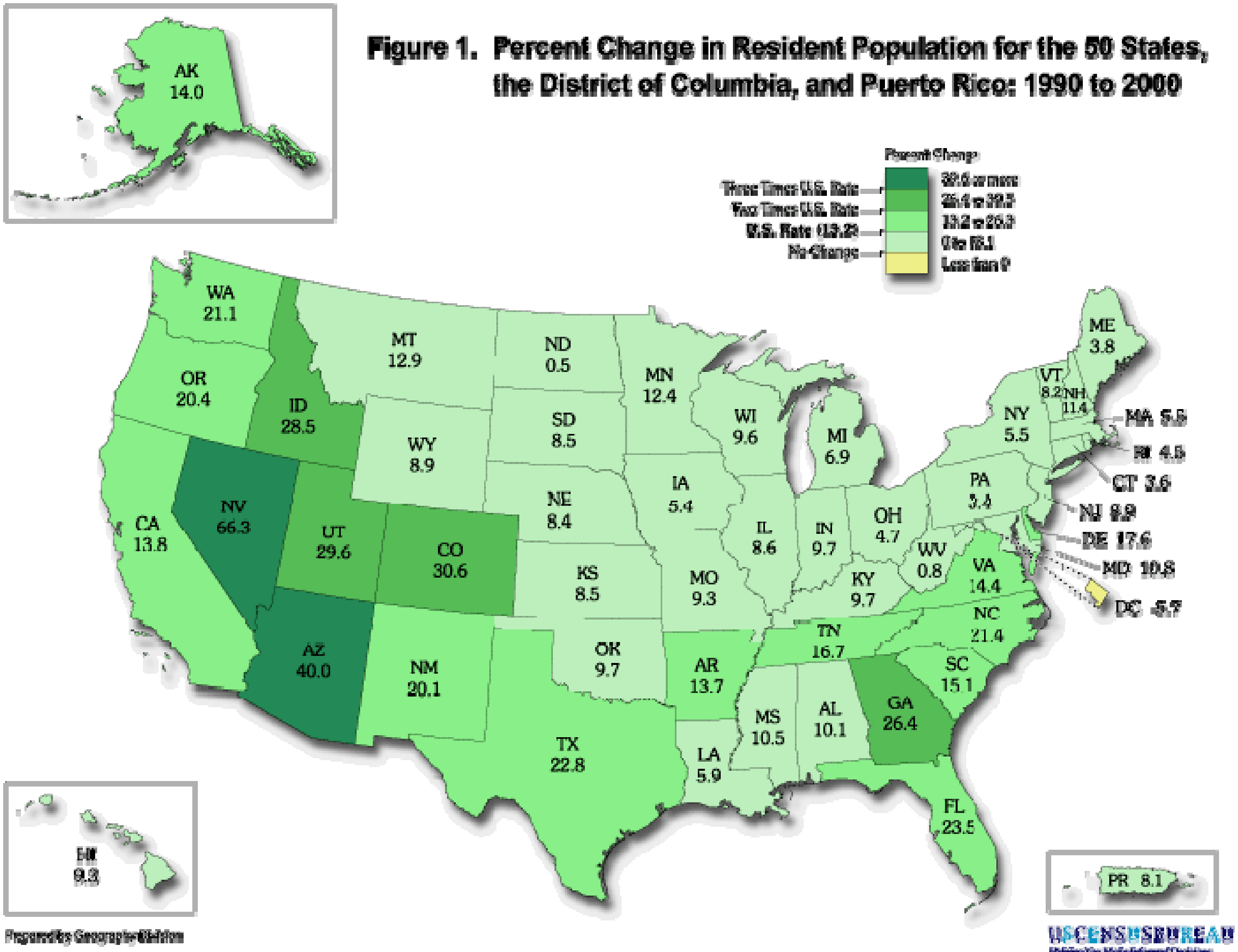
CDL

Human impacts on the world's oceans



Halpern et al. 2008 Science 319.

Human population change



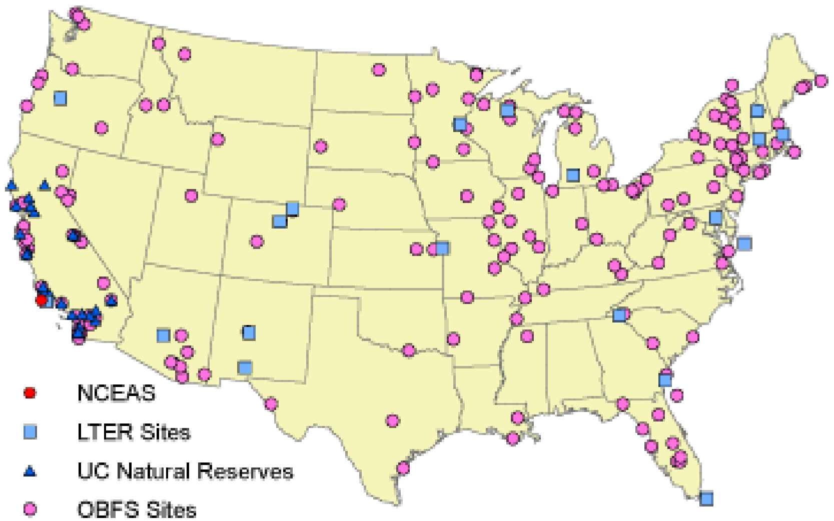
Outline of today's talk

- 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)

