Structure, sharing and preservation of scientific experiment data

Sangmi Lee Pallickara, Beth Plale,
Scott Jensen, Yiming Sun

Computer Science, Indiana University

The Data-*in the Scientific Computing*

Overwhelming amount of data in Computational science and getting more so, from where? And why?

- Evolution of Scientific Model
  - Nested model runs (e.g. data assimilation)
  - Fine Control of models (configuration parameters)
- Improvement of Scientific Experimental environment
  - Finer resolution of observational instruments
  - Streaming continuously from hundreds of sensors and network sources.
  - Large archives
- Sophisticated Collaboration between Scientists
  - More active collaboration (annotation, data sharing) in the Web enabled working environment
- Informatic Technology
  - Data mining

So, is it manageable?

- Computational scientists are reaching their limit on ability to manage data products associated with each of their scientific experiments.
- Common Web-based searching/downloading approaches are not suitable for scientific computing (data modification, interoperating with other services, and sharing with security issues)

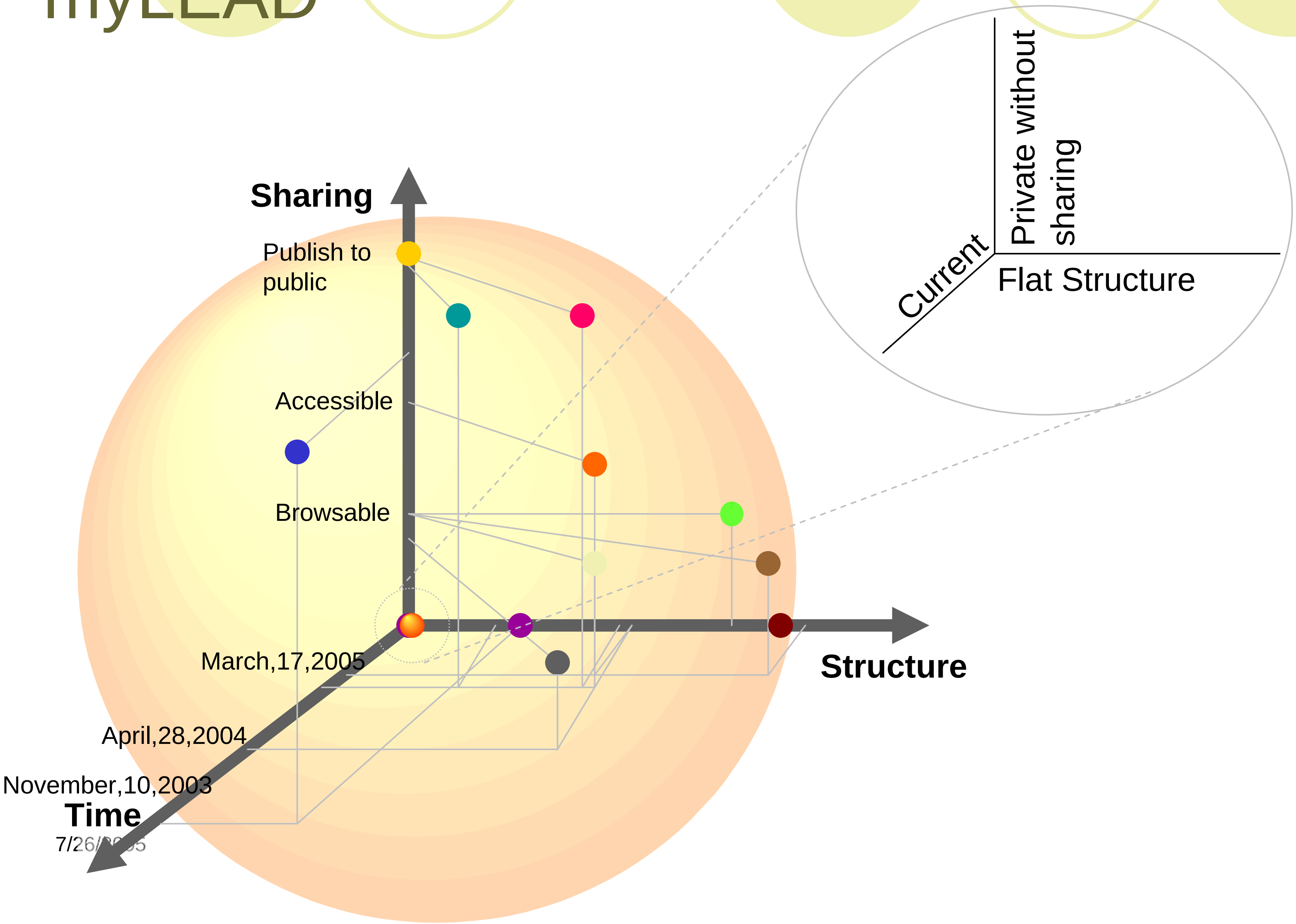
Requirements of the Data Management

- *Total* control over their data products
- The ability to share products but *retain control* over what gets shared, and with whom
- Rich search criteria over the vast information space *without* writing SQL queries.
- Help managing experiment products generated over an extended period of time (i.e., years),
- High level of reliability
- The ability to work locally

myLEAD:
an ‘active’ metadata catalog

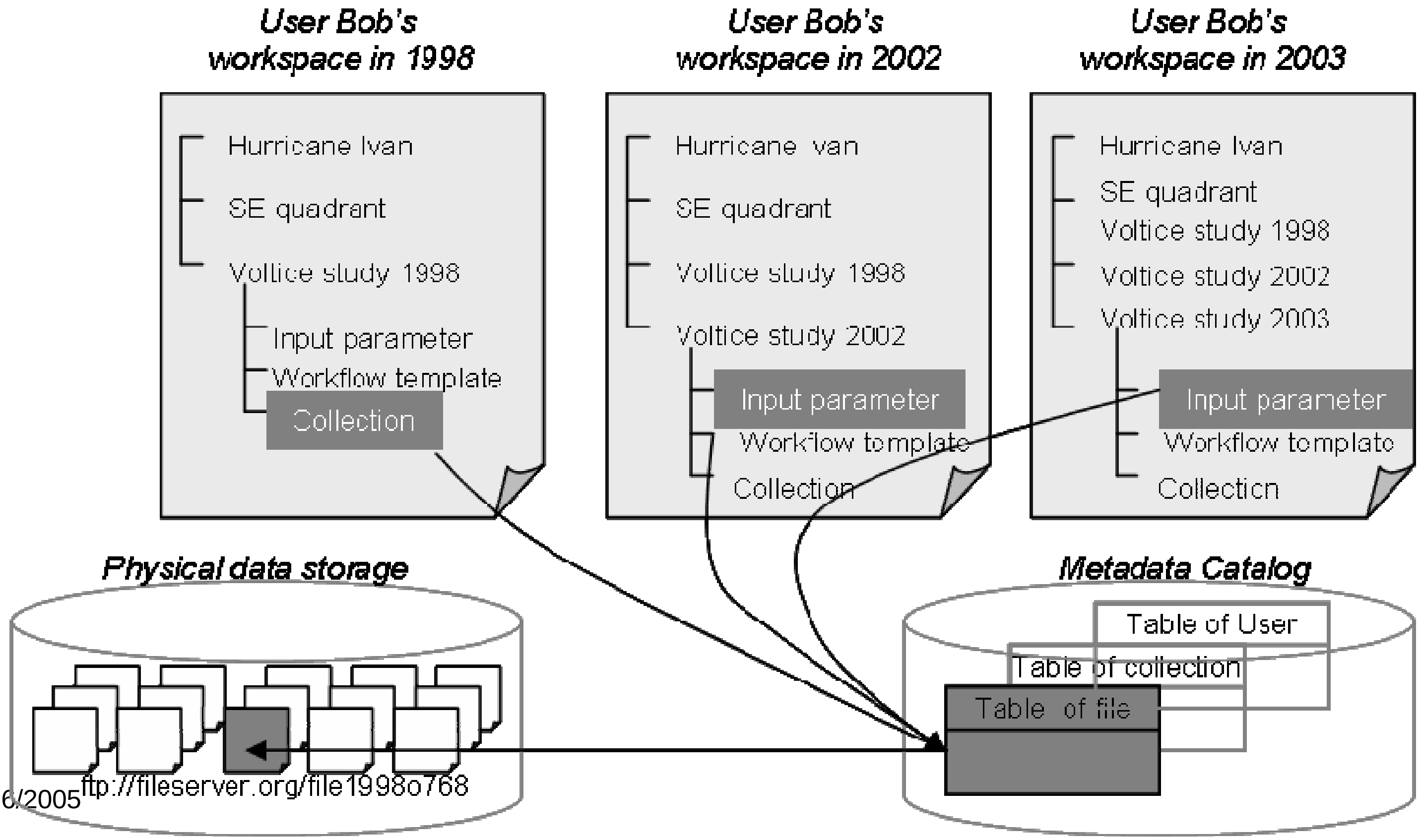
- If we’re going to have half a chance of being widely used, it is going to be us that reaches 3/4’s of the way across the gulf. Our users reach the other 1/4:
  - Easy query “writing”
  - Automated metadata generation
  - Transparent structure management
  - Transparent versioning management
  - Expressive query writing

