National and International Collaborations for Geoinformatics: Challenges and Lessons Learned from Geoinformatics for Geochemistry

W. Christopher Lenhardt (CIESIN - Columbia University), Kerstin Lehnert (LDEO – Columbia University), Sri Vinayagamoorthy (CIESIN - Columbia University), and Steve Goldstein (LDEO – Columbia University)

clenhardt@ciesin.columbia.edu

CIESIN Columbia University

Outline

- Introduction to geochemical and related projects at the Lamont-Doherty Earth Observatory (LDEO) – Columbia University
  - PetDB (Petrologic Database of the Ocean Floor)
  - SedDB (Sediment Geochemistry Database)
  - Earthchem (Advanced Data Management for Solid Earth Geochemistry)
  - SESAR (System for Earth Sample Registry)
- Similarities
- Challenges
- Data to Information systems and beyond
- Next steps

LDEO Projects in Geoinformatics for Geochemistry

PETDB
Petrological Database of the Ocean Floor

SedDB
Integrated Data Management for Sediment Geochemistry

earthchem
ADVANCED DATA MANAGEMENT IN SOLID EARTH GEOCHEMISTRY

SESAR
System for Earth SAmple Registration

Objectives of LDEO Projects

- Maximize the utility of data
  - Build infrastructure that makes data and samples visible and accessible to the broad community
  - Advance the "principle of open access" to data and samples
  - Support the long-term preservation of data (& samples)
    - Provide for persistent archives
    - Ensure comprehensive and accurate documentation
- Support cross-disciplinary approaches in science
  - Facilitate data integration across the Geosciences
    - Technical: interoperability, open access interfaces, better metadata and quality control
    - Cultural:
      - link communities (across related disciplines, nationally, internationally)
      - Facilitate development of relevant expertise

Collaborative Effort

- LDEO
  - Geoscientists
  - Information Technology
  - Data Managers
- CIESIN
  - Information Technology
  - Systems Integration
  - Database Development
  - Data Stewardship
  - Operations

Collaborating Institutions

Harvard (PetDB)

Boston University (SedDB)

Oregon State University (SedDB)

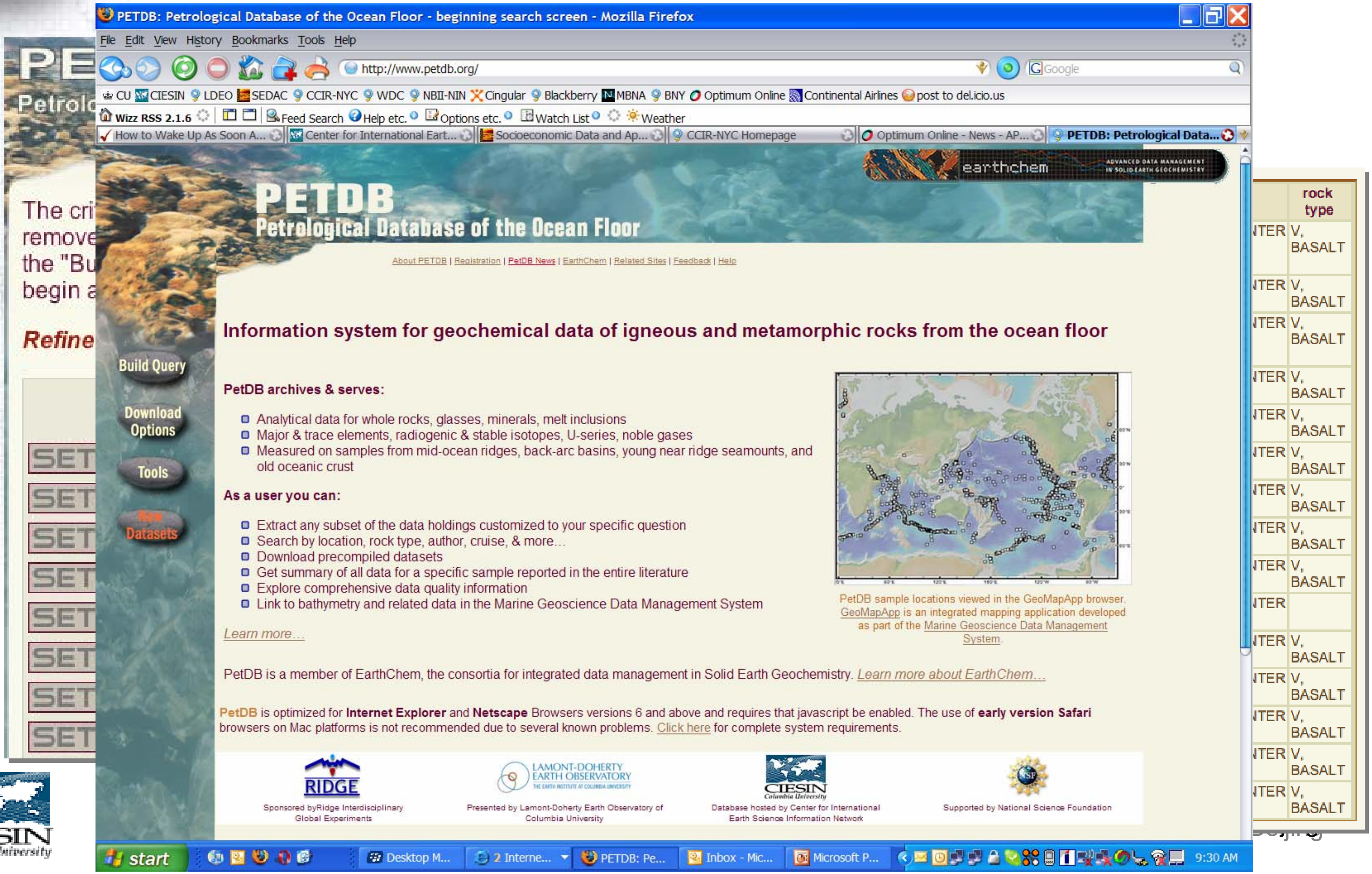
Kansas University (EarthChem)

University of Hawaii (VentDB)

WHOI (VentDB)

Petrological Database of the Ocean Floor (PetDB)

http://www.petdb.org



Reasons for PetDB's Success

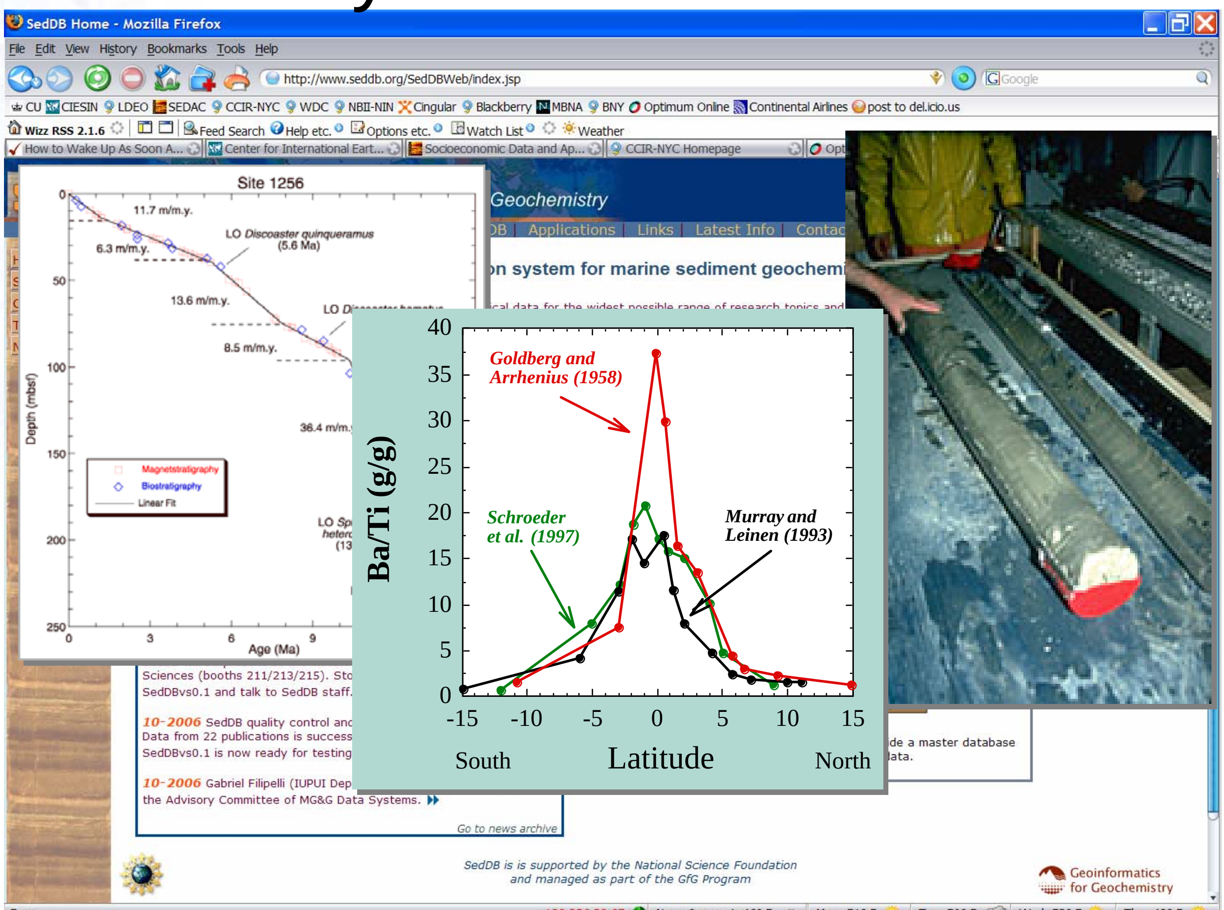
- Technical
  - Design guided by scientists
  - Integrative data model
    - Each individual value searchable through flexible query interface
    - Links & integrates disparate data for individual samples
    - Rich metadata
    - Accessible references
  - User interface with flexible data selection
- Organizational
  - Implementation at professional data center
  - Strong ties with the community
    - Users (science)
    - Professional information technology partners
    - National Science Foundation
- Scientific
  - Has enabled new science