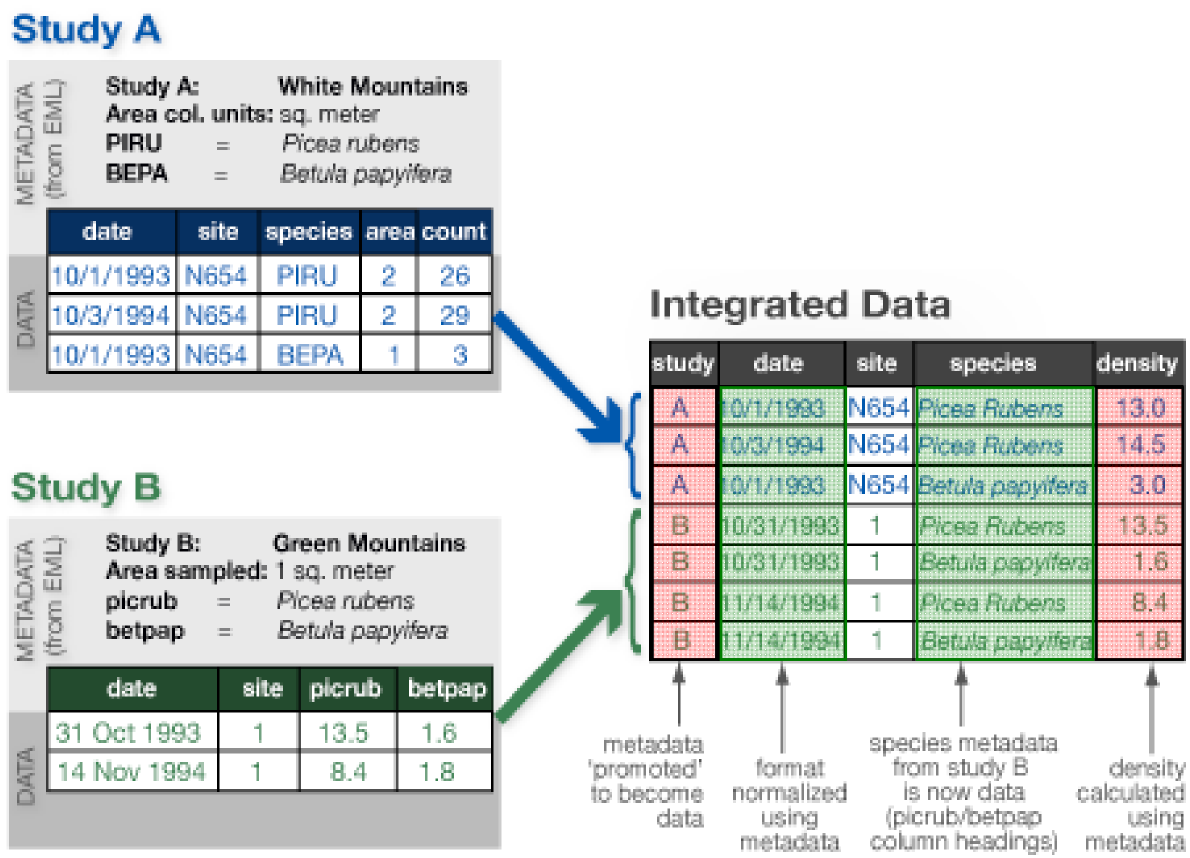


CDL

Data challenge 2: diversity

"the flood of increasingly heterogeneous data"

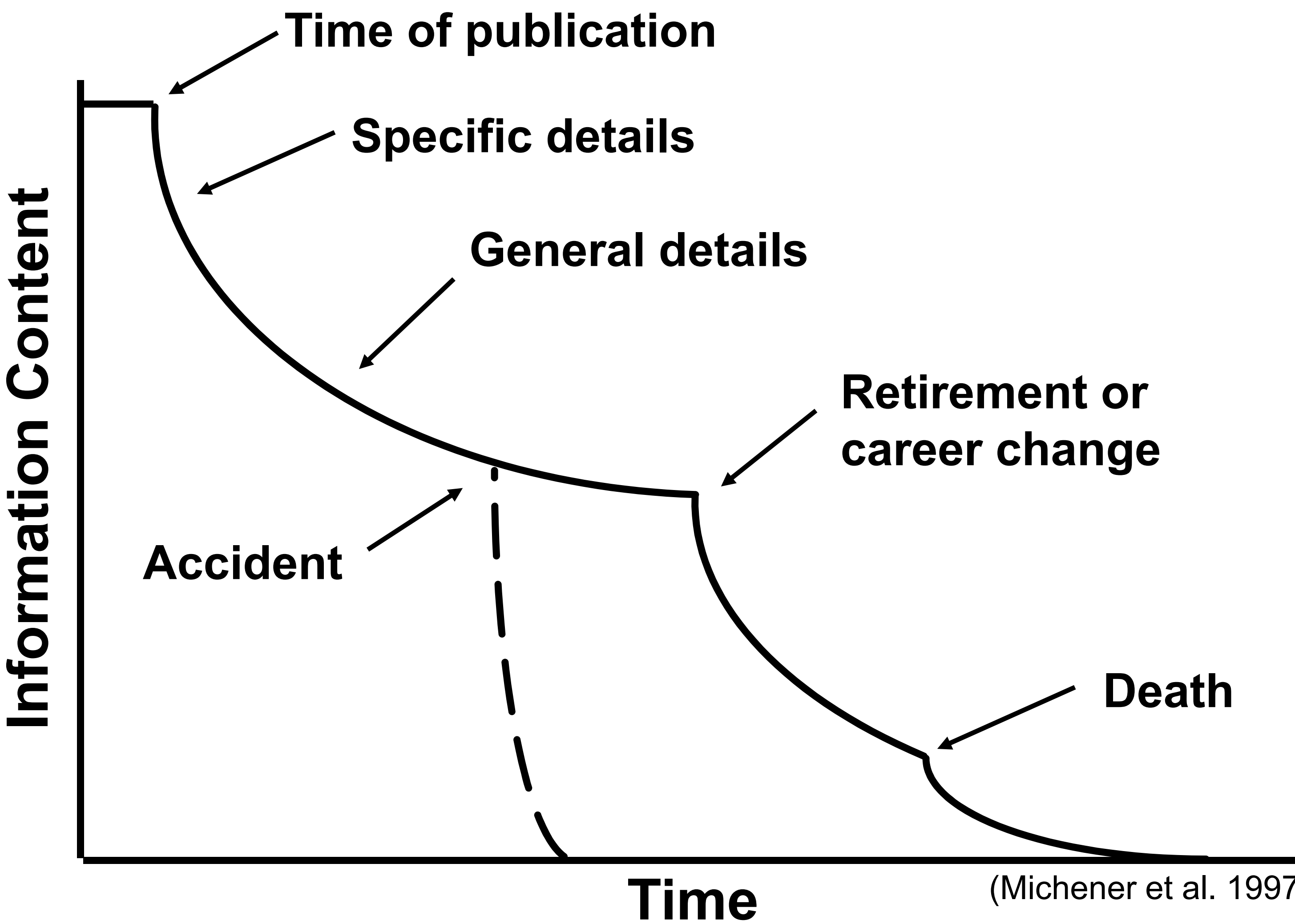
- Data are heterogeneous
  - Syntax
    - (format)
  - Schema
    - (model)
  - Semantics
    - (meaning)



Jones et al. 2007

Data challenge 3: poor practice

"data entropy"

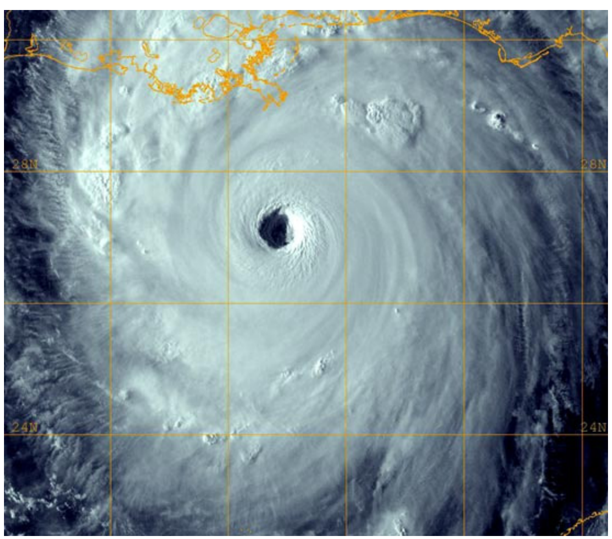


(Michener et al. 1997)

