# Data Subsystem: architectural foundation for storing and serving data

**Beth Plale**
**Indiana University**

**Anne Wilson**
**Unidata**

**Year-2 Site Visit**
**21-22 July 2005**



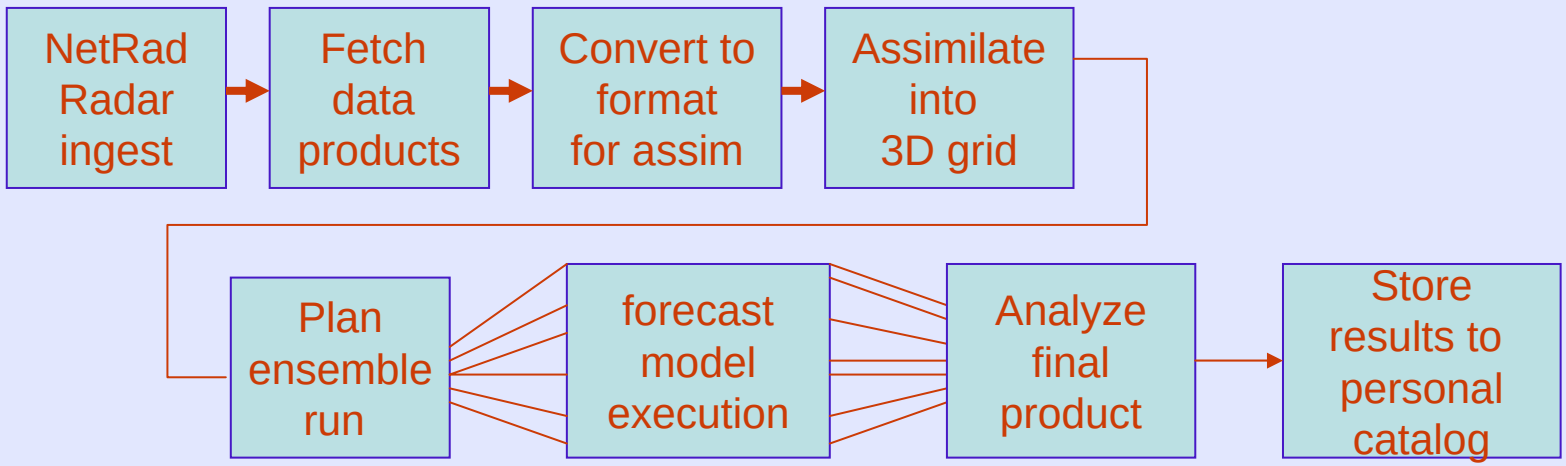
# Outline

- Philosophy of service oriented architecture
  - Motivate understanding of remainder of talk.
- Data subsystem
  - Architecture overview
  - Select component detail
  - Significant subsystem accomplishments
  - Ongoing deployment and research work

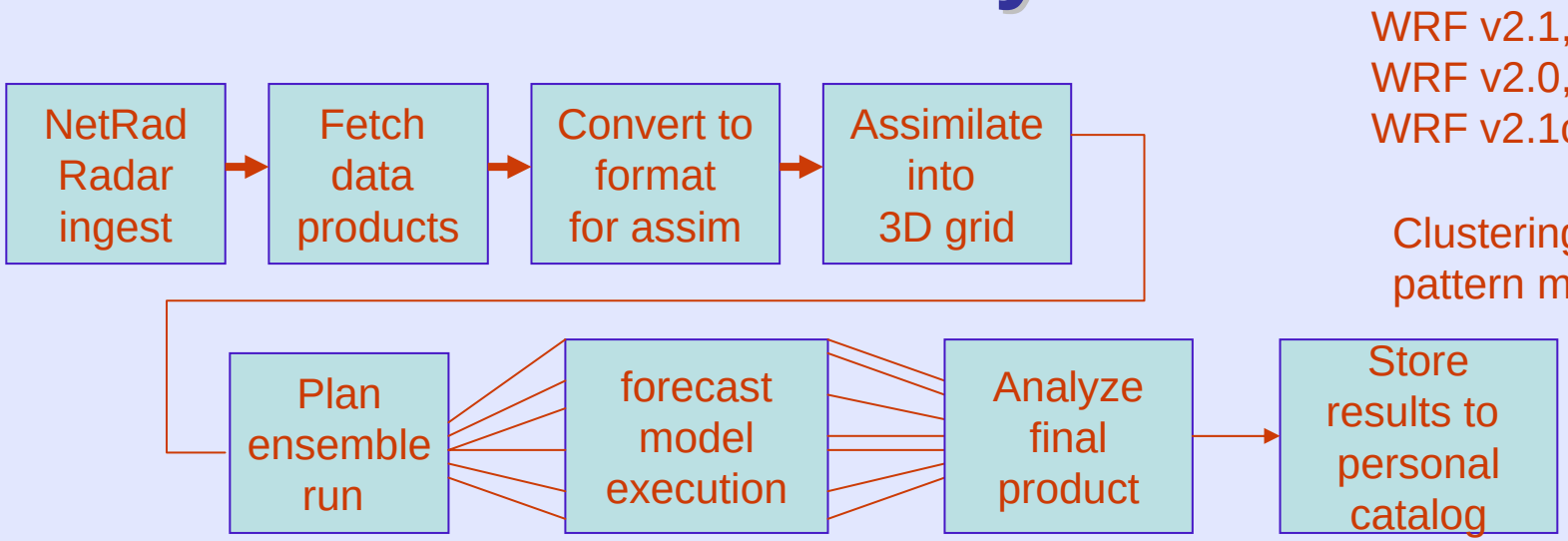


# Service Oriented Architecture Philosophy

- Building large-scale distributed applications of tightly coupled components (hard-wired connections between steps) is straightforward



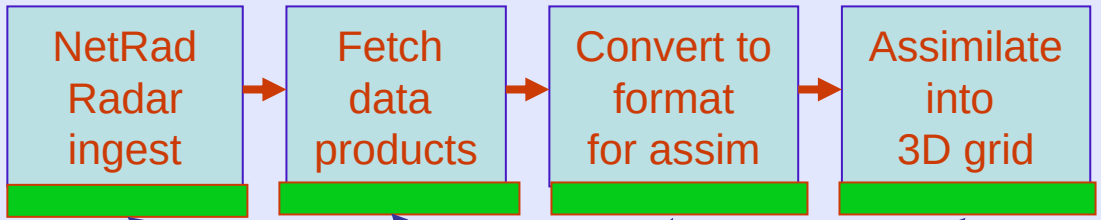
# Problems encountered in building loosely coupled large-scale distributed systems




One type of tool for each step, one version each type, script that glues together is straightforward: **1 path**

