



Databases & Data Infrastructure

Kerstin Lehnert

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE



+ Access to Data is Needed

- to allow verification of research results
- to allow re-use of data

+ “The road to reuse is perilous”⁽¹⁾

- Accessibility
 - Discovery, long-term access, permissions
- Usability
 - understand what was measured and how (materials and methods), computations that were applied, presentation of data (units, symbols, etc.)
 - ability to apply standard tools to all file formats
- Motivation
 - Professional benefits vs effort and economic burden of publication; policies

⁽¹⁾Rees, Jonathan (2010): “Recommendations for independent scholarly publication of data sets.” *Creative Commons Working Paper*

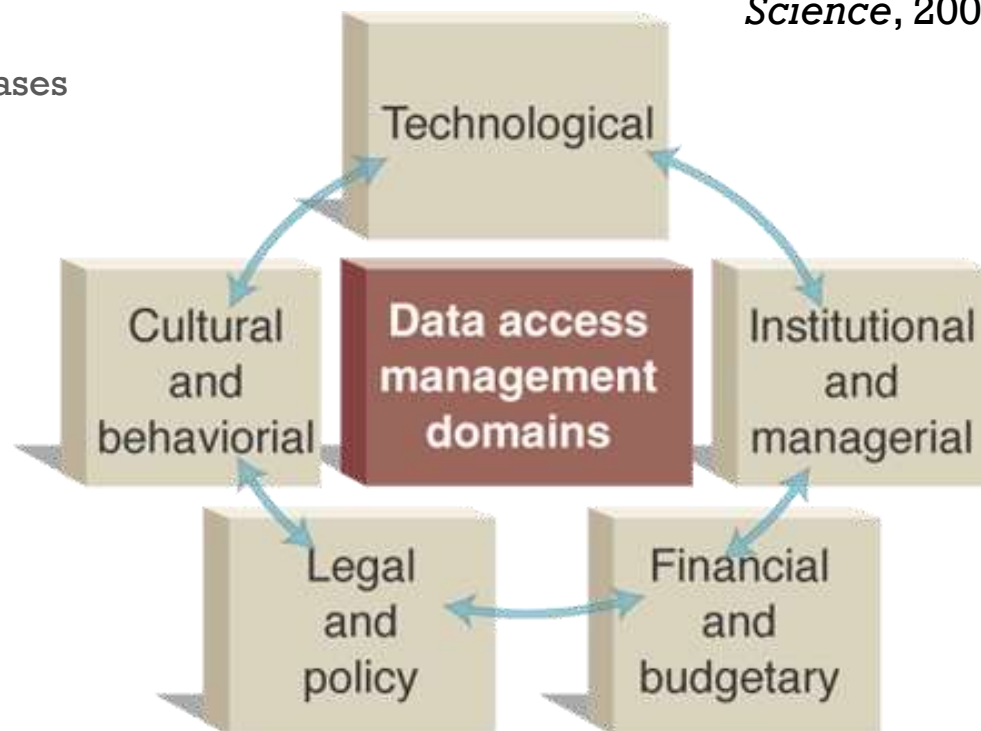
+ From Databases to Data Infrastructure

- Data-driven science creates new requirements for data:
 - Data need to be discoverable.
 - Data need to be persistently and reliably accessible.
 - Data need to be curated and reviewed for quality assurance.
 - Data need to be unambiguously identifiable & located.
 - Data need to be citable.
 - Data need to be interoperable.
- This requires the development of a data infrastructure.
 - Trusted repositories instead of informal databases.

+ From Databases to Data Infrastructure

- Technological Infrastructure
- Workforce
- Management Models
 - Distributed versus centralized databases
 - Control, oversight
- Financial Support
- Legal & Policy Framework
 - Open Access policies
 - Policy enforcement
- Cultural & Behavioral Changes
 - Data sharing
 - Data citation

From: Arzberger et al.,
Science, 2004



+ Where Are We Now?

- Few data repositories fulfill the requirements.
 - National data centers (NCDC, NGDC, NSIDC, etc.)
 - Domain-specific data facilities: IRIS, BCO-DMO, IEDA (MGDS, EarthChem), etc.
- Most databases don't, but at least provide access to data.
 - local, no standards
 - single point of failure
 - not persistent
- Many gaps in coverage
- Too much dark legacy data



From: <http://www.elsevier.com/about/content-innovation/database-linking>

September 10, 2013

+ How Can We Advance ?

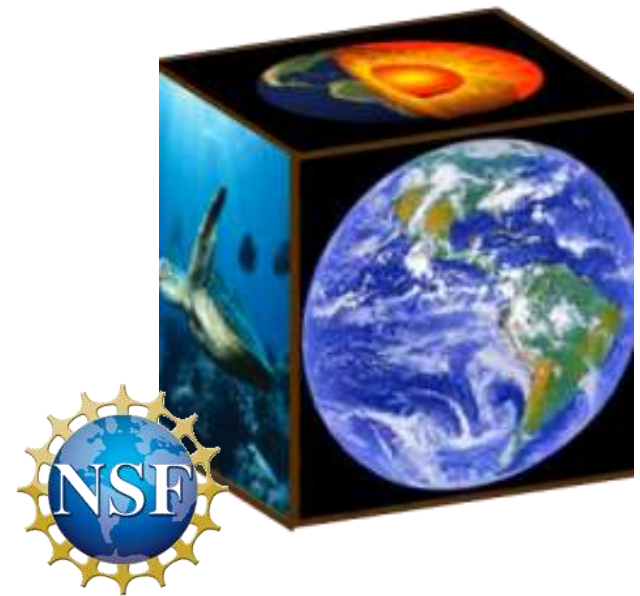
- Infrastructure
 - Sustained and comprehensive repositories
 - Tools and workflows for data management, including data publication
 - Best practices, standards
- Incentives for data sharing
 - Credit (data citation, bibliometrics for data)
 - Better science
- Policy enforcement
 - Funding agencies
 - Publications

