

# Metadata Management, Provenance & Notifications in LEAD

Yogesh L. Simmhan  
Marcus Christie  
Indiana University



LEAD: Linked Environments for Atmospheric Discovery



# MyLEAD Metadata Catalog



LEAD: Linked Environments for Atmospheric Discovery



## Metadata Catalogs

- Large volumes of scientific data require organized storage and search capabilities
- Scientist earlier used file systems & paper-based lab books, but...
  - File systems provide limited metadata capability
  - Notebooks do not scale well
- Allow the user community to store & access important data resources
  - User's personal space on Grid that is shareable
  - Visualize it, publish it, download it, curate it
- Data Discovery
  - Searchable metadata directories



Science Gateways  
2006-06-12

[3]

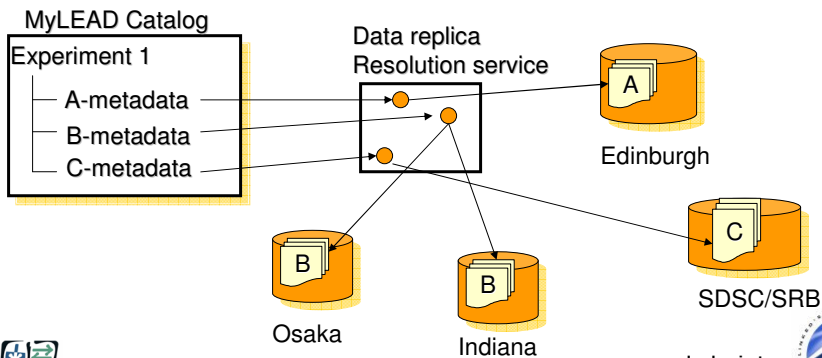
www.leadproject.org

LEAD: Linked Environments for Atmospheric Discovery



## MyLEAD Metadata Catalog

- Organized catalog of personal resource metadata
- Actual data storage is virtualized
  - The data itself may reside anywhere
  - Replicas may exist. Data may move.



Science Gateways  
2006-06-12

[4]

www.leadproject.org

LEAD: Linked Environments for Atmospheric Discovery



## Metadata in a Grid Environment

- Data products in many scientific domains are binary products – difficult to query directly
- Need the ability to find data products without complex query languages
- Queries search for data products based on properties (metadata attributes) of those products
- Service Oriented Architecture.
  - Metadata is communicated via schema-based XML
  - Some properties are complex – beyond name/value
  - Properties must be extensible as models evolve
  - Must be able to handle schema changes, multiple schemas



Science Gateways  
2006-06-12

[5]

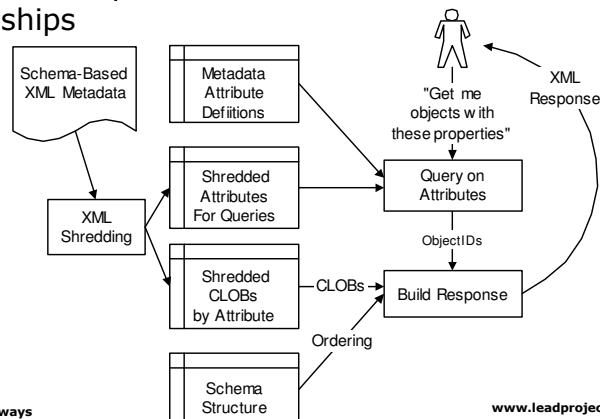
LEAD: Linked Environments for Atmospheric Discovery

www.leadproject.org



## Storing Metadata in MyLEAD

- LEAD XML Metadata Schema (based on FGDC) interface with hybrid XML-Relational storage
- Extensible complex attributes & hierarchical relationships



Science Gateways  
2006-06-12

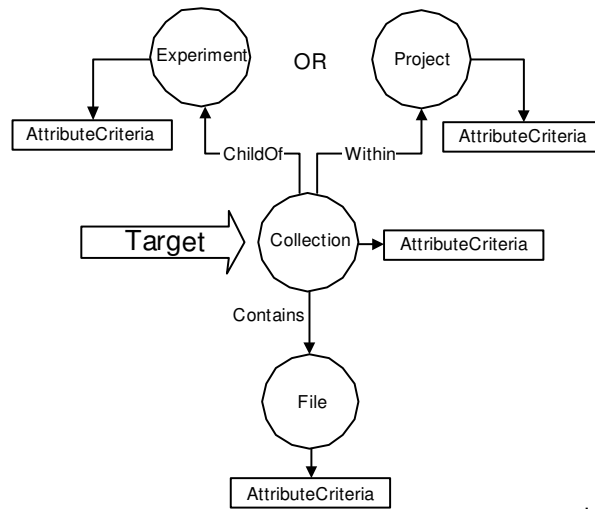
[6]