CIESIN
Columbia University

SedDB

http://www.seddb.org

- Integrated Data Management for Sediment Geochemistry



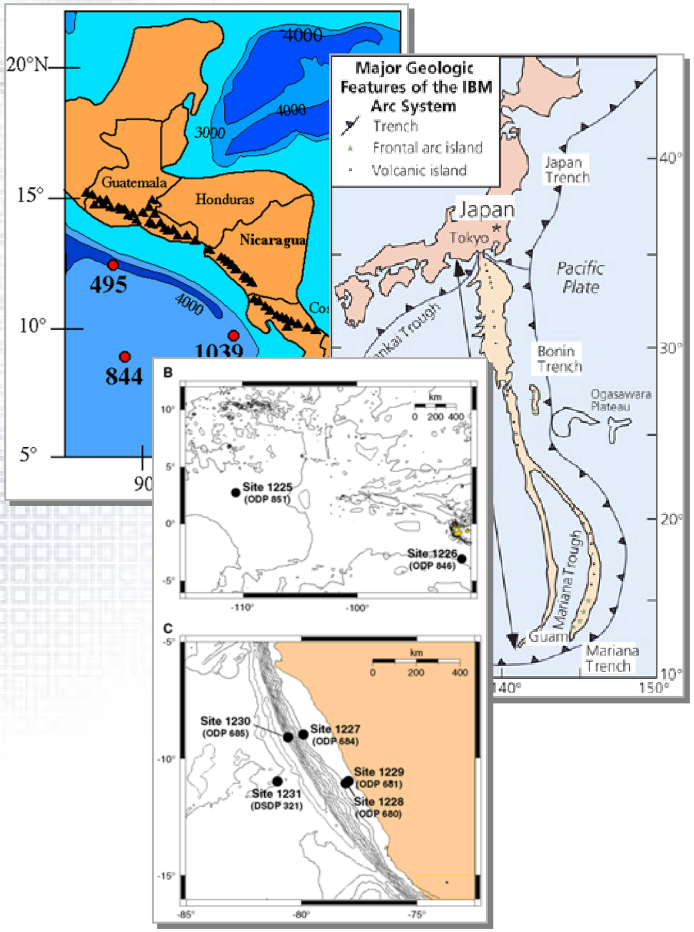
Funding Agency:
NSF (OCE/EAR)

Start Date:
July 2005

Duration:
3 years

Investigators:
K. Lehnert (LDEO)
S. Goldstein (LDEO)
R. Murray (BU)
N. Pisias (OSU)

SedDB



- Apply the concept of PetDB to Marine Sediments
  - Design data model based on PetDB schema
  - Compile complete data sets for 3 test bed areas
  - Build interactive query interface
  - Develop data analysis tools for age model conversion & age-depth correlation
  - Ensure integration with other data (interoperability)

CIESIN Columbia University

Challenges

- Technical
  - Development of additional aspects of the data model (e.g. age models)
  - Optimize interaction with the data for a broad audience ranging from the casual to the expert user
  - Efficiently populate databases with legacy and new data
  - Data quality control
- Organizational
  - Integration/coordination with other geoinformatics efforts
  - Long-term sustainability
  - Workforce ‘under development’
- Cultural
  - “My data syndrome” and data policies
  - Community education (supporting, not competing with science)
  - Standards for data quality assurance & procedures

EarthChem

- Consortium founded 2003 by PetDB, NAVDAT, & GEOROC
  - To nurture synergies among projects
  - To minimize duplication of efforts
  - To share tools and approaches
- Collaborative proposal with D. Walker (Kansas University) funded by NSF EAR&OCE (5 years, start 9/2005) to build an integrated data management and information system for solid earth geochemistry.

The EarthChem Project

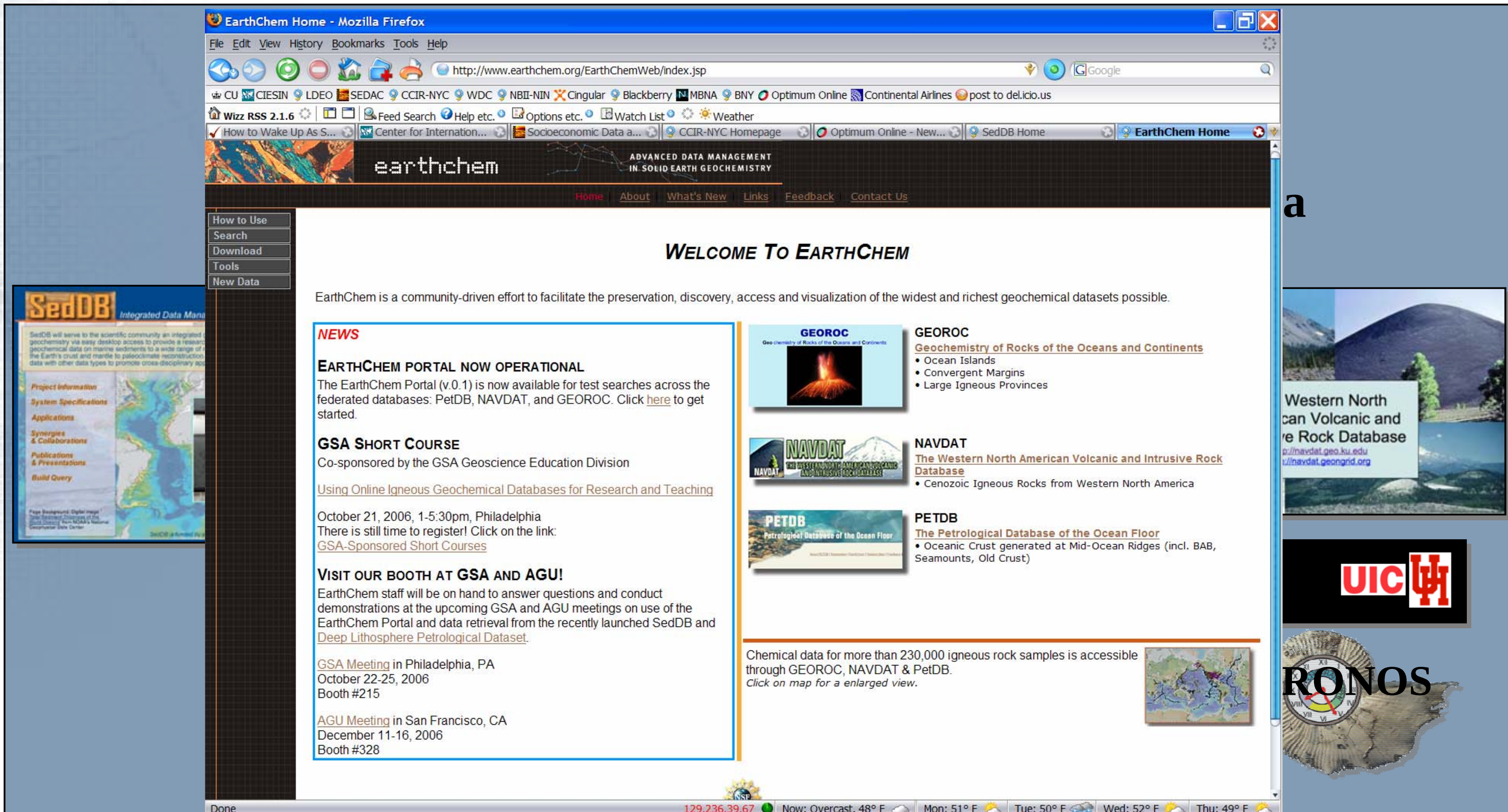
- Build the EarthChem portal as a central access point to a system of federated geochemistry databases (“One-Stop Shop for Geochemical Data”)
- Ensure efficient and continuing update and expansion of data holdings