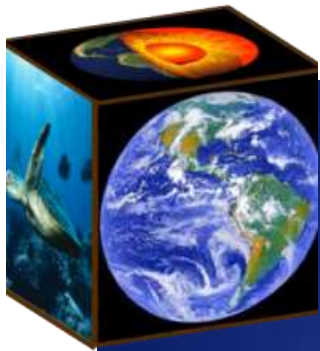
+ Advancing Data Infrastructure

- CIF21 & EarthCube
 - Community building
 - Building Blocks development
 - Interoperability
- Growing number of initiatives & organizations to develop and implement best practices and policies
 - Research Data Alliance
 - CODATA/World Data Systems
 - BRDI
- New approaches to data publication & citation

+ EarthCube

- Transform the conduct of research in geosciences
 - by supporting the development of community-guided cyberinfrastructure
 - to integrate data and information for knowledge management across the Geosciences.





Path to the Vision



EarthCube is
an outcome
AND a process



EarthCube will
require NSF and
broad
community
involvement;
new ways of
doing

Important Features:

- Builds off existing data/modeling systems/cyberinfrastructure investments
- Provides tools/approaches that enhance data discovery, access, and integration
- Addresses serious cyber needs in fields where individual data points and observations are important
- Leverages investments across fields
- Allows for more integrative and interdisciplinary science



+ EarthCube Progress

- 2 year planning and community building phase
 - charrettes
 - community & concept awards
 - domain end-user workshops
- Started initial developments:
 - Research Coordination Networks funded in summer 2013
 - Building Blocks & Conceptual Designs funded in summer 2013
 - CINERGI – Inventory of CI resources
 - Test Enterprise Governance starting Sept 15
 - Closer collaboration of data facilities
 - Web Services Interop Building Block project
 - Consortium of Data Facilities (workshop coming up)
 - The ‘D8’ and/or ‘D20’ concept

+ Guidelines & Best Practices

- data publication
- open access
- data attribution & citation
- data standards
- trustworthiness of repositories

“The Research Data Alliance aims to accelerate and facilitate research data sharing and exchange.”

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE

Home About Working and Interest Groups News & Events Plenary Meetings RDA in the Press

Data Foundation and Terminology WG

The Data Foundation and Terminology Working Group describes a basic abstract data organization model which can be used to derive a reference data terminology that can be used across communities and stakeholders to better synchronize conceptualization, enable better understanding within and between communities and stimulate tool building.

big data Analytics IG
Brokering IG
Citation of Dynamic Data IG
Community Capability Model WG
Data Citation WG
Data Foundation and Terminology WG
Data in Context IG

Home Organization Our Members Services Publications Working Groups

“The WDS supports ICSU's mission and objectives, ensuring the long-term stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders.”

ICSU
WORLD DATA SYSTEM

“... ensuring the long-term stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders.”