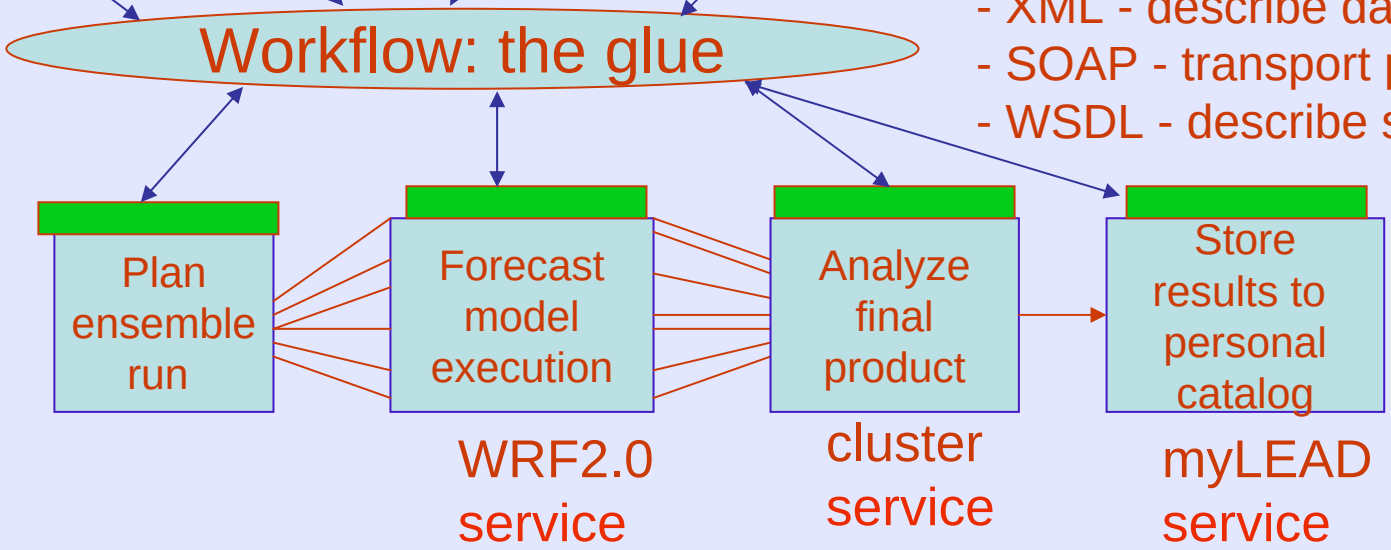
But suppose 8 steps, 2 types of services, 2 versions per service. Script has $2^9$ or **256 paths**.

# Service oriented view for loosely coupled distributed systems




Versions of service interchangeable; all talk same communication protocols and obey explicit boundaries

- XML - describe data set
- SOAP - transport protocol
- WSDL - describe service



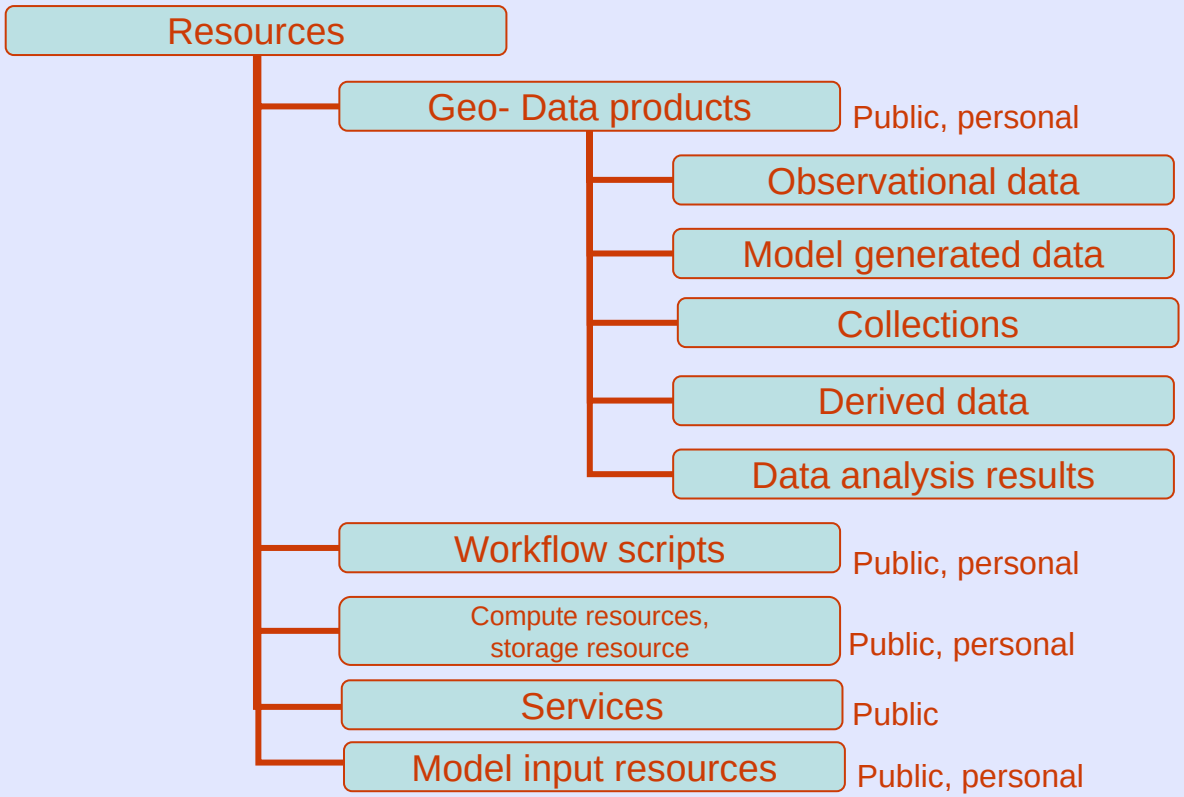
# Data Subsystem

- Philosophy of service oriented architecture
  - Motivate understanding of remainder of talk.

- Data subsystem
  - Architecture overview
  - Select component detail
  - Significant subsystem accomplishments
  - Ongoing deployment and research work



# Categories of data products




**Personal resources**

-- user's experiment products, personal collections, scripts, input config params.

**Public products**

-- data gathered and made accessible by external data providers.

**External products**

-- data not known to resource catalog.

