EARTHCUBE ROADMAP: CROSS-DOMAIN INTEROPERABILITY

http://earthcube.ning.com/group/interop

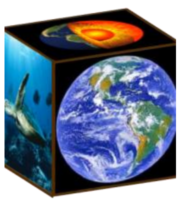
Ilya Zaslavsky
San Diego Supercomputer Center, UCSD

Presentation at Microsoft eScience Workshop; Chicago, October 8, 2012

EarthCube Goal and Outcomes

Goal

To 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.



Outcomes

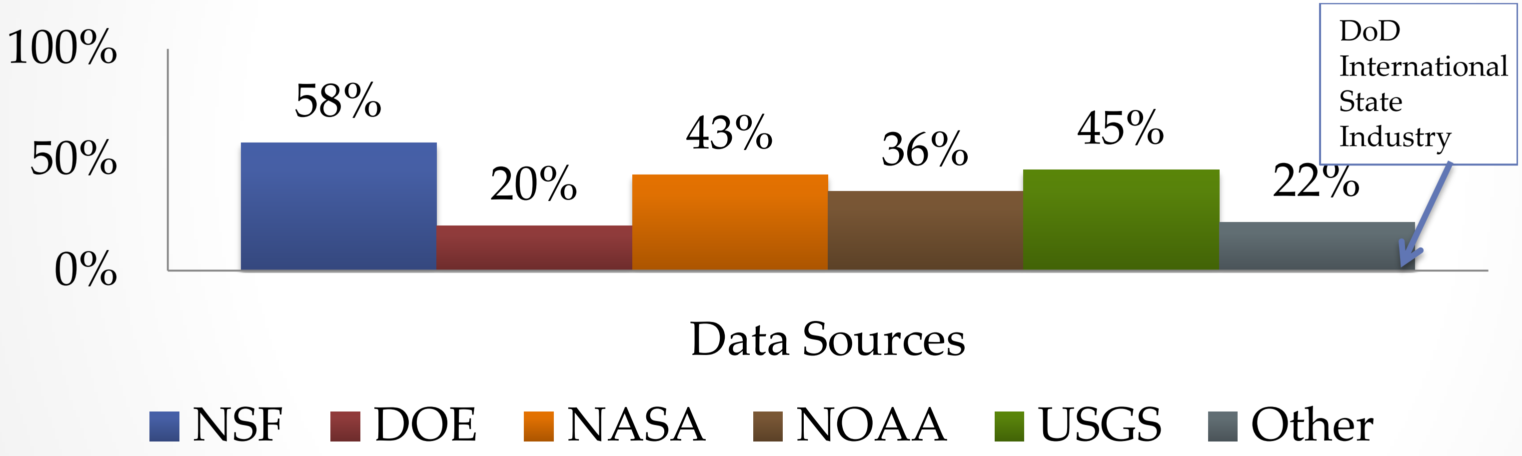
- Transform practices within the geosciences community spanning over the next decade
- Provide unprecedented new capabilities to researchers and educators
- Vastly improve the productivity of community
- Accelerate research on the Earth system
- Provide a knowledge management framework for the geosciences