+ Data Publication: Options



Conventional
publication



Institutional
Repositories



Data Paper



Disciplinary
Repositories

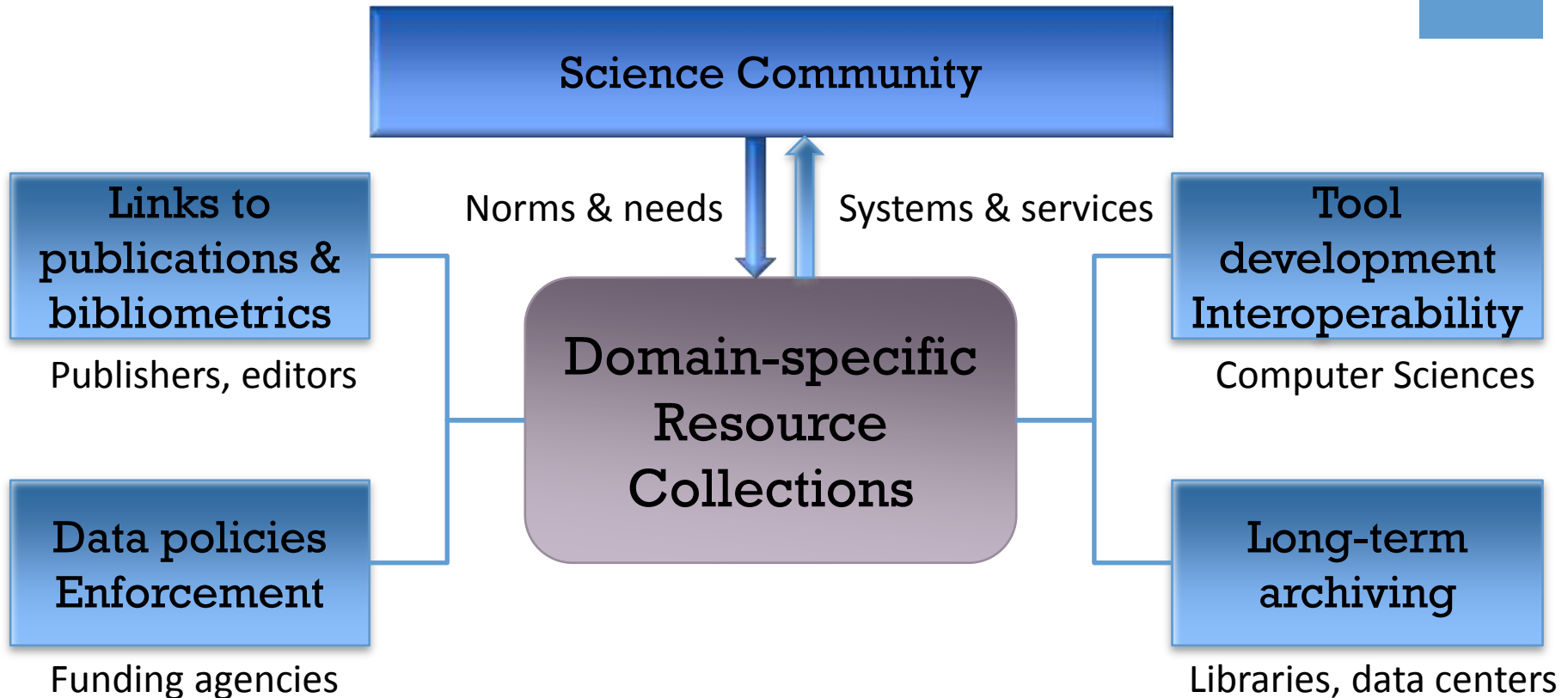
+ Role of Data Repositories



- Ensure long-term preservation
 - Data documentation (catalog metadata)
 - Persistent & unique identification
 - Sustainable infrastructure & business models

- Ensure Usability (Disciplinary Repositories!)
 - Adopt and/or develop community-based standards for documenting:
 - Provenance of data (collection strategies, procedures and underlying assumptions)
 - Data precision, errors, workflows for data quality assurance
 - Comply with standards for data representation (formats, semantics, etc.)
 - QA/QC of datasets and metadata
 - Science-driven tools for data search & access
 - Standards-based interfaces for programmatic access
 - Integrate data for analysis

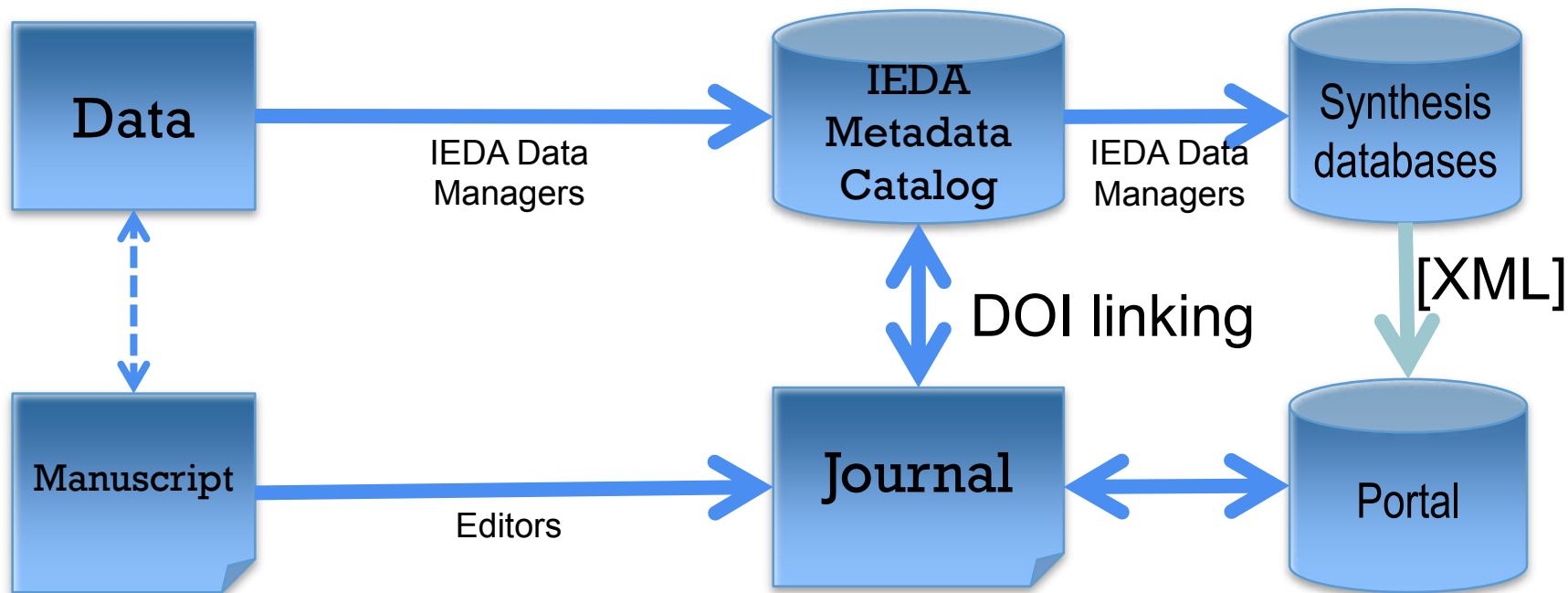
+ Domain-specific Repositories: Linking Stakeholders



+ Data Publication Process

Example: IEDA

Submission → Review → Publication → Integration



+ EarthChem Standards for Data Publication

- Following recommendations of the Editors Roundtable (Policy Statement released in 2009: www.earthchem.org/editors)
 - complete disclosure of data used in a publication
 - full documentation of data provenance & quality (uncertainty)
 - unique identification of samples
 - geospatial & taxonomic information about samples
- Currently reviewing policy statement to align with emerging best practices and new publication capabilities
 - submission of data to repositories as part of editorial process
 - 'data review' by repositories

Dataset Information

Dataset DOI [doi:10.1594/IEDA/100261](https://doi.org/10.1594/IEDA/100261)

DOI to allow proper citation

Dataset Title *Major and trace element geochemical analyses from "The mean composition of ocean ridge basalts" and "Enriched basalts at segment centers: The Lucky Strike (37° 17'N) and Menez Gwen (37° 50'N) segments of the Mid-Atlantic Ridge"*