Geospatial Query GUI | personal Workspace browser | Ask ontology | Viz Client (IDV) | Access interfaces

# Data Subsystem Architecture

**Resource Catalog** LEAD public products and services P | **myLEAD** User's own Information catalog P | **Noesis Ontology** concepts and vocabulary P | **Query Service** query mediation P | Access services

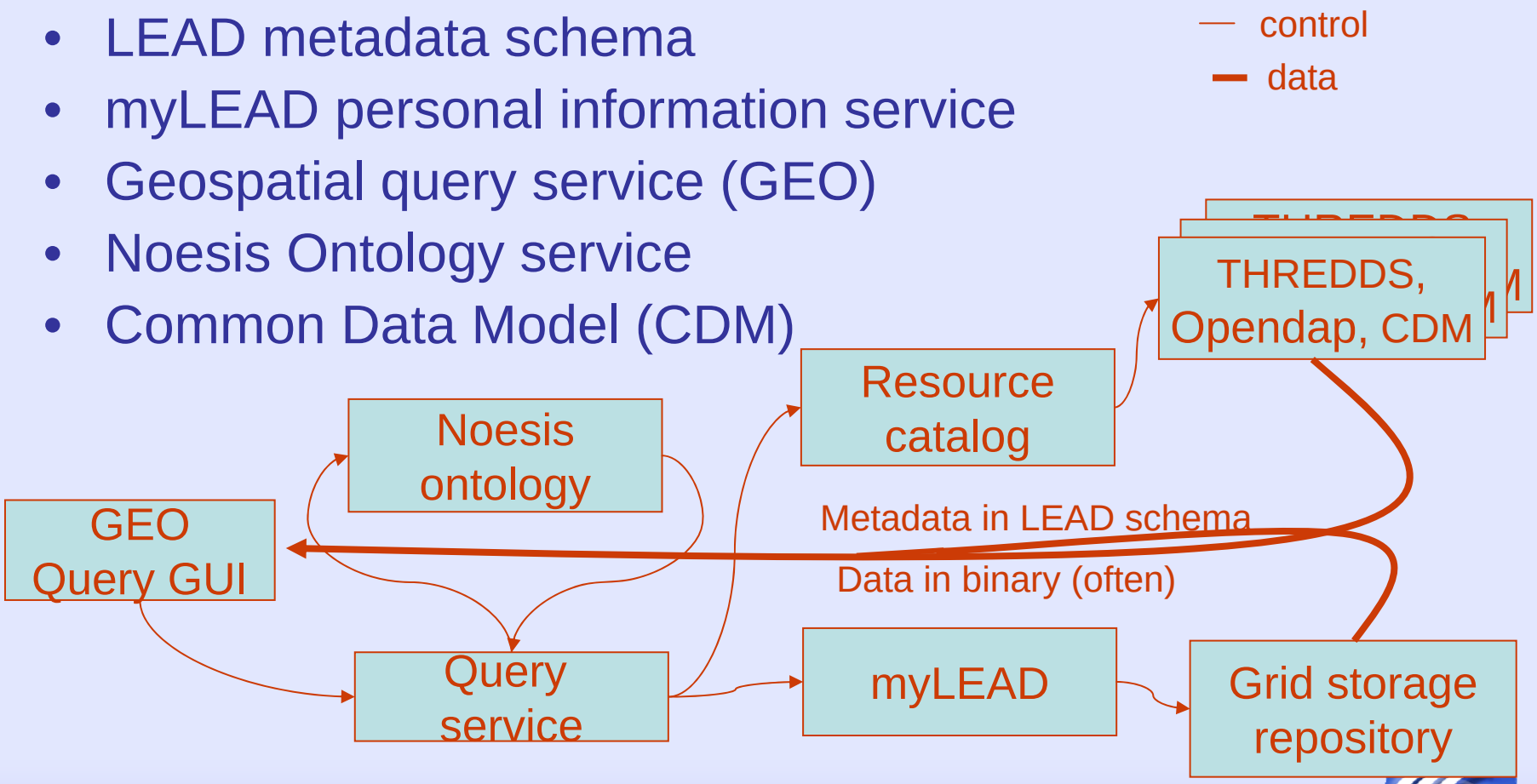
**THREDDS Catalogs** -web browser metadata | **Name Service** -single global naming system | **Automated metadata generation** - a capability | **Stream Service** - from LDM to user's app | Resource services

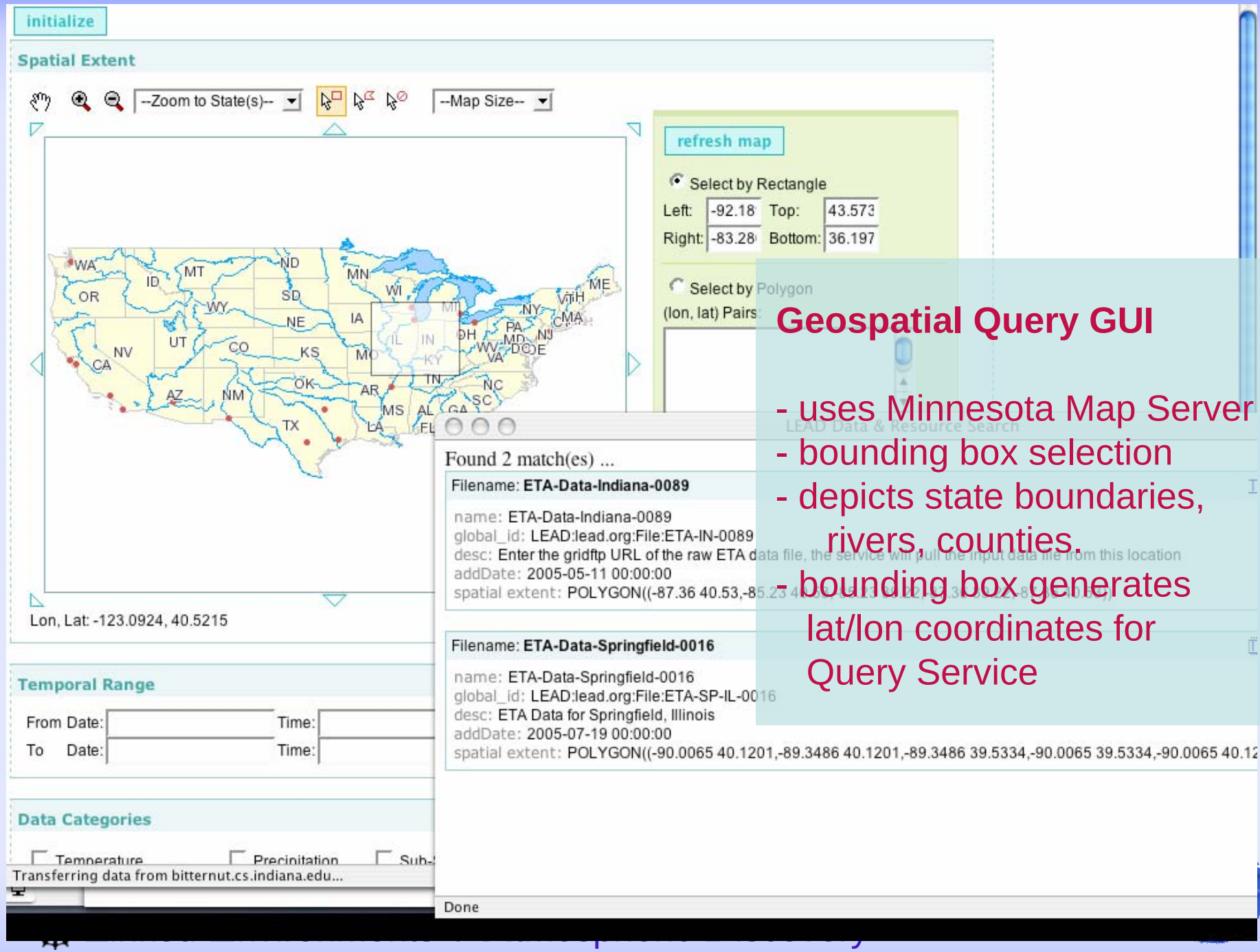
**OPeNDAP** data server | **Unidata Data dissem client (LDM)** | **Grid Storage respository** | **Steerable instruments** - CASA | Resources

