SCIENCE DATA REPOSITORIES (SDRs) ON THE WEB: *AN INITIAL SURVEY*

LAURA MARCIAL

MARCIAL@UNC.EDU

BRAD HEMMINGER

5 March 2010

Rationale

Study

Discussion

Future Work

OBJECTIVES

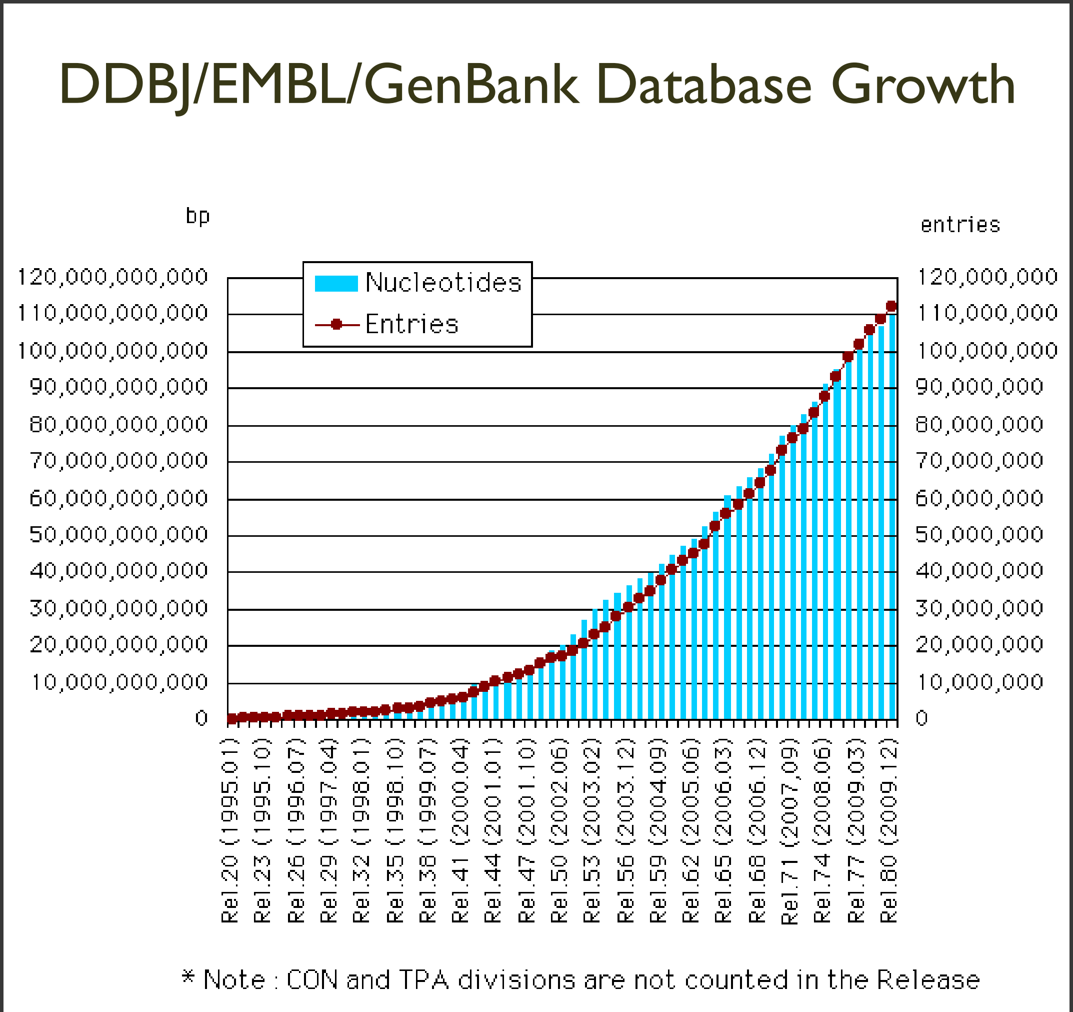
Data
Center
National
Science
Information
Research
University
Network
Archive
Ecological
Genome
Institute
California
Observatory
World
Portal
Repository
Atmospheric
Project
Environmental
Resource
Biodiversity
International
Transplant
Database
Nuclear
Centers
Social
Program
Census
Monitoring
Astronomy
Laboratory
Americas
Human
Survey
Oceanographic
Facility
Analysis
Space
Disease
Tropical
data
Institute's
Forest
Web
Brain
Map
Encyclopedia
Life
System
Synthesis
available
Fuels
West
BioGRID
Institution
Diverse
Nile
ASC
Ensembl
LTER
Quality
Phenology
Visible
ICPSR
NASA/IPAC
HubbleSite
Coastal
WestNile.ca.gov
Scientific
BioSystematic
Craig
AFDC
Hole
Michigan
Melanoma
SkyView
Genbank
Control
NEON
NCEAS
USA-NPN
Store
Joint
Ridge
Geoscientific
CEOS
HEASARC

"Digital data collections are powerful catalysts for progress and for **democratization** of the research and the enterprise."