LEAD: Linked Environments for Atmospheric Discovery

www.leadproject.org



# MyLEAD Query on Hierarchy



Science Gateways  
2006-06-12

TeraGrid

[7]

LEAD: Linked Environments for Atmospheric Discovery

[www.leadproject.org](http://www.leadproject.org)



# Provenance & Notifications



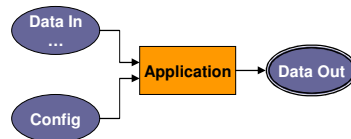
TeraGrid

LEAD: Linked Environments for Atmospheric Discovery



## Data Provenance & Workflow Trace

- **Data Provenance** is metadata on the derivation history of data products
- It provides information on the application used to derive a data product, and the inputs to that application



- **Workflow Trace** is metadata describing the runtime execution of applications composed as workflows
- It describes which applications were run as part of workflow, when & where they ran, and their inputs & outputs



Science Gateways

2006-06-12

TeraGrid

[9]

[www.leadproject.org](http://www.leadproject.org)

LEAD: Linked Environments for Atmospheric Discovery



## Uses of Provenance

- Trace Workflow Execution
  - What services were used during workflow execution?
  - Validate if all steps of execution successful?
- Audit Trail
  - What resources were used during workflow execution?
- Data Quality & Reuse
  - What applications were used to derived data products?
  - Which workflows use a certain data product?
- Attribution
  - Who performed the experiment?
  - Who owns the workflow & data products?



Science Gateways

2006-06-12

TeraGrid

[10]

[www.leadproject.org](http://www.leadproject.org)

LEAD: Linked Environments for Atmospheric Discovery



## Using Notifications to Track Provenance

- ▣ Several Provenance Activities take place during the lifecycle of a workflow
  - Workflow Related (Started, invoking service, ...)
  - Service Related (Invoked, run App, finished...)
  - Data Related (Data Produced, Consumed)
- ▣ Activities are modeled as notifications that are sent by different components
  - Loosely coupled, easy to generate provenance
- ▣ WS-Messenger Notification Broker acts as message bus
- ▣ Provenance service, Workflow composer & Portal (thro' MyLEAD) subscribe to notifications



Science Gateways  
2006-06-12

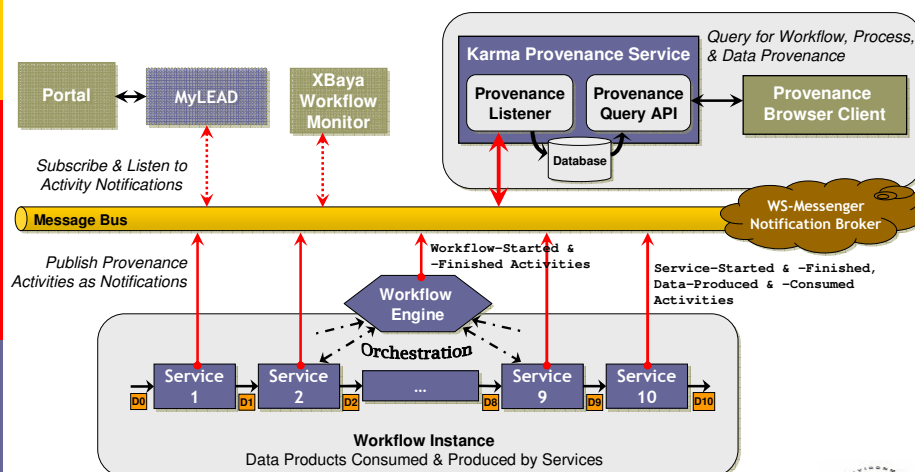
[11]

LEAD: Linked Environments for Atmospheric Discovery

www.leadproject.org



## Karma Provenance Architecture



Science Gateways  
2006-06-12

[12]

LEAD: Linked Environments for Atmospheric Discovery

www.leadproject.org



## Querying Provenance

- Three types of provenance can be queried
- Process provenance
  - When was an application run & in which WF?  
What were its input and output data products
- Data Provenance
  - What service & WF generated this data product & when? Which services & WFs use this & when?
- Workflow Trace
  - What were all the services invoked in this workflow & when? What data were consumed & produced by them?



Science Gateways

2006-06-12

TeraGrid

[13]

[www.leadproject.org](http://www.leadproject.org)

LEAD: Linked Environments for Atmospheric Discovery