Project Components

- Data development
  - Data compilation
  - Data quality control
  - Data maintenance
- Data management
  - Data model development
  - Data loading
- Application development
  - User interfaces
  - Interoperability
  - Tools
- User support
- Outreach
  - Community interaction
  - Web site
  - Presentations, publications
  - Advisory committees
  - Workshops
- Project management

EarthChem Focus: Portal

http://www.earthchem.org



EarthChem Focus: Data Holdings

- Create an infrastructure that ensures efficient and community-based growth of data holdings
  - Data entry by dedicated EarthChem personnel
    - New target datasets identified & prioritized via community outreach & the EarthChem Advisory Committee
  - Facilitate Community Contributions
    - Build on-line data submission capability for future data to encourage direct data contributions by investigators
    - Assist investigators with design, implementation, & population of their own databases
    - Serve these databases via the EarthChem portal
  - Expand federation

EarthChem Focus: Standards

- Promote & implement standards for data management in Geochemistry
  - Ontologies
  - Classification
  - Metadata in publications
    - Analytical information
    - Sample provenance
    - Units
    - Unique sample identifiers (IGSN) ➥ SESAR
  - Data publication & submission
  - (Sample management)

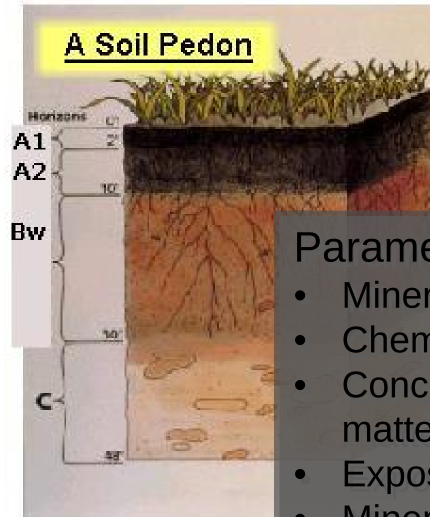
www.geosamples.org

- Providing unique identifiers for Earth samples to allow global sharing, linking, and integration of information and data about these samples.



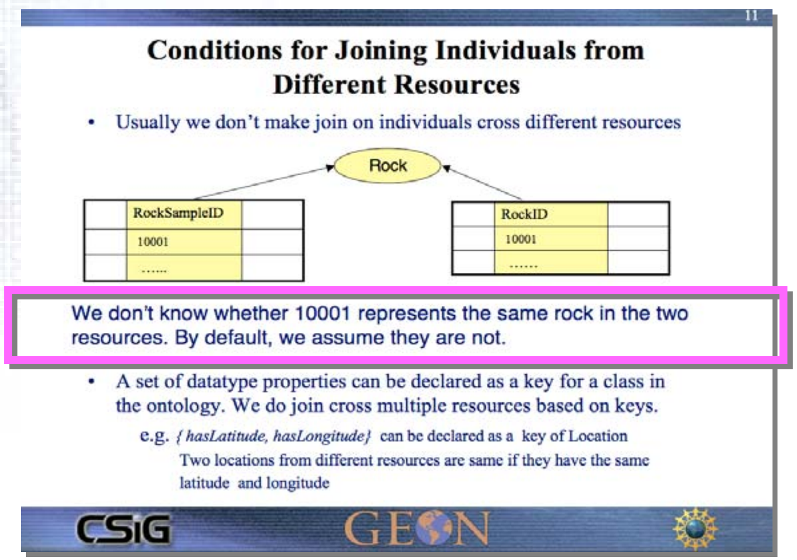

SESAR: Rationale

Many data types are generated by the study of Earth samples. Their usefulness is critically dependent on their integration.



Parameters to be measured:

- Mineralogy
- Chemistry
- Concentration of soil organic matter
- Exposure age
- Mineral surface area



Conditions for Joining Individuals from Different Resources

- Usually we don't make join on individuals cross different resources

We don't know whether 10001 represents the same rock in the two resources. By default, we assume they are not.

- A set of datatype properties can be declared as a key for a class in the ontology. We do join cross multiple resources based on keys.
  e.g. *{hasLatitude, hasLongitude}* can be declared as a key of Location
  Two locations from different resources are same if they have the same latitude and longitude

Kai Lin *(SDSC)*: **"Ontology Based Resource Registration and Integration in GEON"**, Lecture July 2005

Currently, integration of data derived from the same sample, located in distributed systems is obstructed by ambiguous naming of samples.

International Geo Sample Number

IGSN

JDW000001

12-2-01E

IGSN.JDW000001

Unique user code String of (sequential) characters

- Structure
  - String of 9 characters (length limited by use in data publication)
  - First three characters are unique user code (registered with SESAR)
  - Last 5 characters are characters, numbers + letters (one spare character)
  - Allows 2,176,782,336 sample identifiers per registrant
- Managed at a central registry (SESAR)
  - Generated by SESAR or by users.
  - Strict compliance with the IGSN structure required.
- Applied in sample curation, data publication, & digital data management.
- Does not replace personal or institutional names.

CIESIN Columbia University

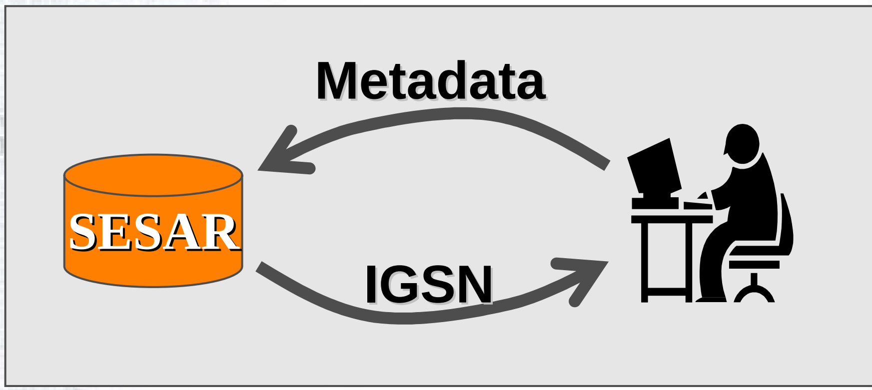
SESAR System for Earth SAmple Registration

NSF

IGSN: Impact

- Ability to link & integrate data for a single sample will
  - advance interoperability among digital data management systems & the development of GeoInformatics
  - help build more comprehensive data sets for samples
  - foster new cross-disciplinary approaches in science
- Ability to unambiguously identify samples will
  - aid preservation and curation, orphaned samples can be identified
  - ensure proper linking of data from samples and subsamples
  - facilitate sample handling and analysis
- Access to a central sample catalog will
  - allow more efficient planning of field & lab projects
  - facilitate sharing of samples
  - facilitate development of sample profiles

Sample Registration



via
- Web site
- Batch loading
- Web services

Review/Edit Individual Specimen Metadata

Sample Name: All125-D3-1 **Collection Method:** Dredge

Collection Metadata: [Edit Collection Metadata]

Ship:	Argo
Cruise:	ARGAMPH
Start Latitude:	44.5 North
Start Longitude:	60.7 East
Stop Latitude:	44.7 North
Stop Longitude:	60.5 East
Min Water Depth:	3200
Max Water Depth:	3300
Location Names:	Mid-Atlantic Ridge
Additional Information on Location:	
Feature:	Spreading Center
Collection Date Start:	
Collection Date End:	

Classification Metadata: [Edit Classification Metadata]

Rock Classification: Level 1:	Igneous
Rock Classification: Level 2:	Mafic, volcanic/subvolcanic
Field Description:	Basalt
Additional Information:	aphyric, altered

Archival Metadata, etc.: [Edit Archival Metadata]