National Science Board [NSB] Report 2005c

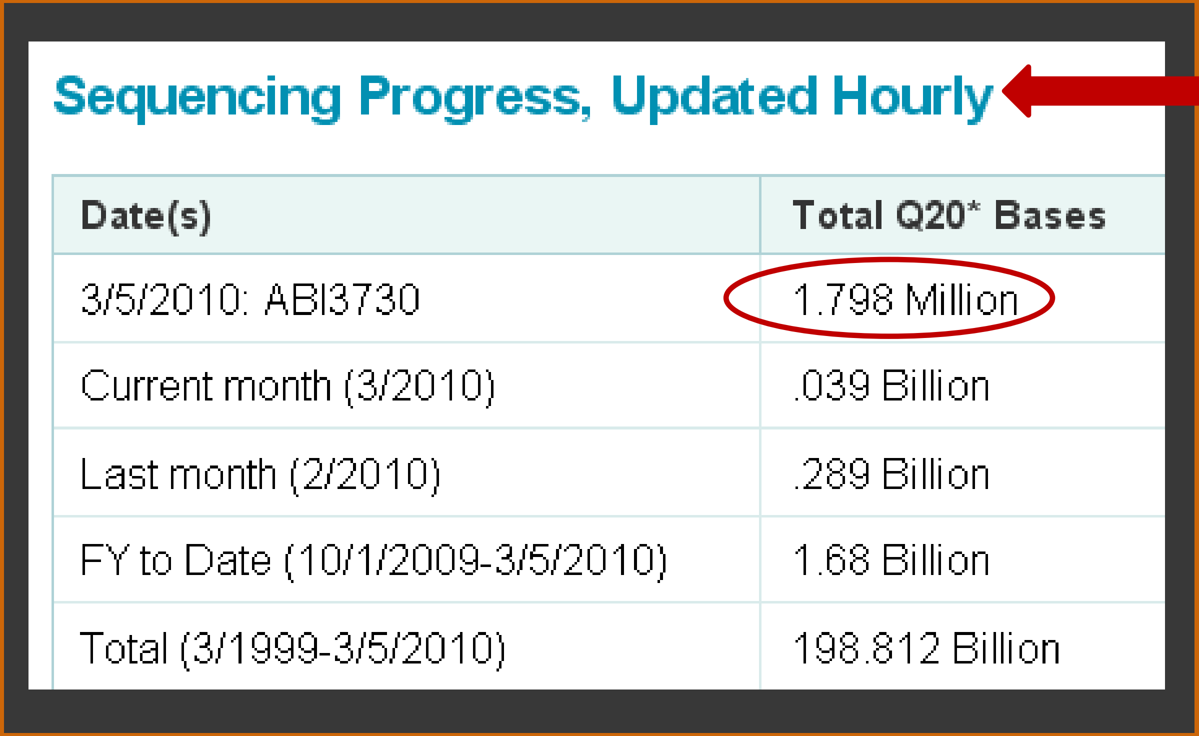
RATIONALE

Total sequencing contributions: 12,000,000,000 base pairs



DDBJ/EMBL/GenBank Database Growth

* Note : CON and TPA divisions are not counted in the Release

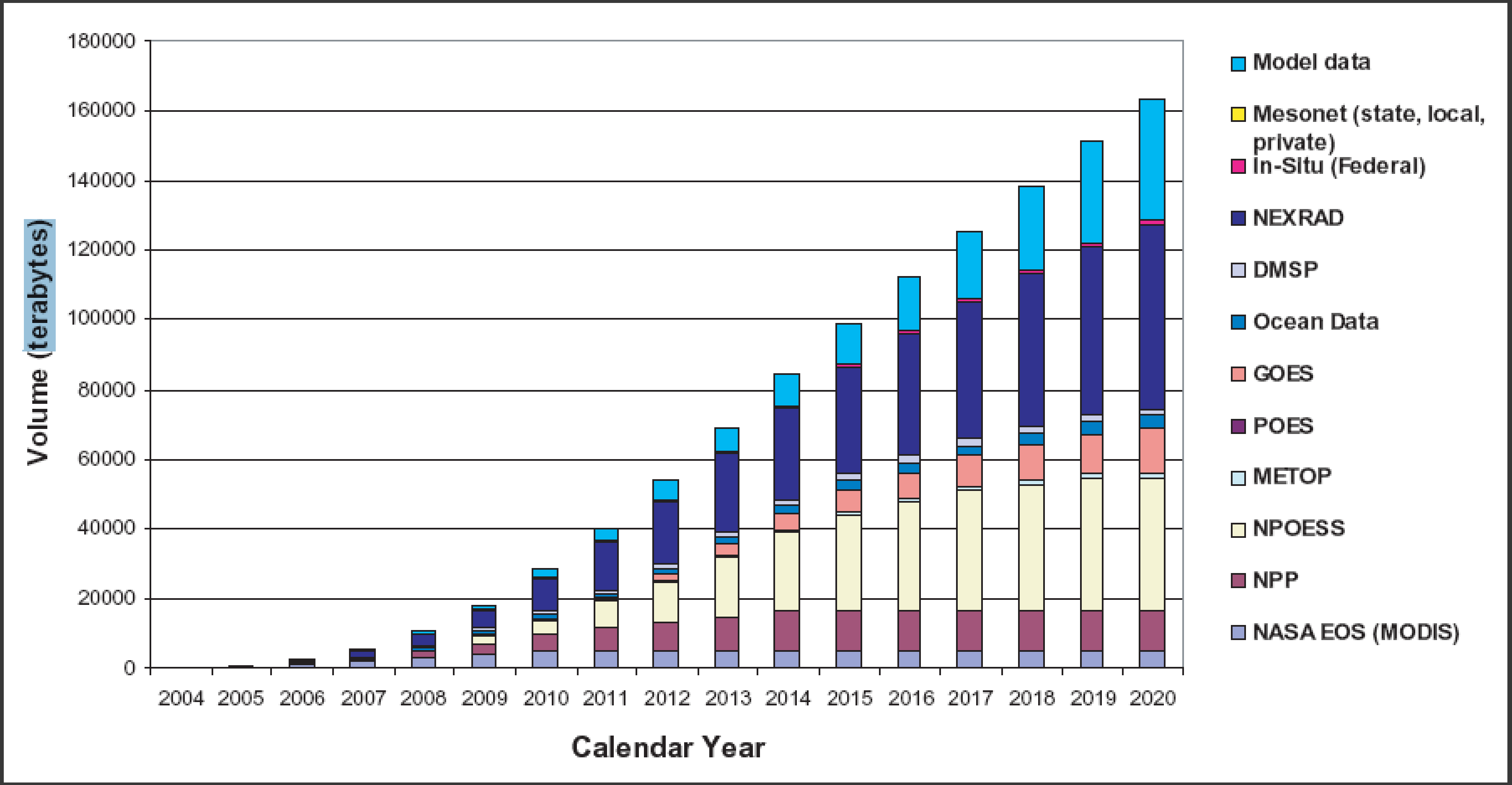


Sequencing Progress, Updated Hourly

Date(s)	Total Q20* Bases
3/5/2010: ABI3730	1.798 Million
Current month (3/2010)	.039 Billion
Last month (2/2010)	.289 Billion
FY to Date (10/1/2009-3/5/2010)	1.68 Billion
Total (3/1999-3/5/2010)	198.812 Billion

Data from the Department of Energy's Joint Genome Institute

JGI generates on the order of 2.3 gigabases of sequence per month or 1 terabyte of data per month



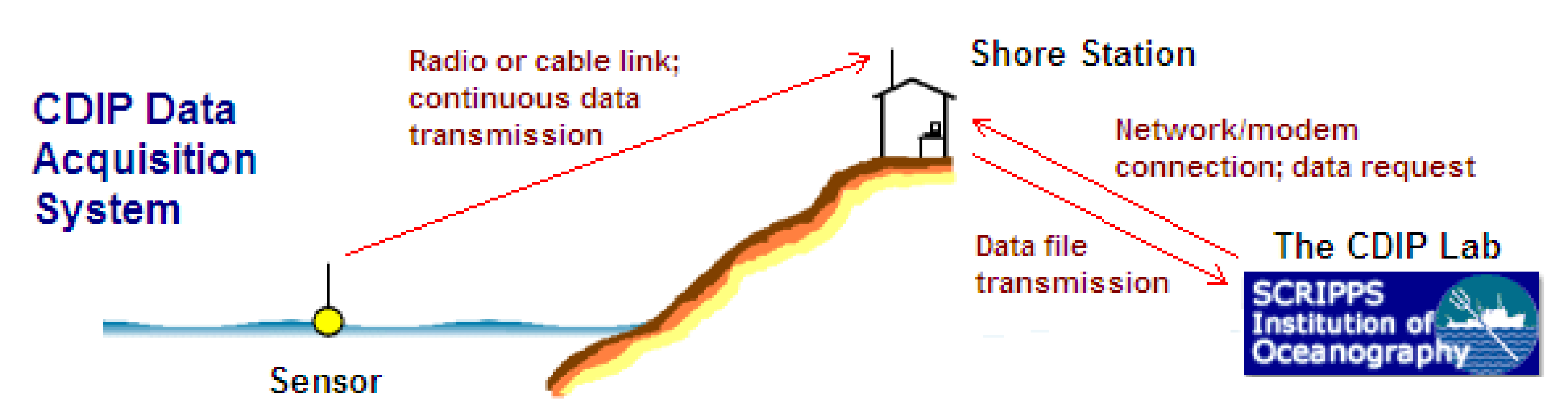
NOAA: Large-array data growth expected over 15 years. Current estimates predict data archive growth to more than 160,000 TB by 2020.

http://www.ngdc.noaa.gov/noaa_pubs/pdf/NOAA_DataManagementReport_Final.pdf

ORIGINS

Hubble Space Telescope: generates10 gigabytes of data per day

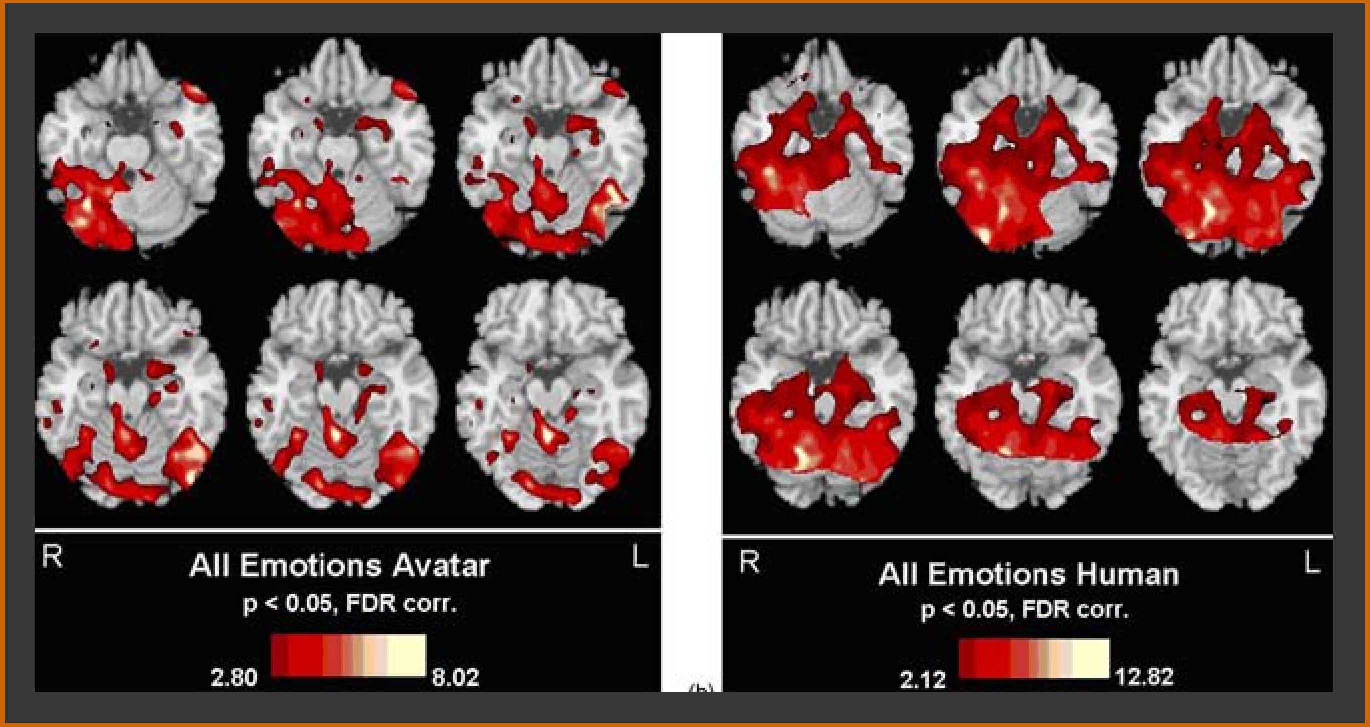
ORIGINS


CDIP Data Acquisition System
Radio or cable link; continuous data transmission
Shore Station
Network/modem connection; data request
Data file transmission
The CDIP Lab
SCRIPPS Institution of Oceanography
Sensor

Coastal Data Monitoring

ORIGINS

Functional Magnetic Resonance Imaging (fMRI)



Amygdala activation at 3T in response to human and avatar facial expressions of emotions.