Poster (P)

# Brief Tour of Capabilities and Component Functionality

- LEAD metadata schema
- myLEAD personal information service
- Geospatial query service (GEO)
- Noesis Ontology service
- Common Data Model (CDM)





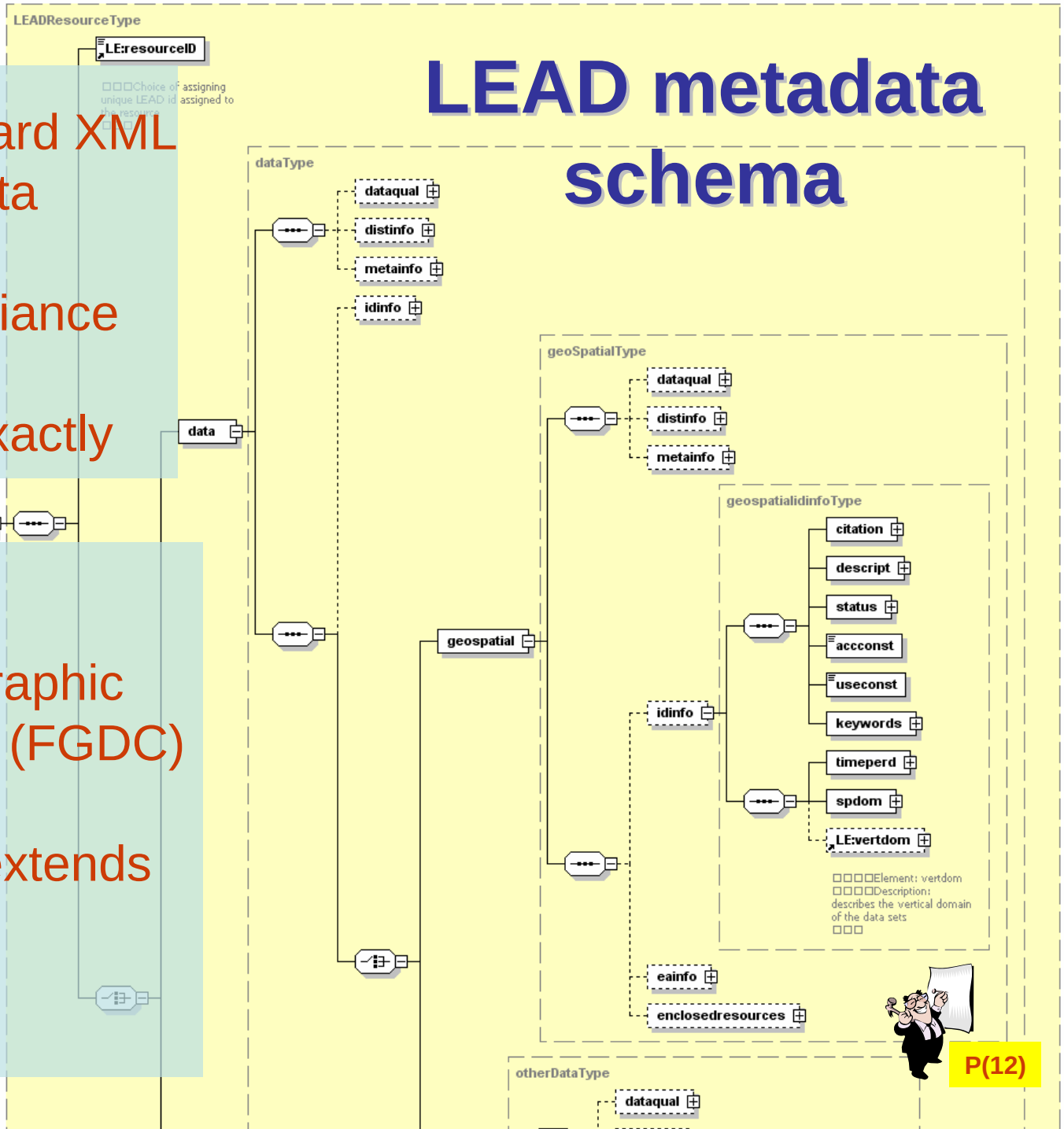

## Geospatial Query GUI

- uses Minnesota Map Server
- bounding box selection
- depicts state boundaries, rivers, counties.
- bounding box generates lat/lon coordinates for Query Service