Slides from Eva Zanzerkia, NSF EarthCube Team

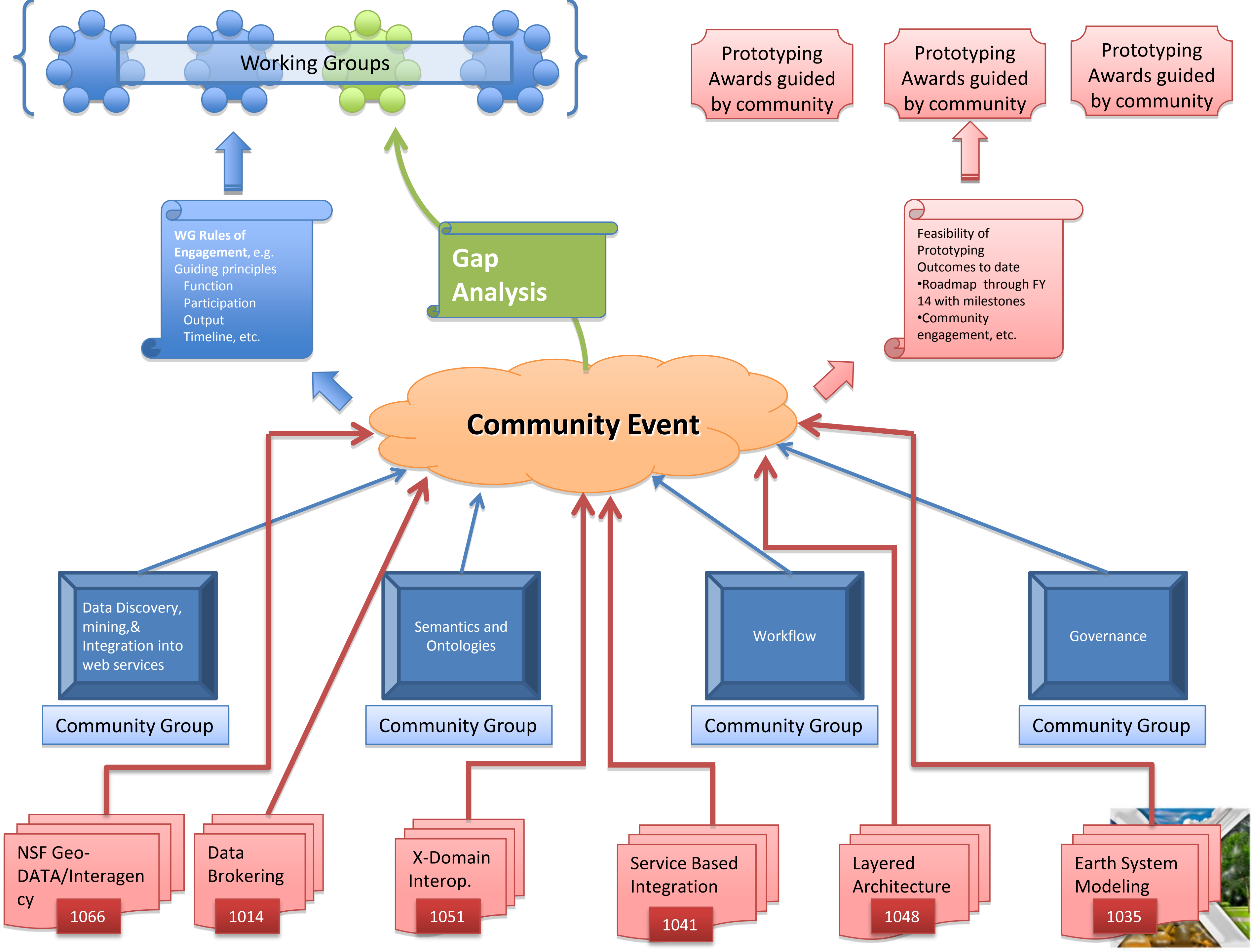
User Requirements Survey Data

- Do you require data or data products from outside your immediate discipline?
  - 52% Yes
  - 14% No

- Where do you obtain your data?



- Is there a place for products of your own work to be archived and accessed?
  - 45% Yes
  - 25% No



Working Groups
Prototyping Awards guided by community
Prototyping Awards guided by community
Prototyping Awards guided by community
WG Rules of Engagement, e.g.
Guiding principles
Function
Participation
Output
Timeline, etc.
Gap Analysis
Feasibility of Prototyping
Outcomes to date
•Roadmap through FY 14 with milestones
•Community engagement, etc.
Community Event
Data Discovery, mining,& Integration into web services
Community Group
Semantics and Ontologies
Community Group
Workflow
Community Group
Governance
Community Group
NSF Geo-DATA/Interagency
1066
Data Brokering
1014
X-Domain Interop.
1051
Service Based Integration
1041
Layered Architecture
1048
Earth System Modeling
1035

WHAT DOES CROSS-DOMAIN MEAN?

- **Datasets collected and compiled by different researchers at different times and places**

- **Application of data to analyses different from those for which they were originally collected.**

Out of context

WHAT IS READINESS FOR CROSS-DOMAIN INTEROPERABILITY?