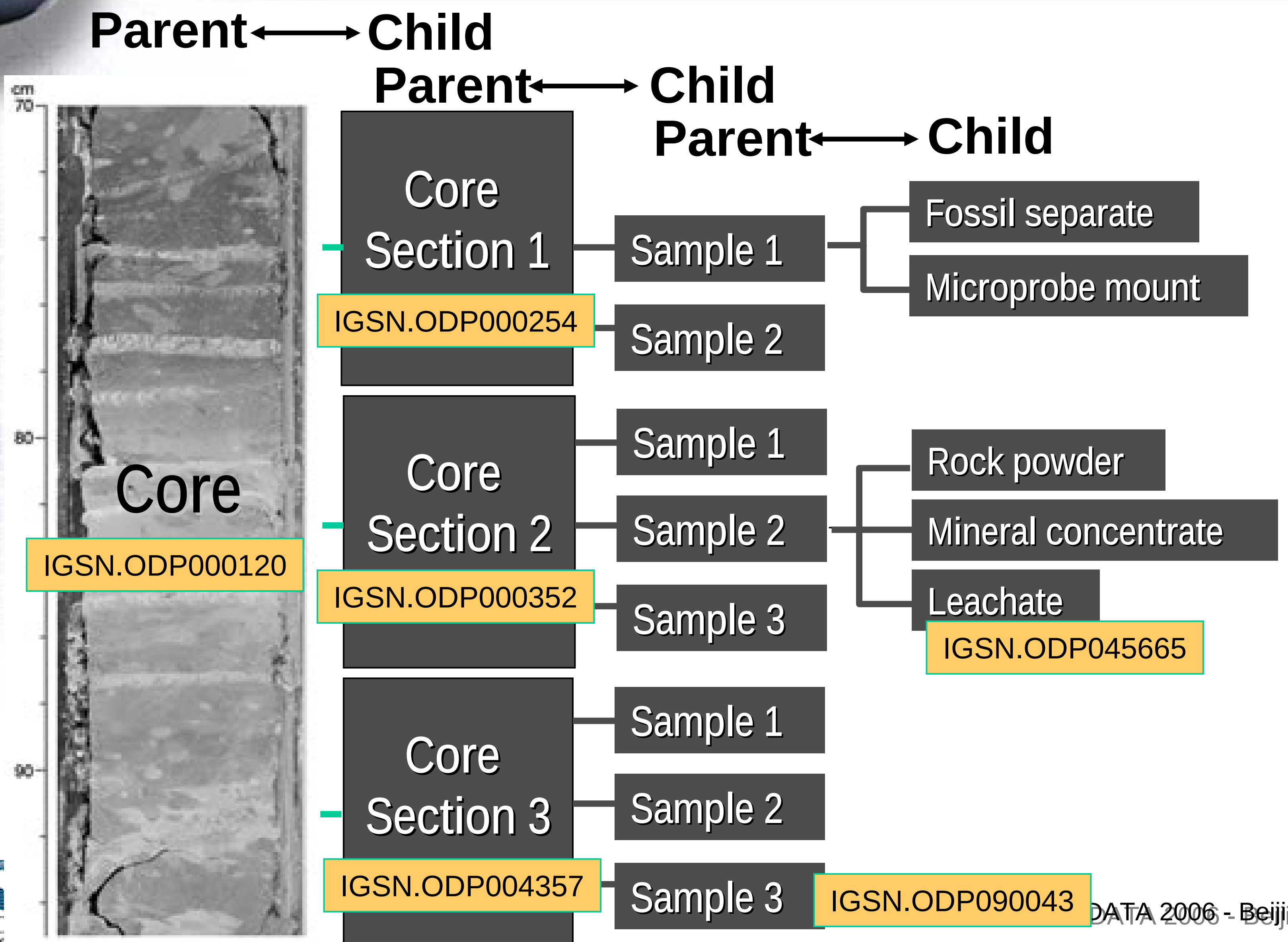
Original Archive
Last Name: Lehnert **First Name:** Kerstin
Institution: Columbia University, Lamont-Doherty Earth Observatory

Latest Known Archive
Last Name: Lehnert **First Name:** Kerstin
Institution: Columbia University, Lamont-Doherty Earth Observatory

[Submit Sample]

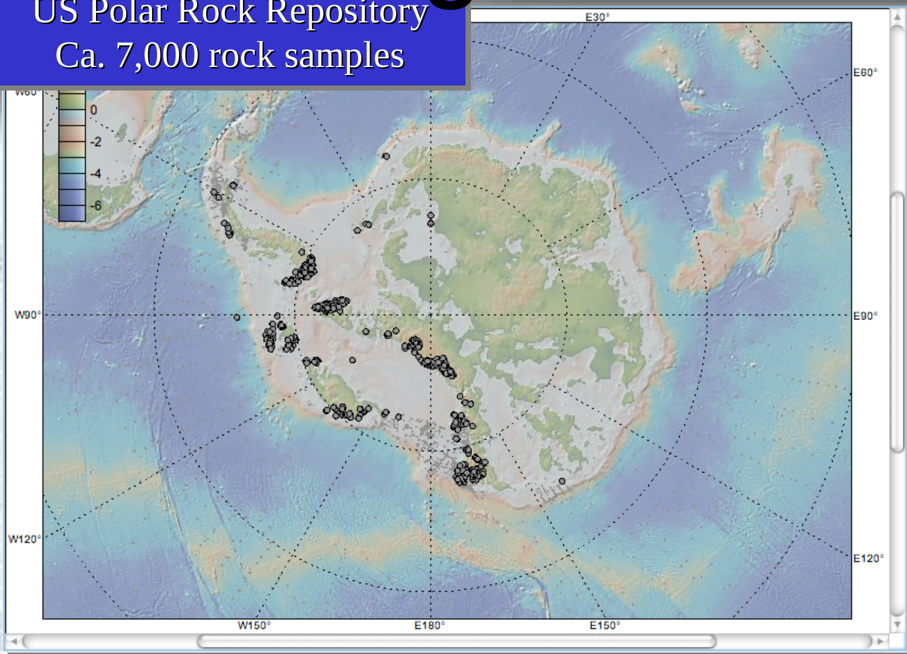
SESAR
System for Earth SAmple Registration

Granularity of Registered Samples

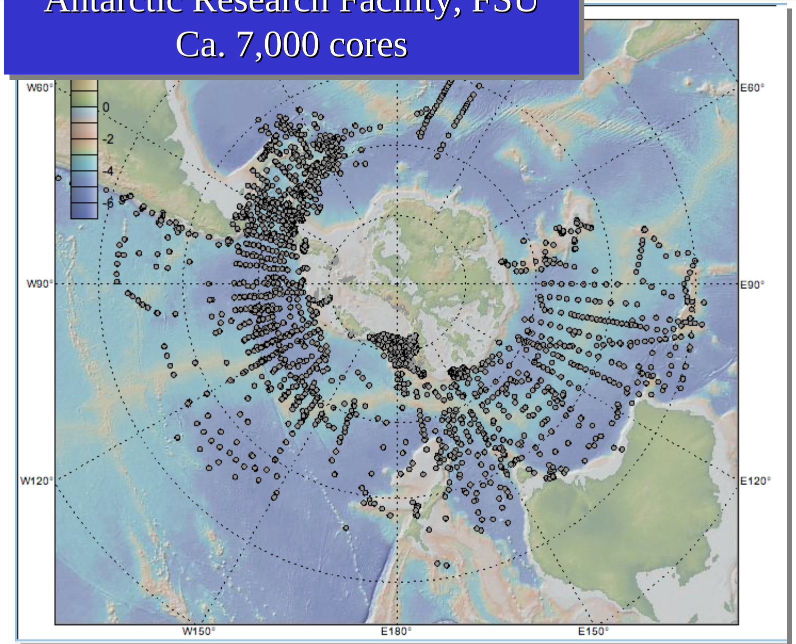


Building a Global Sample Catalog

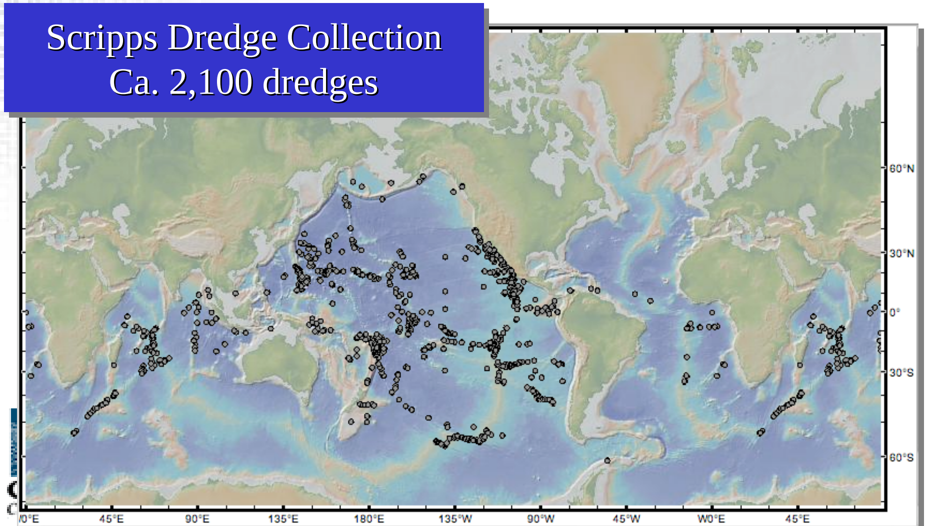
US Polar Rock Repository
Ca. 7,000 rock samples