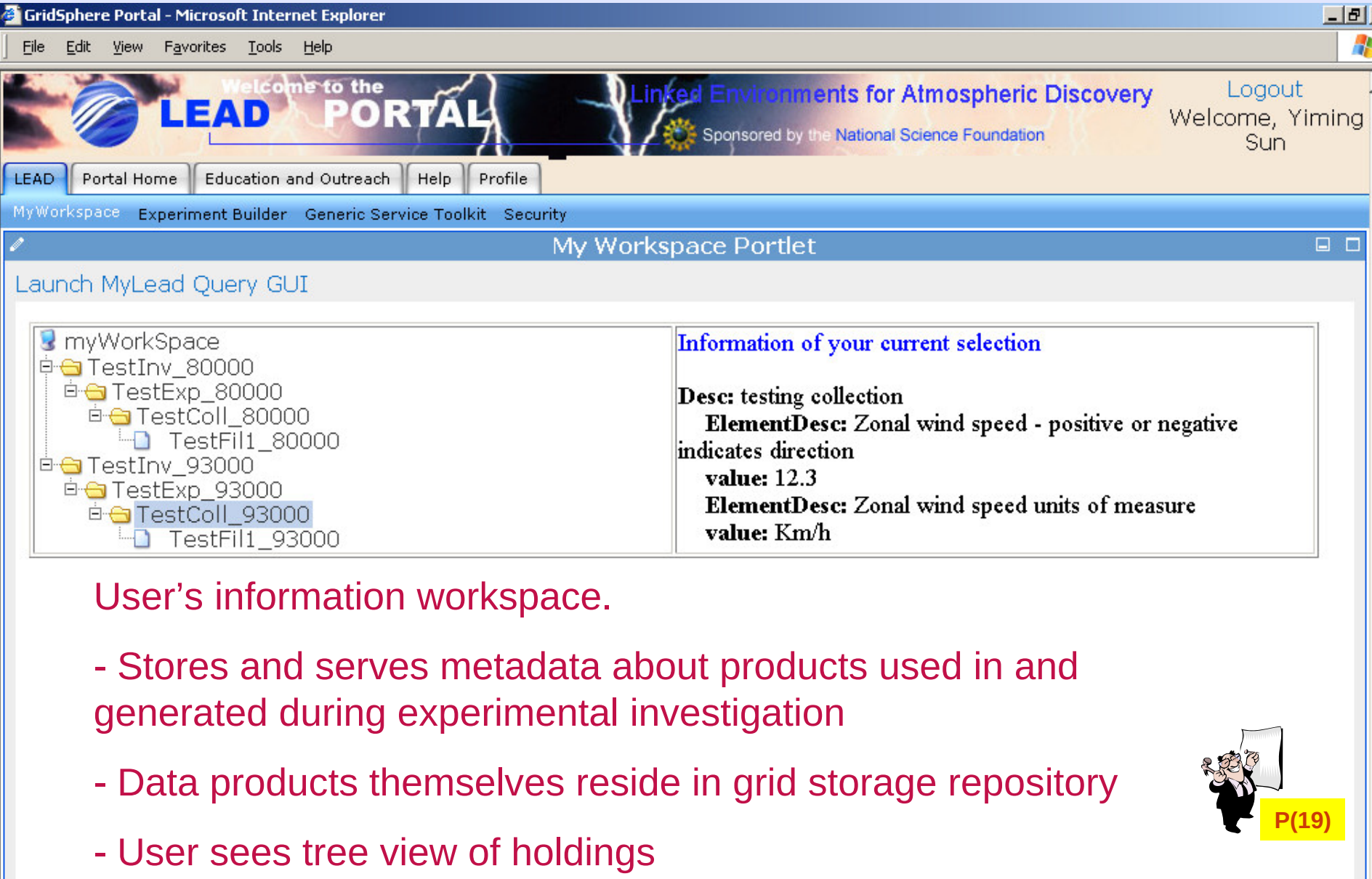
# LEAD metadata schema

-- Needed standard XML description of data products,
-- Needed compliance with standard
-- None suited exactly

-- Federal Geographic Data Committee (FGDC) compliant,
-- LEAD profile extends FGDC schema



# myLEAD personal catalog



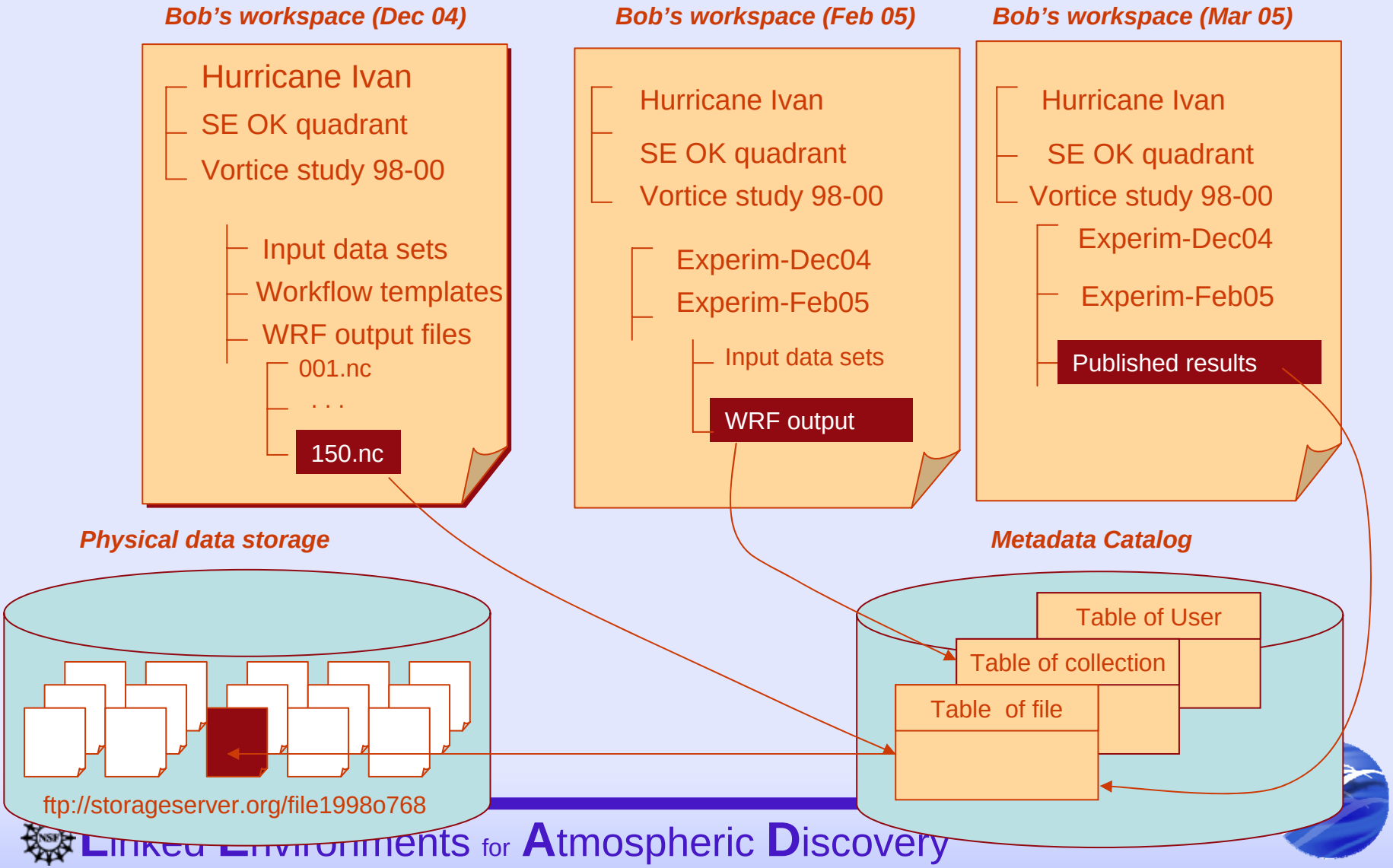

User's information workspace.