- Cross-domain scientists can discover, interpret, access, integrate data from several domains.
- Readiness of product: the enabling technologies that allow that.
- Readiness of process: the process whereby that enabling technology adapts to changes.
- The product will always be evolving; will never be complete:
    - we cannot predict what needs will be.
    - it is not practical to annotate data against all potential needs.
    - often there is a need for data gathered in very specific ways.

Rationale: support discovery, interpretation, data access, data integration – across domains

Levels of interoperability

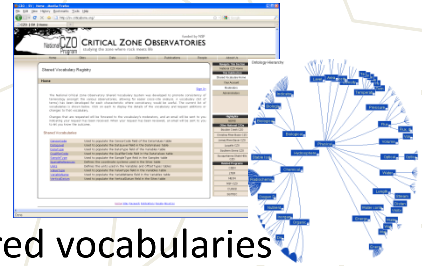
Domain system components

Diverse data types registered in a discovery system

Find and retrieve domain resources: files and file collections, services, documents - by thematic category, type, location

Resource catalogs and data discovery applications

Data available in compatible semantics: ontologies, controlled vocabularies



Shared vocabularies and ontology management

Deeper integration

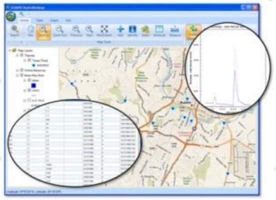
Wider variety of data

Data available via standard service interfaces (e.g. OGC WFS, SOS) but different information models

