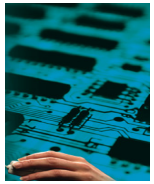
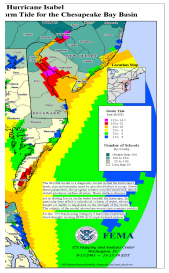
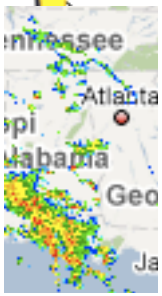






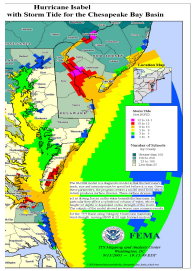
Metadata and Preservation in Geosciences: Issues at Scale

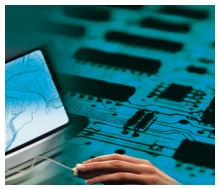
Beth Plale

Director, Data to Insight Center of PTI

School of Informatics and Computing, Indiana University Bloomington

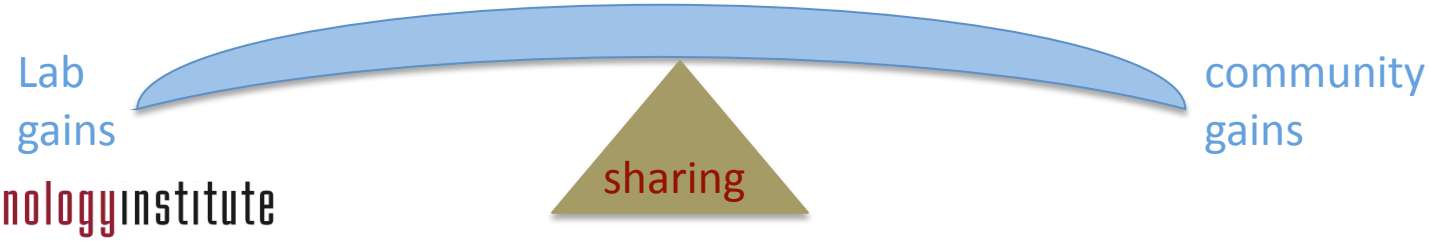
pervasivetechnologyinstitute
AT INDIANA UNIVERSITY



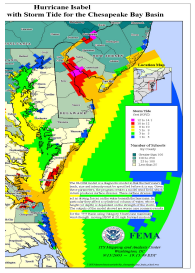


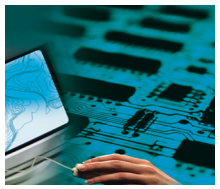
Riches of Data Deluge

- As volume of digital scientific data increases so increases opportunities for new data-driven science
- Funding agencies beginning to mandate that research data products be made publically available
  - "The National Science Foundation is committed to the principle that the various forms of data collected with public funds belong in the public domain."



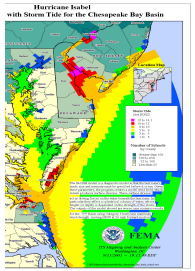
pervasivetechnologyinstitute
AT INDIANA UNIVERSITY

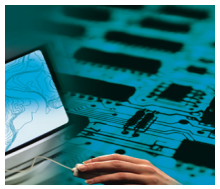




What does it mean “in public domain”?

- It’s available on my department’s server, and I’m happy to share it if you ask.
- It’s available on the web, but you need to know the file name to find it
- It’s available on the web, but only until the machine it sits on gets too old
- *None of these good enough*

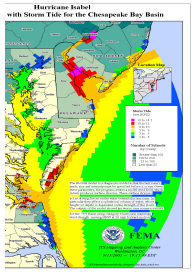


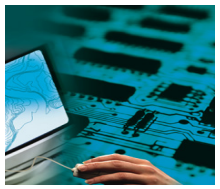


Belonging in public domain

- A collection belonging in public domain should:
  - Be findable ..... **Discovery**
  - Be useable in another's research ..... **Process**
  - Be as useful in 10 years as it is in 10 months ..... **Preservation**

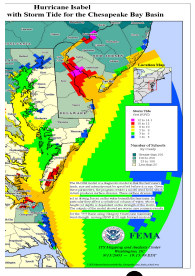
pervasivetechnologyinstitute
AT INDIANA UNIVERSITY

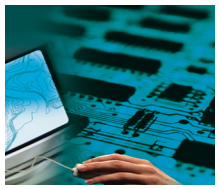




Discovery

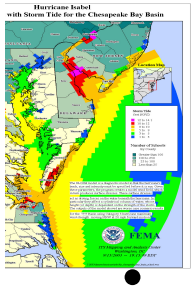
- Legacy solution is to share data by embedding lots of metadata into file names
  - http://lead.unidata.ucar.edu:8080/thredds/dodsC/LEAD/radar2/KVTX/20090914/Level2_KVTX_20090914_1321.ar2v
  - http://lead.unidata.ucar.edu:8080/thredds/fileServer/LEAD/model/NCEP/NAM/CONUS_80km/NAM_CONUS_80km_20090914_1200.grib1
- Good for those initiated into “inner-circle”
- Relying on long file names isn’t enough

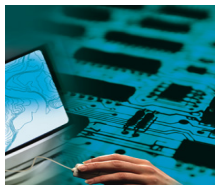




Process: act of operating on data

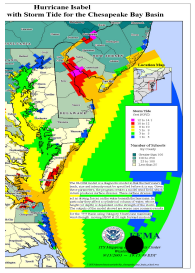
- Regrid array data to match co-ordinate system of some other data set
- Sequence pattern matching
  - e.g., execution trace, DNA sequence
- Statistical analysis
  - e.g., regression analysis, confidence interval
- Inverted index for full text search

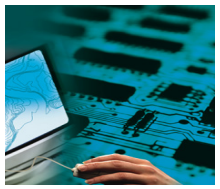




Preservation : long term perspective

- Deals with issues of:
  - Quality of metadata about data, its accuracy, ownership. This is curation and it is expensive.
  - Media longevity – media on which digital data are stored
  - Environmental conditions – conditions in which media are stored and used. Storage on which digital files are held
  - Software and hardware requirements – tools needed to ensure longevity and usability of digital data and metadata
  - Workflow requirements – policies and practices needed to ensure longevity and usability of digital data, metadata, and tools





Preservation

- From perspective of researcher
  - Researcher needs to provide more than bag of files all with long file names, which are an incomplete and inadequate form of metadata
  - Researcher expects to retain ownership, receive credit
- Curation is expensive
  - Open problem to reduce cost of curation

Portal/web access ⟷ Metadata ⟷ Data

Discovery

Portal/web access

↕

Database management system

Name: wrfout_d01_2009-03-05_12:00:00
GUID:urn:uuid:b419247e-876d-4842-b463-e79fc50aea3b
Owner:/O=LEAD Project/OU=portal.leadproject.org/
OU=cs.indiana.edu/CN=plale/EMAIL=plale@cs.indiana.edu
Create time: …

↕

File system (e.g., Data Capacitor)

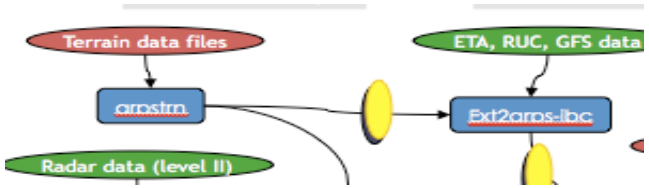
b419247e-876d-4842-b463-e79fc50aea3b

<u>Discovery solution shown:</u>

Metadata in database; data in files on disk

<u>Alternate solution:</u>

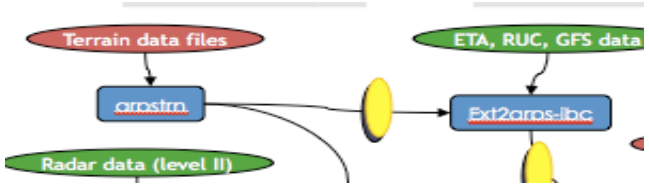
Metadata and data both in database



pervasivetechnologyinstitute
AT INDIANA UNIVERSITY

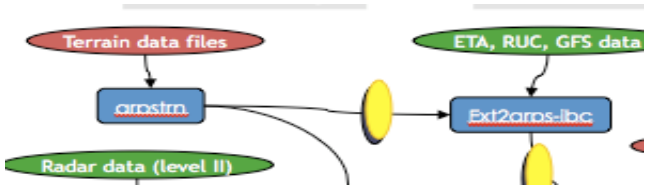
Platforms for discovery

- Databases do Discovery very well, so using databases to store metadata makes sense.
- RDF (semantic) stores also good at storing metadata, particularly complex relationship between data objects
- Preservation platforms support discovery, but are strongest at curation and life cycle workflows

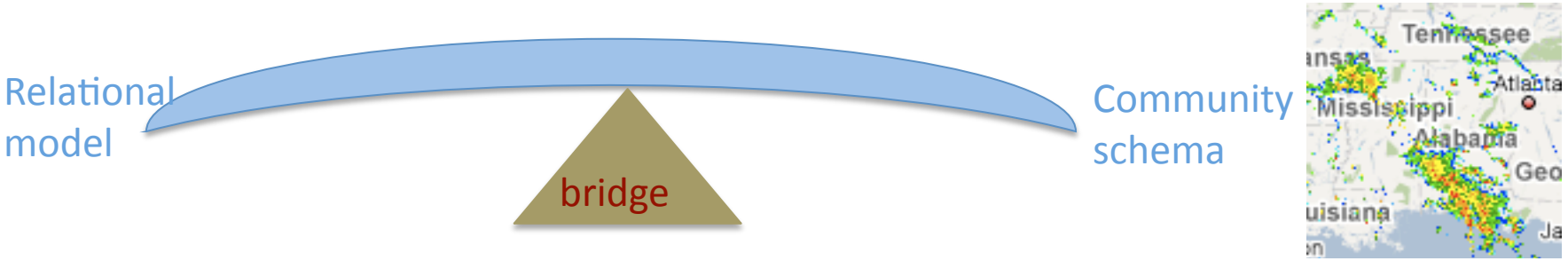


Platforms for Process

- … depends on Process.
  - Text applications don't use relational DBMSs
  - Regriding array data using SQL operations is nearly impossible. Earth scientists want science-specific operations as primitive operation in database.
  - Bio community uses RDF stores to store data. But column store (data warehouse servers) better at some RDF workloads.