- Stores and serves metadata about products used in and generated during experimental investigation

- Data products themselves reside in grid storage repository

- User sees tree view of holdings

# myLEAD research goals: transparent structure (through agent), privacy and sharing



# Noesis Ontological Smart Search

Noesis - A Smart Search Tool for Earth Science Resources

Search

Search Resources

Search Results

*Noesis is defined as the cognition process.*

Noesis Search Tool is resource gathering service for Earth Science. Given a s term, noesis uses a domain ontology as its knowledge base to collect all the rel resources. It has been developed at Information Technology and Systems C University of Alabama in Huntsville as part of the Linked Environment Atmospheric Discovery (LEAD) project.

Noesis - A Smart Search Tool for Earth Science Resources

Pressure

Search

UAH

Search Resources

Search Results

Categories

*5 Results found!*

All Searches:5

Ontology Service: 5

**1. Pressure**

***Definition:***A type of stress characterized by uniformity in all directions. As a measurable on a surface, the net force per unit area normal to that surface exerted by molecules rebounding from it.

http://amsglossary.allenpress.com/glossary/search?id=pressure1

**2. Hydrostatic Pressure**

***Definition:***The pressure in a fluid in hydrostatic equilibrium.

http://amsglossary.allenpress.com/glossary/search?id=hydrostatic-pressure1

**3. Total Pressure**

***Definition:***The sum of the static pressure and the dynamic pressure when these concepts are applicable.

http://amsglossary.allenpress.com/glossary/search?id=total-pressure1

**4. Atmospheric Pressure**

***Definition:***The pressure exerted by the atmosphere as a consequence of gravitational attraction exerted upon the "column" of air lying directly above the point in question.

http://amsglossary.allenpress.com/glossary/search?id=atmospheric-pressure1

**5. Static Pressure**

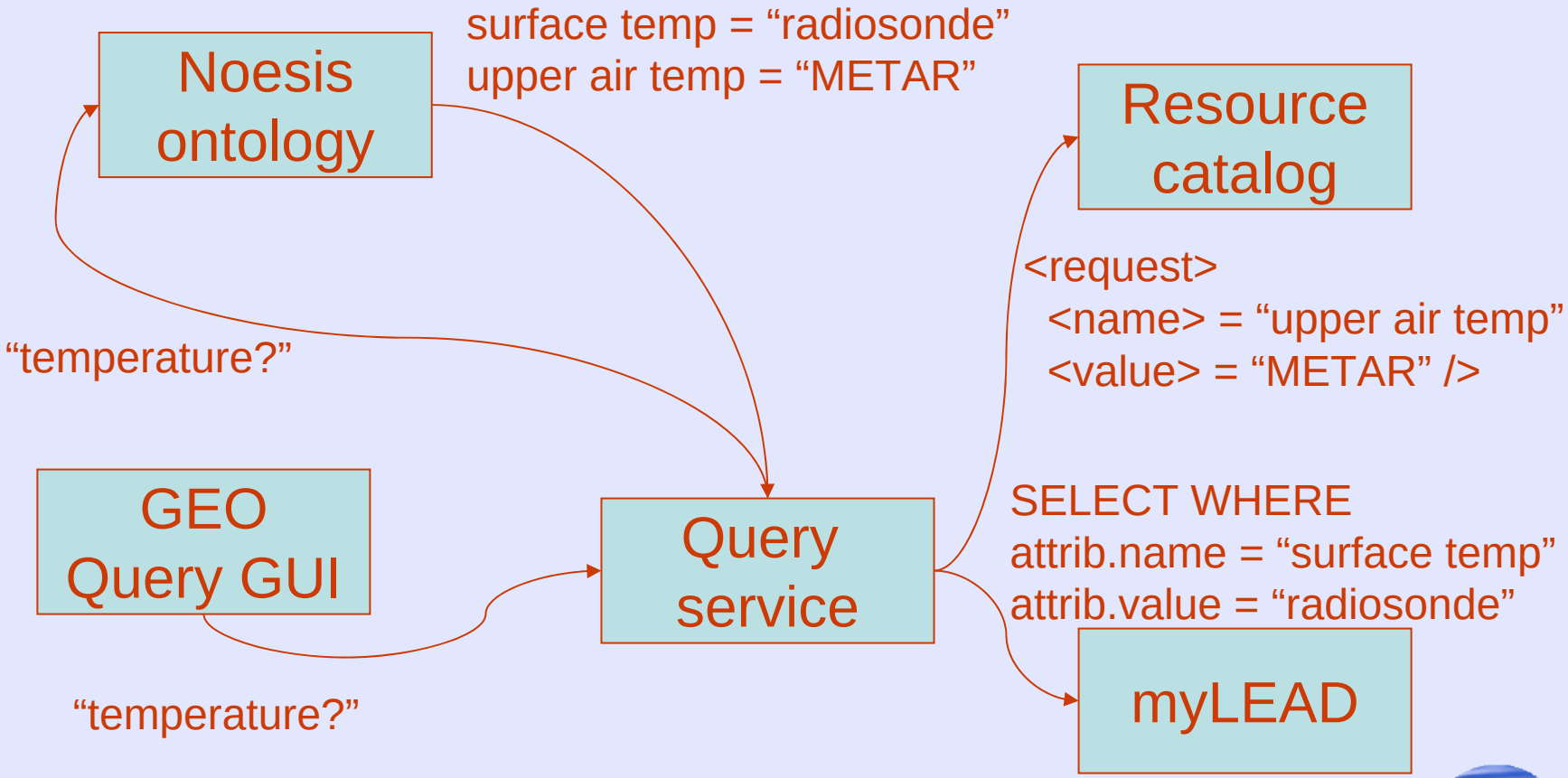
***Definition:***In engineering fluid mechanics, the pressure in a homogeneous incompressible fluid in steady flow along a level streamline at points other than the stagnation point.

http://amsglossary.allenpress.com/glossary/search?id=static-pressure1

Stores relationships between domain specific concepts and terms

Poster (8)