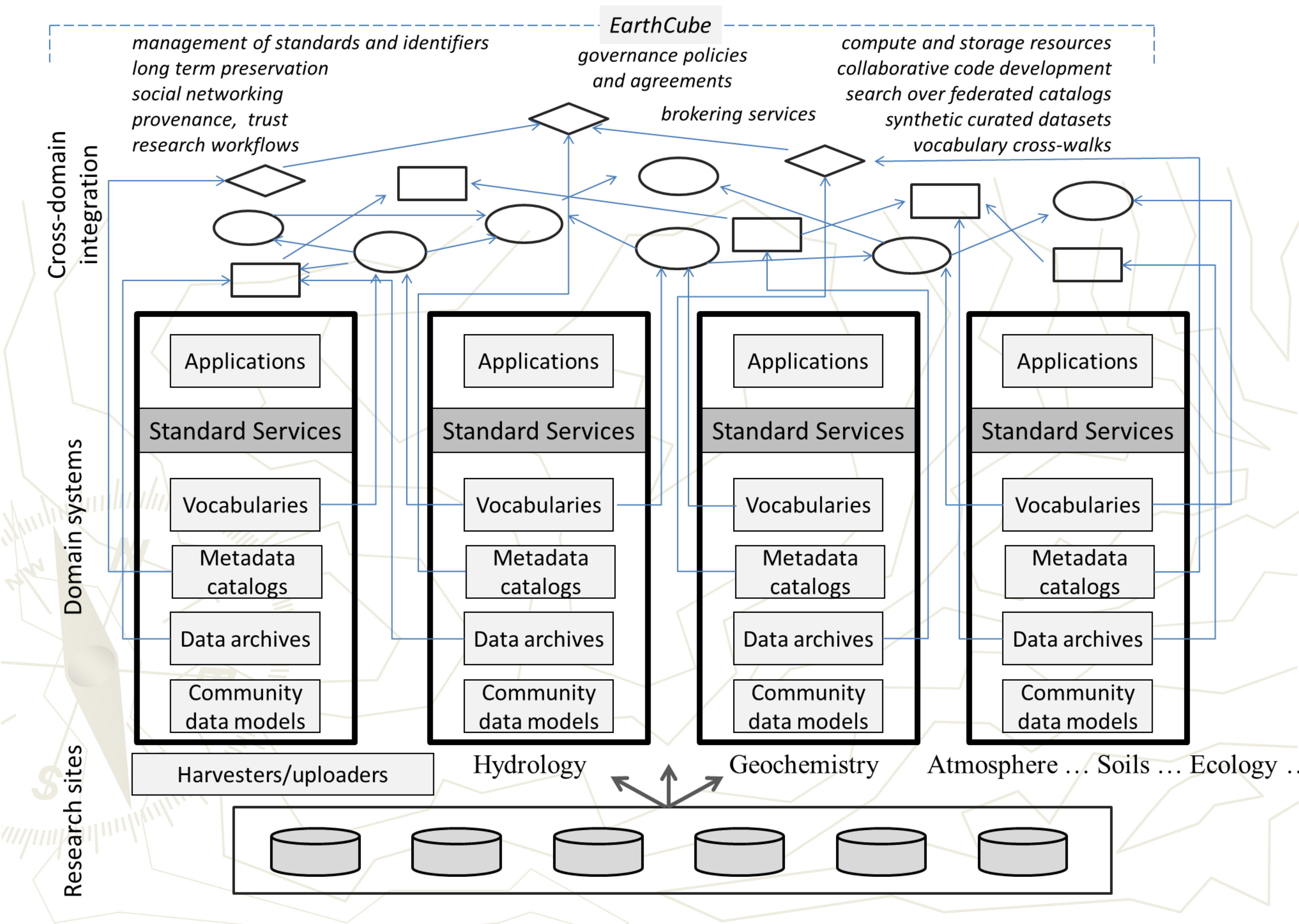
Service administration

Compatibility at the level of domain information models and databases

Well-understood data with formal information models available via standard services



Research applications


EarthCube
management of standards and identifiers
long term preservation
social networking
provenance, trust
research workflows
governance policies
and agreements
brokering services
compute and storage resources
collaborative code development
search over federated catalogs
synthetic curated datasets
vocabulary cross-walks
Cross-domain integration
Domain systems
Research sites
Applications
Standard Services
Vocabularies
Metadata catalogs
Data archives
Community data models
Harvesters/uploaders
Hydrology
Geochemistry
Atmosphere … Soils … Ecology …

CONTRIBUTE TO INVENTORIES!

- System architectures
- CI components
- Model catalogs
- Use Cases

- http://earthcube.ning.com/group/interop/page/community-inventories

READINESS ASSESSMENT 1

	Catalog Metadata
M1	Has a data listing
M2	Uses minimal metadata standard, such as Dublin Core
M3	Uses metadata standard, such as FGDC, or INSPIRE
	Catalog Search
S1	Search Interface
S2	Search API, not following a standard
S3	Complies with Opensearch API
S4	Complies with OGC CSW API
	Catalog Harvest
H1	Has a harvest API
H2	OAI API
H3	OGC CSW API

	Vocabulary – Control and Access
V1	Uses controlled terminology
V2	Community Managed Terminology
V3	SPARQL
	Vocabulary -- Representation
T1	Listing of terminology, such as web pages
T2	Uses ontology or SKOS

	Information Model Conceptual
C0	Unspecified
C1	Domain/Conceptual Model using UML
C2	Domain/Conceptual Model using UML based on OGC or ISO standards
	Information Model as XML
X1	XML Format. Schema may not be specified
X2	Xml Schema
	Information Model as SQL
S1	Provides an SQL Schema

	Data Access API
A1	Bulk download
A2	Static URL
A3	Web Service
	Data Query API
Q1	Simple query subset
Q2	Complex query
Q3	Processing Subset

Also evaluated: processing services; visualization services; community consensus efforts; identifier persistence

READINESS ASSESSMENT 2

	Catalog			Vocabulary		Data Access		Processing Services		Visualization Services	Information Model			Identifiers	Consensus Effort
	Metadata	Search	Harvest	Vocabulary	Terminology	Access	Query	Remote Execution	Local Execution	Visualization Services	Conceptual	XML	SQL	Persistence	
GEOSS-- Global Earth Observation System of Systems	3	4	3	3	2	3	3	1,2		2	2	2		1	2
GEON --GEOsciences Network	3	2,4		2,3	2	1	2	1		3	0	1		1	
SONET-- Scientific Observations Network	3			3	2	3	2				2	2			1
EOSDIS --Earth Observing System Data and Information System	3	3	1		1	3	2	1			2	2		3	2
CZO--Critical Zone Observatory		2	1		2	3	1				1	1	1	0,1	1