Dataset Language English

Dataset Type Collection

Author(s) Gale, Allison

Abstract or Description Supplementary tables from "The mean composition of ocean ridge basalts" and "Enriched basalts at segment centers: The Lucky Strike and Menez Gwen segments of the Mid-Atlantic Ridge". Tables include: (1) Major and trace element analyses and isotopic analyses of a global compilation of mid ocean ridge basalts. (2) Major and trace element compositions of glasses and glassy basalts from KP-2,3,5 and PO-1

Data Type(s) Chemistry:Rock

Subject/Keywords *ocean ridge basalts, major element analyses, trace element analyses, isotopic analyses*

Related Publication(s) (citation) Gale, A., C. A. Dalton, C. H. Langmuir, Y. Su, and J.-G. E. Schilling (2013), The mean composition of ocean ridge basalts, *Geochem. Geophys. Geosyst.*, doi:10.1029/2012GC004334.

Gale, A., S. Escrig, E. J. Gier, C. H. Langmuir, and S. L. Goldstein (2011), Enriched basalts at segment centers: The Lucky Strike (37° 17'N) and Menez Gwen (37° 50'N) segments of the Mid-Atlantic Ridge, *Geochem. Geophys. Geosyst.*, 12, Q06016, doi:10.1029/2010GC003446.

Primary Publication DOI [doi:10.1029/2012GC004334](https://doi.org/10.1029/2012GC004334)

Link to publications

Related Funding Award(s) [0752281](#), [1061264](#)

Link to funding source

Date Released/Published 05/01/2013


Last Updated 03/06/2013 2:54 PM

+ Linking Data & Publications

Show thumbnails in outline

Abstract
Keywords

1. Introduction



2. Geological background and sample selection

2.1. The 9–10°N region and the Siqueiros Transform

2.2. The 12–14°N region

3. Analytical techniques

Table 1a

Table 1b

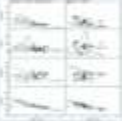
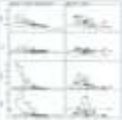
4. Results

4.1. 9–10°N and the Siqueiros Transform

4.2. 12–14°N

5. Discussion

5.1. Magma crystallization and contamination


Earth and Planetary Science Letters

Volume 251, Issues 3–4, 15 November 2006, Pages 209–231

The effects of variable sources, processes and contaminants on the composition of northern EPR MORB (8–10°N and 12–14°N): Evidence from volatiles (H₂O, CO₂, S) and halogens (F, Cl)

Petrus J. le Roux^{a, *}, Steven B. Shirey^b, Erik H. Hauri^c, Michael R. Perfit^d, John F. Bender^d

^a Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road NW, Washington, DC 20015, USA

^b Danish Lithosphere Centre, Øster Voldgade 10, 1350 Copenhagen K, Denmark

^c Department of Geological Sciences, University of Florida, P.O. Box 112120, Gainesville, FL 32611, USA

^d Department of Geography and Earth Sciences, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

<http://dx.doi.org/10.1016/j.epsl.2006.09.012>, How to Cite or Link Using DOI

[Permissions & Reprints](#)

Abstract

New volatile (H₂O, CO₂, S), halogen (F, Cl) and trace-element data for selected MORB glasses are reported from two geologically and geophysically well-studied regions on the East Pacific Rise (8–10°N and 12–14°N) with distinct differences in spreading rate and magma supply. Sample locations include on-axis and young off-axis eruptions, as well as off-axis fissures, abyssal hills and pillow mounds. H₂O, F, S and trace-element concentrations increase with decreasing MgO content, displaying over-enriched liquid lines of descent consistent with combined fractional and in-situ crystallization. A negative correlation between CO₂/Nb and MgO indicates simultaneous degassing and magma crystallization, while broadening of this correlation to lower CO₂/Nb at constant MgO indicates shallow degassing and CO₂ loss during magma transport to the seafloor.

Excess Cl concentrations and associated high Cl/Nb and Cl/K ratios of some northern EPR MORB result from variable pre-eruption contamination by high-salinity brines derived from supercritical phase separation of seawater within deeply-rooted hydrothermal circulation systems. In the faster-spreading

biobibliographic information


Citing and related articles

Applications and tools

Data for this Article

More information on this application

Data for this article is available at the following data repositories:

 IEDA EarthChem
32 extracted samples

eReader Formats

"The effects of variable sources, processes and contaminants on the composition of northern EPR MORB (8–10°N and 12–14°N): Evidence from volatiles (H₂O, CO₂, S) and halogens (F, Cl)"

article is available in eReader formats:
[ePUB](#) [Mobipocket](#)

[About eReader Formats](#)

Add Apps | Help

ADVERTISEMENT

EVENTS YOU MAY BE INTERESTED IN

9th International Symposium on the Cretaceous System
1–5 Sep 2013
Ankara, Turkey

Linking Data & Publications

Article

DOI:	10.1016/j.epsl.2006.09.012
Title:	THE EFFECTS OF VARIABLE SOURCES, PROCESSES AND CONTAMINANTS ON THE COMPOSITION OF NORTHERN EPR MORB (8-10 DEG N AND 12-14 DEG N): EVIDENCE FROM VOLATILES (H ₂ O, CO ₂ , S) AND HALOGENS (F, Cl)
Journal:	EARTH PLANET SCI LETT
Author:	LE ROUX, P J; PERFIT, M R; BENDER, JOHN F; SHIREY, S B; HAURI, E H
Pub. Year:	2006

Sample Locations:



Click to Enlarge

Data for this article

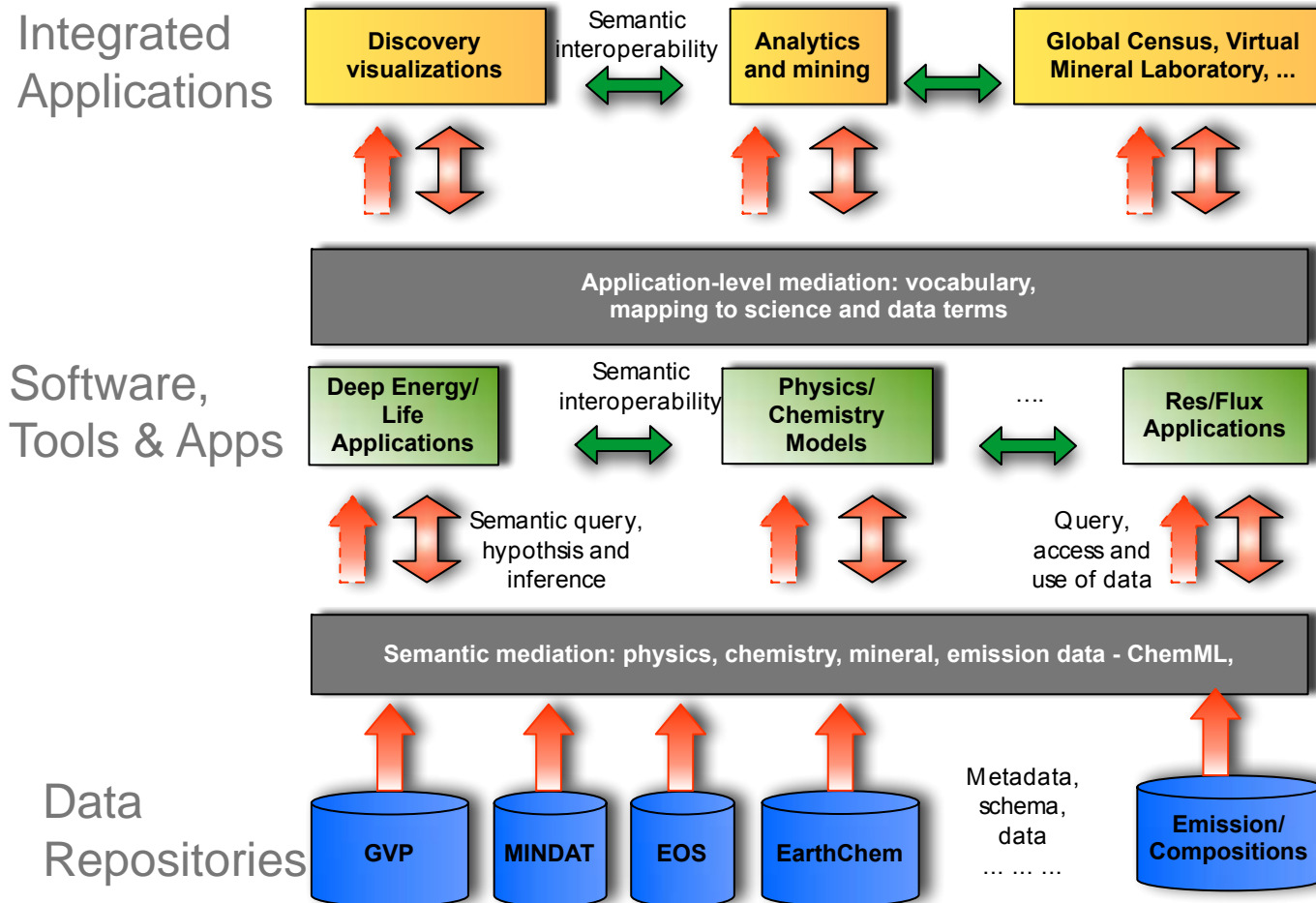
Records:	32
Source Database(s):	PETDB
Data @ EarthChem Portal:	Download Data

SAMPLE ID	SOURCE	DETAIL	LATITUDE	LONGITUDE	LOC PREC	MIN AGE	AGE	MAX AGE	MATERIAL	TYPE	COMPOSITION	ROCK NAME
ALV2490-003	PETDB	DETAILS	9.53	-104.25	0.02				GLASS	VOLCANIC	MAFIC	BASALT
ALV2489-005	PETDB	DETAILS	9.53	-104.22	0.02				GLASS	VOLCANIC	MAFIC	BASALT
ALV2489-002	PETDB	DETAILS	9.53	-104.23	0.02				GLASS	VOLCANIC	MAFIC	BASALT
ALV2490-010	PETDB	DETAILS	9.53	-104.26	0.02				GLASS	VOLCANIC	MAFIC	BASALT
ALV2759-005	PETDB	DETAILS	9.82	-104.31	0.02				GLASS	VOLCANIC	MAFIC	BASALT
ALV2390-005	PETDB	DETAILS	8.293	-104.023	0.001				GLASS	VOLCANIC	MAFIC	BASALT
ALV2768-006	PETDB	DETAILS	9.8333	-104.27	0.0001			1	GLASS	VOLCANIC	MAFIC	BASALT
ALV2768-004	PETDB	DETAILS	9.8333	-104.26	0.0001			1	GLASS	VOLCANIC	MAFIC	BASALT

+ Multi-Disciplinary Data Science Implementation

Slide: Courtesy of Peter Fox, RPI (July 2012)