# Noesis Ontological Smart Search as service



# Subsystem-wide Drivers

- Data and query access transparency
- Extensibility

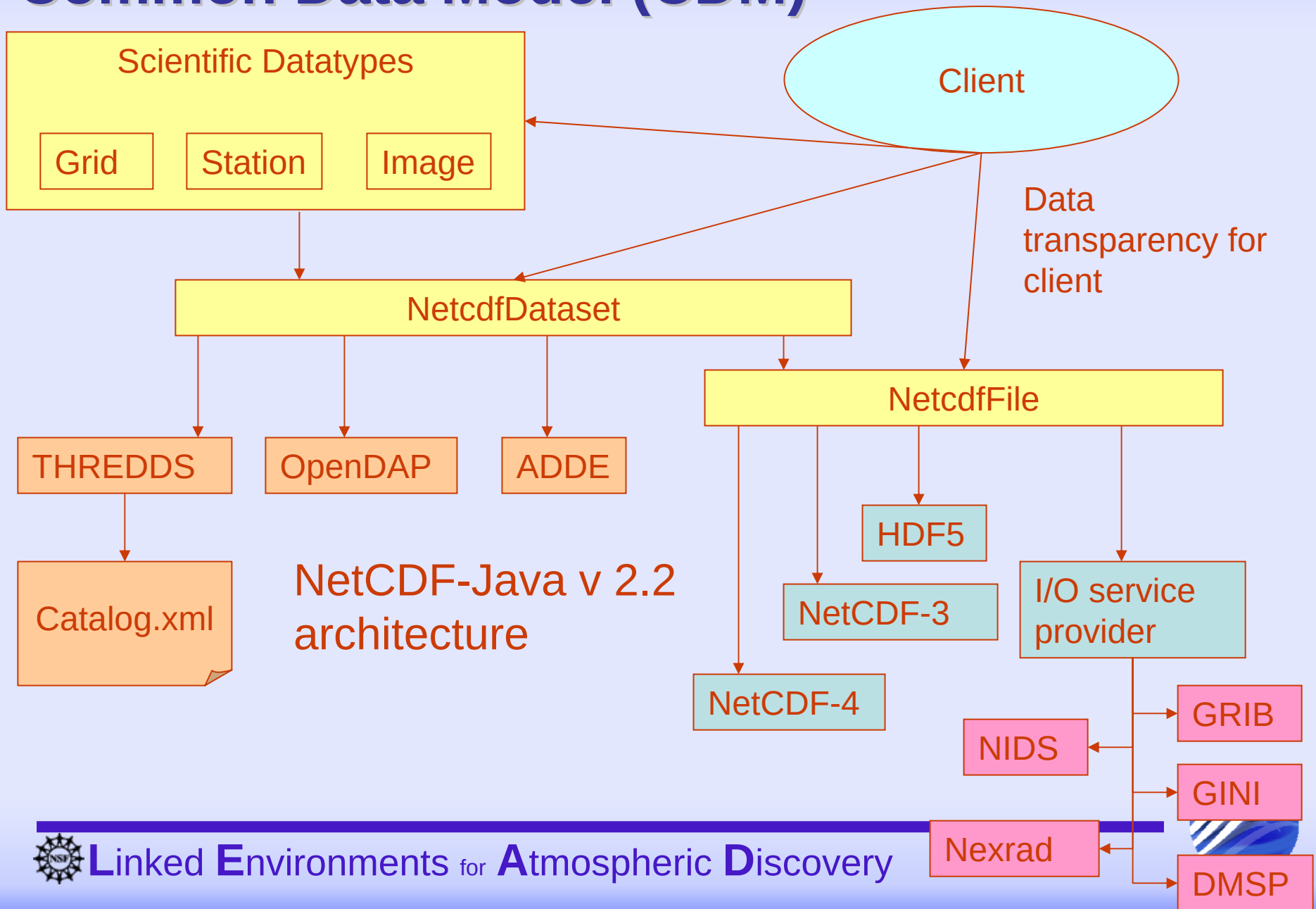


# Query and Data Access Transparency

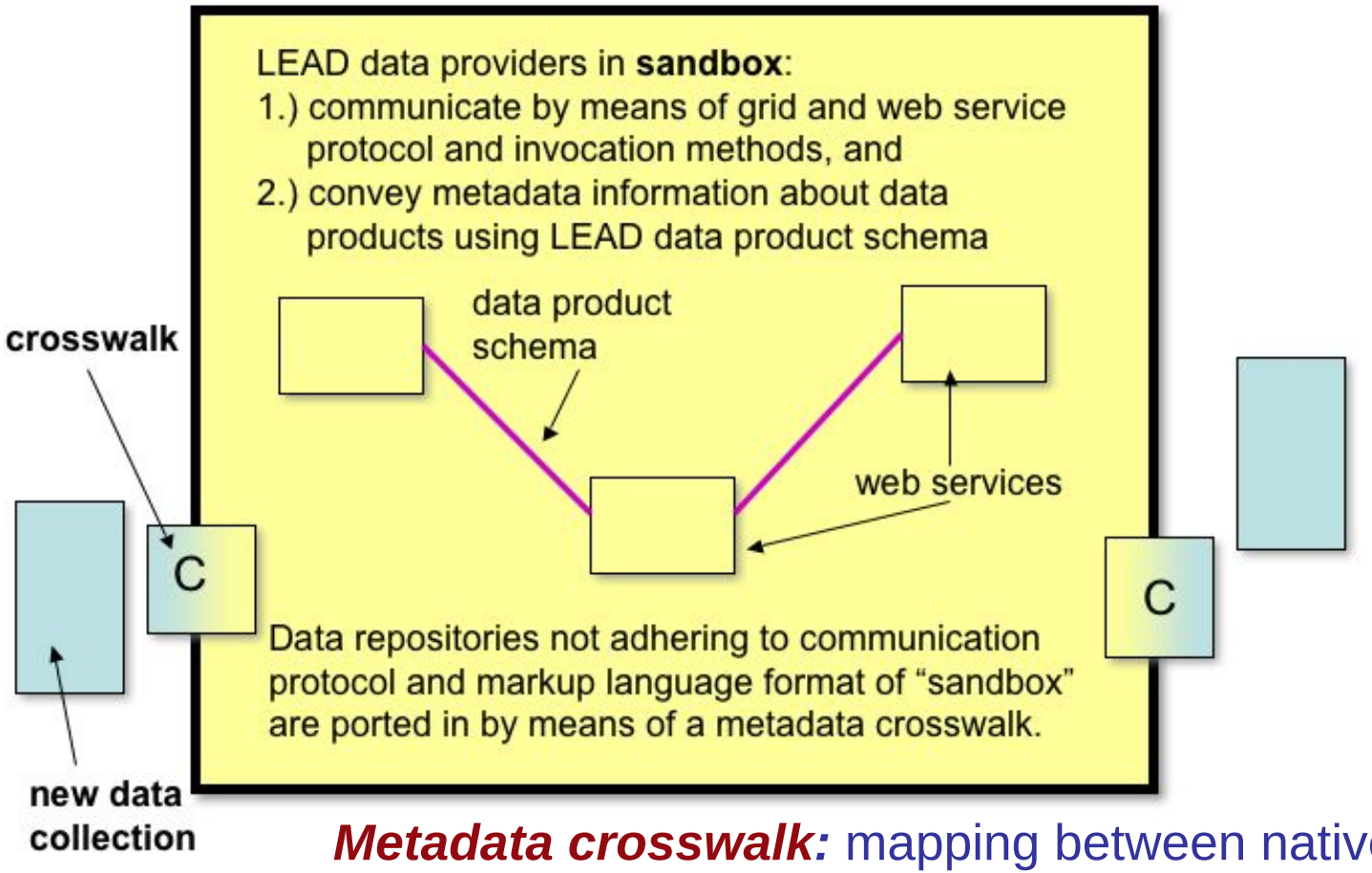
- Hide differences in data representation and way in which resources accessed by users.
  - Query on high-level application domain concepts,
  - retrieve results across heterogeneous data products and servers.
- Human and component integration required:
  - *GEO GUI and Query service* - IU
  - *Common vocabulary* - Millersville, UAH, OU
  - *Noesis ontology* - UAH
  - *myLEAD, Resource Catalog* - IU
  - *Automated metadata generation* - Unidata
  - *LEAD metadata schema* - UAH, IU, Unidata, NCSA
  - *Common data model* - Unidata