READINESS ASSESSMENT 3

	Catalog			Vocabulary		Data Access		Processing Services		Visualization Services	Information Model			Identifiers	Consensus Effort
	Metadata	Search	Harvest	Vocabulary	Terminology	Access	Query	Remote Execution	Local Execution	Visualization Services	Conceptual	XML	SQL	Persistence	
SSIS -- Spatial Information Services Stack	3	4	3	3	2	3	3	1		3	2	2	1	2	1
USGS WaterSmart				3	3	3	2			3	2	2		1	
CUAHSI HIS --CUAHSI Hydrologic Information System		2	1		2	3	1				1	1	1	0,1	1
USGIN --United States Geoscience Information Network	3	4	3			3	1			2	2	2		1	
DataOne	3	2	1			3	1			3	0			2	1
OpenTopograhy	3	1				1	2	1	1	3				3	
OOI -- Ocean Observatories Initiative	2	2	1		2	1,2,3	2	1	1		1	1		2	2
VOEIS -- Virtual Observatory and Ecological Informatics System	2	2		2	1	3	1	1	1	3	1			1	

MODEL CATALOGS

- AS PROXIES OF USE CASES

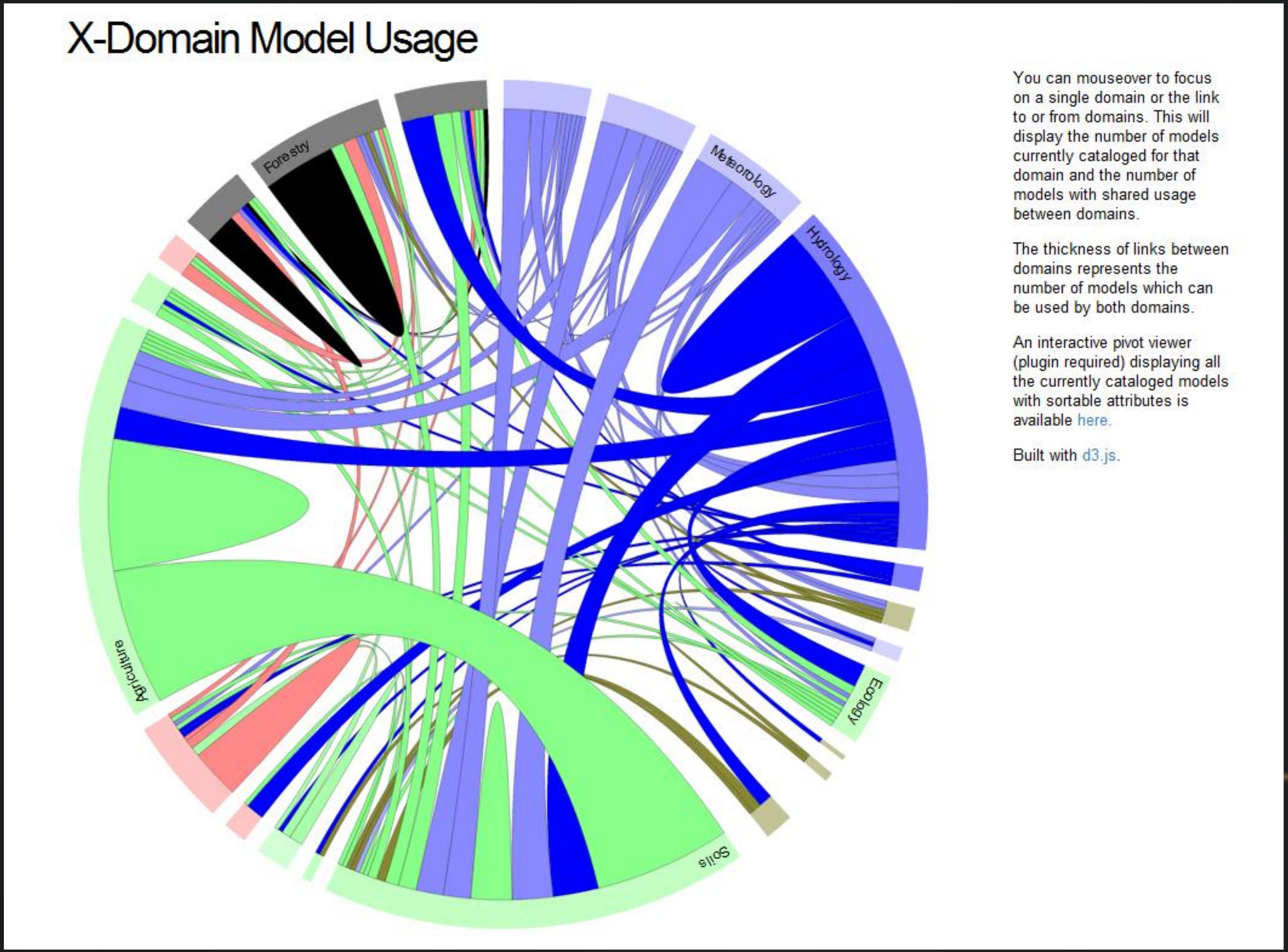
Not an isotropic EarthCube!