Schematic for Deep Carbon Virtual Observatory and

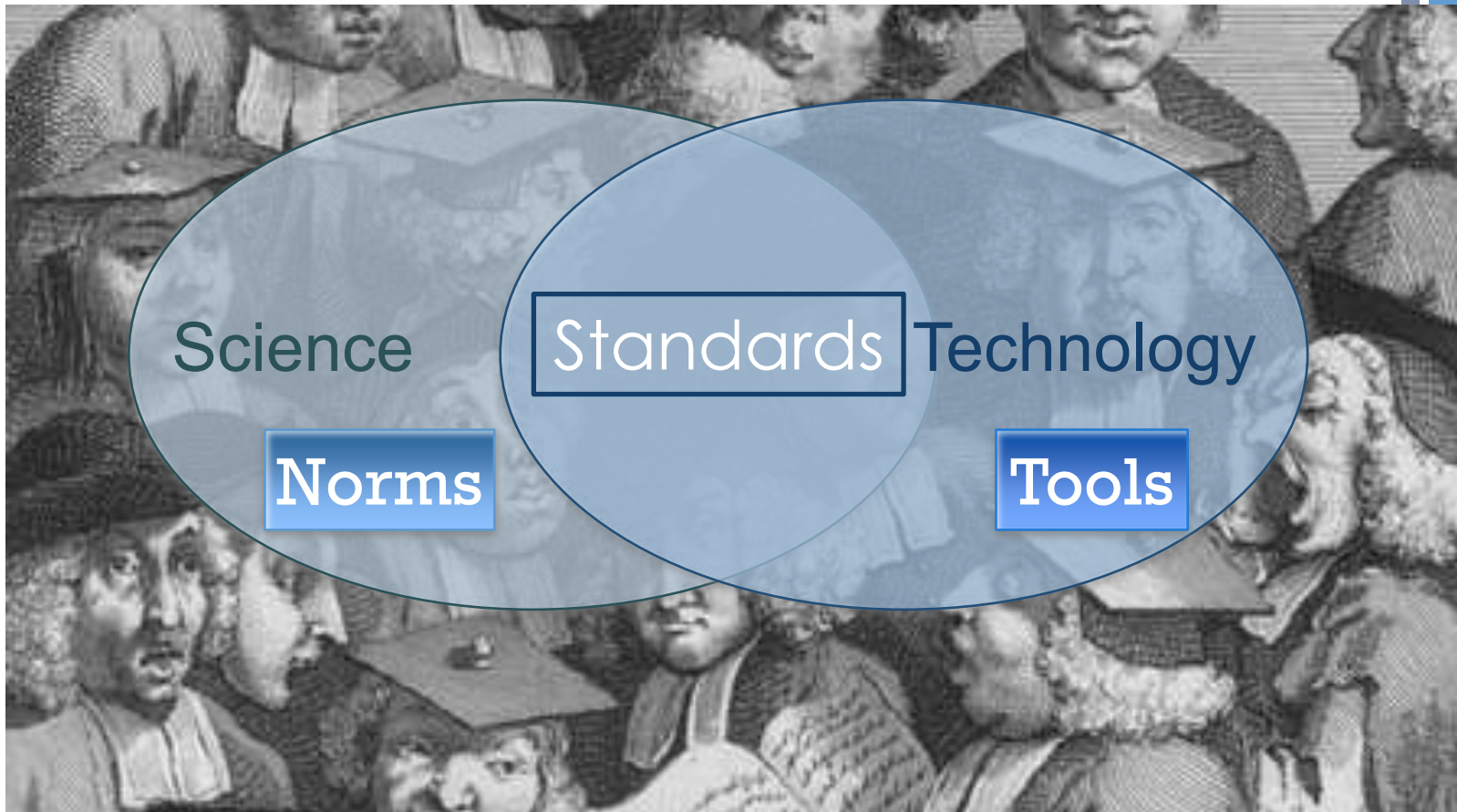


+ What makes data useful?

“Knowing that I can **trust** the numbers.”



+ Data Quality Standards





WEB OF KNOWLEDGE™

THOMSON REUTERS

ABOUT PRODUCTS & TOOLS BENEFITS & RESOURCES TRAINING & SUPPORT NEWS & EVENTS CONTACT US

Site Search [SEARCH](#)

Products and Tools Multidisciplinary Data Citation Index

THE
**DATA CITATION
INDEX™**
CONNECTING THE DATA TO
THE RESEARCH IT INFORMS

What is it?
VIEW VIDEO



DataCite

Helping you to find,
access, and reuse data

Why cite data?

We believe that you should cite data in just the same way that you can cite other sources of information, such as articles and books. Data citation can help by:

- enabling easy reuse and verification of data
- allowing the impact of data to be tracked
- creating a scholarly structure that recognises and rewards data producers