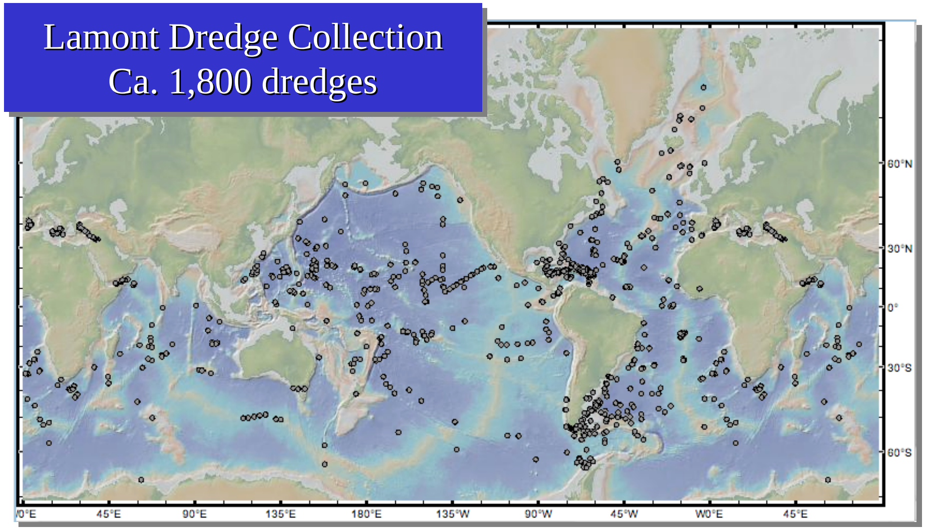
Antarctic Research Facility, FSU
Ca. 7,000 cores



Scripps Dredge Collection
Ca. 2,100 dredges



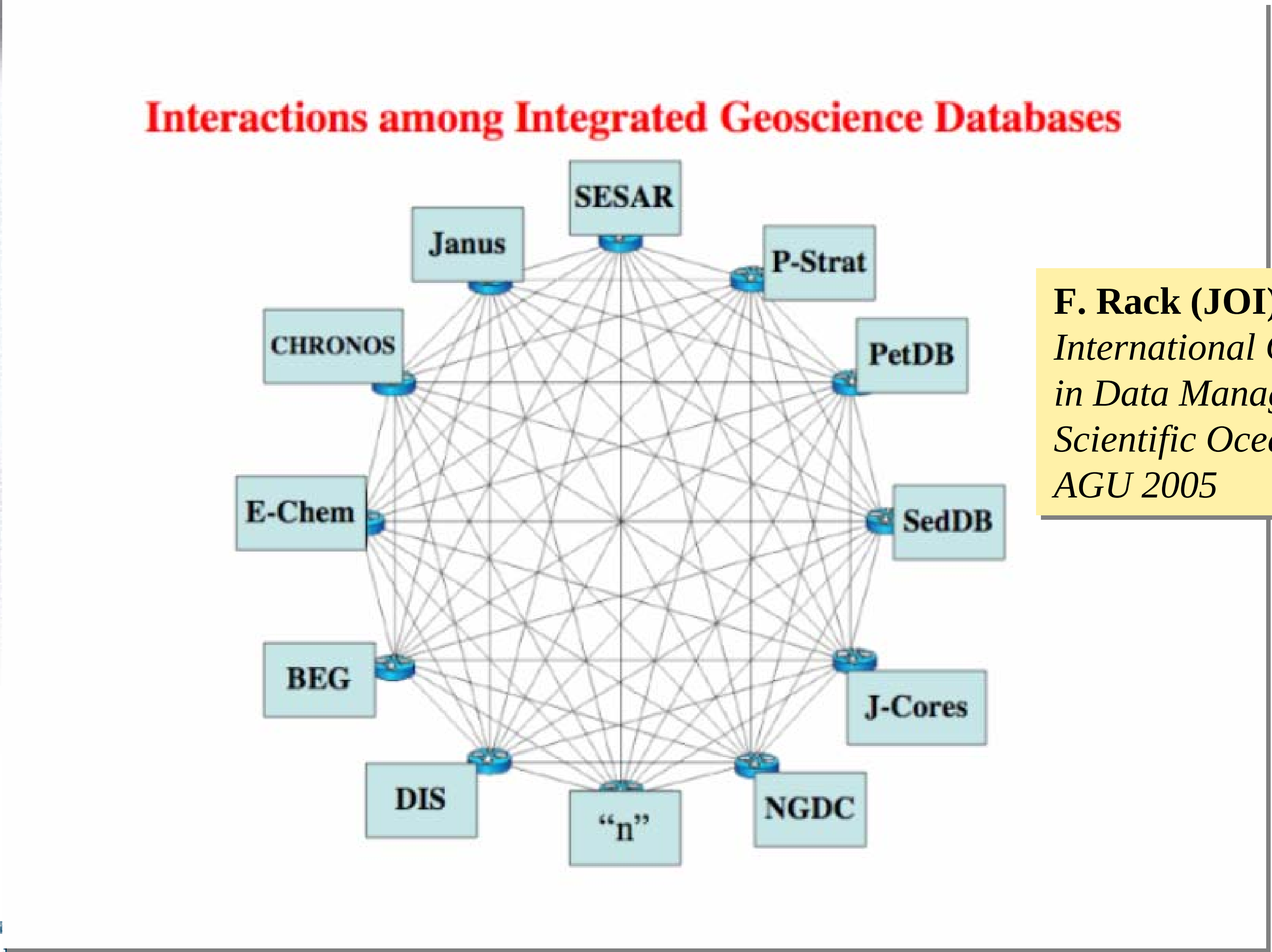
Lamont Dredge Collection
Ca. 1,800 dredges



Similarities

Many related sources of data and information in the field of geochemistry

Many potential interactions



F. Rack (JOI):
International Collaboration in Data Management for Scientific Ocean Drilling, AGU 2005

Commonalities Across Geochemical Data

- Small volumes
- Complex ‘background information’ (=metadata)
- Diversity of acquisition methods
- Sample-based
- Producer is ‘owner’

Summary of Lessons Learned

- Data is the foundation
- Science is the driver
- Development of information systems essential
  - Data capture and access
  - Data stewardship
  - Knowledge capture
- Community participation is essential
- Outreach is essential
  - Vertical and horizontal

CIESIN Columbia University

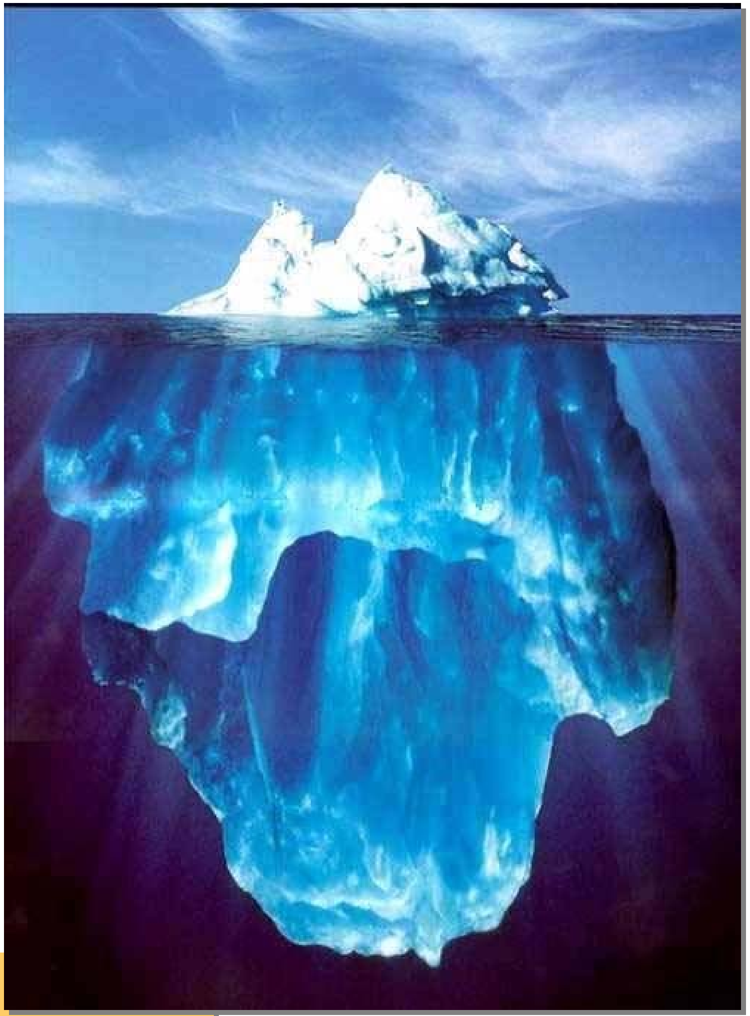
General Trajectory

- Data to ‘information systems’
- Develop and enhance the growing cyberinfrastructure for geoinformatics
- Expand both the data, the systems, interoperability, *AND* participation to move towards a geoinformatics science commons