Envisioning Personal Workspace with myLEAD

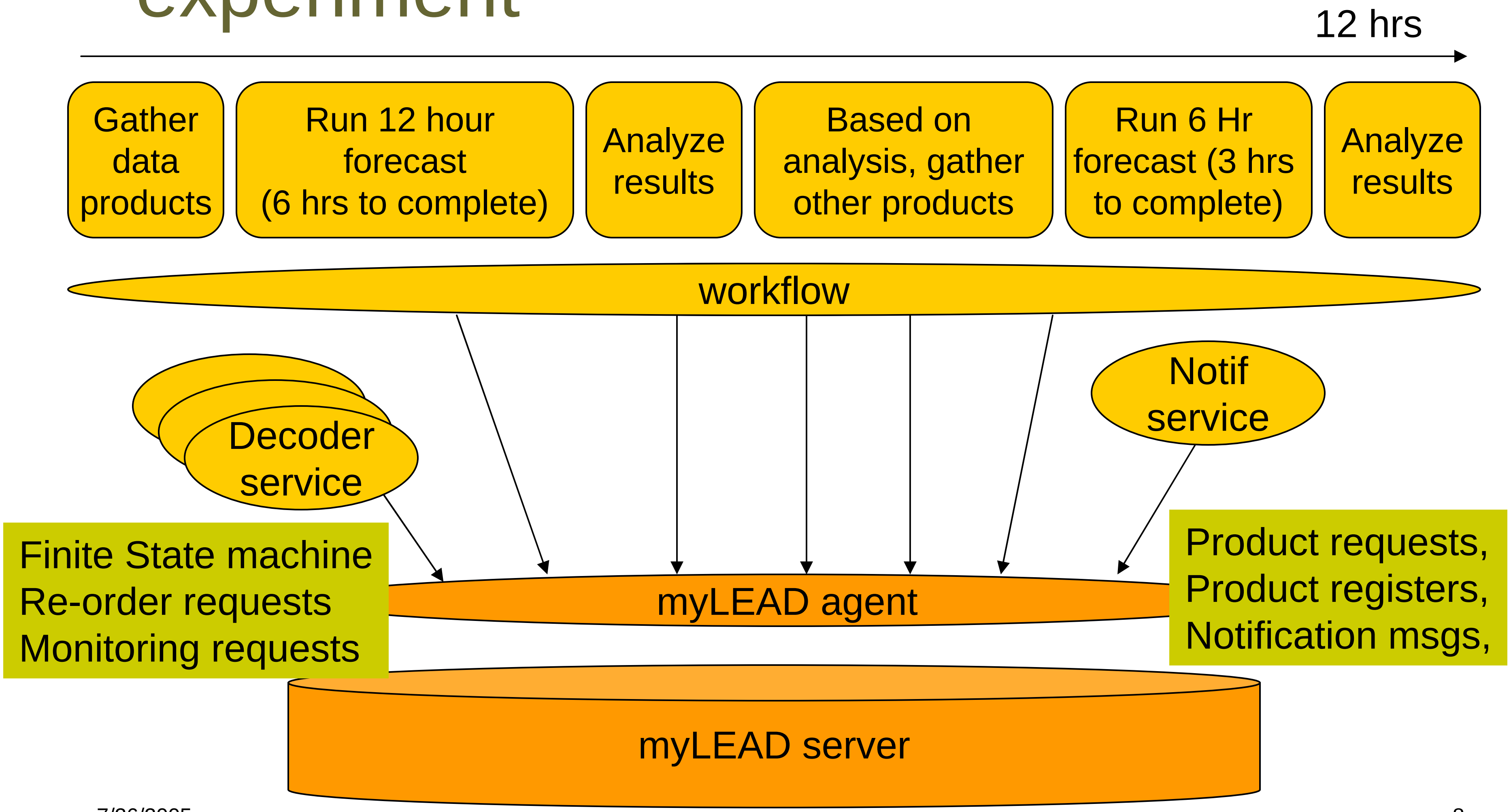


Structure I.
Providing Structural Transparency

- Flexible but interoperable structure
- Structural Transparency



Structure II. Creating structure in database that mirrors structure of experiment



12 hrs

Gather data products

Run 12 hour forecast (6 hrs to complete)

Analyze results

Based on analysis, gather other products

Run 6 Hr forecast (3 hrs to complete)

Analyze results

workflow

Decoder service

Notif service

Finite State machine
Re-order requests