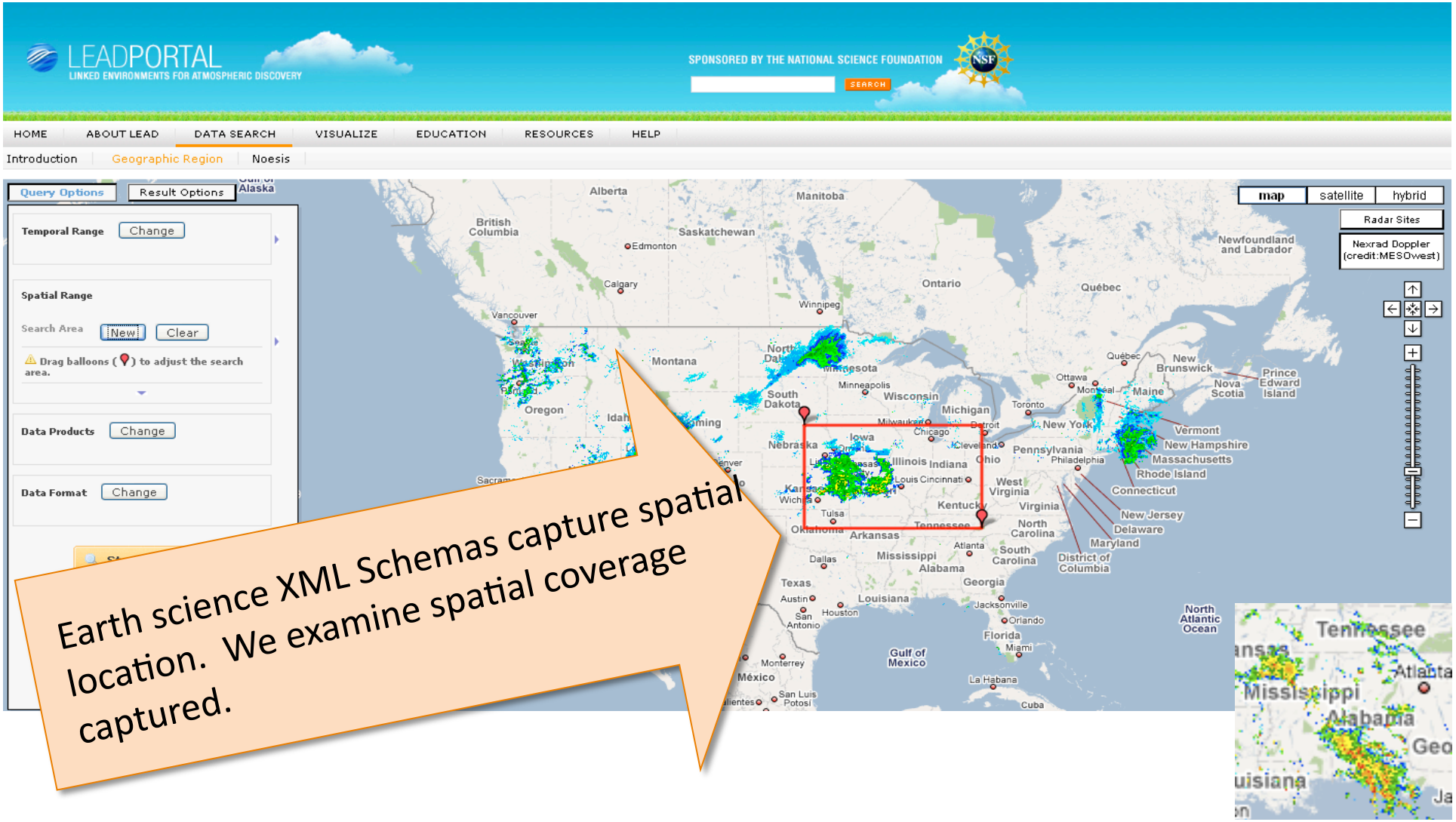


Additional issue scientists face when sharing: community schemas.



XMC Cat is a metadata catalog with wizard that builds a bridge between community XML Schema and general relational metadata data model

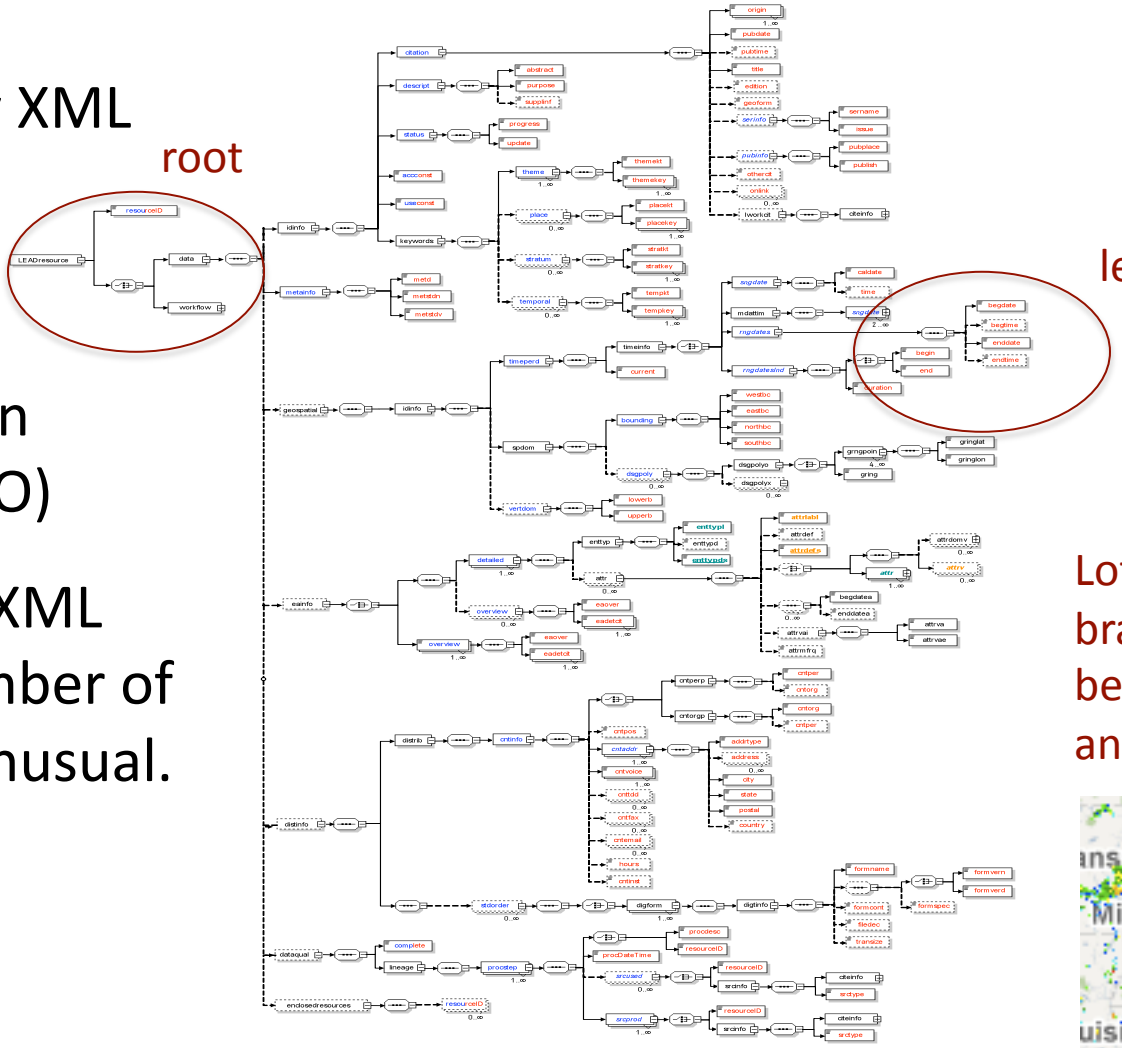
pervasivetechnologyinstitute
AT INDIANA UNIVERSITY


LEADPORTAL
LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY
SPONSORED BY THE NATIONAL SCIENCE FOUNDATION
SEARCH
HOME
ABOUT LEAD
DATA SEARCH
VISUALIZE
EDUCATION
RESOURCES
HELP
Introduction
Geographic Region
Noesis
Query Options
Result Options
Temporal Range
Change
Spatial Range
Search Area
New
Clear
Drag balloons ( ) to adjust the search area.
Data Products
Change
Data Format
Change
map
satellite
hybrid
Radar Sites
Nexrad Doppler (credit:MESOwest)
Earth science XML Schemas capture spatial location. We examine spatial coverage captured.

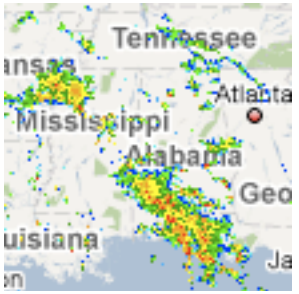
Typical community XML Schema

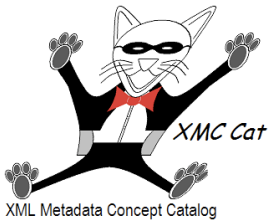
- Profile of common standard (FGDC/ISO)

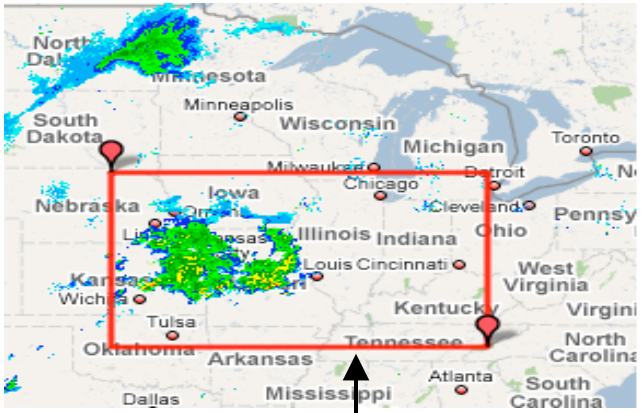
- Composed of 11 XML Schema files – number of schema files not unusual.