CIESIN Columbia University

Challenges

- How to get the word out
- How to expand participation
- How to promote standardization and interoperability globally



Most of the technology exists
Challenges are cultural and organizational

L. Allison, SedDB Workshop 2004

CIESIN
Columbia University

Urgency to act

- Increasing data volumes
  - Need systems to support data management.
- Large-scale scientific questions
  - Need access to global data compilations.
- New cross-disciplinary approaches
  - Need integration of data with broader Geoscience data set.
- Decreasing funding
  - Need to maximize utility of data (and samples).

CIESIN
Columbia University

Next steps

- Continue outreach
- Invite participation and collaboration
  - Collaborators
  - Data integration
  - Linkages across systems
- Propose a CODATA task group on geoinformatics?

CIESIN
Columbia University

Thank you.

謝謝

Backup Slides

NSF-OCI

CIESIN
Columbia University

Geoinformatics

= Science + Data + Cyberinfrastructure + Data Stewardship

Transforms into…

"Science Commons" for geochemical data

Cyberinfrastructure =

- "new research environments in which advanced computational, collaborative, data acquisition, and management services are available to researchers through high-performance networks"

Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure (Atkins et al. 2003)

Cyberinfrastructure

is the organized aggregate of technologies that enable us to access and integrate today's information technology resources –

data, computation, communication, visualization, networking, scientific instruments, expertise

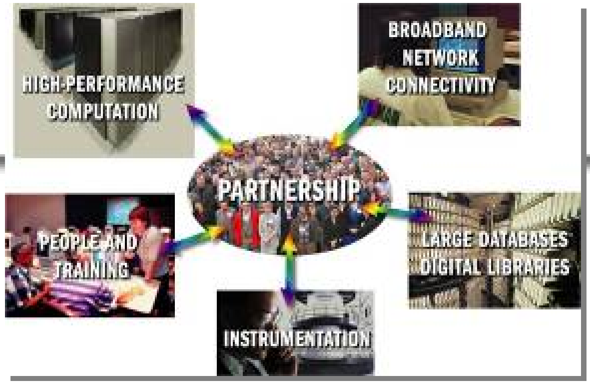
– to facilitate science, engineering, and societal goals.

Cyberinfrastructure =

- "Like the physical infrastructure of roads, bridges, power grids, telephone lines, and water systems that support modern society, "cyberinfrastructure" refers to the distributed computer, information and communication technologies combined with the personnel and integrating components that provide a long-term platform to empower the modern scientific research endeavor."

Access News Release: "National Science Foundation Releases New Report from Blue-Ribbon Advisory Panel on Cyberinfrastructure," 02.03.03 David Hart

CI Components



The cyberinfrastructure should include

- grids of **computational** centers, some with computing power second to none;
- comprehensive **libraries of digital objects** including programs and literature;
- multidisciplinary, well-curated federated **collections of scientific data**;
- thousands of **online instruments** and vast **sensor arrays**;
- convenient **software toolkits** for resource discovery, modeling, and interactive visualization;
- the **ability to collaborate** with physically distributed teams of people using all of these capabilities.

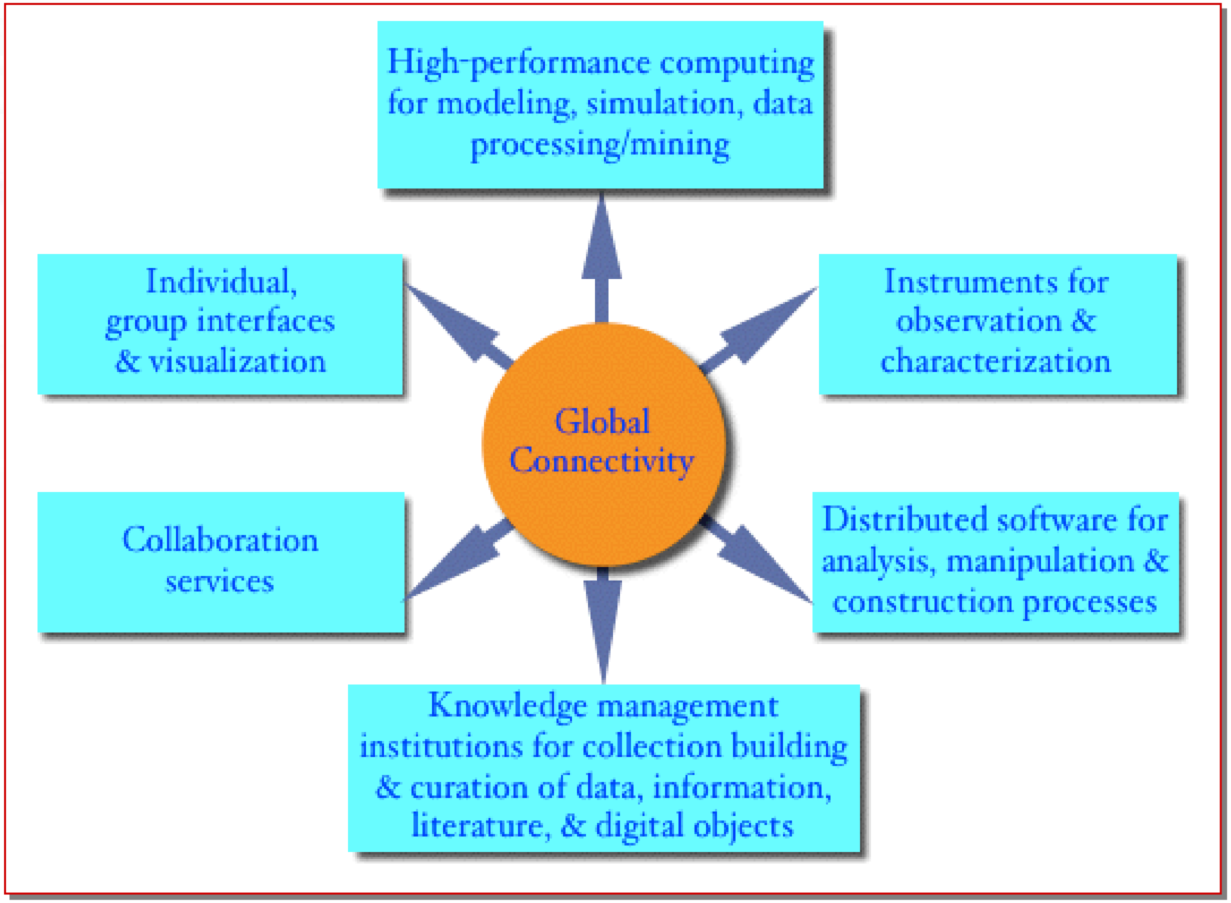
Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure (Atkins et al. 2003)

Geoinformatics

CYBERINFRASTRUCTURE FOR THE EARTH SCIENCES

- **Geoinformatics is the application of computer technologies and methodologies to scientific results with spatial-temporal coordinates**.
- Geoinformatics encompasses efforts to promote collaboration between computer science and the geosciences to solve complex scientific questions.

Components of Geoinformatics



K. Droegemeier, S. Graves, & J. Orcutt: "Geo-CI"; NSF-CI workshop 2003

Required Geoinformatics Components

- Interoperability of diverse databases on diverse systems
- Distributed, web-based, web services
- Access to data, tools & Computational resources
- Security

- Large data sets
- Complex data sets
- Data input – ease vs complexity
- Remote sensing & sensor arrays
- Real-time digital field technologies
- Capture analogue "legacy" data
- Data storage and curation

- Band width
- Computational resources: high performance, mid-level, desktop; grids, etc.
- File transfer protocols, etc.