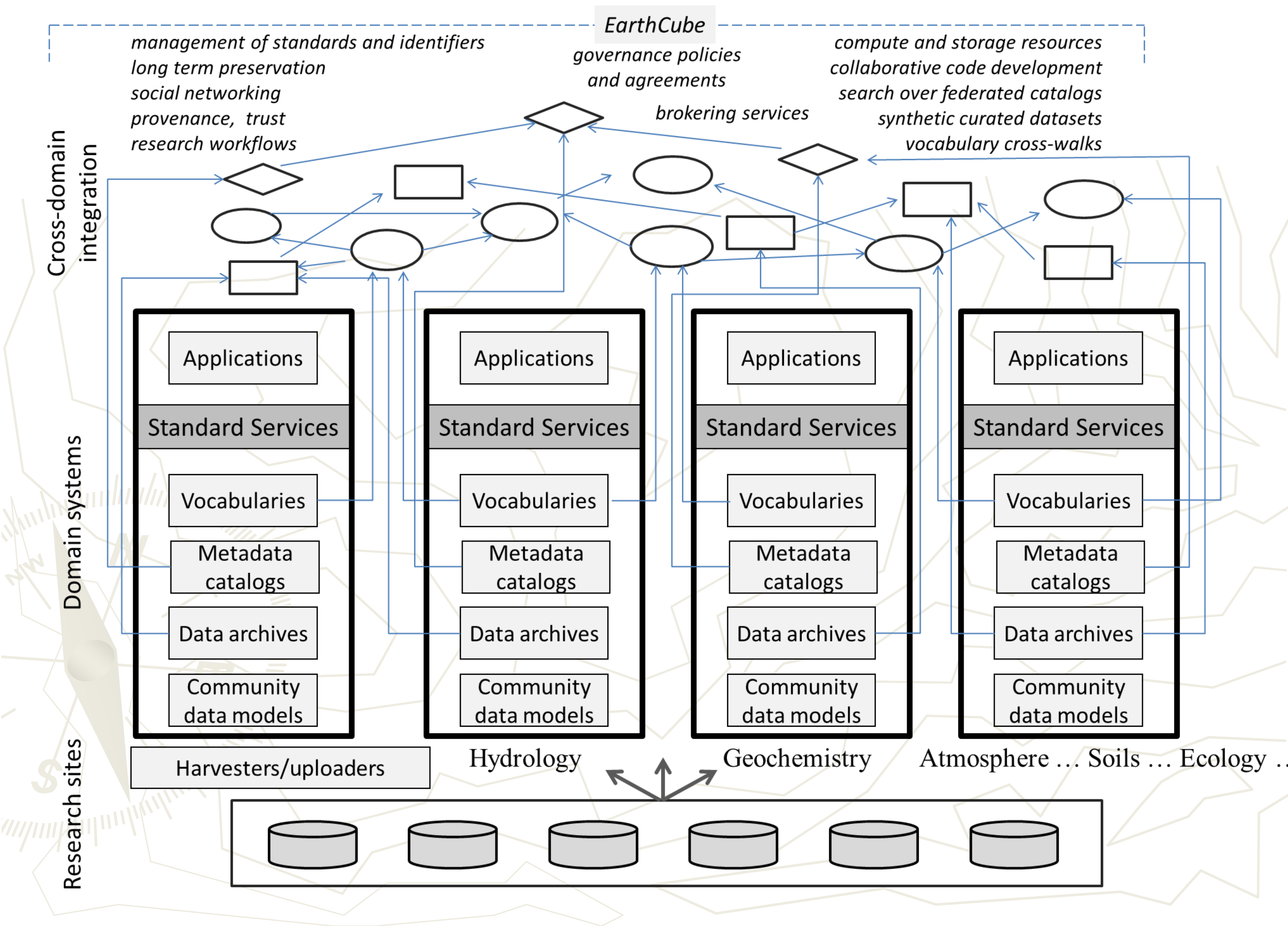
- **Manual parsing of models catalogued by CSDMS**
- **Environmental models assembled by the TESS project**
  - http://maxim.ucsd.edu/tessmodels3/
- **ESMF coupled models**
  - http://maxim.ucsd.edu/esmfmodels/
- **CSDMS-compliant, OpenMI-compliant and ESMF-compliant model components**
  - http://maxim.ucsd.edu/NoaaModelsList/
- **To be added:**
  - **EPA models**
  - **NOAA models**