Data Citation

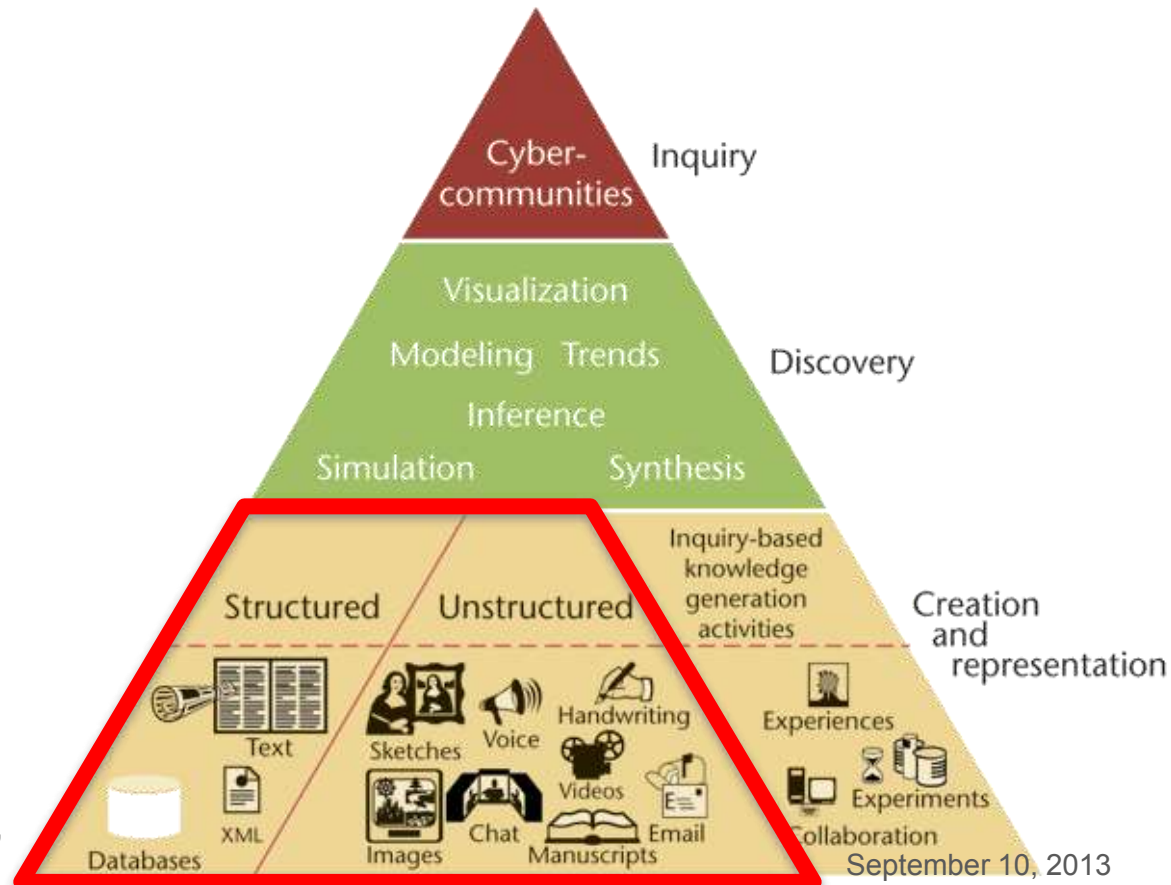
+ Polar Data Infrastructure

- Don't reinvent
 - Many data types already have well-established repositories, standards, best practices, community governance
 - Integration of polar data into appropriate disciplinary repositories will augment their quality and usage
 - Polar data are diverse, difficult to cover all data types with the appropriate level of expertise
- Fill obvious gaps
 - Leverage existing data infrastructure
 - Follow standards for data publication, metadata, and repository trustworthiness

+ The CI Vision

■ Enable new forms of scholarship that are

- information-intensive
- data-intensive
- distributed
- collaborative
- multi-disciplinary



From [Elmagarmid et al. \(2008\)](#):
 “Community-Cyberinfrastructure-Enabled
 Discovery in Science and Engineering”

+ Opportunities for Polar CI

- Leverage EarthCube developments
 - Build capabilities that can be used by other communities
 - Adopt & adapt developments that are useful
- Use EarthCube Resources
 - Stakeholder Alignment survey
 - Lists of existing resources
 - Gap analyses
 - Use cases / science scenarios
 - Social network (EarthCube MatchMaker)
- Experiences of community building & governance

CZO Disciplines

Big Data

Data Management/
CyberInfrastructure

Long Tail Data

Sample-based

Biology/Molecular

Biology/Ecology

Hydrology

Water Chemistry

Biogeochemistry

Climatology/
Meteorology

Sensor-based

Geochemistry/
Mineralogy

Modeling/
Computational Science

Soil Science/Pedology

Geospatial Grids & Vectors

Geology/Chronology

GIS/Remote Sensing

Geophysics

Geomorphology

Categorical

Engineering/Method
Development

Outreach/
Education Research

+ Recurring Themes of CI Gaps

- **Data (& samples)**: *access, coverage, integration, standards*
- **Models**: *dynamic, shared, linked*
- Interdisciplinary **conceptual frameworks**
- **Data analysis tools**: *visualization, multivariate analysis, statistics*
- **Data management support**: *workflows, software, education*
- **Knowledge**: *limitations and uses of data and models across, within and between disciplines*
- **Community**: *collaboration, shared knowledge of existing resources*



SciVerse ScienceDirect Hub ScienceDirect Scopus Applications Beate Specker Logot Go to SciVal Suite

Home + Recent Actions Publications Search My settings My alerts Help

Download PDF Export citation Jump to References More options...

Journal of South American Earth Sciences Volume 47, November 2013, Pages 12–31

Provenance, volcanic record, and tectonic setting of the Paleozoic Ventania Fold Belt and the Claromecó Foreland Basin: Implications on sedimentation and volcanism along the southwestern Gondwana margin

Luciano Alessandretti^a, Ruy Paulo Philipp^a, Farid Chemale Jr.^b, Matheus Philipe Brückmann^a, Gustavo Zvirtes^a, Vinicius Mattê^a, Victor A. Ramos^c

^a Instituto de Geociências, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, Porto Alegre 91509-900, RS, Brazil
^b Instituto de Geociências, Universidade de Brasília, Campus Universitário Darcy Ribeiro, 70910-900, DF, Brazil
^c Instituto de Estudios Andinos, CONICET, FCEyN, Universidad de Buenos Aires, Argentina

<http://dx.doi.org/10.1016/j.jsames.2013.05.006>, How to Cite or Link Using DOI

Permissions & Reprints

Highlights

- The Curamalal and Ventana Groups deposited at a passive margin environment.
- Upper part of the Pillaahuincó Group contains geochemical and petrographical signatures of active margin.
- The age of 288 ± 15 Ma was obtained for tuffs in the upper half of the sequence.
- Regional correlation of southern margin of Gondwana is presented.