- Real-time collaboration
- Data mining / pattern recognition
- Tools development and maintenance: numerical, statistical, visual
- Online workspace, software, and tutorials
- Community and computational models and Collaboratories
- Model-data fusion
- Skills, career paths and reward structures
- Intellectual property and academic credit
- E-Journals

CIESIN Columbia University

***W. Snyder, K. Lehnert, & J. Klump:* "**Building an International Collaboration for Geoinformatics", 2005

CI Challenges

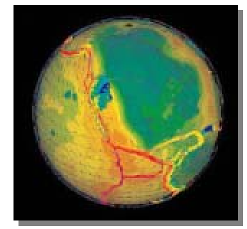
- "The challenge of Cyberinfrastructure is to integrate relevant and often disparate resources to provide a useful, usable, and enabling framework for research and discovery characterized by broad access and "end-to-end" coordination."

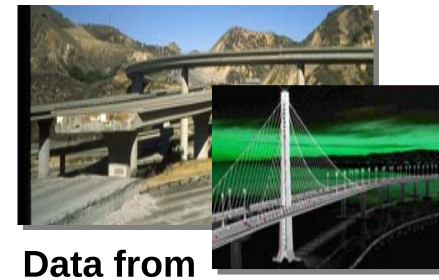
Fran Berman, Director San Diego Supercomputer Center

SBE/CISE Workshop on Cyberinfrastructure for the Social Sciences

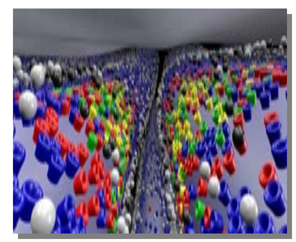
Data: The Foundation of Geoinformatics

- Data comes from everywhere
  - Scientific instruments
  - Experiments
  - Sensors and sensor-nets
  - New devices (personal digital devices, computer-enabled clothing, cars, …)
- And is used by everyone
  - Scientists
  - Consumers
  - Educators
  - General public
- Data Cyberinfrastructure environment must support unprecedented diversity, globalization, integration, scale, and use



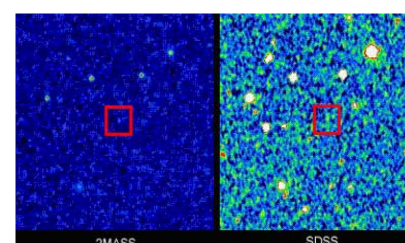


Data from sensors



Data from simulations

Data from instruments



Data from analysis

F. Berman (SDSC): *The Emerging Cyberinfrastructure: Opportunities & Challenges, Pardee Symposium 2004*

Preserving the legacy

- *"The science community has invested vast resources – intellectual and financial - into our present state of knowledge that is bound up in the data it was generated from.* ***These legacy data are an incredibly valuable resource on which new theories, new discoveries, new knowledge will be based in the future - if they remain available to the community.*** *Due to limited accessibility, we have under utilized these data in the past, and we are at significant risk of losing them altogether. Capturing legacy data therefore has to be an essential part of Geoinformatics development."*

W. Snyder & K. Lehnert: *White Paper to the NSF, January 2006*

Geoinformatics builds on DATA

"The National Science Board (NSB) recognizes the growing importance of these digital data collections for research and education, their potential for broadening participation in research at all levels, the ever increasing National Science Foundation (NSF) investment in creating and maintaining the collections, and the rapid multiplication of collections with a potential for decades of curation."

'Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century'

National Science Board Report, September 2005

CIESIN
Columbia University

NSB Report

- Recommendations to NSF
  - Develop clear technical and financial strategy
  - Create policy for key issues consistent with the technical and financial strategy
    - Community oversight for data collections
    - Data policies for data generating projects
    - Education & training for using data collections
    - Recognition for ‘data scientists’

CIESIN
Columbia University

Digital Data Collections: Benefits

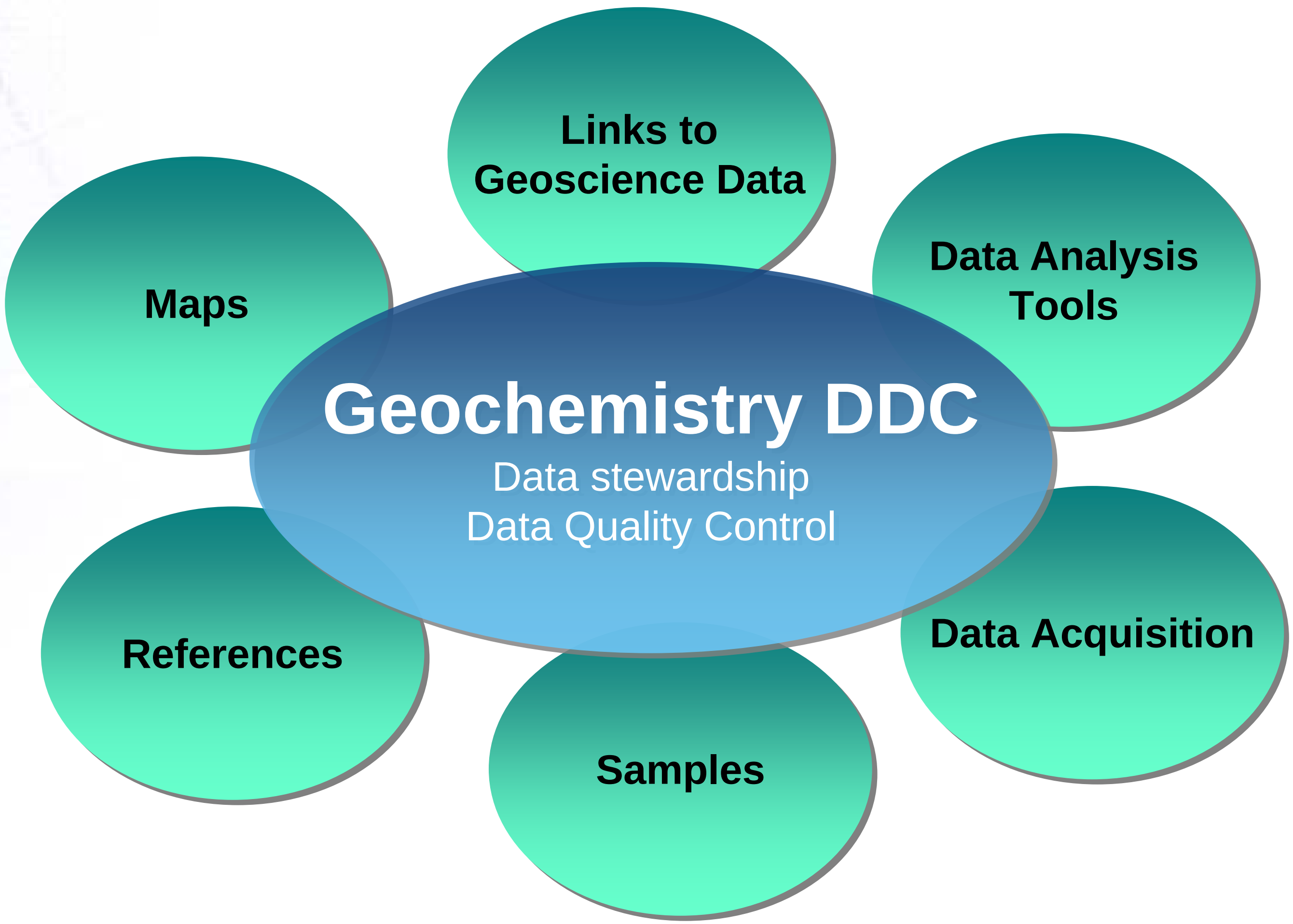
- "Are equally accessible to study at all levels"
- "Serve as an instrument for performing analysis"
  - with an accuracy that was not possible previously
  - from a perspective that was previously inaccessible (by combining information in new ways)

'Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century'

National Science Board Report, September 2005

DDC need to be "Information Systems" rather than "Data Libraries".