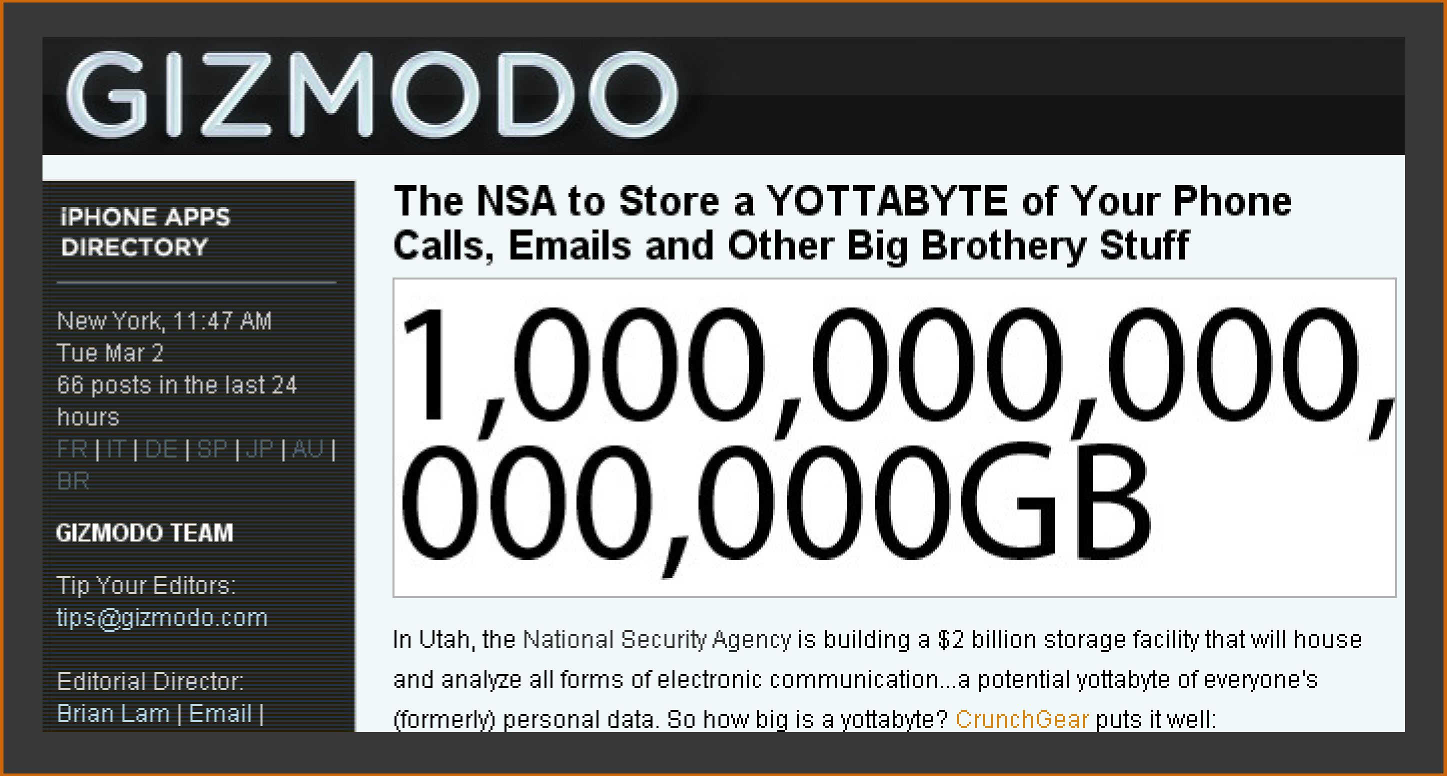
http://evolution.anthro.univie.ac.at/institutes/urbanethology/projects/simulation/fmri/index.html

So, what is happening with all of the Closed Circuit TV (CCTV) data generated every day?

ORIGINS

Clearly, we are entering the yottabyte (YB) era:

1,000,000,000,000,000,000,000,000 (one septillion) bytes



GIZMODO

iPHONE APPS DIRECTORY

New York, 11:47 AM
Tue Mar 2
66 posts in the last 24 hours
FR | IT | DE | SP | JP | AU | BR

GIZMODO TEAM

Tip Your Editors:
tips@gizmodo.com

Editorial Director:
Brian Lam | Email |

The NSA to Store a YOTTABYTE of Your Phone Calls, Emails and Other Big Brothery Stuff

1,000,000,000,000,000,000GB

In Utah, the National Security Agency is building a $2 billion storage facility that will house and analyze all forms of electronic communication...a potential yottabyte of everyone's (formerly) personal data. So how big is a yottabyte? CrunchGear puts it well:

RATIONALE

At least many thousands of SDRs

Often start as government projects

Keys to success are elusive

Highly heterogeneous

Highly domain specific

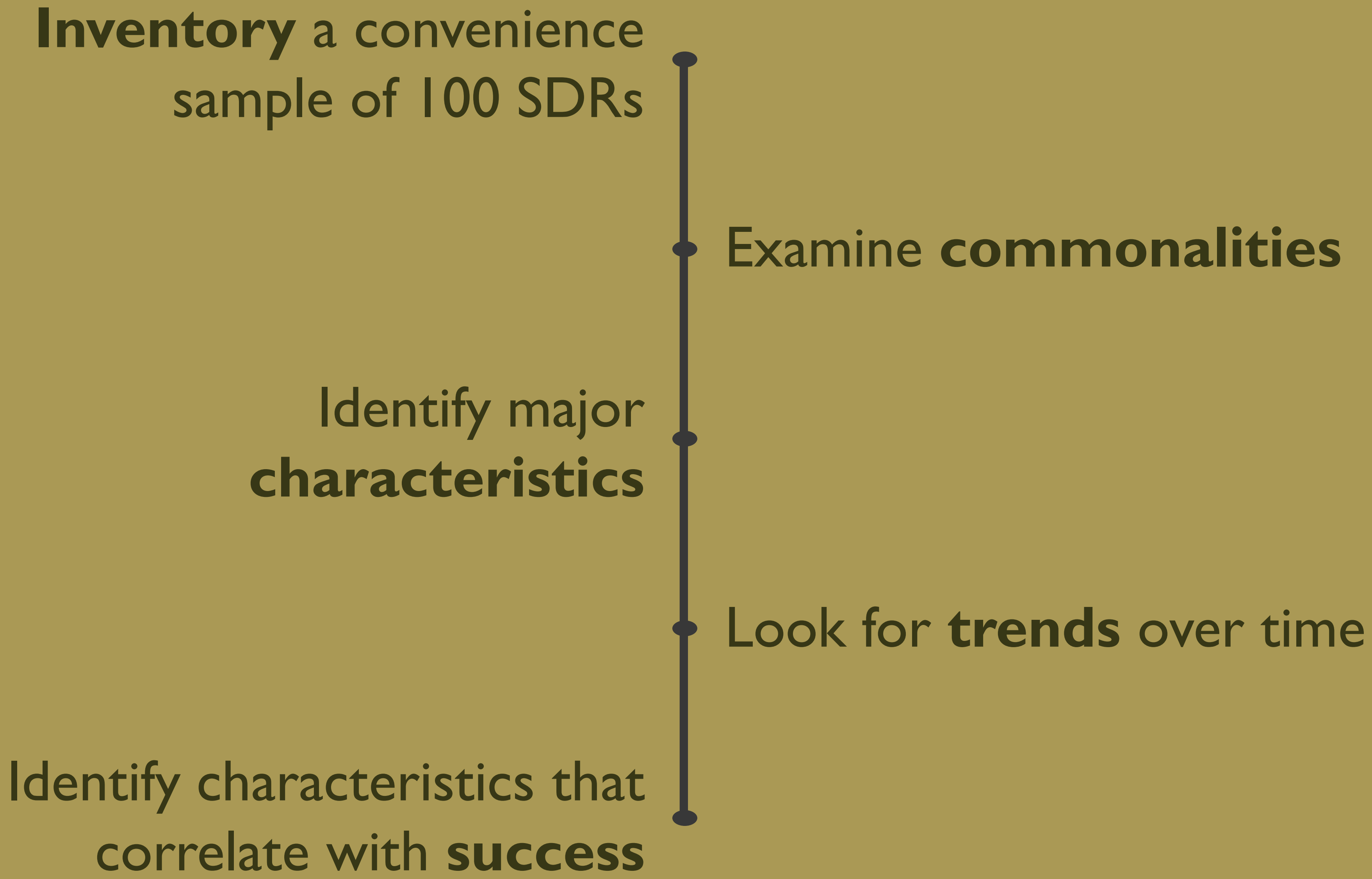
Using
impact
salient
policy
multiple
software
similar
Information
discussed
Google
change
identifies
discovered
regression
fundamental
determine
provided
corrections
comparison
eyes
combination
searches
institutional
domain
undergoing
consulting
sponsorship
potentially
information
particularly
relationship
contracts
development
organizational
composition
reviewed
feedback
Characteristics
Repositories
data
preservation
NSB
websample
work
form
supported
users
stewardship
export
determining
services
ingest
SDR
business
administrators
investigation
Four
study
success
groups
policies
holding
role
support
Science
websites
Data
sustainability
guided
analysis
membership
type
important
observations
National
whether
cluster
logistic
identified
framework
identify
exists
grants
characteristics
group
critical
repositories
examined
comes
management
understand
sponsors
inferential
standpoint
size
conducted
understanding
made
science
archive
attempt
elements
explored
recognized
SDRs
better
using
refinement
Board
four
curation

> "I found it interesting to read your survey results and see what information you inferred about the KNB. It points out areas that we need to improve upon in terms of communication from our web presence."