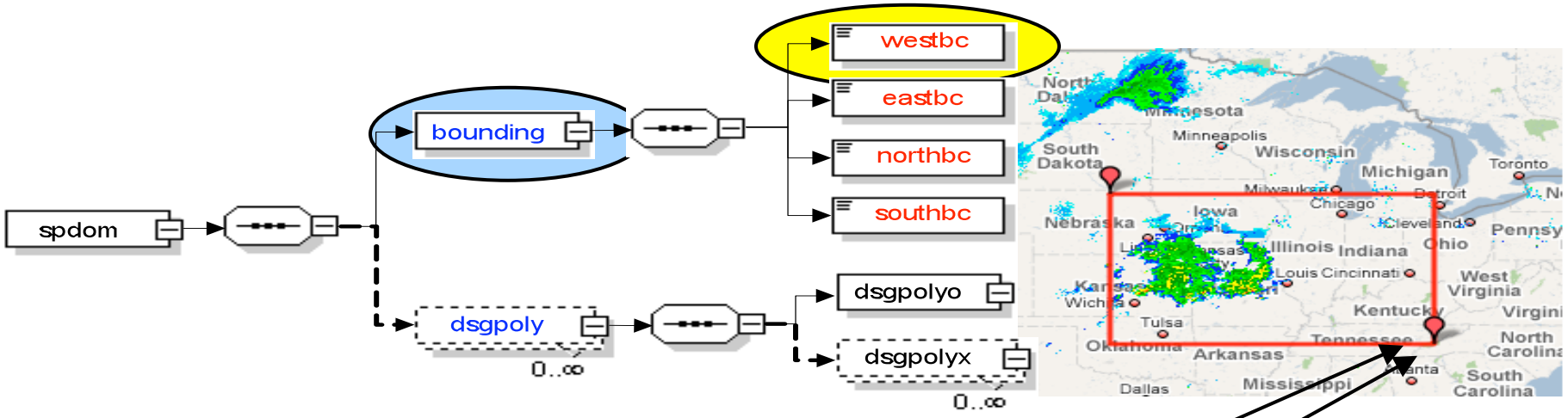
Lots of branches between root and leaves







Spatial coverage of data object described by in one of the schemas: IdentificationInformation.xsd



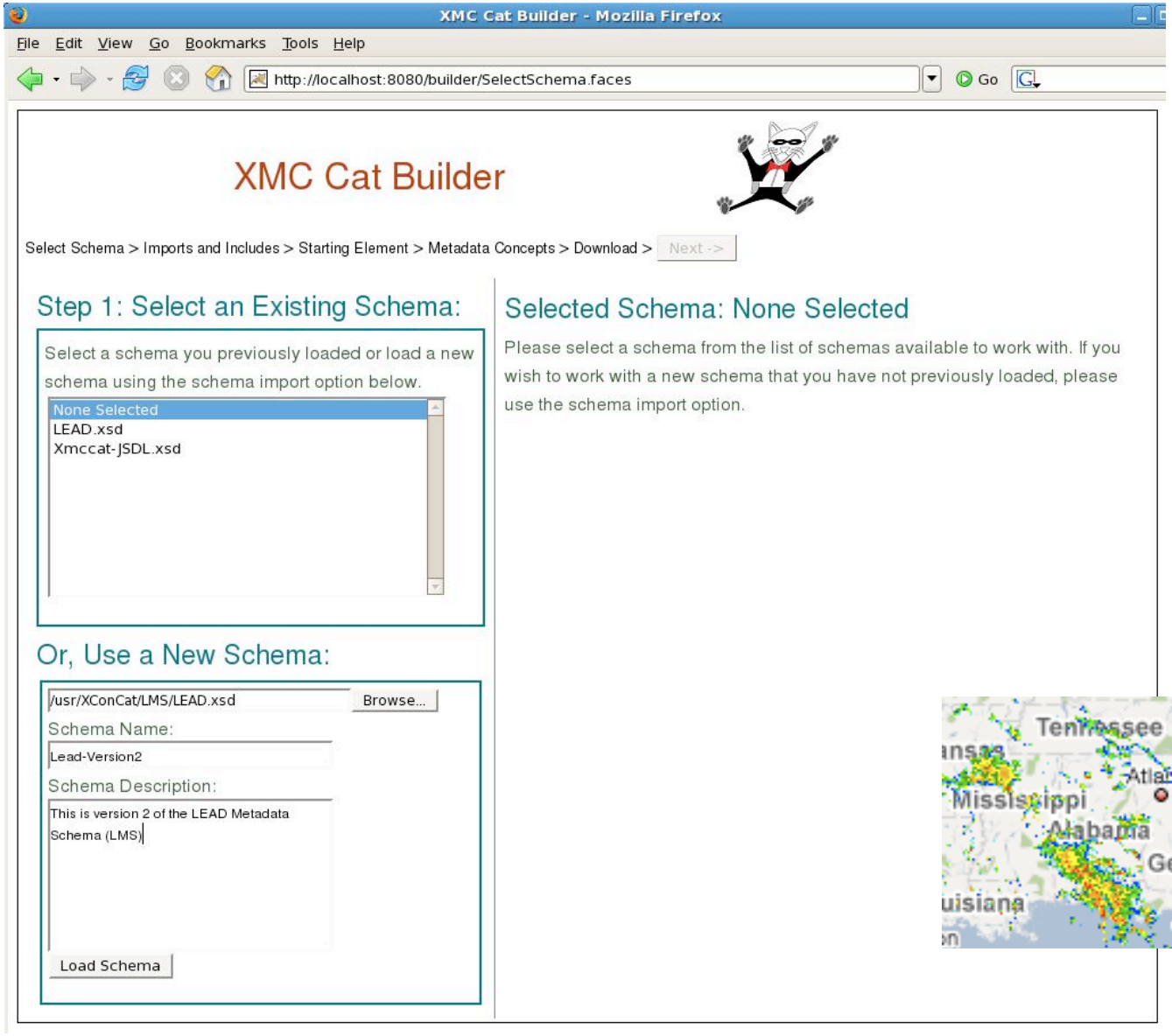

```
<xs:element name="bounding" type="boundingType“/>

<xs:complexType name="boundingType">
   <xs:sequence>
      <xs:element ref="westbc"/>
      <xs:element ref="eastbc"/>
      <xs:element ref="northbc"/>
      <xs:element ref="southbc"/>
   </xs:sequence>
</xs:complexType>
```

Need to get this XML into a relational database and query it using geospatial support provided by DBMS

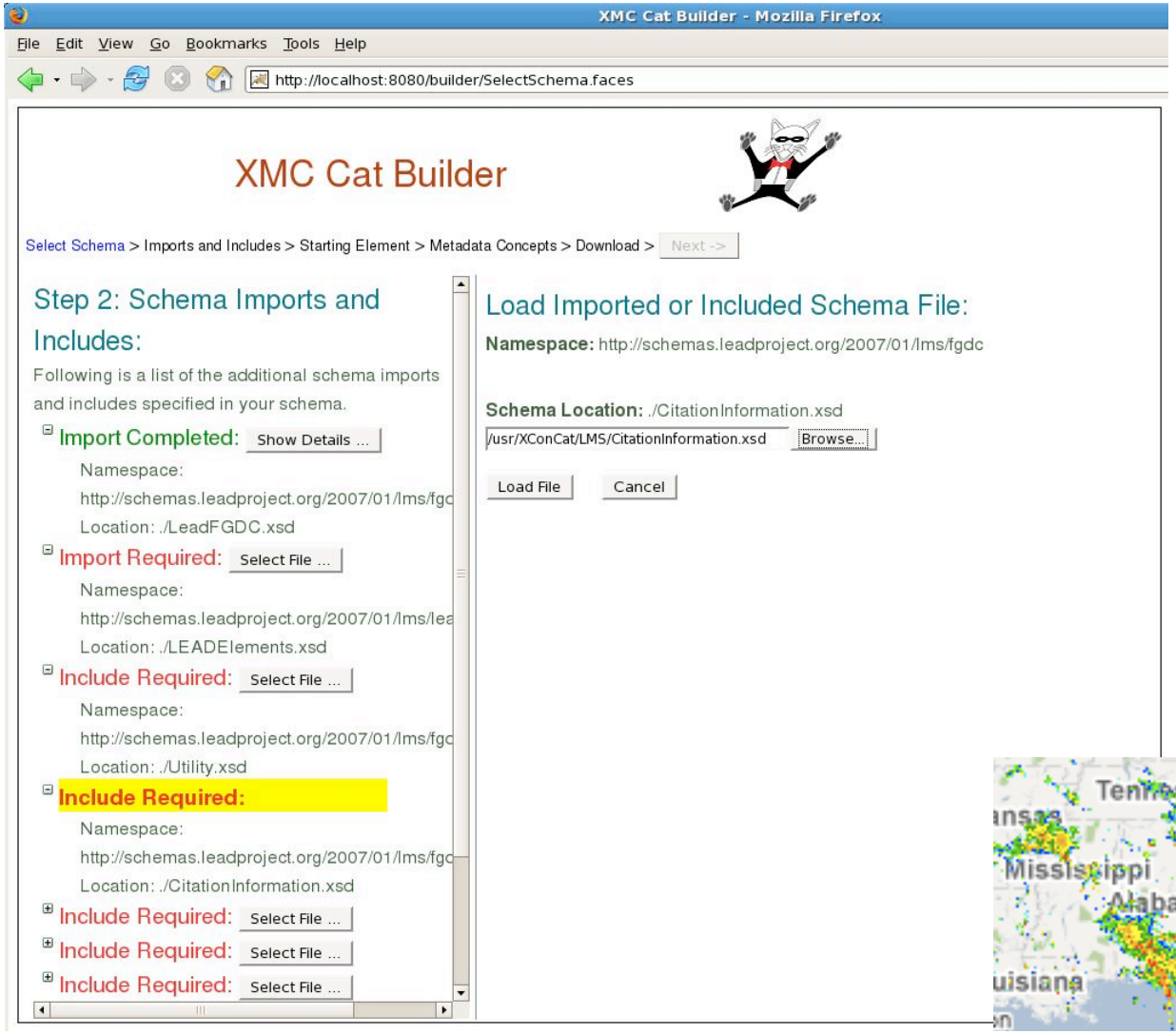
To publish your collection in community "compliant" XML, you

1. Download the XMC Cat catalog, and
2. Walk through a wizard that lets XMC Cat set itself up with your community schema.