Monitoring requests

myLEAD agent

Product requests,
Product registers,
Notification msgs,

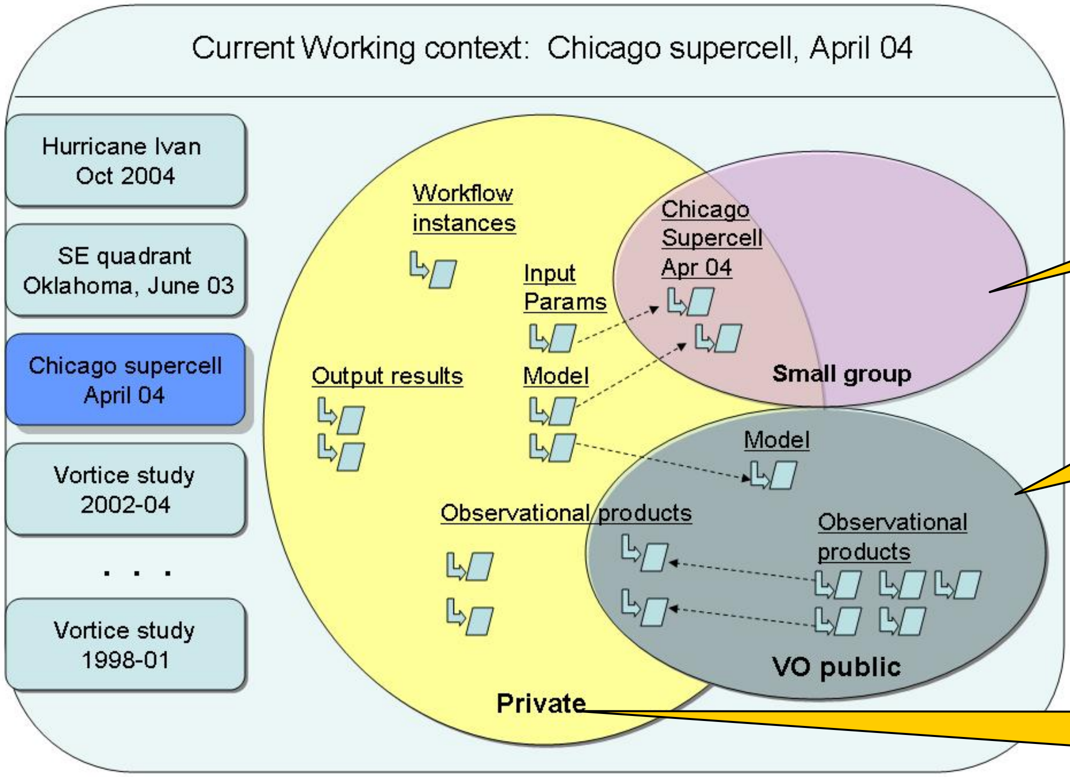
myLEAD server

Sharing I. Supporting Data Sharing

- Flexible sharing between individuals, groups, and individuals vs. groups.
- Flexible depth of sharing:
  - Depth-0: participant (P) is unaware that experiment data (E) owned by user (U) exists
  - Depth-1: P is aware that E exists
  - Depth-2: P can search E
  - Depth-3: P can browse the content of E
  - Depth-4: P can access E and its contents
  - Depth-5: P can remove and write E