# Common Data Model (CDM)

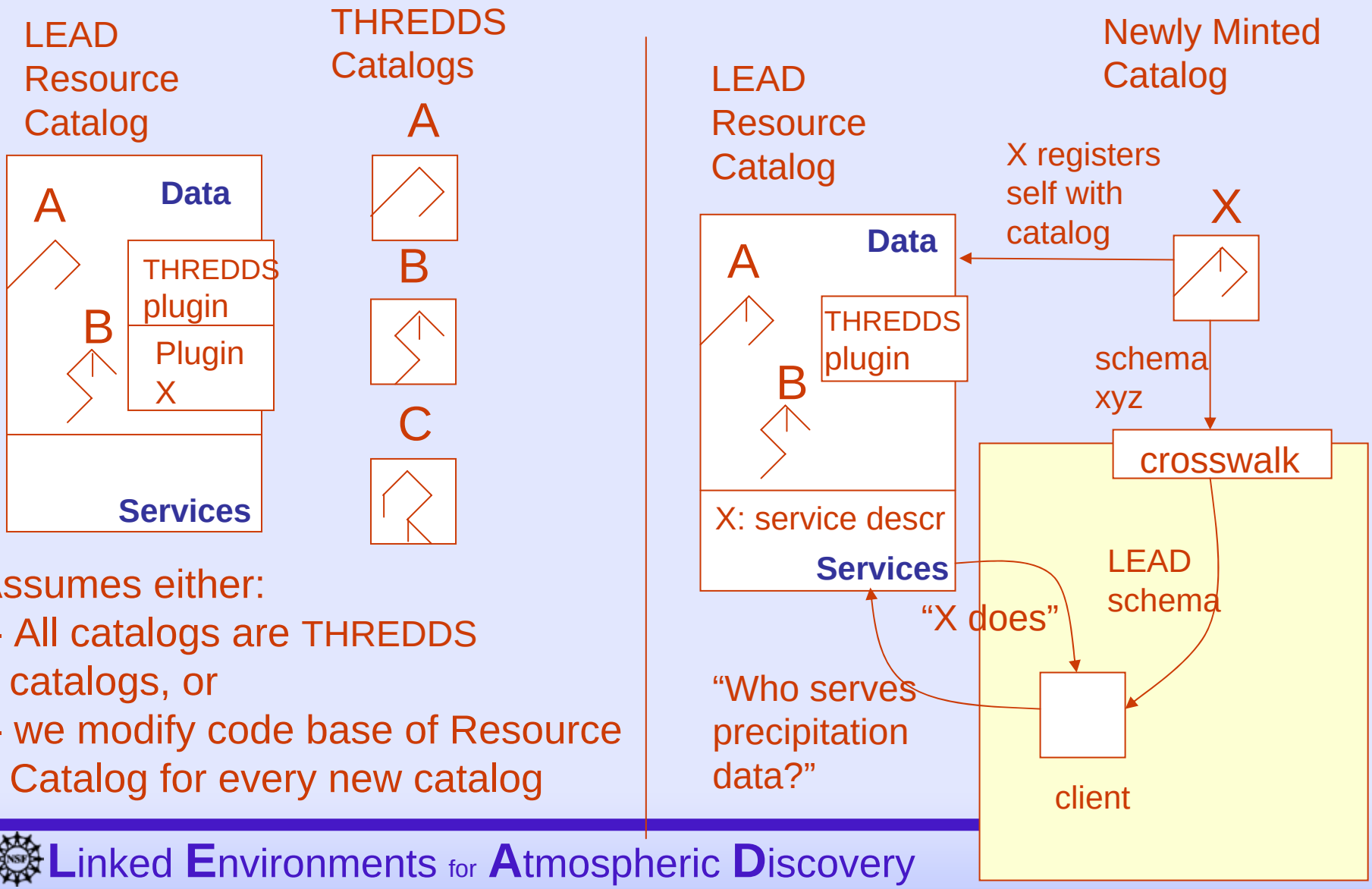


# Conceptual support of expandable architecture: sandbox and crosswalk




***Metadata crosswalk:*** mapping between native interface schema supported by external collection and LEAD metadata schema.

# Adding new catalogs: current vs. future schemes



# Subsystem Accomplishments

- LEAD metadata schema -
  - 12 month highly cooperative effort (3 group-level F2F meetings, agreed upon standards compliance, agreed upon content)
  - V1 released Summer 05
- Subsystem level requirements document
  - Spring 05 effort
  - V1 released June 05
- Interoperability and integration
  - myLEAD, resource catalog integrated with workflow orchestration
  - Portal, query service, ontology, resource catalog and myLEAD - high (concept) level query access transparency
  - Metadata generation - leveraging Unidata Common Data Model
- myLEAD v0.3alpha publicly released open source May 2005



# Year 3 Deployment Goals

Integration of 4 components

**Resource Catalog**
LEAD public products and services

**myLEAD**
User's own experiment products

**Noesis Ontology**
concepts and vocabulary

**Query Service**
query mediation

**THREDDS Catalogs**
-web browser metadata

**Name Service**
- unique ID all products

**Automated metadata generation**
- a capability

**Stream Svc**
- response to weather, stream to app

Stream svc deployment

**OPeNDAP**
data server

**Unidata Data dissem client (LDM)**

**Grid Storage respository**

**Steerable instruments**
- CASA

deployment

# Ongoing Research Goals

- myLEAD
  - Sharing with peers,
  - Versioning experiments through time,
  - Publishing experiment products as LEAD public resource
- Automated metadata generation
  - How much can be accomplished (attribute names only or values as well?) and at what cost?
  - Leverage Common Data Model (CDM) for tool support?
- Provenance - capturing provenance on the fly.
- Noesis Yellow Pages to data catalogs
- Semantic mediation
- Ontology browsing
- Performance scalability of myLEAD and LEAD resource catalog



# Questions?