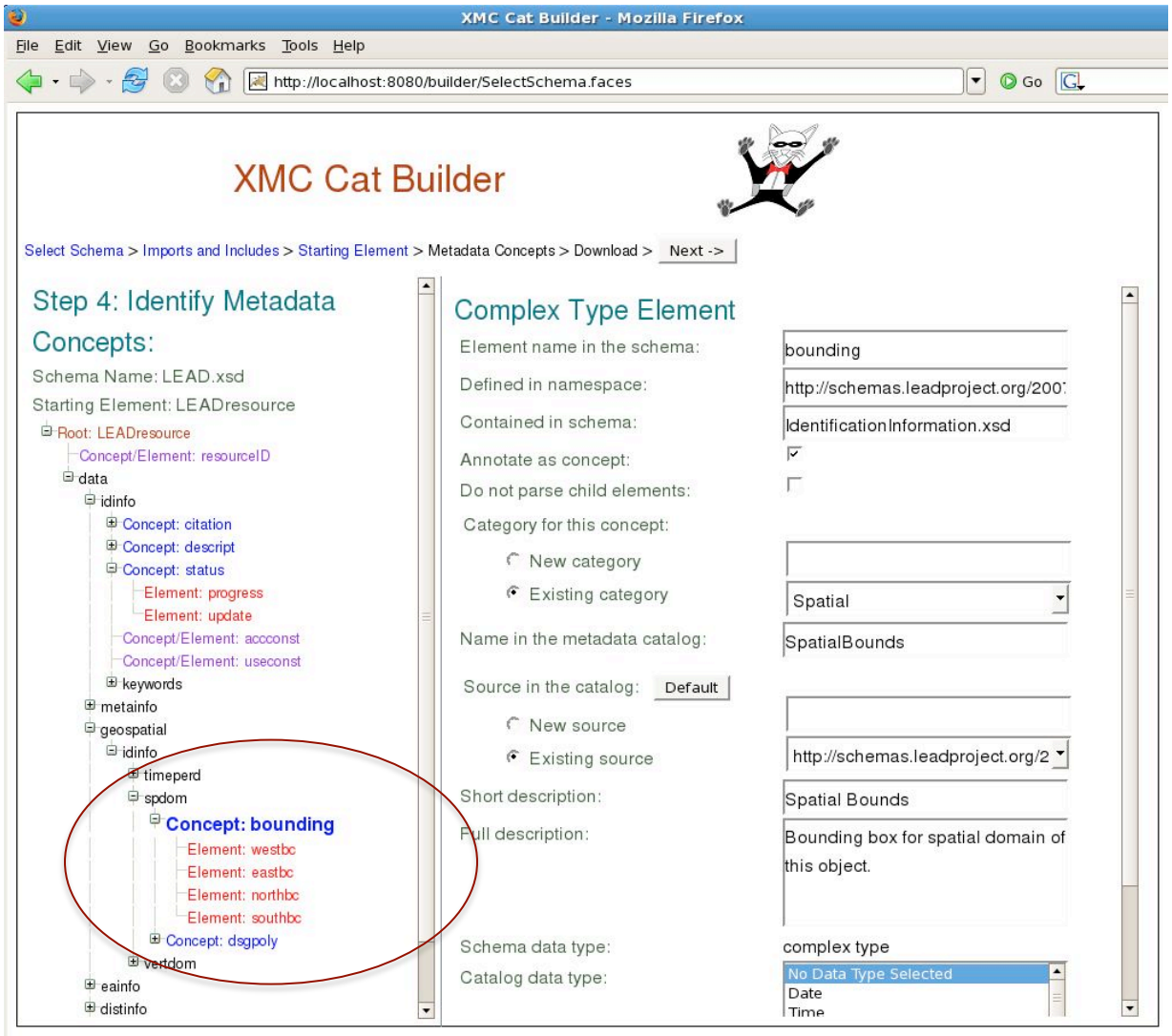
From top-level schema, Wizard can figure out the dependencies (knows that all 11 schemas are needed).

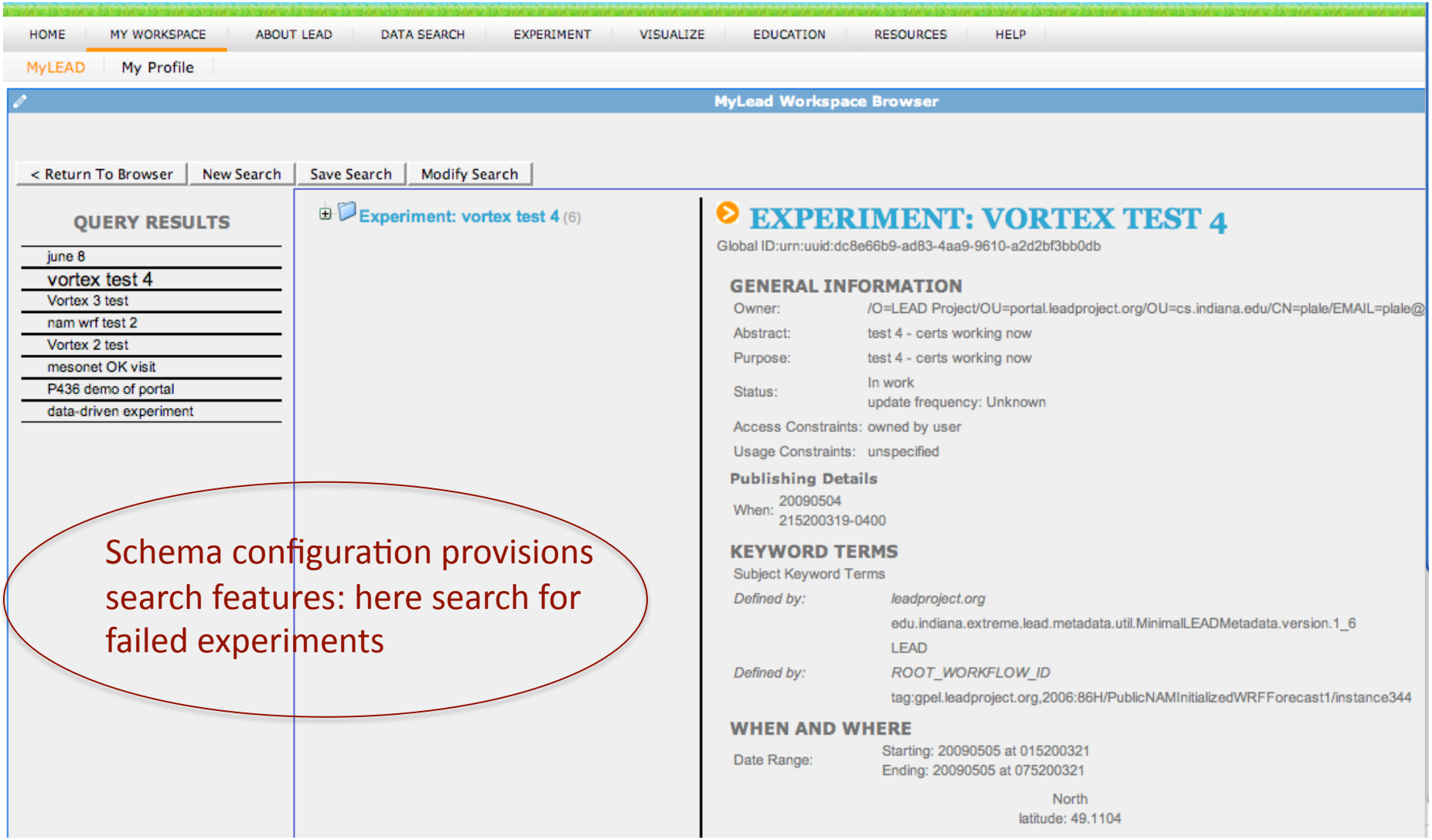
Wizard follows schema dependencies to make sure administrator gets them all identified.



Wizard then asks user to point out major concepts in schema and starting point in schema tree – the Root.

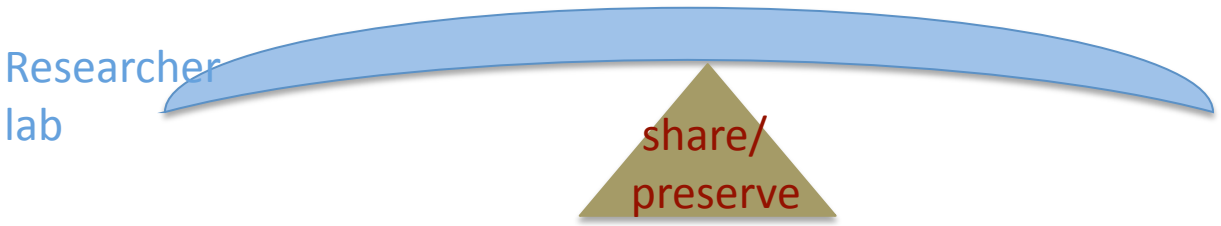
There's our bounding box again, now it is a major concept so will automatically have a search built.

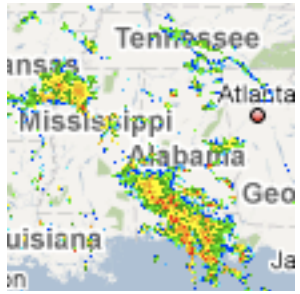




Schema configuration provisions search features: here search for failed experiments

Additional issue scientists face when sharing over long term (preservation): metadata and annotations.

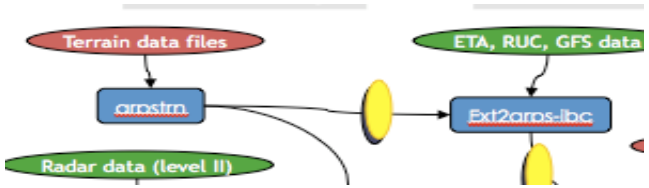




Could chuck bag of files over fence but then curation cost are very high. Open question on how cost can be reduced.