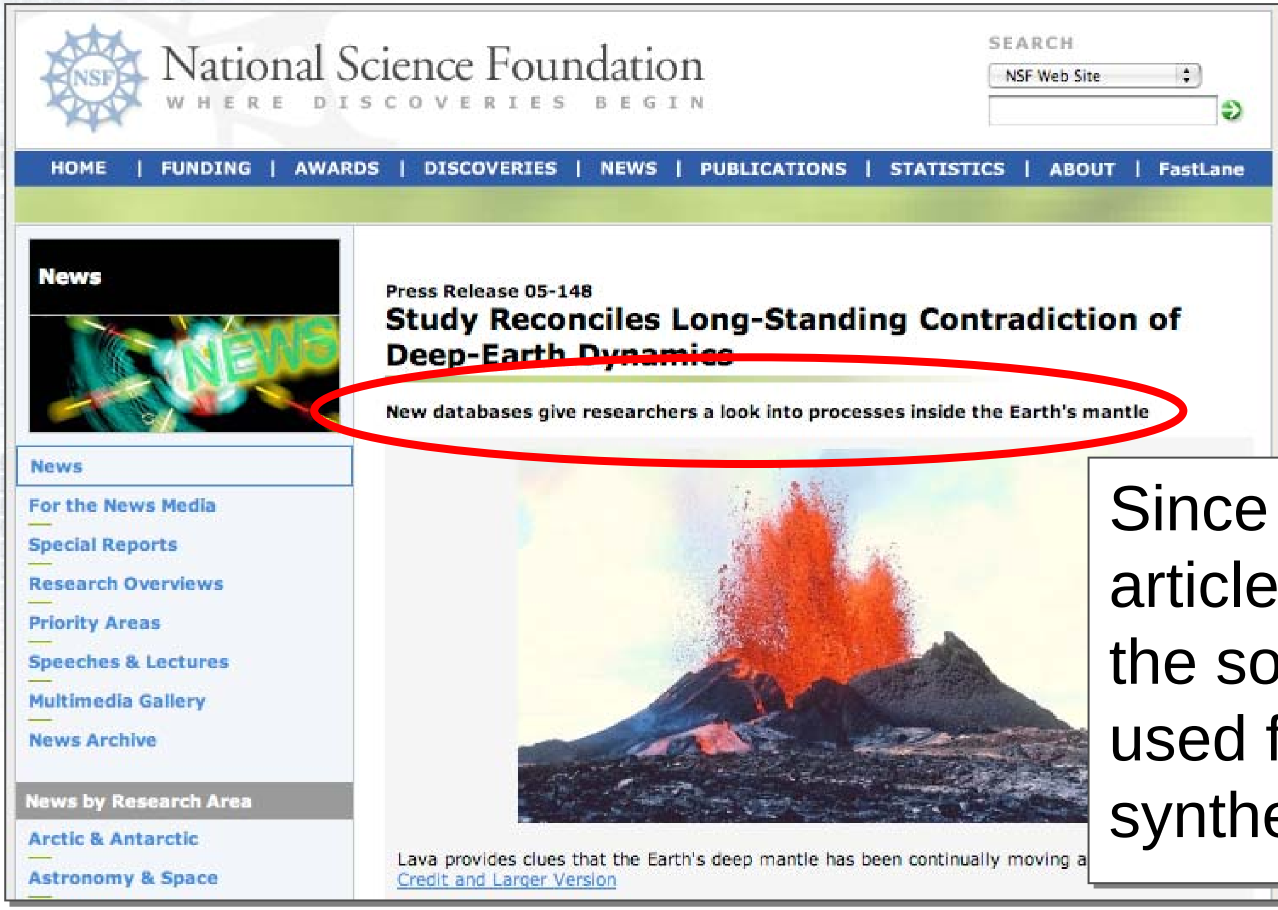
Information Systems in Geochemistry



Benefits of Information Systems

- Advance scientific discovery
- Maximize utility of the Geochemical data set in science & education
- Allow data integration & visualization across the Geosciences
- Enhance data quality control

Impact on Science



Since 2002, ca.100 articles cite PetDB as the source of data sets used for comparison or synthesis.

CIESIN Columbia University

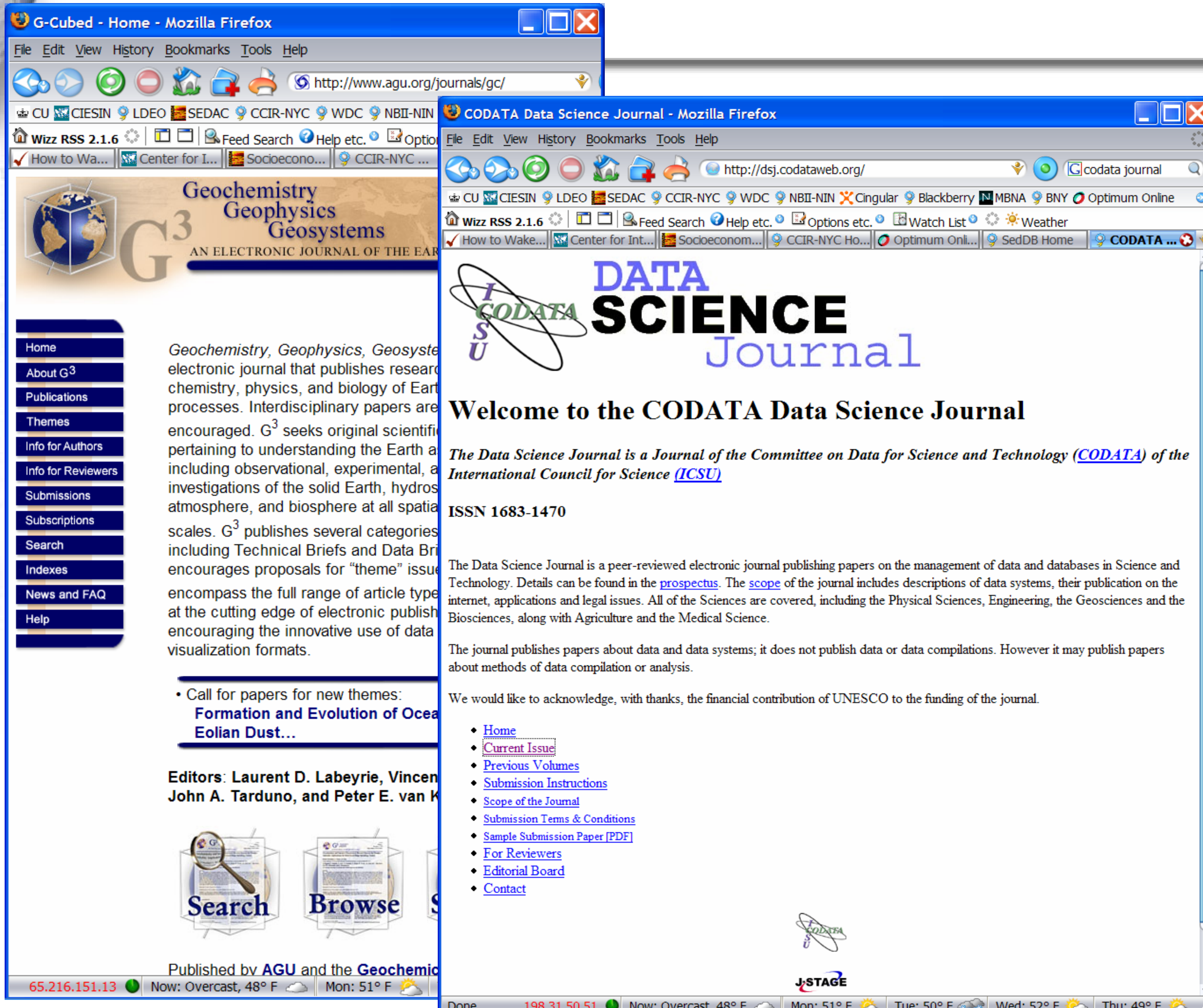
User Survey 2005

- “More than just a timesaver, these databases make it possible to address both global and regional questions that I would otherwise never bother to attempt. The amount of time saved is such that countless ideas cross from the realm of the totally impractical for a busy working scientist into the realm of easy to squeeze into a spare half hour.” (Paul Asimov, CalTech)
- “I think these online databases are absolutely necessary to ensure some level of access to geochemical data for all. I cannot imagine a more efficient way to compile and distribute this data.” (Garrett Ito, U Hawaii)
- “I use both GEOROC and PETDB regularly and have used them in 2 or 3 publications. I consider them critical for advancing isotope geochemistry.” (Don DePaolo, UC Berkeley).
- “It has been hugely helpful in both my research & teaching activities. One recent paper I have published in Journal of Petrology was on MORB, & I cited PETDB extensively.” (Claude Herzberg, Rutgers Univ)

A User’s Vision

- “… in theory the best thing would be one big Geo-database where all different types of geochemical reservoirs are included and all analytical tools as well and where you can search for either regions or reservoir type or method...

ok that’s a big goal.”



CIESIN
Columbia University