--Matt Jones, Knowledge Network for Biocomplexity

STUDY

GOALS



Inventory a convenience sample of 100 SDRs

Examine **commonalities**

Identify major **characteristics**

Look for **trends** over time

Identify characteristics that correlate with **success**

TIMELINE

In 2007-2008, **identified 100 SDRs** through Google searches

In 2009, site profiles were sent to site administrators for **review and comment**

2007

2008

2009

2010

Initial review was done to **refine salient characteristics**

50 characteristics were captured, **17 of which were analyzed** using cluster analysis

CHARACTERISTICS of the 50

GENERAL

- Scientific Domain
- Research, Community or Reference
- Holding Size
- Information

BUSINESS

- Governmentally based
- Business Type
- Memberships or Subscriptions

DATA DETAILS

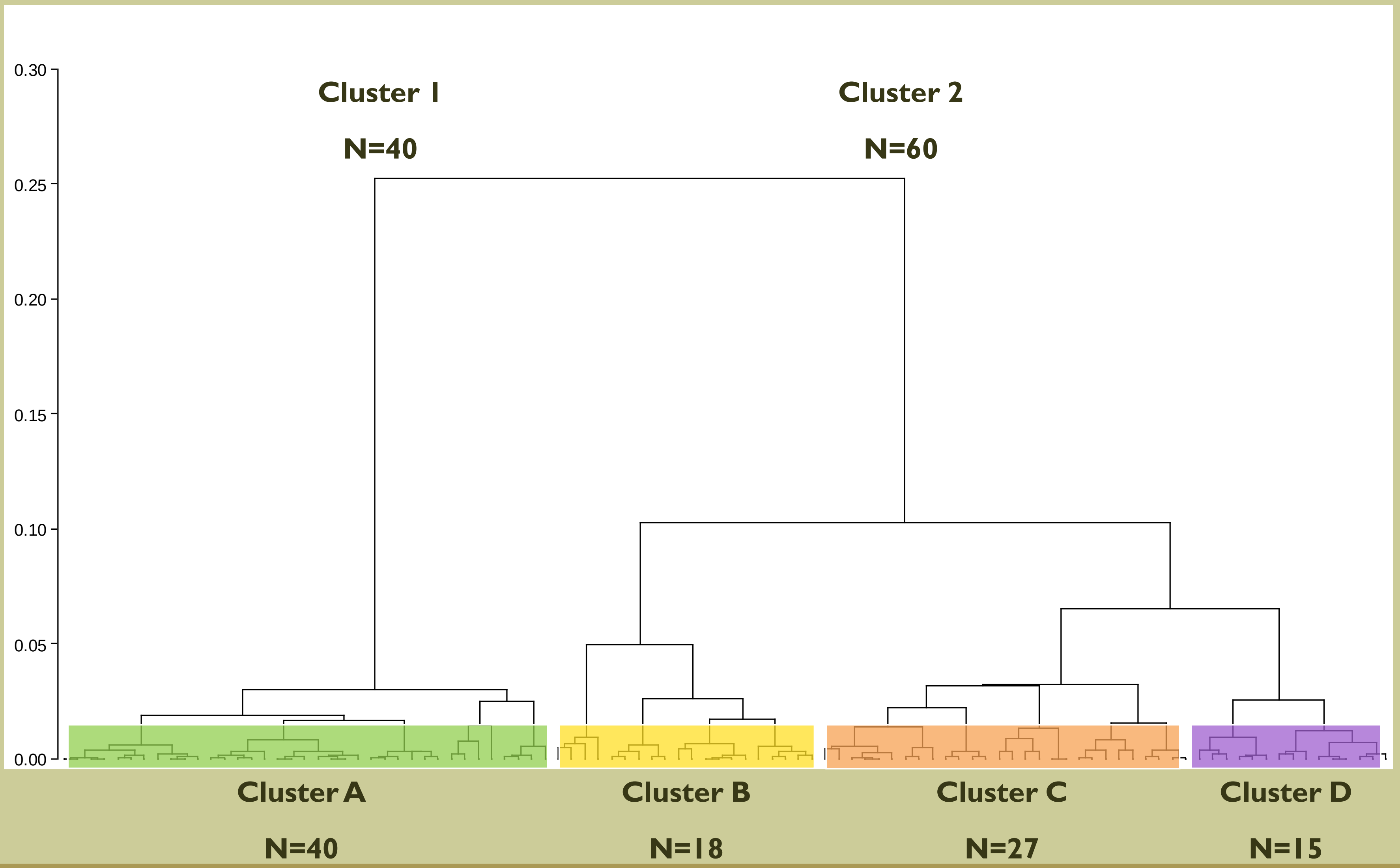
- Deposits and Access
- Representation
- Ingest Methods
- Metadata
- Preservation
- Additional Services
- Usage Statistics

ANALYSIS

The 17 characteristics suitable for analysis and their data type

#	Characteristic	Type
1	Natural Science	Binary
2	Science Area	Nominal
3	Virtual	Binary
4	Holding Size	Ordinal
5	Research/Community/Reference	Nominal
6	Centralized/Distributed	Binary
7	Instrument Based	Binary
8	Business Type	Nominal
9	Subscription or Membership	Binary
10	How Based	Nominal
11	Multi-Sponsored	Binary
12	Grants & Contracts	Binary
13	Accept Submitted Data	Binary
14	Registration Required	Ordinal
15	Free in the Public Domain	Ordinal
16	Preservation Policy	Binary
17	Portal	Binary

CLUSTER RESULTS



Cluster 1

N=40

Cluster 2

N=60

Cluster A

N=40

Cluster B

N=18

Cluster C

N=27

Cluster D

N=15

CLUSTER A

Agency for Healthcare Quality and Research

Multimission Archive at STScI (MAST)

Alternative Fuels Data Center (AFDC)

NASA Langley Atmospheric Science Data Center

Atlantic Oceanographic and Meteorological Laboratory (AOML) Environmental Data Server or ENVIDS

NASA/IPAC Infrared Science Archive (IRSA)

Atmospheric Radiation Monitoring (ARM) Data Centers

National Ecological Observatory Network (NEON)

Carbon Dioxide Information Analysis Center (CDIAC)

National Nuclear Data Center Nuclear Data Portal

Centers for Disease Control and Prevention Data and Statistics

National Space Science Data Center

Climate and Environmental Retrieval and Archive (CERA) for the WDCC

Natural Resource and GIS Metadata and Data Store of the National Park Service

Chandra data archive

Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)

Comprehensive Epidemiological Data Resource (CEDR)

Planetary Data System (PDS)

Controlled Fusion Atomic Data Center (CFADC)

Renewable Resource Data Center (RReDC)

DNA Data Bank of Japan (DDBJ)

Solar Data Analysis Center (SDAC) at NASA Goddard Space Flight Center

DOE Joint Genome Institute's (JGI) Genome Web Portal

SkyView

DOE's Energy Information Administration (EIA)

Smithsonian Tropical Research Institute's (STRI) Center for Tropical Forest Science (CTFS)

European Southern Observatory (ESO) Archive Facility

U.S. Transuranium and Uranium Registries (USTUR)

Genbank

United States Census Bureau

Geodata.gov

US National Virtual Observatory (NVO)

NASA’s High Energy Astrophysics Science Archive Research Center (HEASARC)

US Transplant -- Scientific Registry of Transplant Recipients

HubbleSite Gallery

Visible Human Project®

NOAA's Integrated Coral Observing Network (ICON)

World Data Center (WDC)

Integrated Monitoring Network

World Data Center (WDC) for Biodiversity and Ecology

CLUSTER B

BioSystematic Database of World Diptera (BDWD)
CalSurv, the California Vectorborne Disease Surveillance System
Ecological Society of America's Ecological Archives
European Molecular Biology Laboratory - European Bioinformatics Institute or EMBL-EBI
Encyclopedia of Astronomy and Astrophysics
Ensembl
International Council for Science : Committee on Data for Science and Technology
Iubio
J. Craig Venter Institute
Jaspar
Journal of Applied Econometrics (JAE) Data Archive
National Center for Ecological Analysis and Synthesis (NCEAS) Data Repository
NC One Map
Spec Patterns
The BioGRID
The Sanger Institute

ACE Science Center (ASC)

Antarctic Glaciological Data Center (AGDC)

Astronomy Digital Image Library

Brain biodiversity bank at Michigan State University

Bugwood Network

Center for International Earth Science Information Network (CIESIN)

Chesapeake Bay Environmental Observatory (CBEO) Portal

Coastal Data Information Program (CDIP) of the Scripps Institution of Oceanography, University of California at San Diego