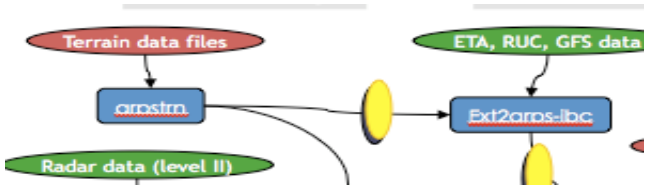
Provenance of art works

- Trace of history of work of art from moment it was made until it comes into a collection.
- Impartial and authoritative information on authenticity, ownership, theft, and other artistic, legal, and ethical issues concerning art objects.



It can be impossible to determine complete provenance for work of art:

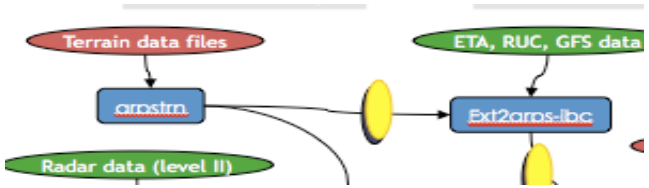
- provenance records often reflect an owner or former owner's wish for anonymity.
- ephemeral nature of historical records, which are often lost or destroyed over time

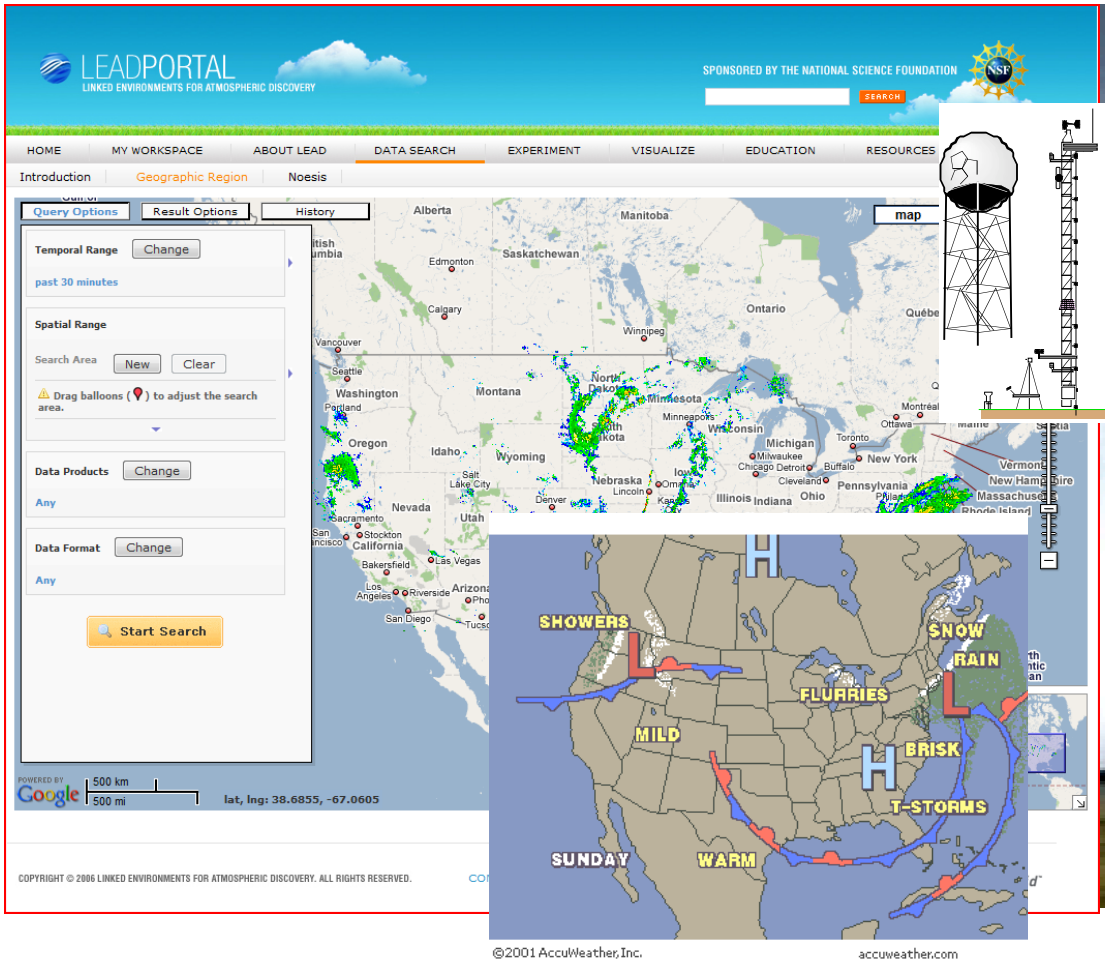


Saint John the Baptist in the Wilderness;
circa 1535 Moretto da Brescia Italy,
Brescia 1498–1554

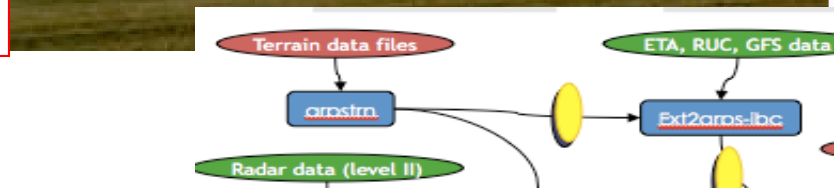
Provenance:

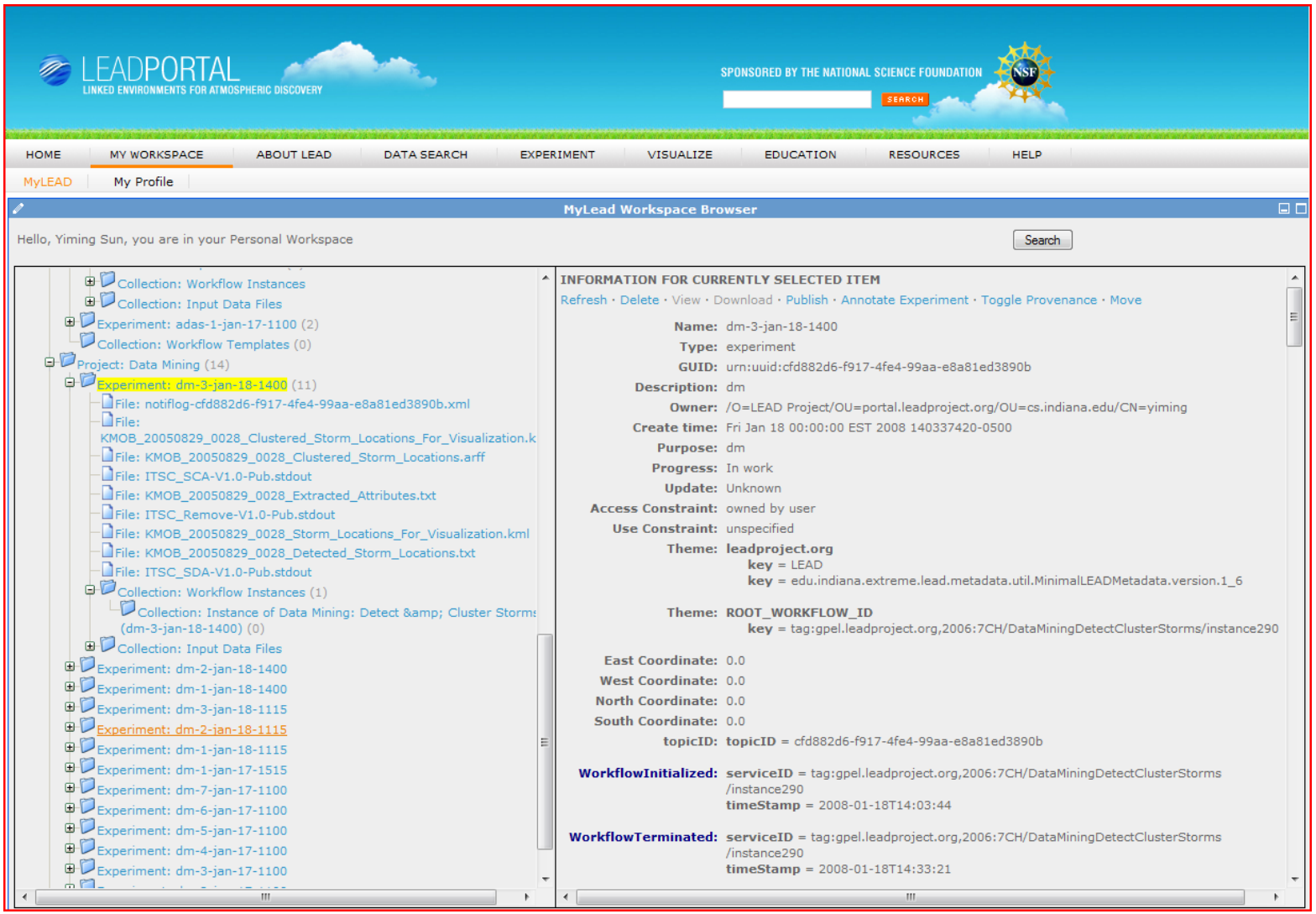
Oskar Bondy [d. 1944], Vienna[1]
[Jacob M. Heimann, Beverly Hills];[2] Philip Yordan, Beverly Hills (until 1951),[3] gift 1951; to Los Angeles County Museum of Art