- **See http://earthcube.ning.com/group/interop/page/community-inventories**
- **Currently ~ 750 models; 800 to be added soon**

PATHWAYS THROUGH EARTHCUBE

- http://maxim.ucsd.edu/crossdomain/Models.html

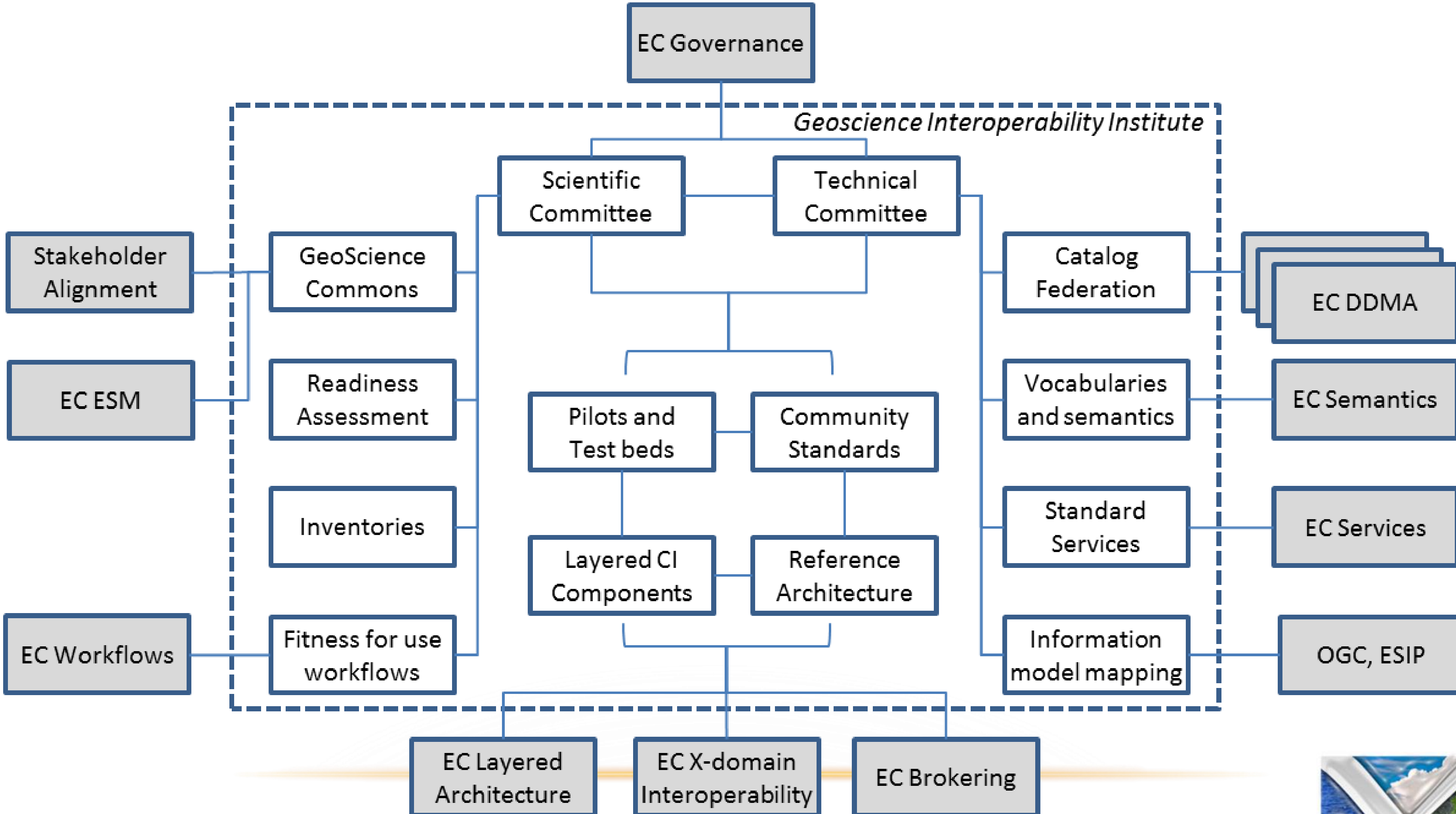



EarthCube
management of standards and identifiers
long term preservation
social networking
provenance, trust
research workflows
governance policies
and agreements
brokering services
compute and storage resources
collaborative code development
search over federated catalogs
synthetic curated datasets
vocabulary cross-walks
Cross-domain integration
Domain systems
Research sites
Applications
Standard Services
Vocabularies
Metadata catalogs
Data archives
Community data models
Harvesters/uploaders
Hydrology
Geochemistry
Atmosphere … Soils … Ecology …

GEOSCIENCE INTEROPERABILITY INSTITUTE

- Cross domain resource center; representatives from various communities to insure enterprise integration
- Virtual or real organization
- Support multi-disciplinary, interoperability test beds
- Represent EarthCube in collaborations, e.g. Belmont Forum
- Interoperability readiness assessments
- Education and engagement programs
- Resource inventories
- Workshops/ consensus meetings
- Visiting scholars
- Plumbing + forward-looking projects
  - Interacting, partnering, involving funders

EC GOVERNANCE AND GII



GENERAL CHALLENGES

- **How to capture measurement intent and context, add semantics – without additional burden for data publishers**
- **How to establish trust across domains**
  - **From the weather radar use in hydrologic modeling example:**
    - Wide variety of types; no standard metadata or data format; complex vocabularies; access issues. Also: cross-domain use issues:

Problem	Cause	Potential Solution
Hail contamination	Assumes high rainfall rate	Use of dual-pol, QC