Sharing II. Flexible sharing of the Data Product

- user interface to information space showing current experimental context and levels of sharing of various data products

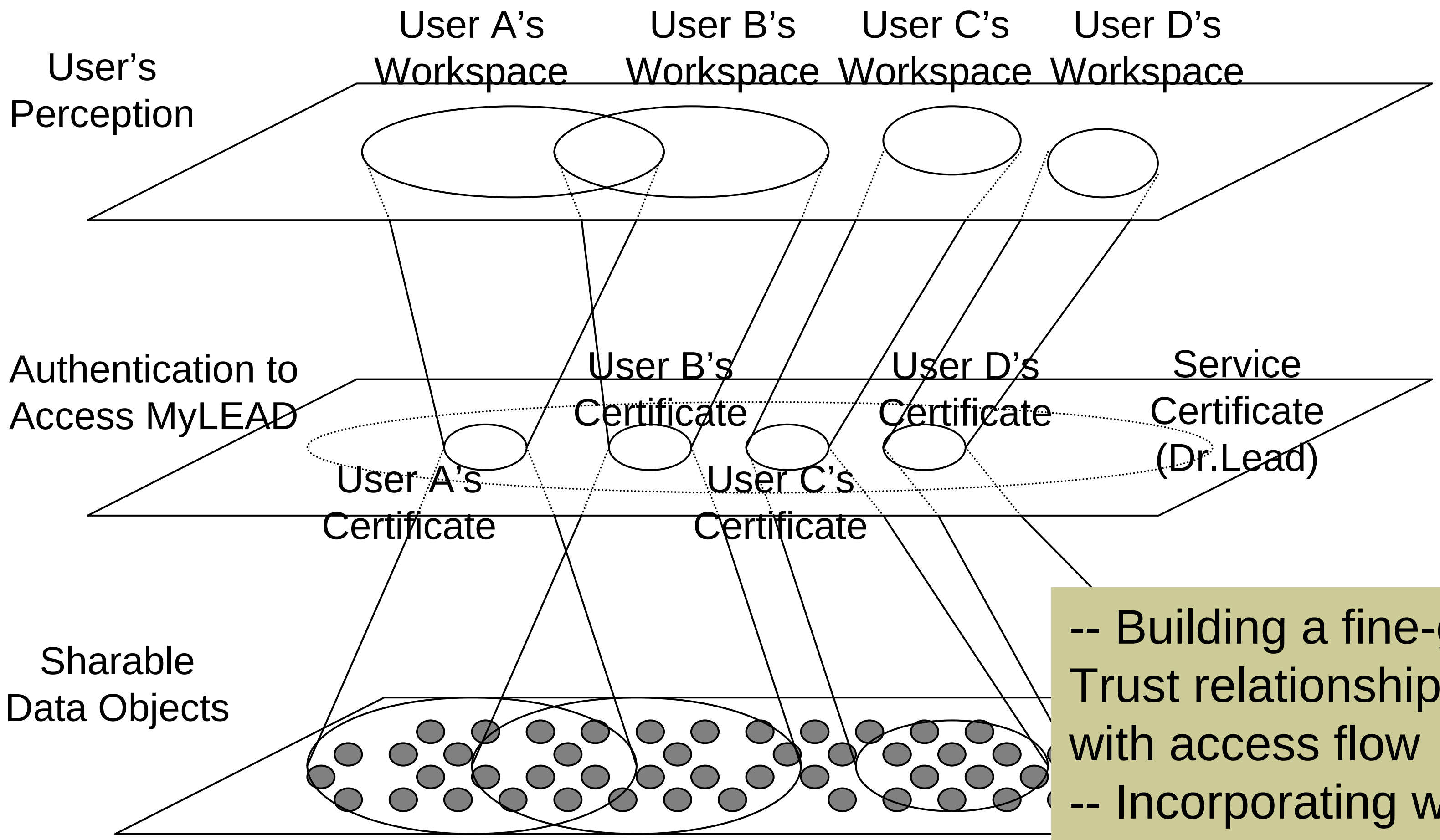


Limited Data Access (e.g. Browse or copy)