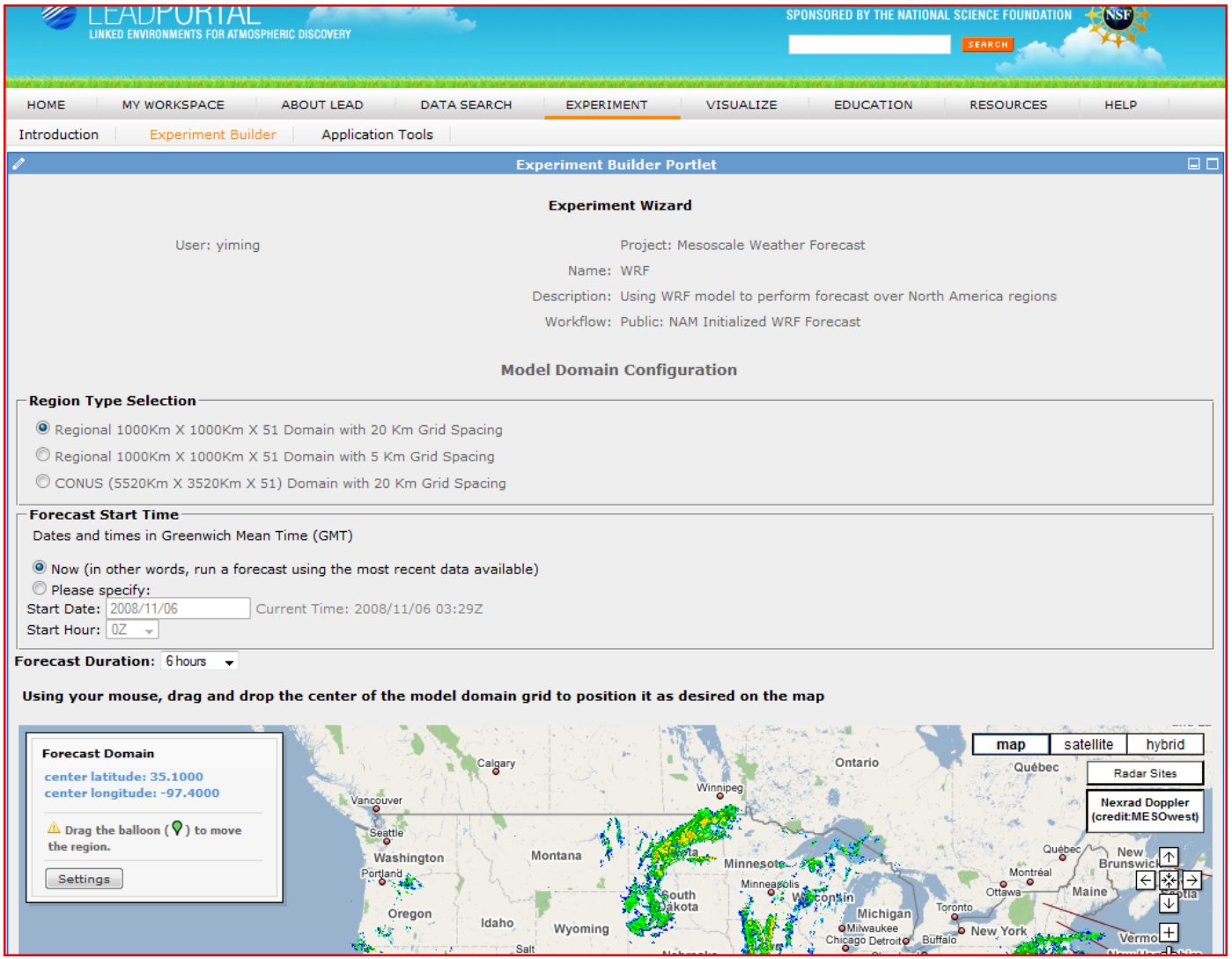


What is Provenance for Digital Scientific Data?

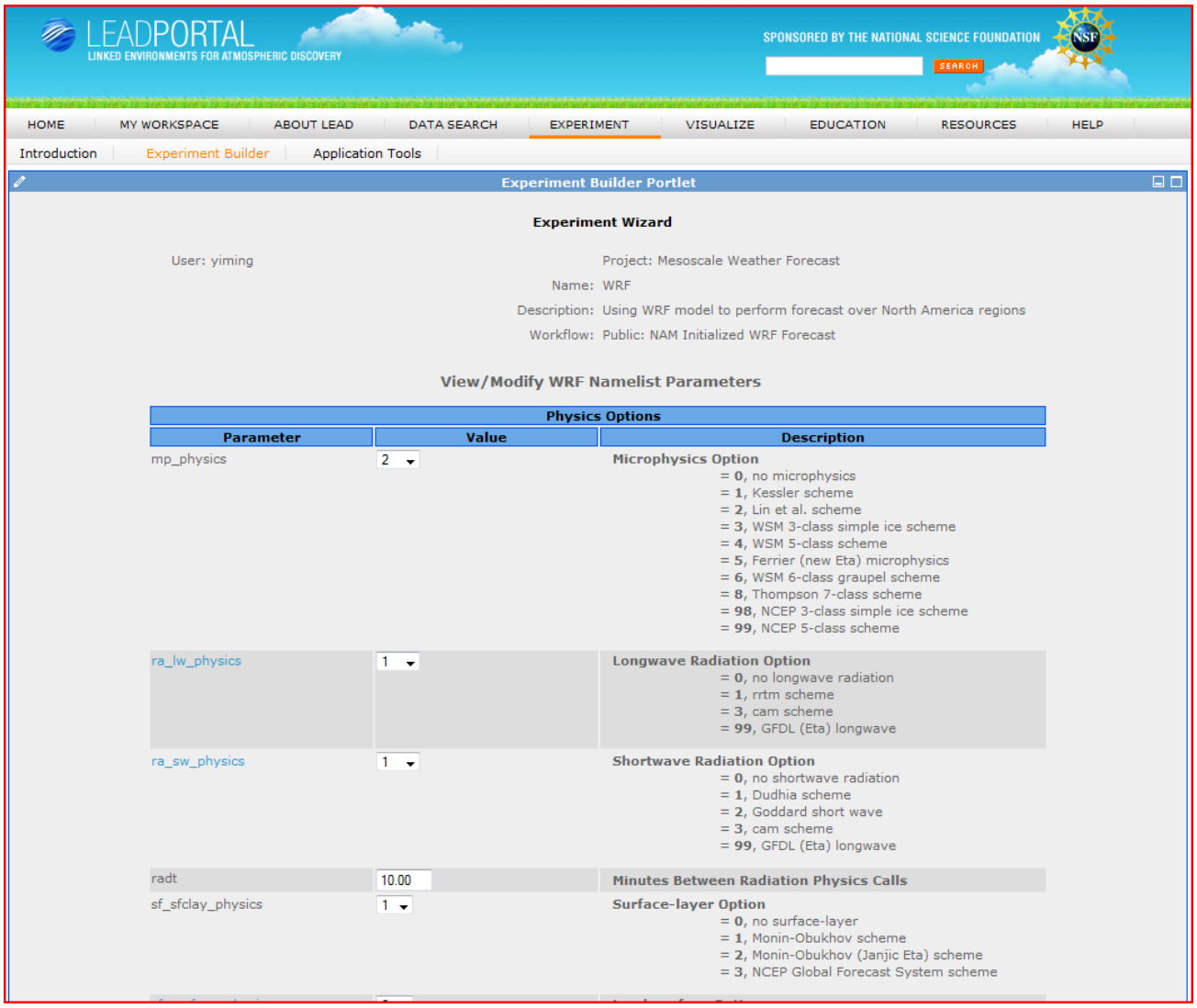




Science Gateway: User workspace captures metadata and activity through time



Science Gateway: Wizard guides scientist through workflow configuration

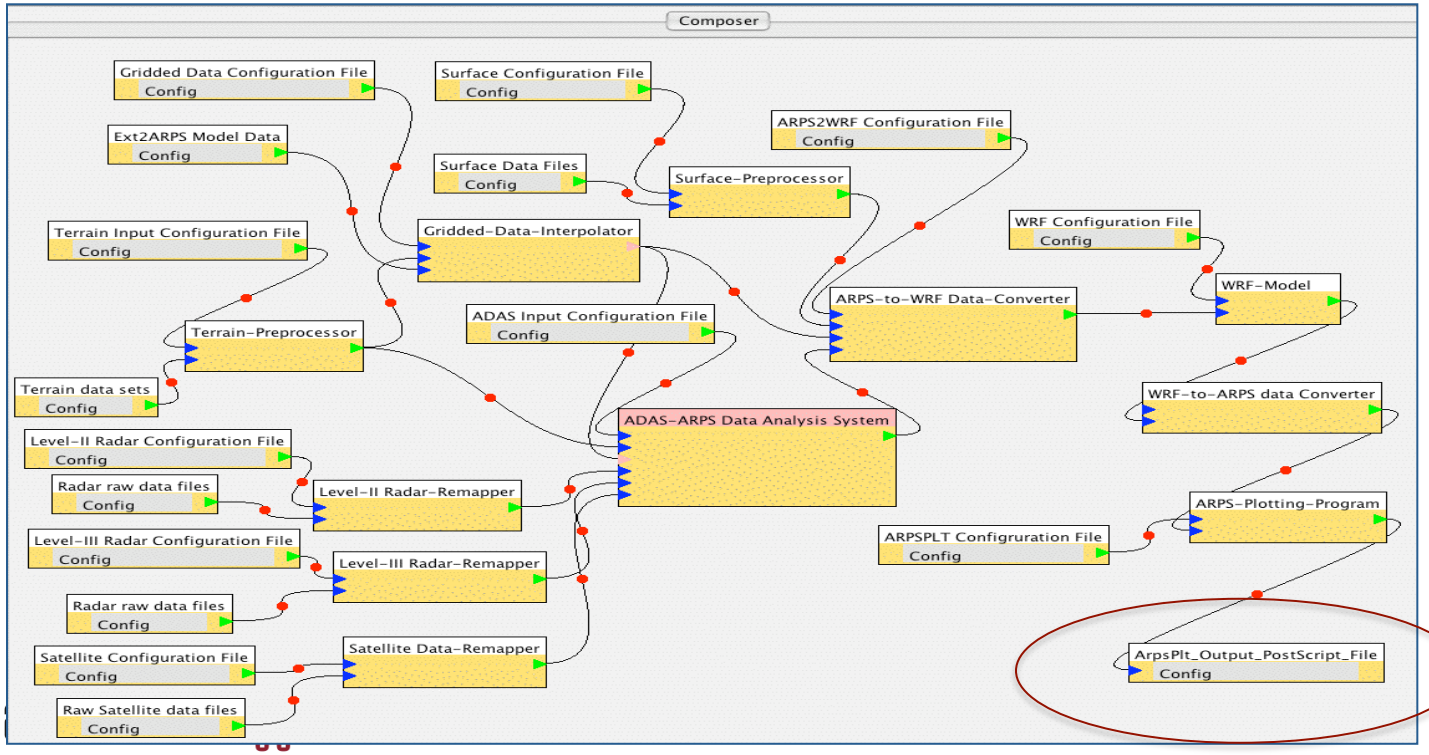


Science Gateway: Configuration includes physics parameters, making workflows better suited to high level experts.