Cornell University Geospatial Information Repository

Forestry Images

Henry A. Murray Research Archive MRA)

IAU Minor Planet Center

Inter-university Consortium for Political and Social Research (ICPSR)

IQSS Dataverse network

LTER Network

McIDAS

Melanoma Molecular Map Project

Repository for Archiving, Managing and Accessing Diverse Data (RAMADDA)

Socioeconomic Data and Applications Center (SEDAC)

Space Science and Engineering Center (SSEC) Data Center, University of Wisconsin-Madison

The Howard W. Odum Institute for Research in Social Science

The USA National Phenology Network (USA-NPN)

Thematic Realtime Environmental Distributed Data Services (THREDDS) Data Server

Unidata Program at the University Corporation for Atmospheric Research (UCAR)

University of California Santa Cruz Genome Bioinformatics

Woods Hole Oceanographic Institute Data Center

World Data Center for Human Interactions in the Environment

CLUSTER C

CLUSTER D

Amphibian Ark Team Portal

Discover Life in America's Great Smoky Mountains National Park's All Taxa Biodiversity Inventory

Encyclopedia of Life

fMRI Data Center

Global Biodiversity Information Facility

Knowledge Network for Biocomplexity (KNB)

Mouse Genome Informatics

NEEScentral

Netlib

Ocean Biogeographic Information System (OBIS)

Paleobiology Database

PANGAEA® - Publishing Network for Geoscientific and Environmental Data

Tree of Life Web Project

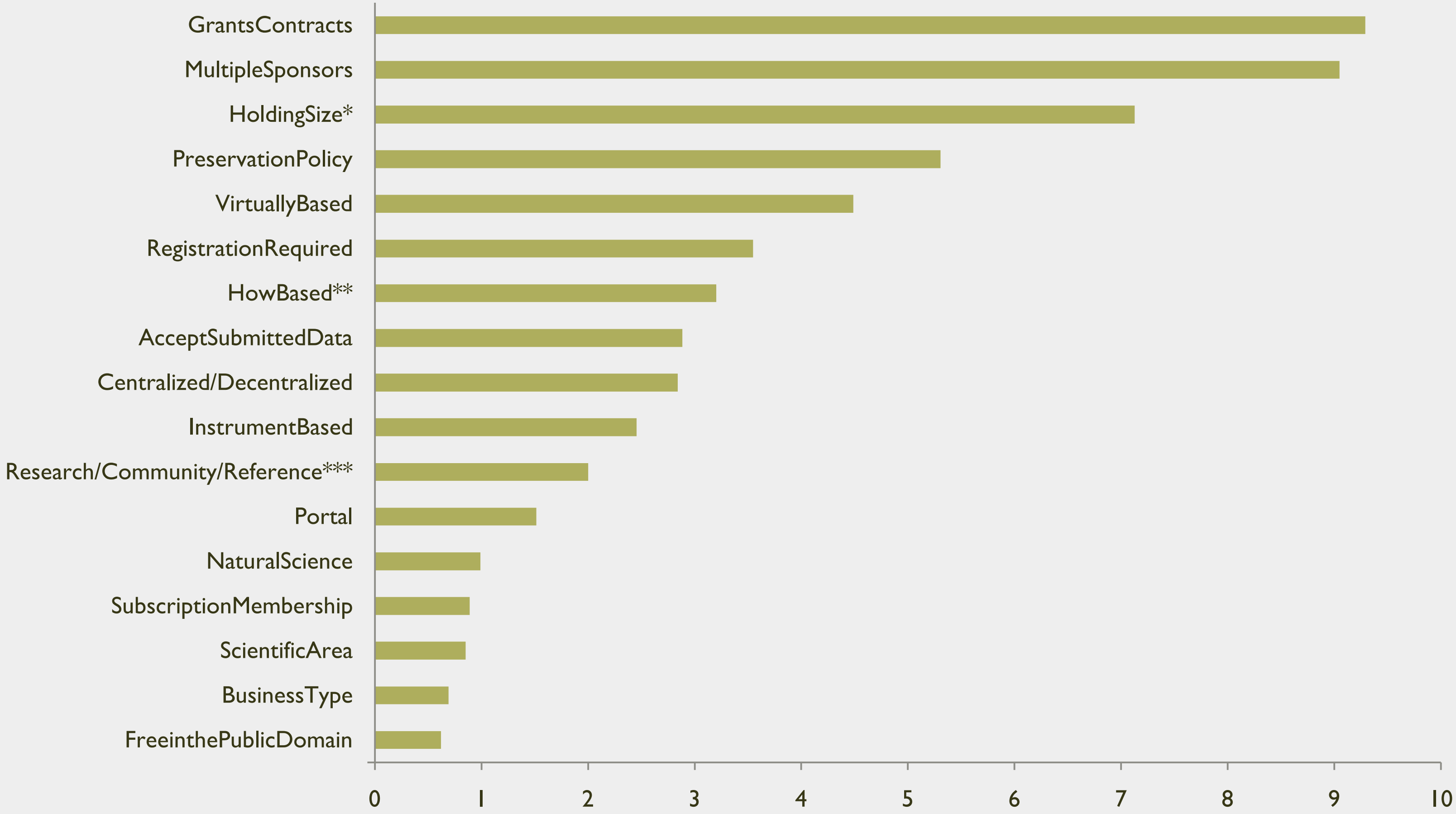
Treebase, Treebase2

VegBank, a vegetation plot database

LOGISTIC REGRESSION

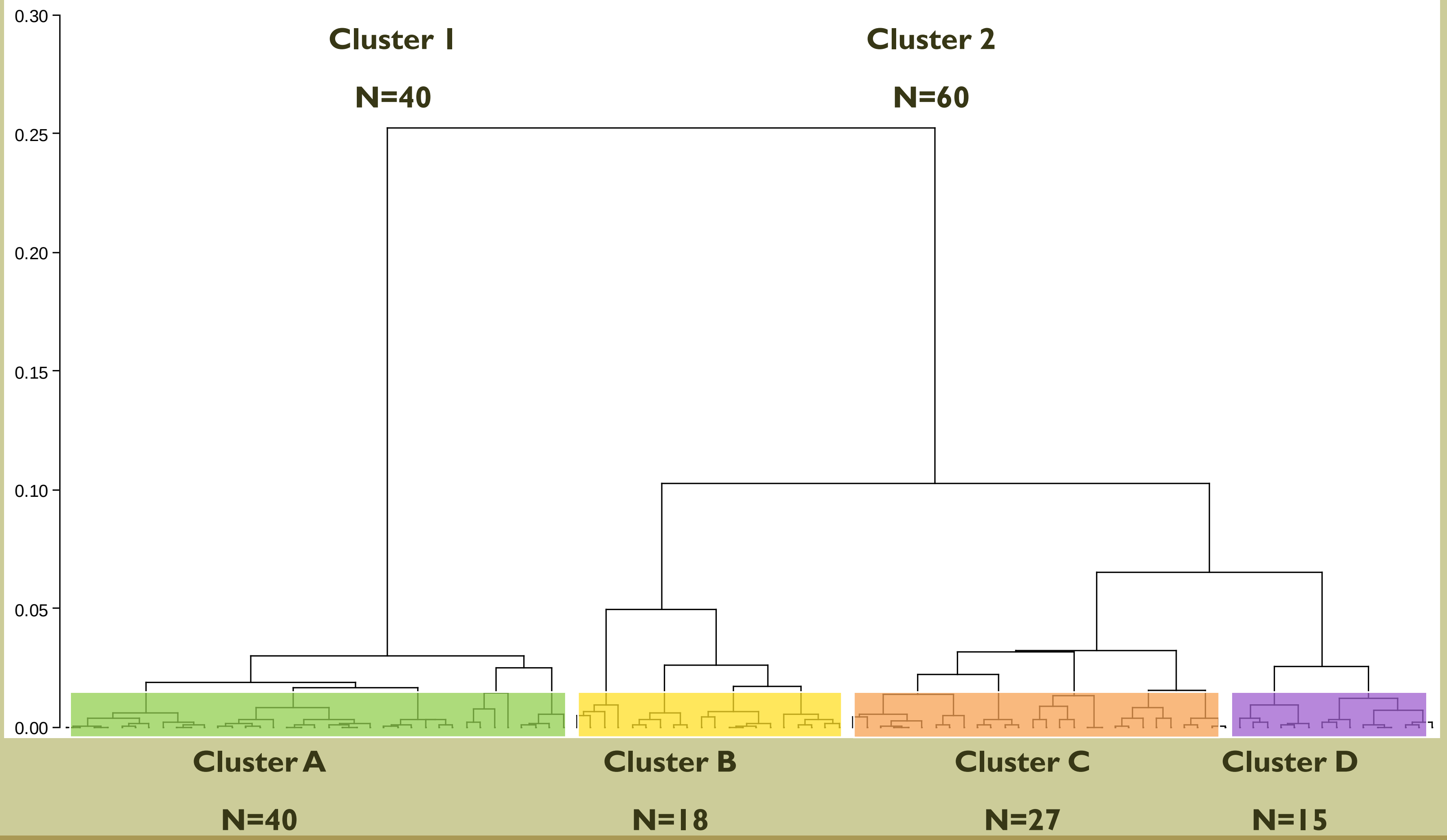
Relative contribution of variables

(measured using simple logistic regression Wald Chi-Square/df)