Data challenge 4: loss

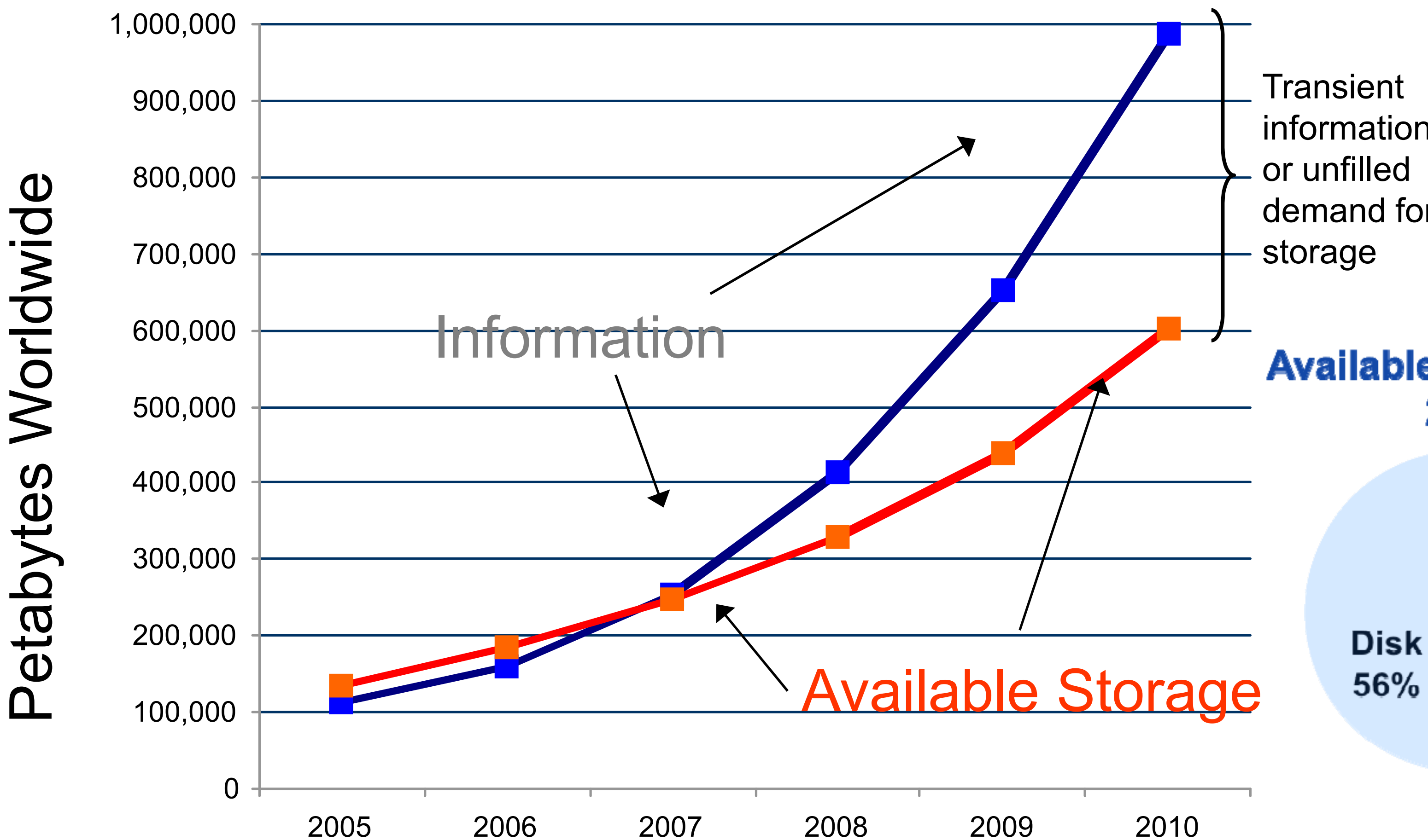
- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing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

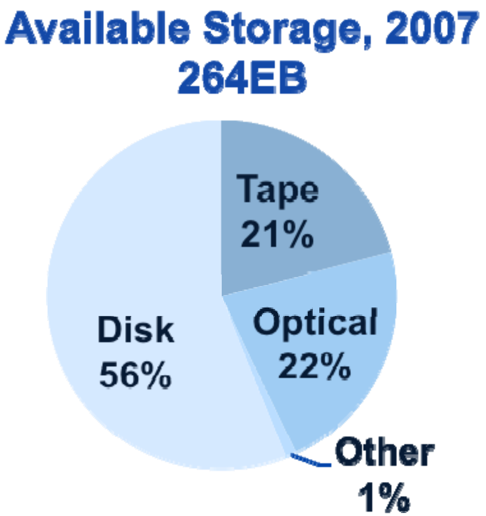
Source: S. Abrams, CDL



CDL

Data challenge 4: more loss





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

CDL

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.

CDL

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

CDL

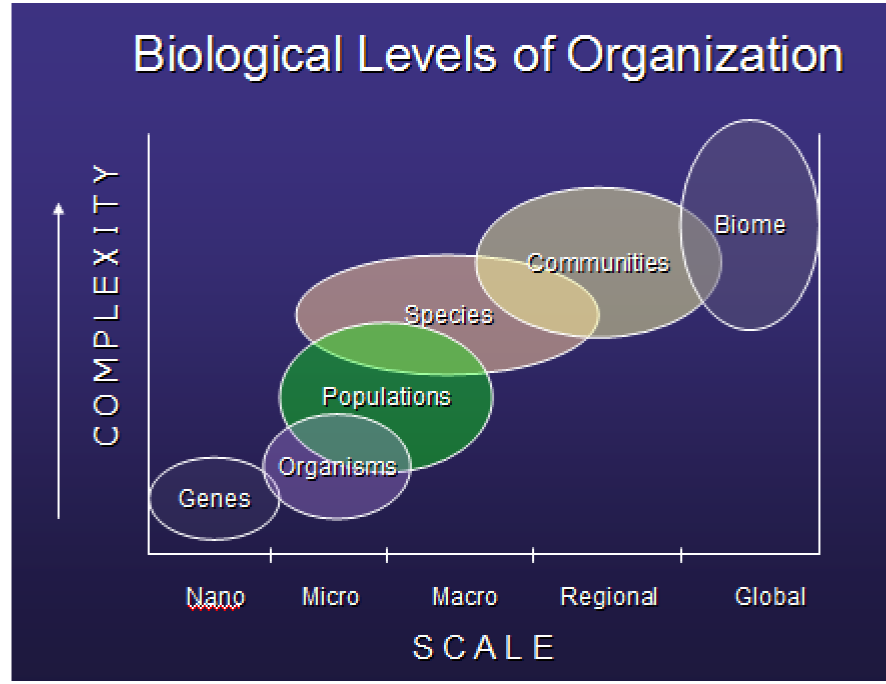
Data Observation Network for Earth

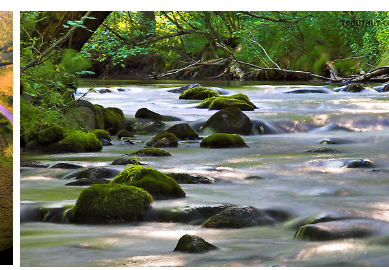
- The goal of DataONE is to enable new science through universal access to data about life on earth by:
  - engaging the scientist in the data preservation process
  - supporting the full data life cycle,
  - encouraging data stewardship and sharing
  - promoting best practices
  - engaging citizens