Typical e-Science Experiment

Weather forecast using WRF in LEAD : role of provenance?

Pre-Processing | Assimilation | Forecast | Visualization

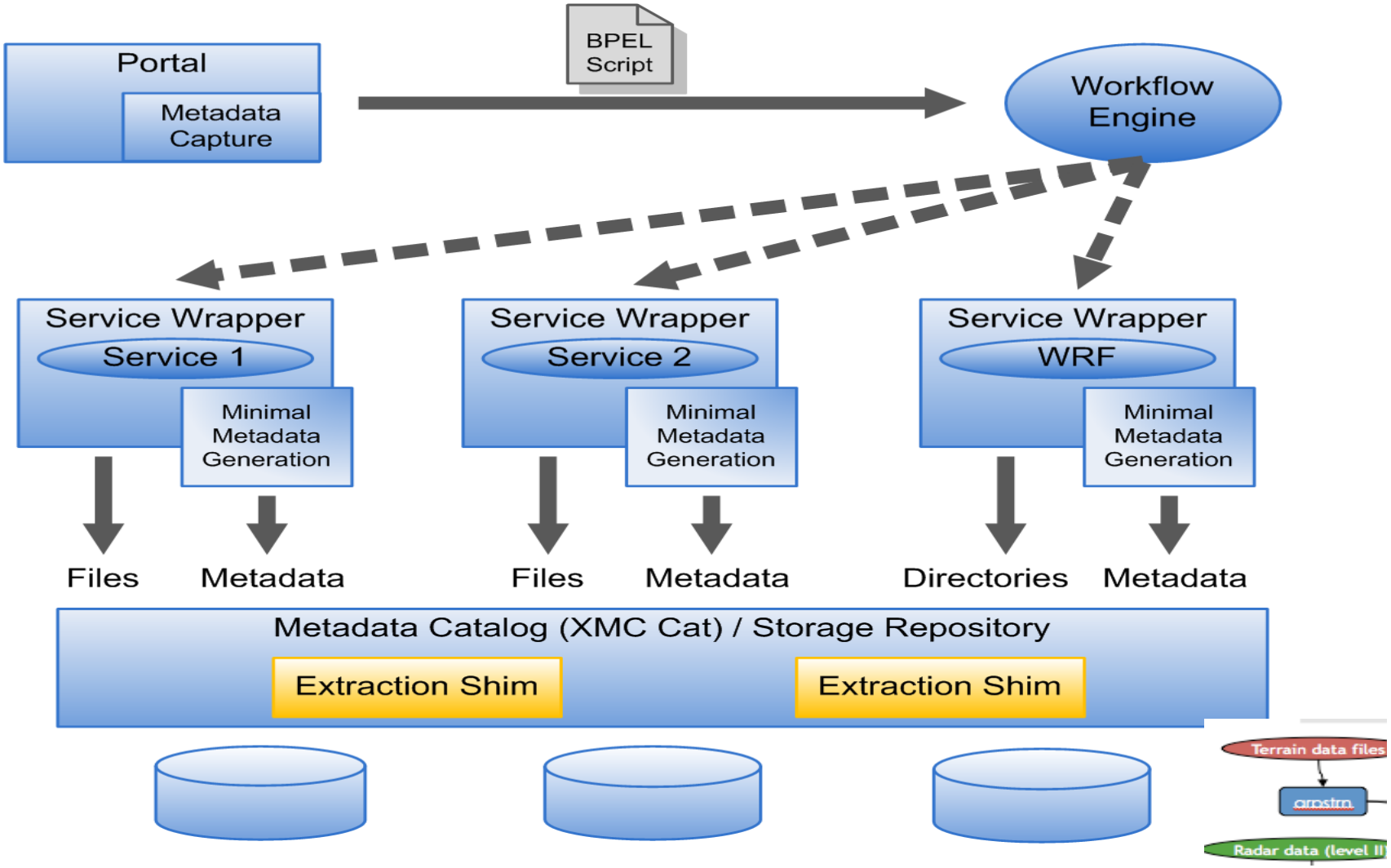


Give me provenance of this product

AT INDIANA UNIVERSITY

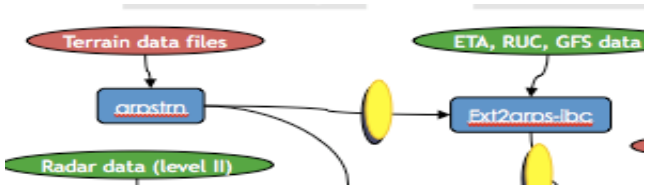
[]

Automated capture

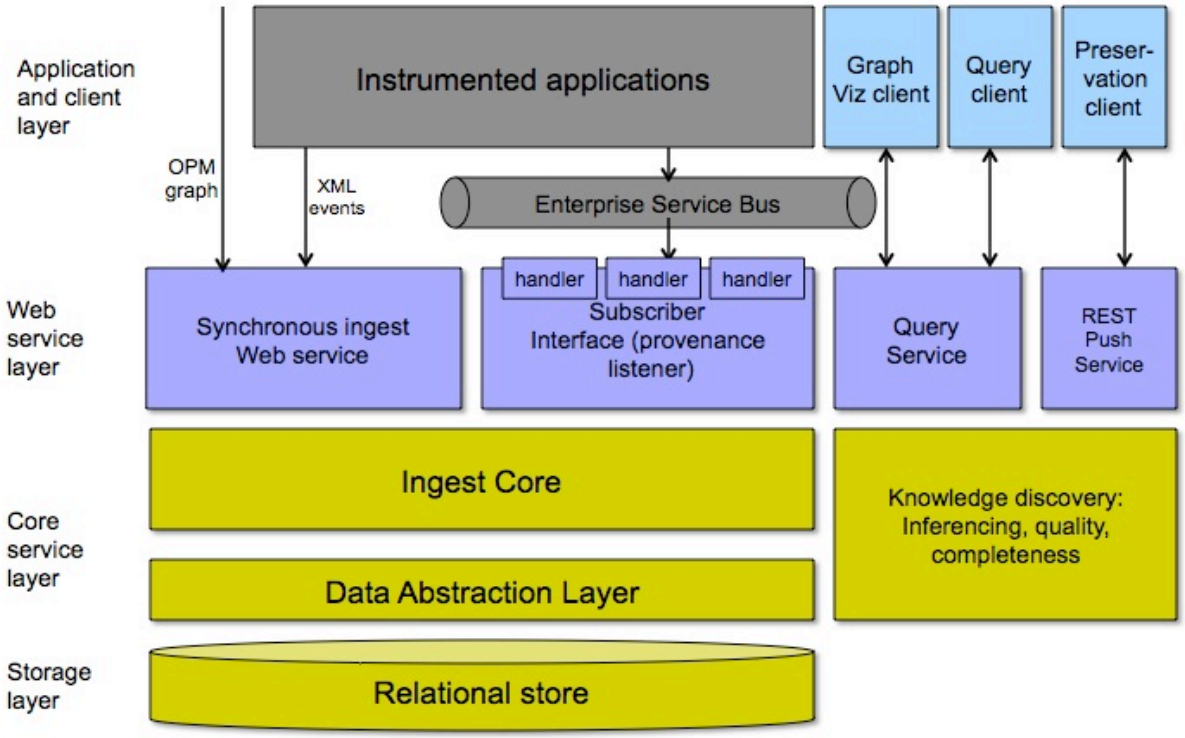


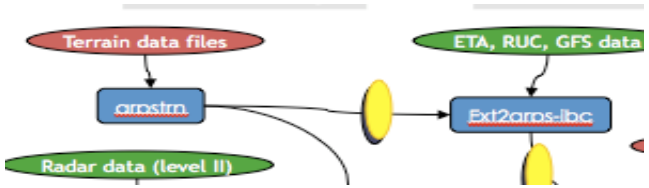
Forms of provenance capture

User Annotation	**Scavenging**	**Full Provenance Instrumentation**
Low application burden	Low application burden	High application burden
High human burden	Low human burden	Low human burden (high developer setup cost)
Error rates and omissions can be high leading to incomplete information	Incomplete information	Complete information