GROUP COMPOSITION

Variables	Cluster A: 'Governmental'	Cluster B: 'Medicine/Small'	Cluster C: 'University'	Cluster D: Community 'Biology'
Grants Contracts	**No**	Mixed	Yes	Yes
Multiple Sponsors	**No**	Yes	Yes	Mixed
Holding Size	**Large**	**Small**	Mixed	**Moderate**
Preservation Policy	Yes	Mixed	Yes	**No**
Virtually Based	No	**Mixed**	No	No
Registration Required	No	No	**Mixed**	No
How Based	**Governmental**	Mixed	**University**	Mixed
Accept Submitted Data	**Mixed**	Yes	Yes	Yes
Centralized/Distributed	Mixed	Mixed	mixed	**Distributed**
Instrument Based	**Mixed**	No	No	No
Res/Com/Ref	Research	Mixed	Research	**Community**
Portal	Mixed	**No**	mixed	Mixed
Natural Science	Yes	Yes	Yes	Yes
Subscription Membership	No	No	No	No
Scientific Area	Mixed	**Medicine**	Mixed	**Biology**
Business Type	Federal Center	Mixed	**University**	**Partnership**
Free in the Public Domain	Yes	Yes	Yes	Yes

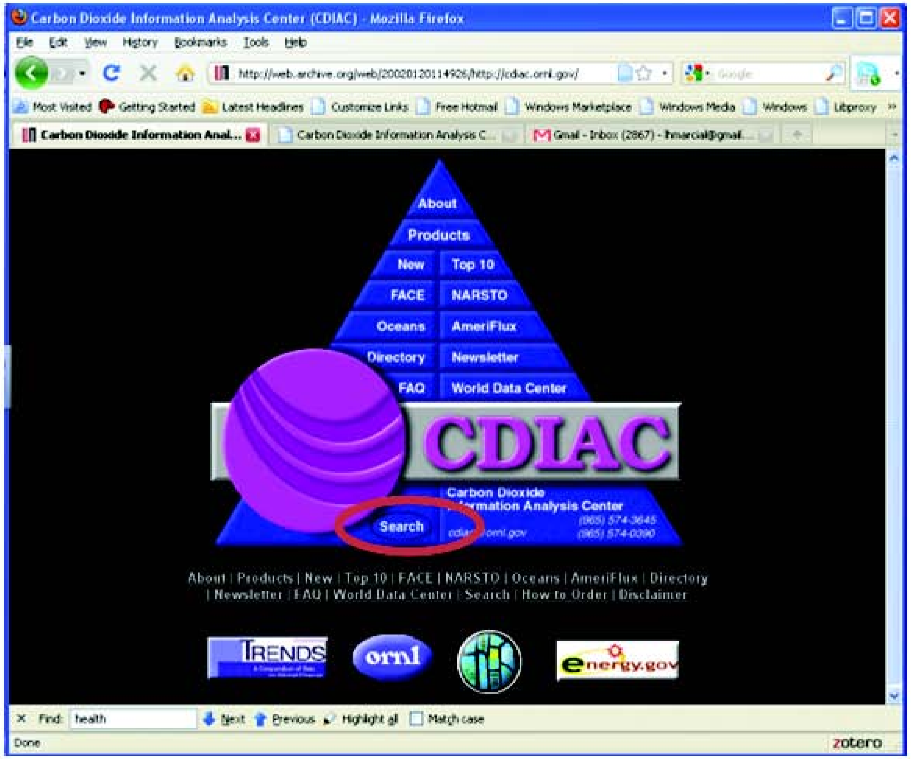


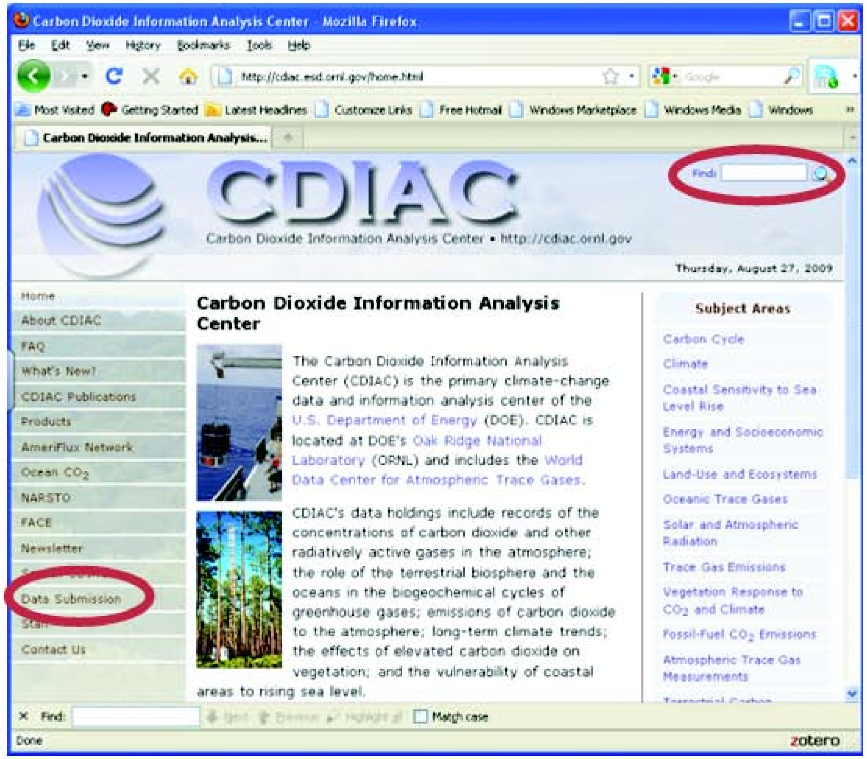
CLUSTER RESULTS

If we are all about studying success
and
SUCCESS = performance over time

How can we study SDRs over time?

ALTERNATIVE SAMPLING


Carbon Dioxide Information Analysis Center (CDIAC) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://web.archive.org/web/20020120114926/http://cdiac.ornl.gov/
Most Visited Getting Started Latest Headlines Customize Links Free Hotmail Windows Marketplace Windows Media Windows Libproxy
Carbon Dioxide Information Anal...
Carbon Dioxide Information Analysis C...
Gmail - Inbox (2867) - lhmarcial@gmail...
About
Products
New
Top 10
FACE
NARSTO
Oceans
AmeriFlux
Directory
Newsletter
FAQ
World Data Center
CDIAC
Carbon Dioxide Information Analysis Center
(865) 574-3645
(865) 574-0390
Search
cdiac@ornl.gov
About | Products | New | Top 10 | FACE | NARSTO | Oceans | AmeriFlux | Directory | Newsletter | FAQ | World Data Center | Search | How to Order | Disclaimer
TRENDS
ornl
energy.gov
Find: health
Next
Previous
Highlight all
Match case
Done
zotero


Carbon Dioxide Information Analysis Center - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://cdiac.esd.ornl.gov/home.html
Most Visited Getting Started Latest Headlines Customize Links Free Hotmail Windows Marketplace Windows Media Windows
Carbon Dioxide Information Analysis...
Find:
CDIAC
Carbon Dioxide Information Analysis Center • http://cdiac.ornl.gov
Thursday, August 27, 2009
Home
About CDIAC
FAQ
What's New?
CDIAC Publications
Products
AmeriFlux Network
Ocean CO2
NARSTO
FACE
Newsletter
Data Submission
Staff
Contact Us
Subject Areas
Carbon Cycle
Climate
Coastal Sensitivity to Sea Level Rise
Energy and Socioeconomic Systems
Land-Use and Ecosystems
Oceanic Trace Gases
Solar and Atmospheric Radiation
Trace Gas Emissions
Vegetation Response to CO2 and Climate
Fossil-Fuel CO2 Emissions
Atmospheric Trace Gas Measurements
Find:
Next
Previous
Highlight all
Match case
Done
zotero

Carbon Dioxide Information Analysis Center

The Carbon Dioxide Information Analysis Center (CDIAC) is the primary climate-change data and information analysis center of the U.S. Department of Energy (DOE). CDIAC is located at DOE's Oak Ridge National Laboratory (ORNL) and includes the World Data Center for Atmospheric Trace Gases.

CDIAC's data holdings include records of the concentrations of carbon dioxide and other radiatively active gases in the atmosphere; the role of the terrestrial biosphere and the oceans in the biogeochemical cycles of greenhouse gases; emissions of carbon dioxide to the atmosphere; long-term climate trends; the effects of elevated carbon dioxide on vegetation; and the vulnerability of coastal areas to rising sea level.