- One of two DataNet awardees recommended for funding by NSF

Data types

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

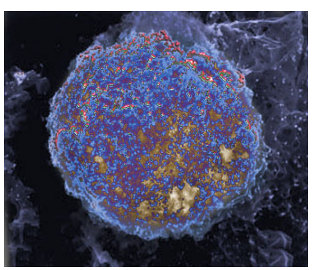


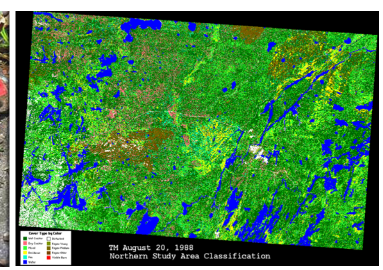


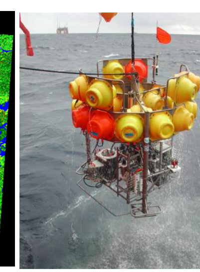
DataONE

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








CDL

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)
nbii National Biological Information Infrastructure	Biodiversity, taxonomic, ecological	BDP, DwC, DC, OGIS
DAAC for biogeochemical dynamics, Distributed Active Archive Center, Oak Ridge National Laboratory	Biogeochemical dynamics, terrestrial ecological Earth observation imagery	DIF, BDP, ECHO
The US Long Term Ecological Research Network	Ecological, biodiversity, biophysical, social, genomics, and taxonomic	EML
Avian Knowledge Network	Avian populations and molecular biology	DC
ALA Atlas of Living Australia	Biological and taxonomic	DC subset
South African Environmental Observation Network	Biophysical, biodiversity, disturbance, and Earth observation imagery	EML
TERN Taiwan Ecological Research Network	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

CDL

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)

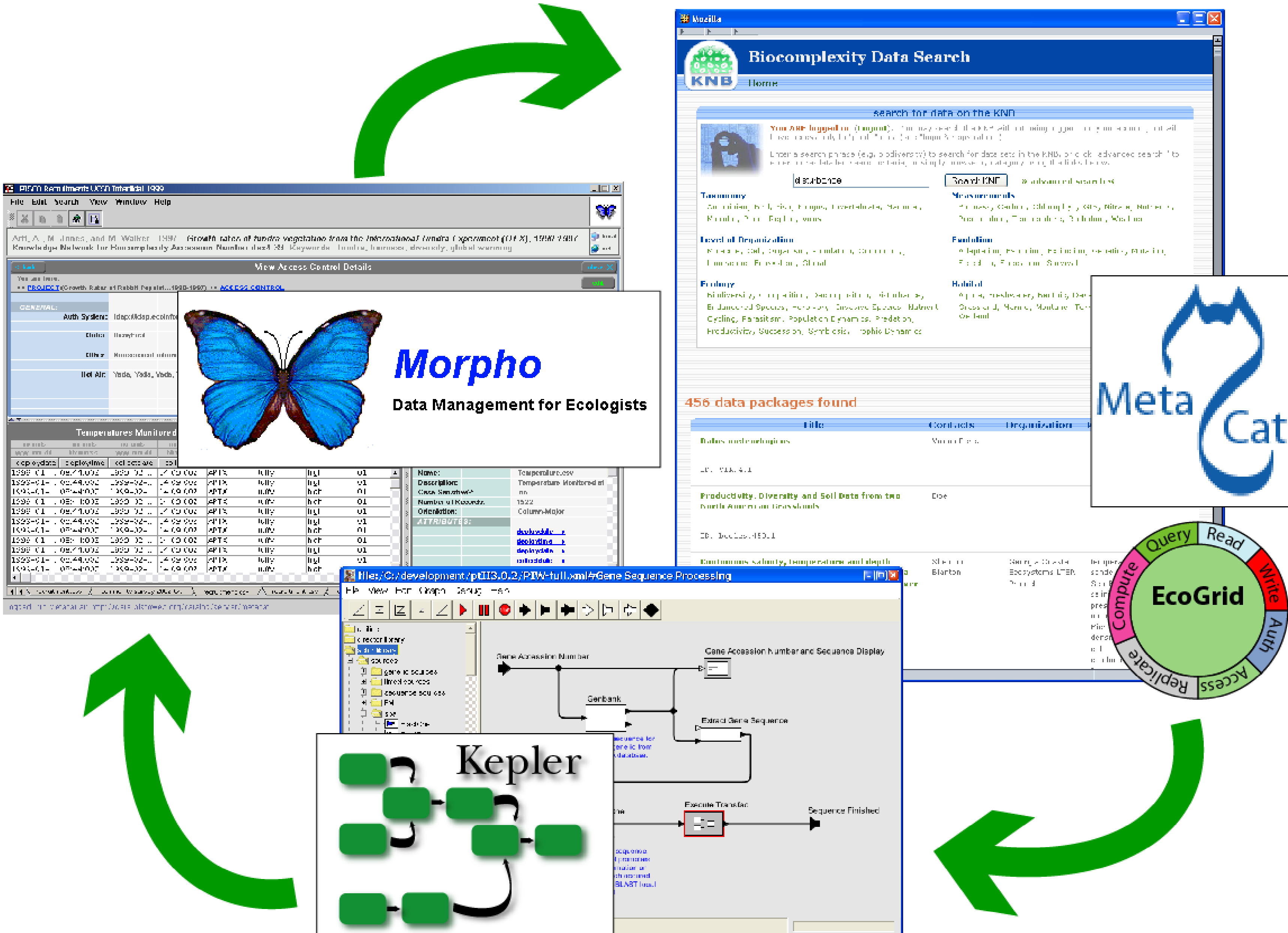
MODIS and ASTER Products (LPDAAC) (38)

National Phenology Network (USANPN) (29)