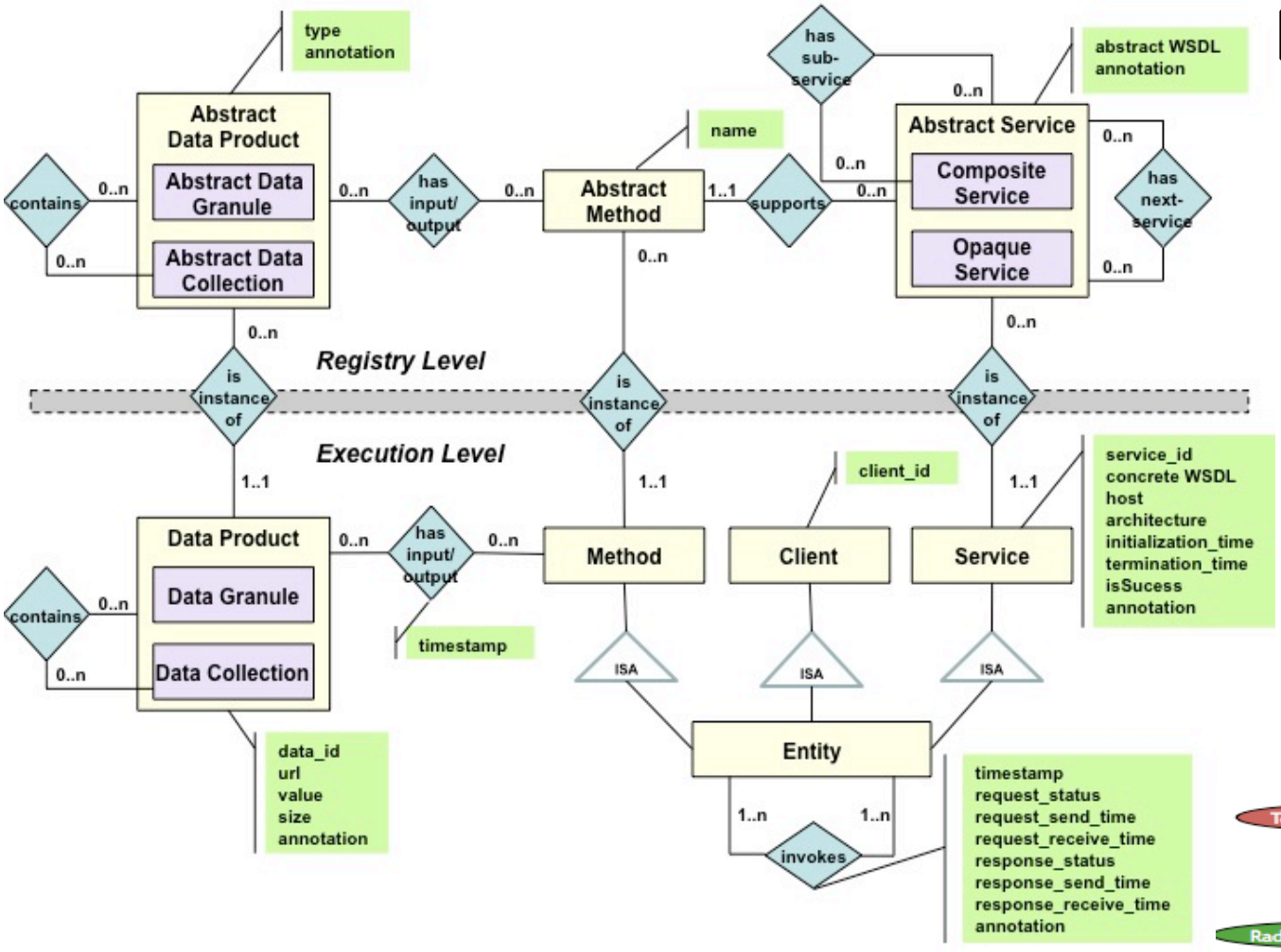


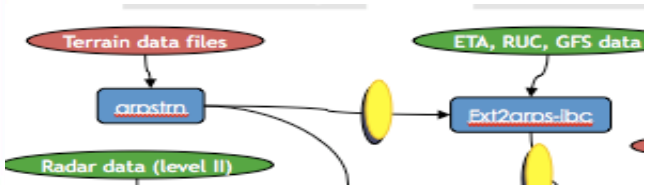
Architecture for Provenance Capture, Storage, and Analysis and Discovery



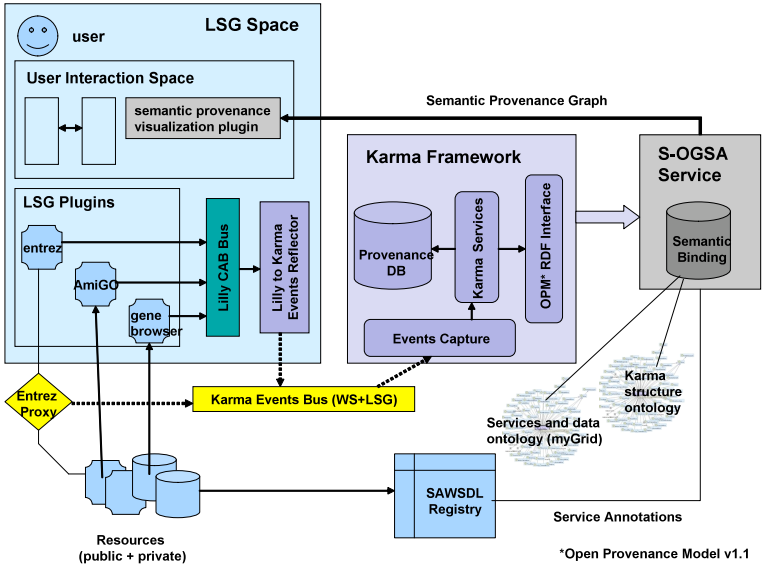


Data model for provenance representation

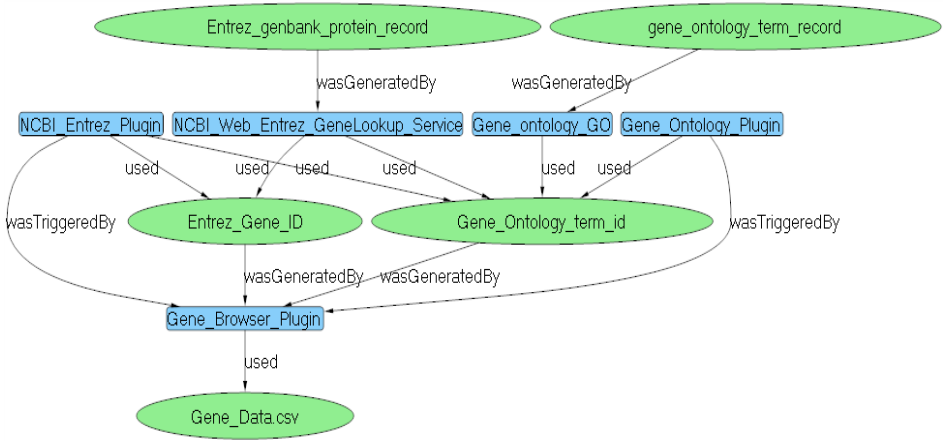




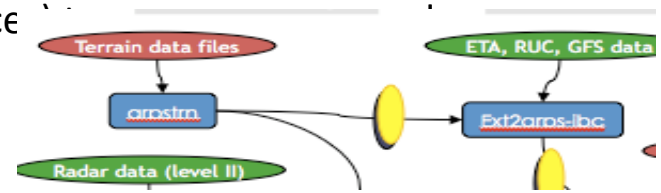
Application to the Life Science Grid

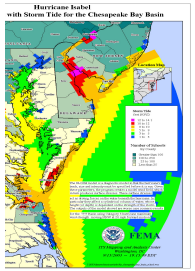


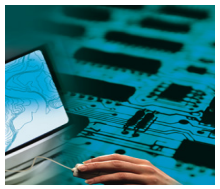
- Life Science Grid (LSG): open source cyberinfrastructure for drug discovery



- Karma: captures where and how an artifact originates, and processes previously applied to an artifact.
- S-OGSA adds semantic annotations (services and user preference
- Collaborators: Eli Lilly and University of Manchester.



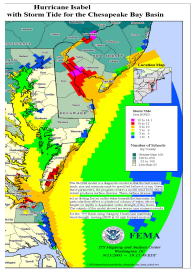


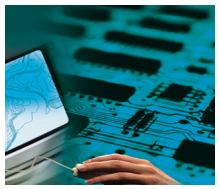


Data To Insight Center Scientific Mission

The mission of Data to Insight Center is to bring *first class research* in data-driven technology to *interdisciplinary problems* and to *provide value* to the citizens and business of the State of Indiana and beyond.

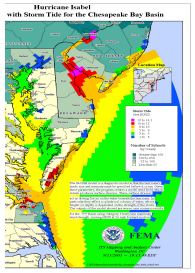
pervasivetechnologyinstitute
AT INDIANA UNIVERSITY

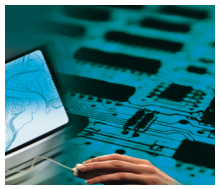




Center Goals

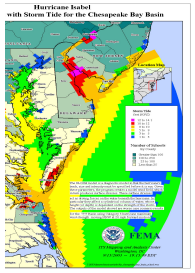
1. Establish open structures for involvement in Center by IU faculty and staff; and beyond.
2. Aid in the creation and growth of interdisciplinary research through identification of and channeling effort channeling into ***research thrusts***
3. Perform basic and applied research within or across thrusts with focus on large scale projects
4. Harden selected data-focused technology solutions that have broader use potential (translational research)
5. Prototype new data-focused technology with opportunity for broader use potential
6. Train students, postdocs

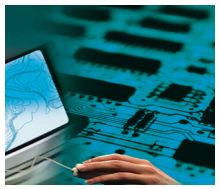




Purpose of Center Thrusts

- The Data To Insight Center adopted an organizational model of using thrusts as a way of focusing effort.
  - A *thrust* is an area of research focus that concentrates efforts and provides clarity in pursuit of opportunities.
- Three thrusts identified for 2009-2011.
- Thrusts are reviewed every 2 years to assess their continued viability to stimulate.
- The Center may also engage in additional thrusts that are shared across one or more of the PTI centers.



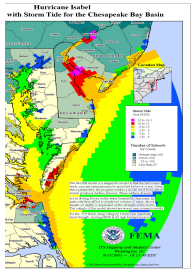


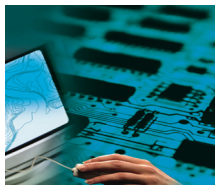
2009-2011 Thrusts

Scientific data preservation

- Long term preservation of scientific data collections

- *Sustainability, climate and the environment*
  - Sustainability science deals with interactions between natural and social systems and how those interactions conserve the planet's life support system. Climate and the environment are broadly defined as encompassing atmospheric, oceanographic, hydrologic, ecological, and earth physical systems.
  - Data are geo-located social, observational, and derived data. Includes but not limited to visualization; data provenance and preservation; leverages campus strength in human-oriented sustainability
- *Data at scale*
  - IU's significant storage and compute resources create opportunities for understanding how to deal with large-scale data, that is, data at scale.
  - Includes data mining over large scale data sets, visualizations and intelligent user interfaces that connect users and their goals (scale in user diversity); challenges in integration and interoperability across databases. Data from devices distributed across a population. Incentives and barriers to data sharing.





Researcher Engagement in Center

Seminar Series:

- Jointly (DSI and DLP) host several speakers in ‘09-’10 year.
- Distribute talks at Lindley Hall, Library, Innovation Center; extend welcome to IUPUI and host venue there.

- Fellows Program
  - Proposal driven; contribute to center’s thrusts and goals
  - 2010
- Consortia affiliation

Thank you.

Beth Plale
plale@indiana.edu

XMC Cat
Scott Jensen scjensen@indiana.edu
Karma
Bin Cao bincao@indiana.edu and Girish Subramanian

http://www.dataandsearch.org/dsi/research

pervasivetechnologyinstitute
AT INDIANA UNIVERSITY