WAYBACK MACHINE

Email
within
Ingest
Statement
Portal
vehicle
Sciences
contact
Instrument
Type
Name
FileFormatTransmittedCode
data
Natural
submitted
Governmental
Information
Collection
Services
FileFormatAcceptedCode
updtd
Centralized/Distributed
Res/Ref/Com
Fees
Sponsors
accepted
S/M
policy
Multiple
based
Export
areas
Submission
Use
Methods
Science
Description
Preservation
organization
size
Url
Copyright
Area
Details
type
Subscription/Membership
Registration
Inception
FPD
REV-R/C/R
required
Business
Processing
details
history
formats
Funding
Supported
Restrictions
transmitted
Structure
Management
Holding
collections
Rights
file
Attribution
Gov/DFed/Univ
Accept
science
category
G/C
Access
Virtual
sponsors
business
Data
Partners
Brief
Import
model
Date

"I am interested in your preservation policy line. We don't have a policy explicitly listed, though we do hope and aim to make the data permanently preserved. Could you provide me with some examples of preservation policies so that we might create one?"

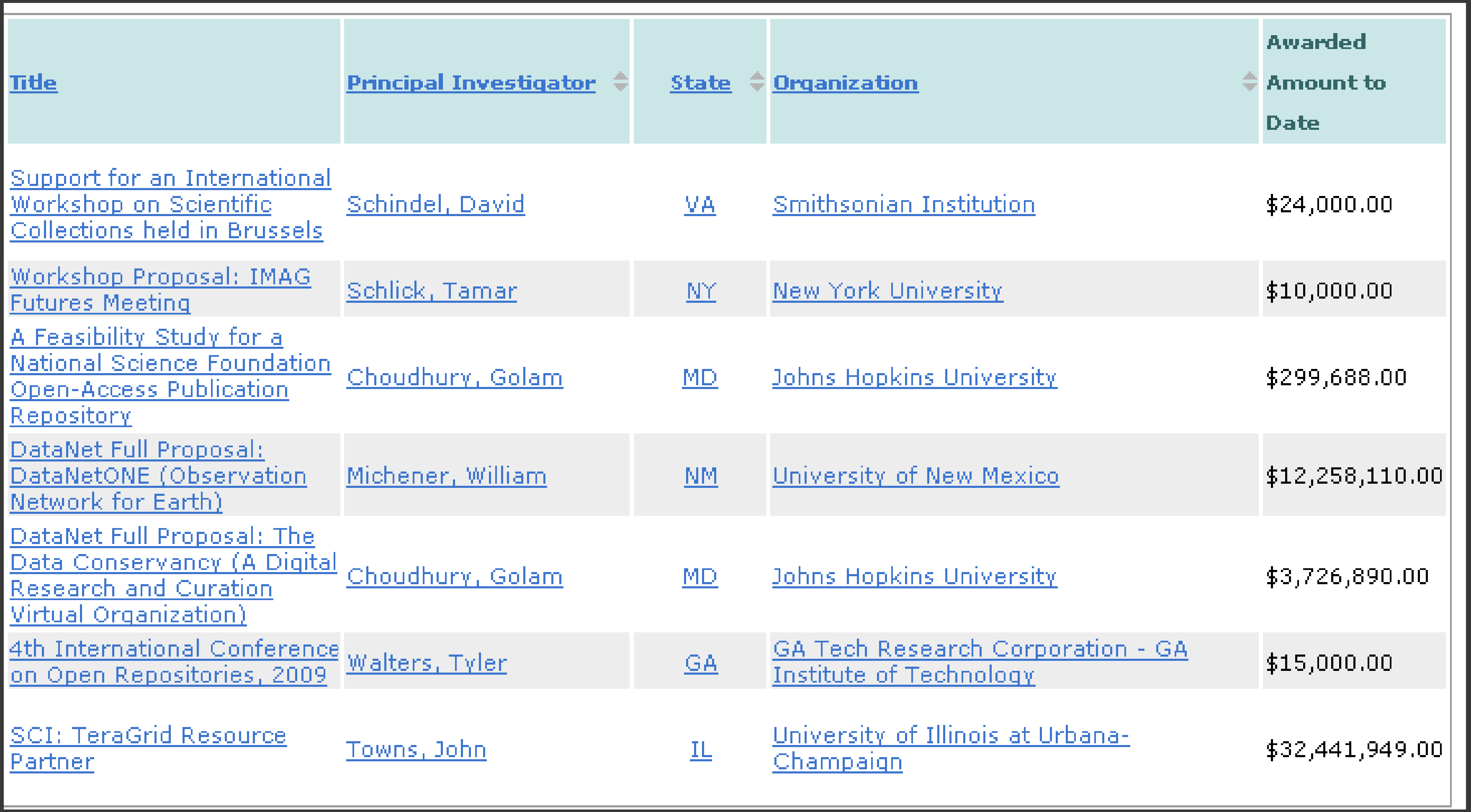
--Michael Lee, VegBank

DISCUSSION

Preservation Policy = any mention of long term data storage

Although preservation ranked as the fourth most important variable (taken independently) in defining group membership, what we did not find was at least as important as what we did.

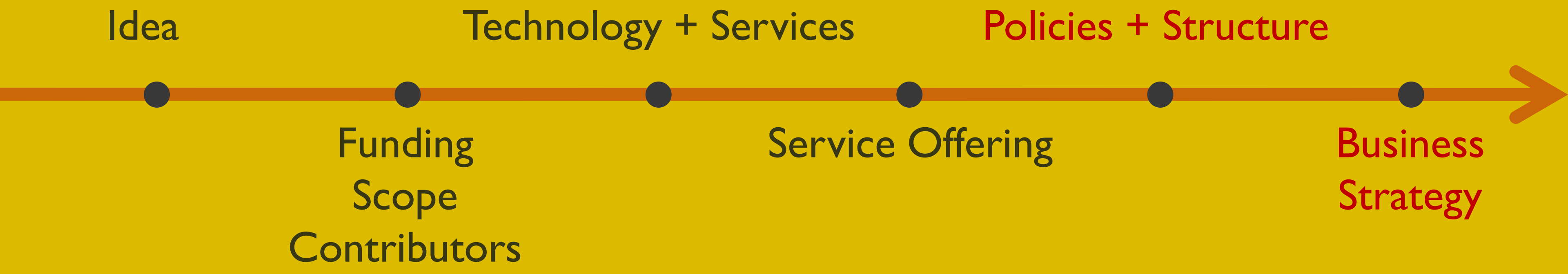
PRESERVATION



Title	Principal Investigator	State	Organization	Awarded Amount to Date
Support for an International Workshop on Scientific Collections held in Brussels	Schindel, David	VA	Smithsonian Institution	$24,000.00
Workshop Proposal: IMAG Futures Meeting	Schlick, Tamar	NY	New York University	$10,000.00
A Feasibility Study for a National Science Foundation Open-Access Publication Repository	Choudhury, Golam	MD	Johns Hopkins University	$299,688.00
DataNet Full Proposal: DataNetONE (Observation Network for Earth)	Michener, William	NM	University of New Mexico	$12,258,110.00
DataNet Full Proposal: The Data Conservancy (A Digital Research and Curation Virtual Organization)	Choudhury, Golam	MD	Johns Hopkins University	$3,726,890.00
4th International Conference on Open Repositories, 2009	Walters, Tyler	GA	GA Tech Research Corporation - GA Institute of Technology	$15,000.00
SCI: TeraGrid Resource Partner	Towns, John	IL	University of Illinois at Urbana-Champaign	$32,441,949.00

DATANET

Emerging environments (observed) pattern:



Idea

Funding
Scope
Contributors

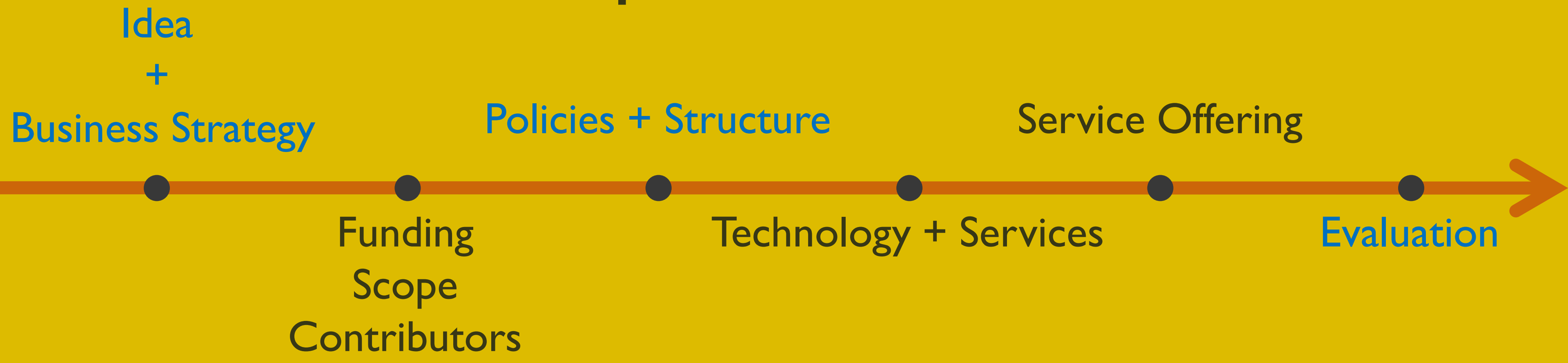
Technology + Services

Service Offering

Policies + Structure

Business
Strategy

Mature environments pattern:



Idea
+
Business Strategy

Funding
Scope
Contributors

Policies + Structure

Technology + Services

Service Offering

Evaluation

EVOLUTION/ECOLOGY

***Research*:** products of one or more focused research projects and typically contain data that are subject to limited processing or curation. These collections are generally small and/or project specific.