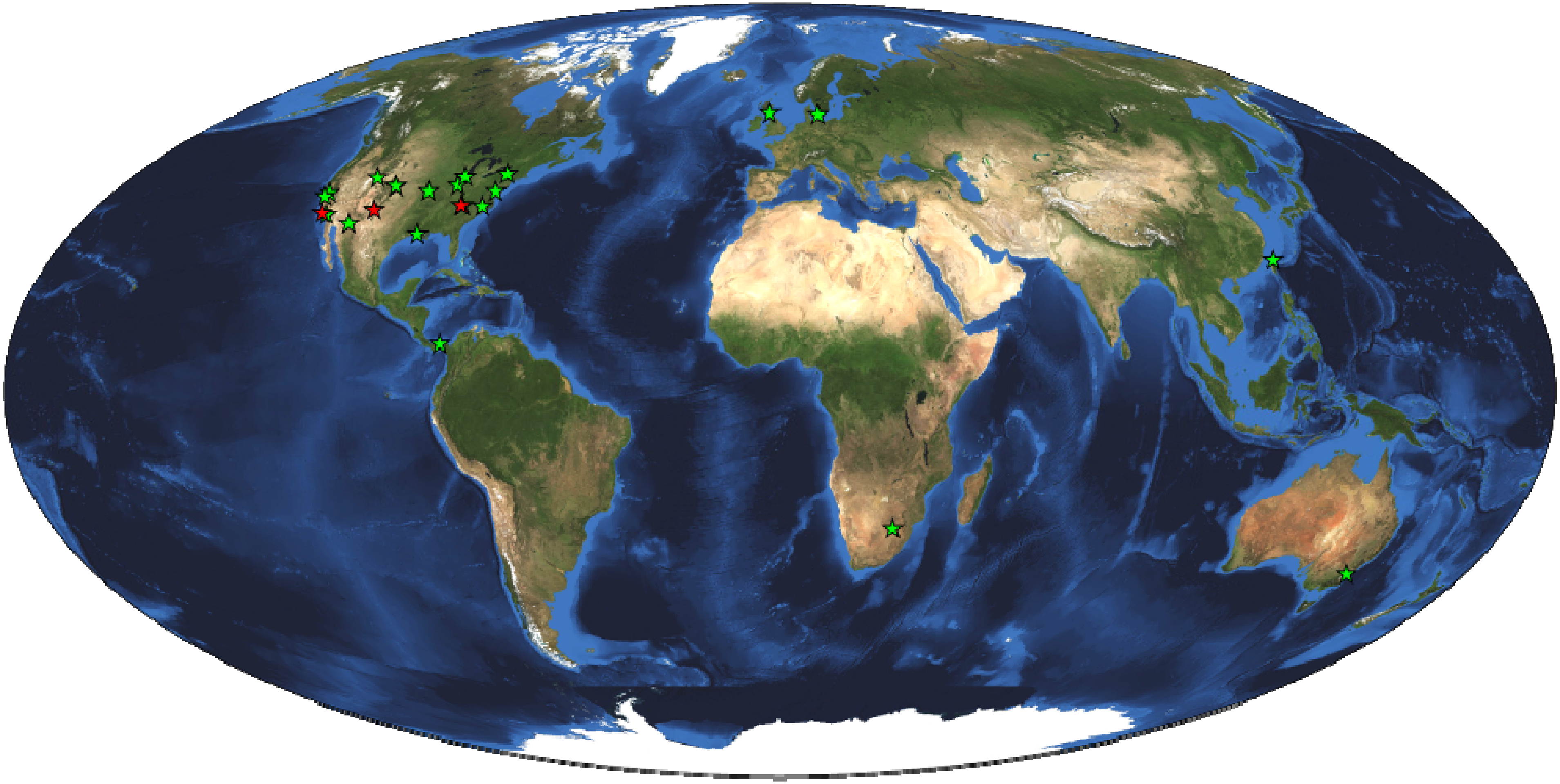
CDL

DataONE

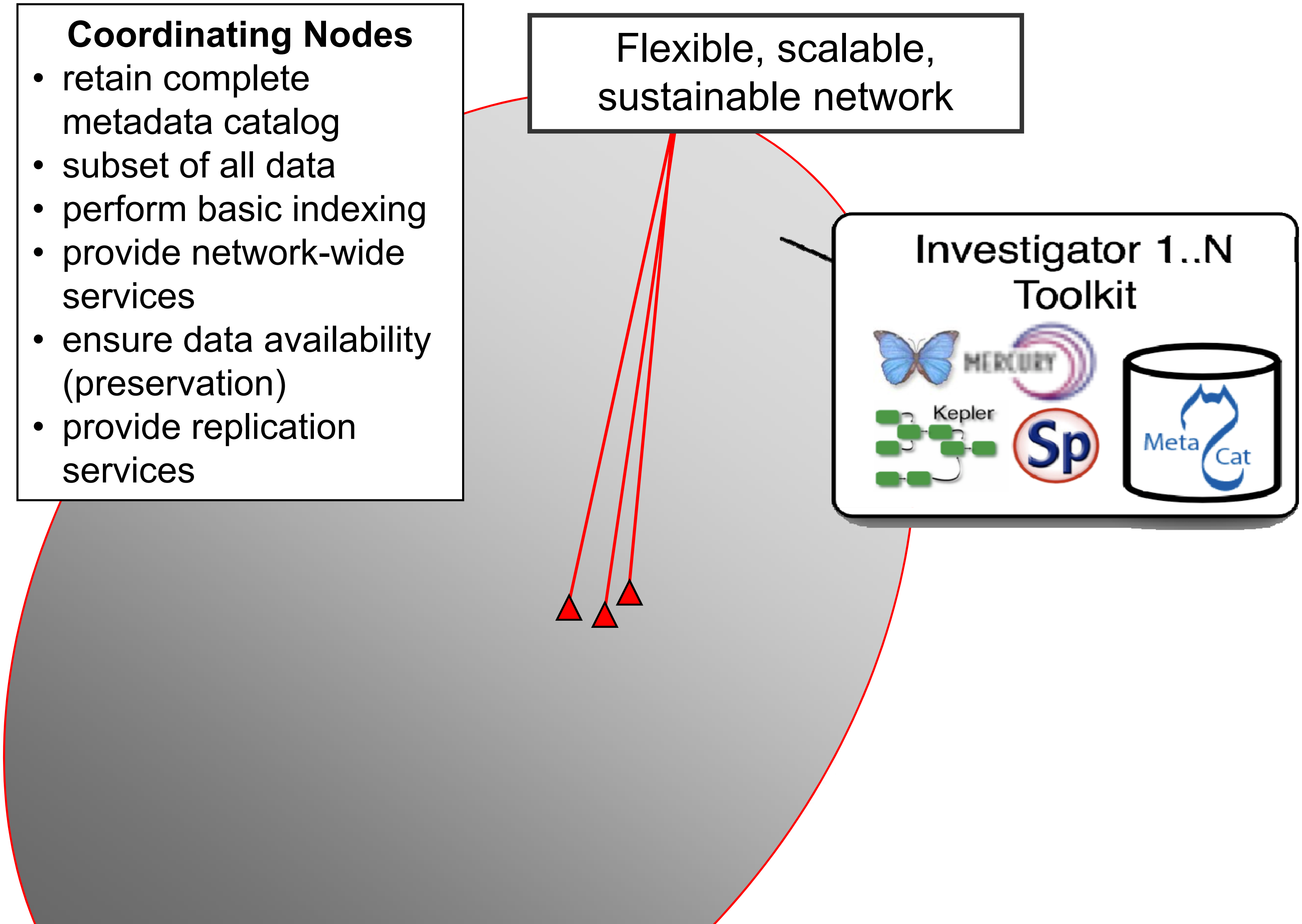
Existing cyberinfrastructure: tools



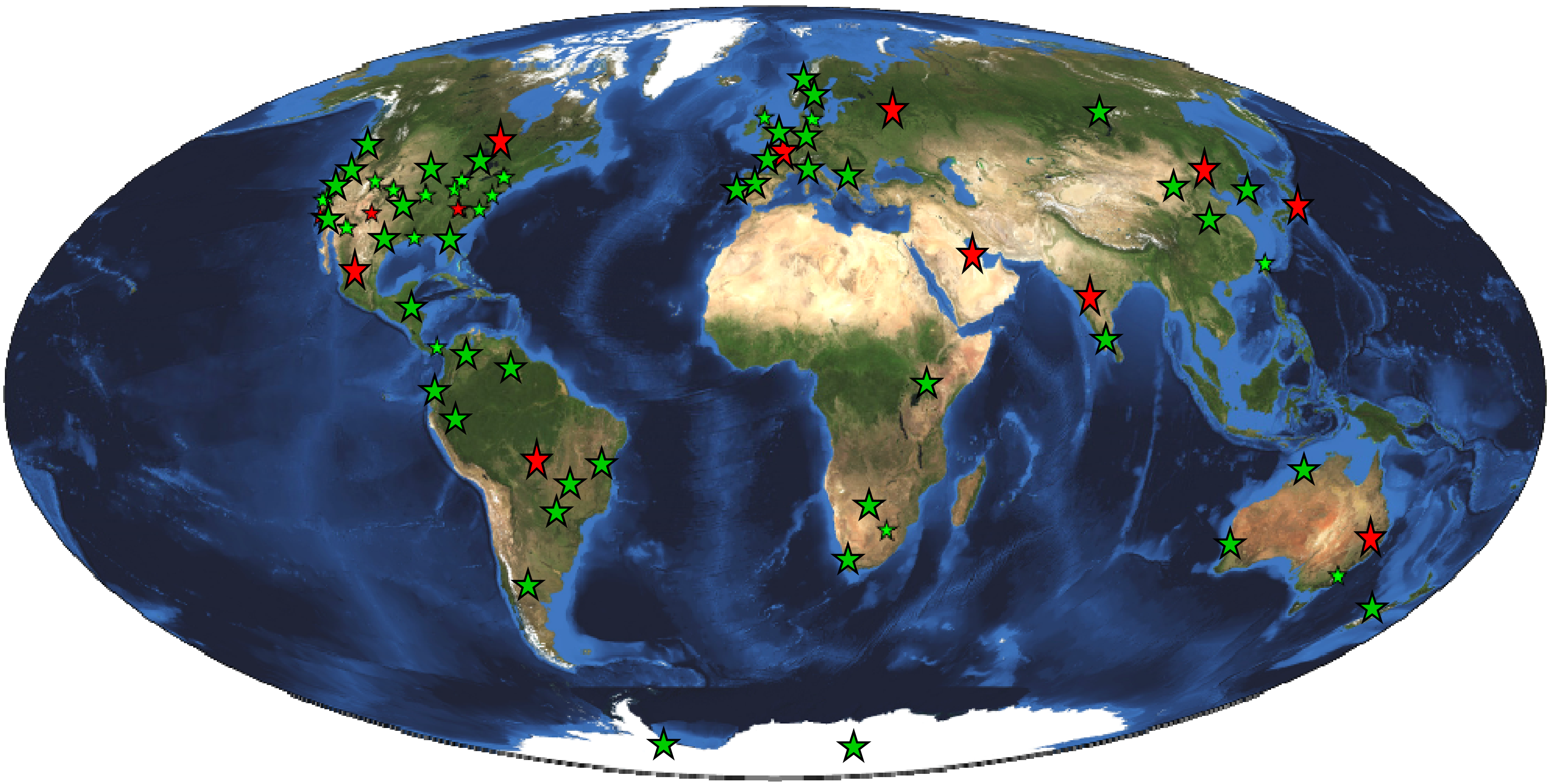
DataONE Building new global cyberinfrastructure



New distributed framework



DataONE Building new global cyberinfrastructure



CDL

DataONE management and partners

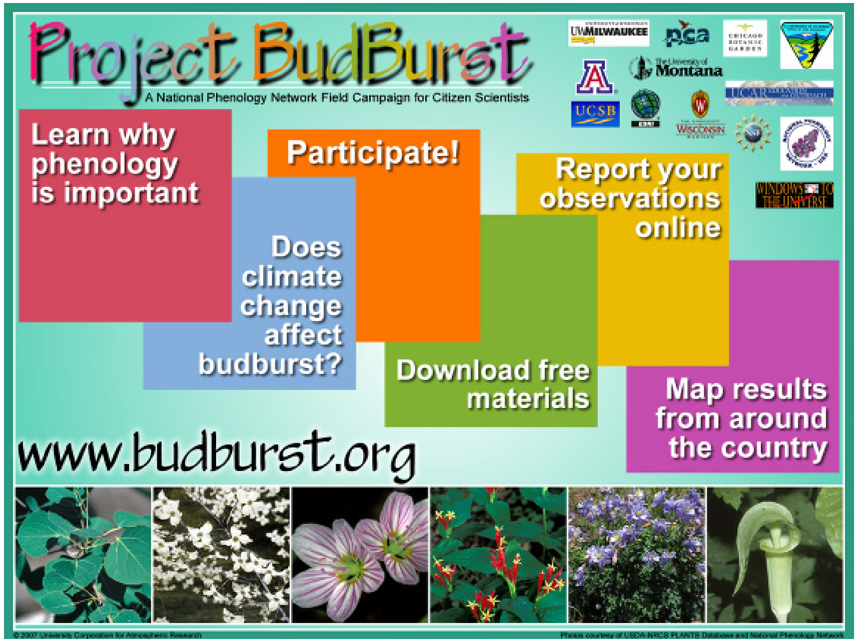
William Michener, University of New Mexico
Suzie Allard – University of Tennessee
Bob Cook – Oak Ridge National Laboratory DAAC
Patricia Cruse – California Digital Library
Mike Frame – USGS, National Biological Info. Infrastructure
Matt Jones – University of California Santa Barbara
Steve Kelling – Cornell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

DataONE Partners plus Kepler-CORE and SEEK/KNB Teams

We welcome your involvement!