Abstract

This study focuses on the provenance, volcanic record, and tectonic setting of the Paleozoic Ventania System, a geologic province which comprises the Cambro-Devonian Ventania Fold Belt and the adjoining Permo-Carboniferous Claromecó Foreland Basin, located inboard the deformation front. The Ventania Fold Belt is formed of the Curamalal and Ventana groups, which are composed of a variety of tectonic units that

Search ScienceDirect Search

Bibliographic information Citing and related articles Applications and tools

IEDA data for this article

Legend: NAVDAT, PciDF, GEOMOC, USG2, GANBENT, Sst/DE, MalPst/DE

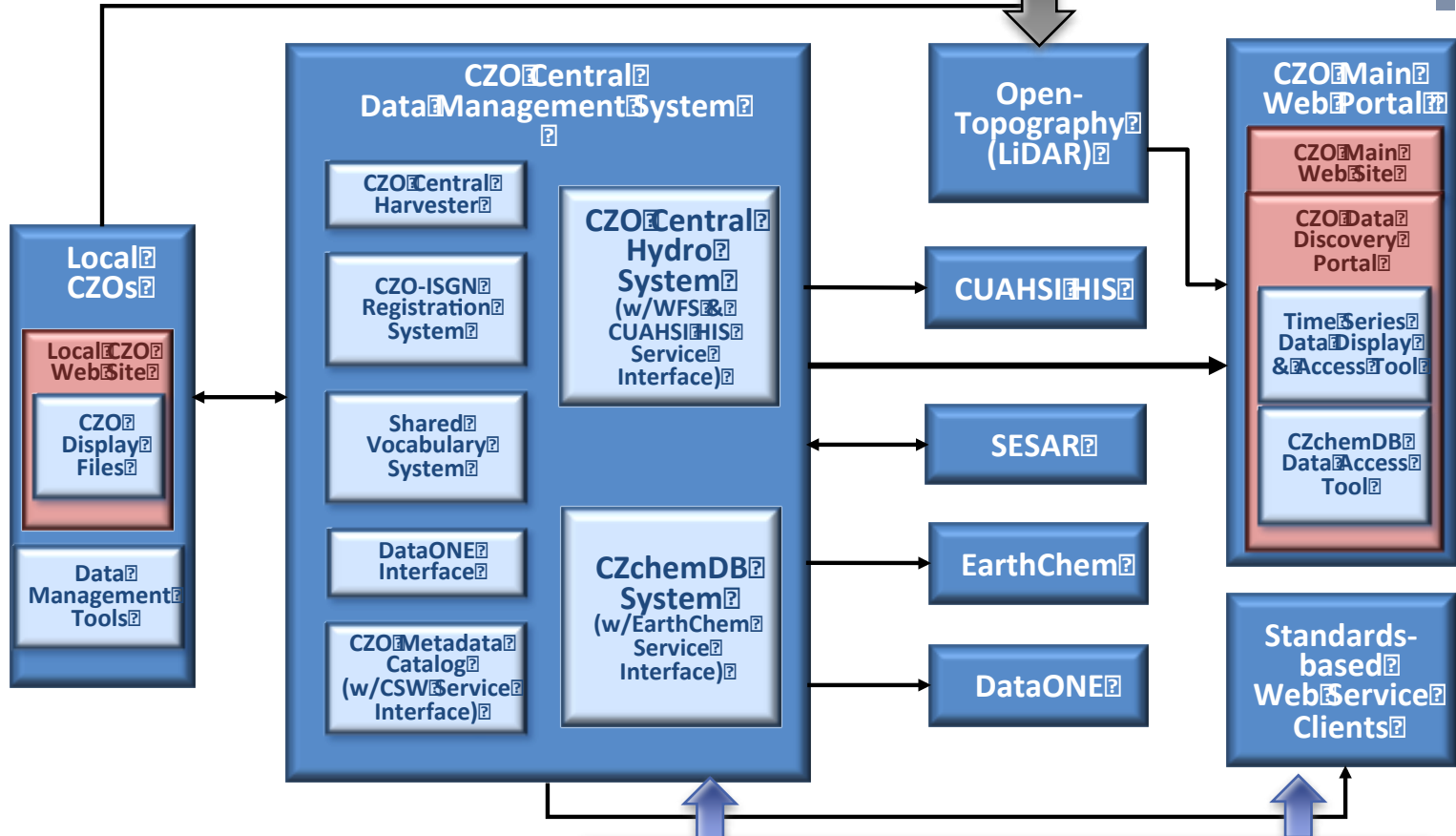
TAS Diagram

Powered by IEDA Workspace

Enriched Links (under development)

+ CZOData II Architecture

Leverage existing systems & capabilities



Build integrative components

Local CZOs

CZO Central Coordination Functions

CZO Central Data Repositories

Non-CZO Integrated Data & Discovery Sites

Standards-based Web Service Clients
September 10, 2013

+ Data Infrastructure for Polar Sciences

- Data Diversity: A unique situation?
 - Many disciplines
 - Big Data versus Long Tail
 - International

+ Disciplinary Repositories

- Ensure Usability
 - Develop/promote community-based data reporting standards
 - Provenance of data, data precision, errors, etc.
 - Work with publishers, editors, professional societies
 - Align with other data & interoperability standards
 - Provide services for persistent data identification (DOI), data attribution and citation, long-term archiving, etc.
- Advance Access
 - provide science-driven tools for data search & access
 - provide programmatic interfaces for cross-disciplinary use
 - links to publications

+ Data Infrastructure

Acquisition

Access

Analysis

Archiving

+ The Foundation: Data

- Open access to a global, distributed knowledge base of scientific data and information, including legacy data
- Seamless integration of data
 - within disciplines
 - across disciplines
 - with tools (visualization, analysis) and models