***Community data collections*:** serve a single science or engineering community. They are generally intermediate in size and supported in a somewhat more distributed fashion by the community served.

***Reference data collections*:** serve large segments of the scientific and education community. These are generally broad and/or multidisciplinary as well as long lived.

A TYPOLOGY

Grants and Contracts

Multiple Sponsors

Holding Size

Preservation Policy

How Based

Business Type

Scientific Area

FRAMEWORK

Characteristic	Institutional Repository	Science Data Repository
Holdings Management	IRs have a high degree of similarity in terms of management of holdings.	SDRs are dissimilar, often highly domain specific, to each other in terms of holdings.
Handling Procedures	Homogeneity of handling procedures both within and among repositories (DRIVER, 2008)	Heterogeneity of handling procedures, perhaps necessary to degree of specialization within a domain, often seemingly due to lack of standardization.
Base	Institutionally based (DRIVER, 2008)	Typically domain based, though increasingly cross cutting making the call for standardization more critical.

ENVIRONMENT

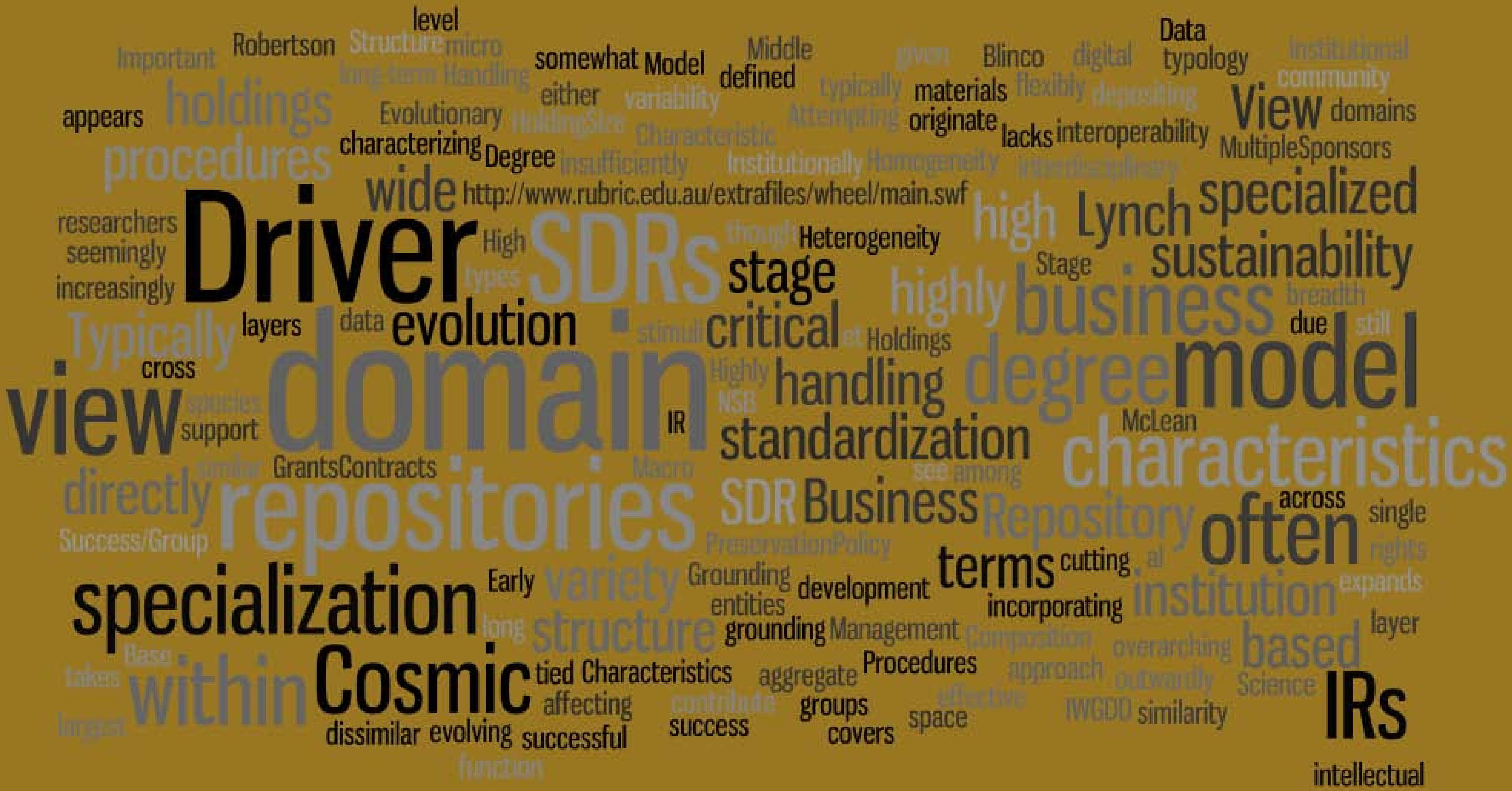
Characteristics of Success/Group Composition

DRIVER (2008):

- Business of digital repositories,
- Stimuli for depositing materials into repositories, intellectual property rights,
- Data curation, and
- Long-term preservation

SDRs:

- *GrantsContracts*
- *MultipleSponsors*
- *HoldingSize*
- *PreservationPolicy*


level
Data
Important
Robertson
Structure
micro
somewhat
Model
Middle
given
Blinco
digital
typology
Institutional
long-term
Handling
defined
typically
materials
flexibly
depositing
community
holdings
either
variability
Attempting
originate
View
domains
appears
Evolutionary
HoldingSize
Characteristic
lacks
interoperability
procedures
characterizing
Degree
insufficiently
Institutionally
Homogeneity
interdisciplinary
MultipleSponsors
wide
http://www.rubric.edu.au/extrafiles/wheel/main.swf
high
Lynch
specialized
researchers
Driver
SDRs
though
Heterogeneity
seemingly
High
stage
Stage
sustainability
increasingly
types
highly
business
breadth
Typically
layers
data
evolution
stimuli
critical
et
Holdings
due
still
cross
domain
Highly
handling
degree
model
view
species
NSB
support
IR
standardization
McLean
similar
GrantsContracts
Macro
see
among
characteristics
directly
repositories
SDR
Business
Repository
across
often
single
Success/Group
PreservationPolicy
rights
specialization
Early
variety
Grounding
terms
cutting
al
development
entities
institution
expands
long
structure
grounding
Management
Composition
incorporating
layer
Base
overarching
based
takes
within
Cosmic
tied
Characteristics
aggregate
Procedures
approach
outwardly
Science
contribute
affecting
effective
IRs
largest
groups
space
IWGDD
similarity
dissimilar
evolving
successful
success
covers
function
intellectual

> “The format of this form forces us to pigeonhole ourselves in a way that is not accurate or useful. Sorry I can't be of more help.”
>
> --Matthew LaPoint, J. Craig Venter Institute

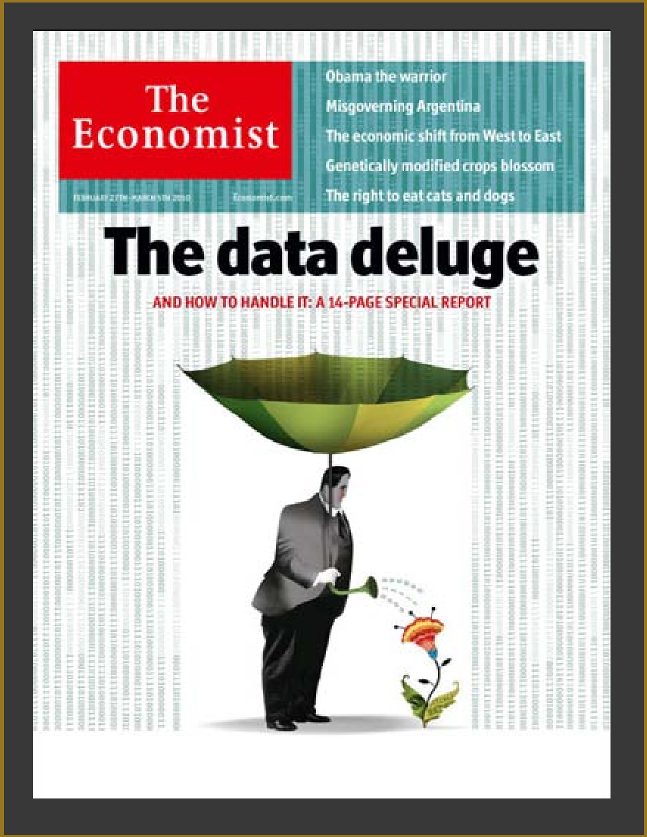
FUTURE WORK

Looking back, the key to moving ahead is

LONGITUDINAL EVALUATION

FUTURE WORK

nature
THE BITER BIT
TROPICAL CYCLONES
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
SCIENCE IN THE PETABYTE ERA


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

GETTING THE WORD OUT

QUESTIONS?

THANK YOU!