DataONE

Engaging citizens in science



eBird

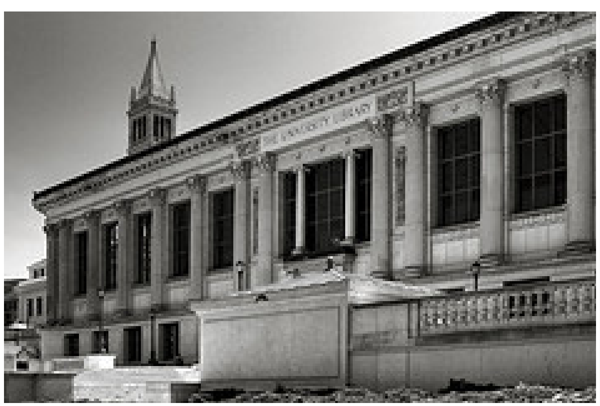
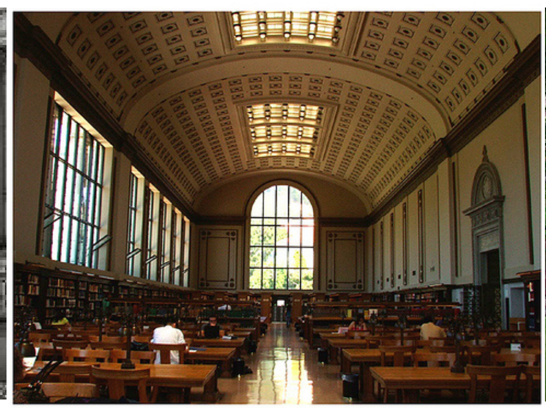
www.CitizenScience.org

USA npn National Phenology Network

CDL

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?

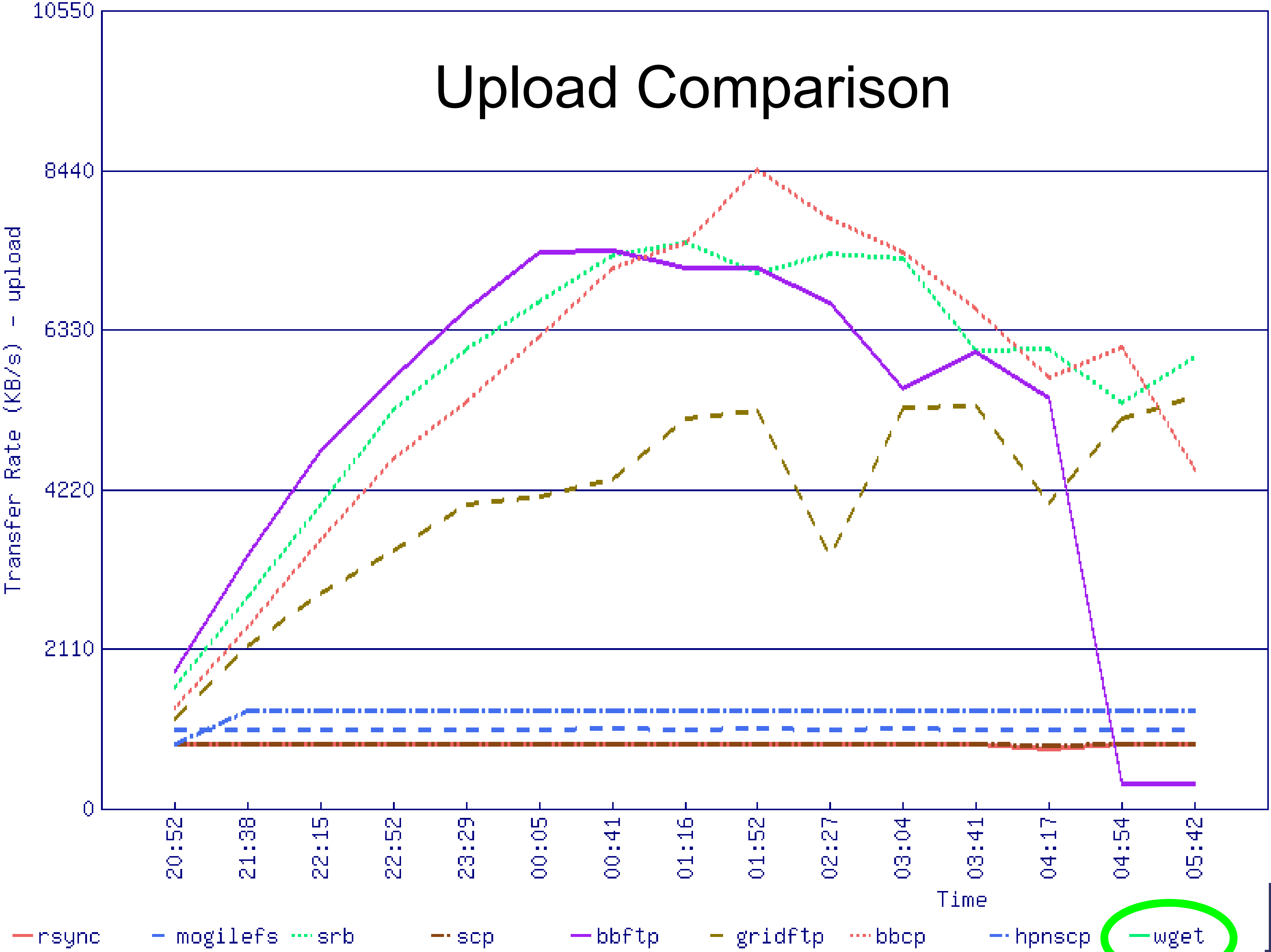
- Unbundled alternative to monolithic systems with single archival “culture”; avoiding the deadly embrace
- Low barrier, low commitment tools
  - Leverage native operating system file handling tools

- 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

CDL



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

CDL

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

CDL

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

CDL

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

CDL

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)

CDL

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

CDL

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)
```

CDL

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

```
C: [opens session]
C: GET http://ark.cdlib.org/ark:/13030/ft167nb0vq? HTTP/1.1
C:
S: HTTP/1.1 200 OK
S: Content-Type: text/plain
S: THUMP-Status: 0.5 200 OK
S:
S: erc:
S: who:   Stanton A. Glantz and Edith D.  Balbach
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