Bright band	Ice at mid-levels biases dBZ	Real-time QC, 2 radar beams
Ground clutter	Wind farms, blockage	Use of Neural Net, velocity
Radar attenuation	High-frequency radars	Real-time QC model, fix
Anomalous propagation	High stable environment	Use of Level 1, velocity
Velocity de-aliasing	High velocity returns	Real-time QC
Radar calibration	Poor maintenance	Post QC
Over/under estimation below beam	Radar too far from area of interest; undersampled	Improved radar sampling; additional surface input
Poor time sampling	Radar 5-min volume sampling	Improved temporal sampling
Evapotranspiration under beam	Lack of surface information	Additional surface data
Spatial interpolation	Polar to Cartesian coordinates	Interpolation algorithm
Use of Reflectivity	Does not measure rain directly	Calibration against surface data

GENERAL CHALLENGES

- Third-party brokering versus engaging domain systems in standards support
- Supporting both loose coupling and tight integration for high performance
- What is the model of “fitness for use”, and how to assess “quality” of interoperability (i.e. beyond supported interop levels)
- Prioritization of cross-domain science questions?
- How to engage different research and education communities?
- How to reconcile policies and data lifecycles across domains?
- How would data providers get credit for re-use?

QUESTIONS TO ALL

- How is cross-domain interoperability different from intra-domain?
- How do we move to a self-sustainable cross-domain CI for broad groups of researchers?
- How do we know if we accomplished interoperability?
- What technologies look most promising and what are the low hanging fruit?
- How do we balance needs of large data crunchers vs myriad of small research projects across domains?
- To make cross-domain CI a reality, should we mandate global standards or allow local domain-specific protocols?