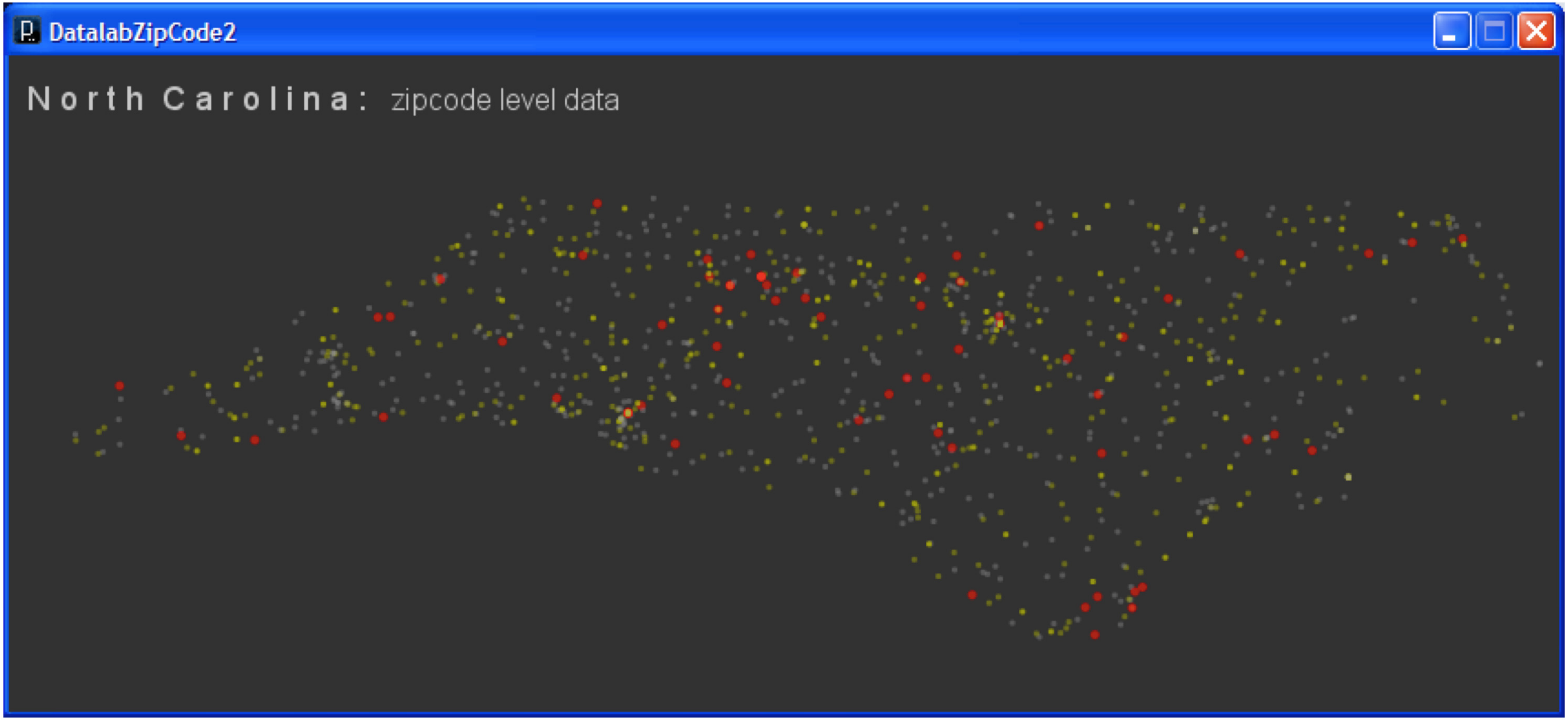
http://www.cdlib.org/inside/diglib/ark/thumpspec.pdf

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

CDL

A sample Flex application plotting data retrieved from DataLab (Kevin Gamiel)

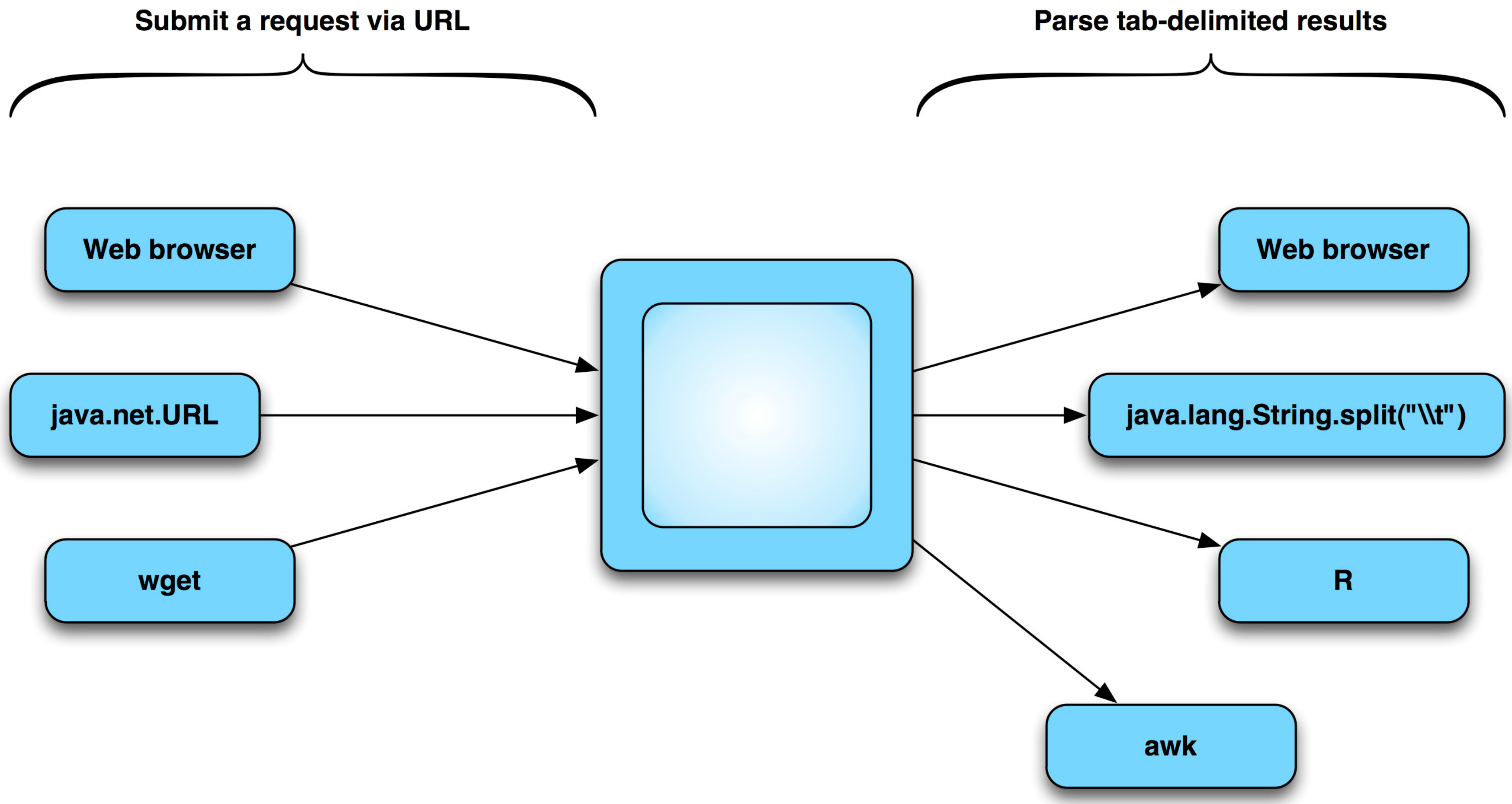
Sample Java visualization by ZIP code



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

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

Integrating THUMP DataLab extensions



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

Source: Nassib Nassar

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
```

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

Why not use XML?

In MathML:

```
<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>
```

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

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!

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