Complete Data Access

User's private Data Area

Sharing III. Building Fine-grained Trust scheme



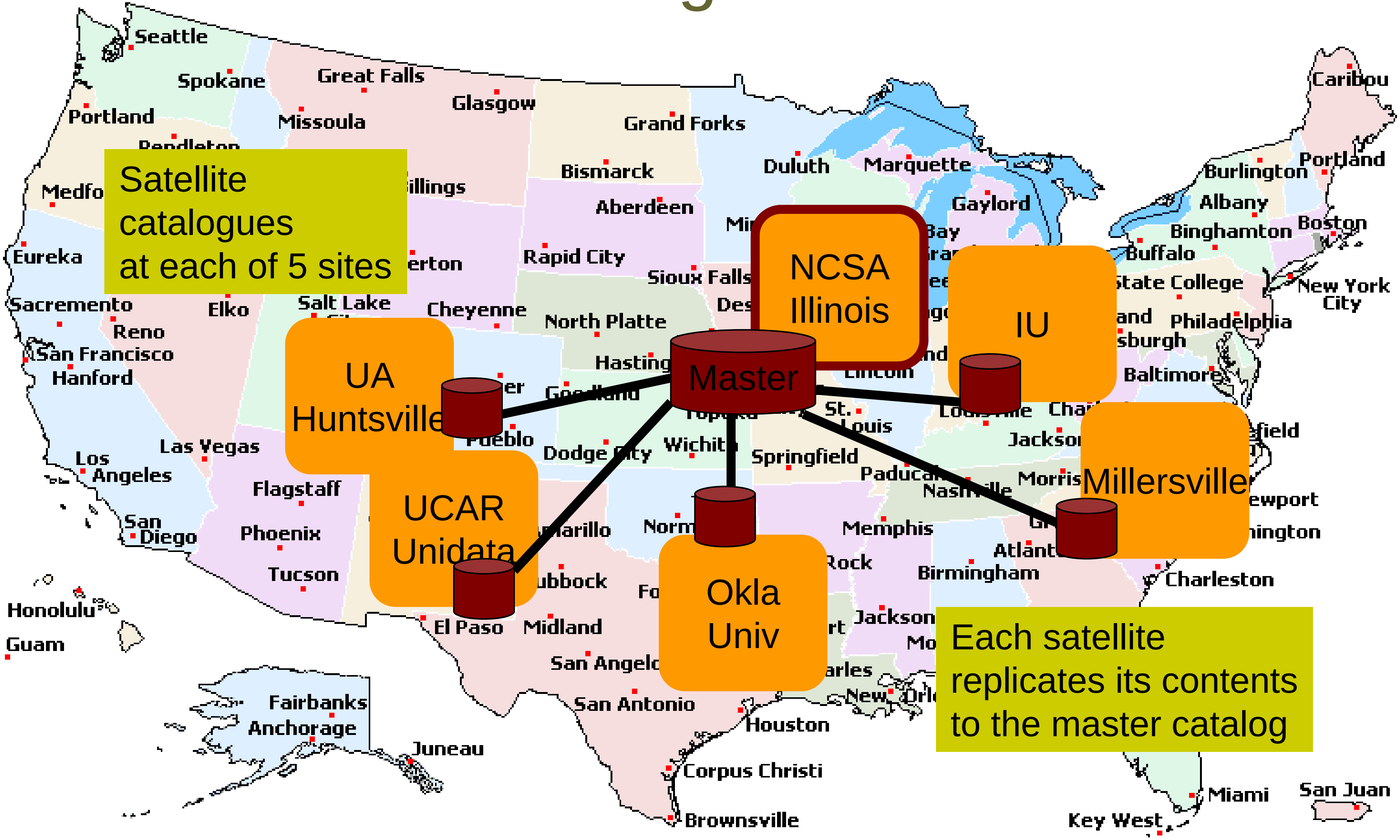
-- Building a fine-grained Trust relationship along with access flow
-- Incorporating with GSI and adapting the access Control.

7/26/2005

Preservation

- Versioning the data objects along with time frame based on user’s decision
- Scientific experiments are repeated until the scientist is satisfied with the result
- Mark with *Landmark* for useful data product
- Archive data product

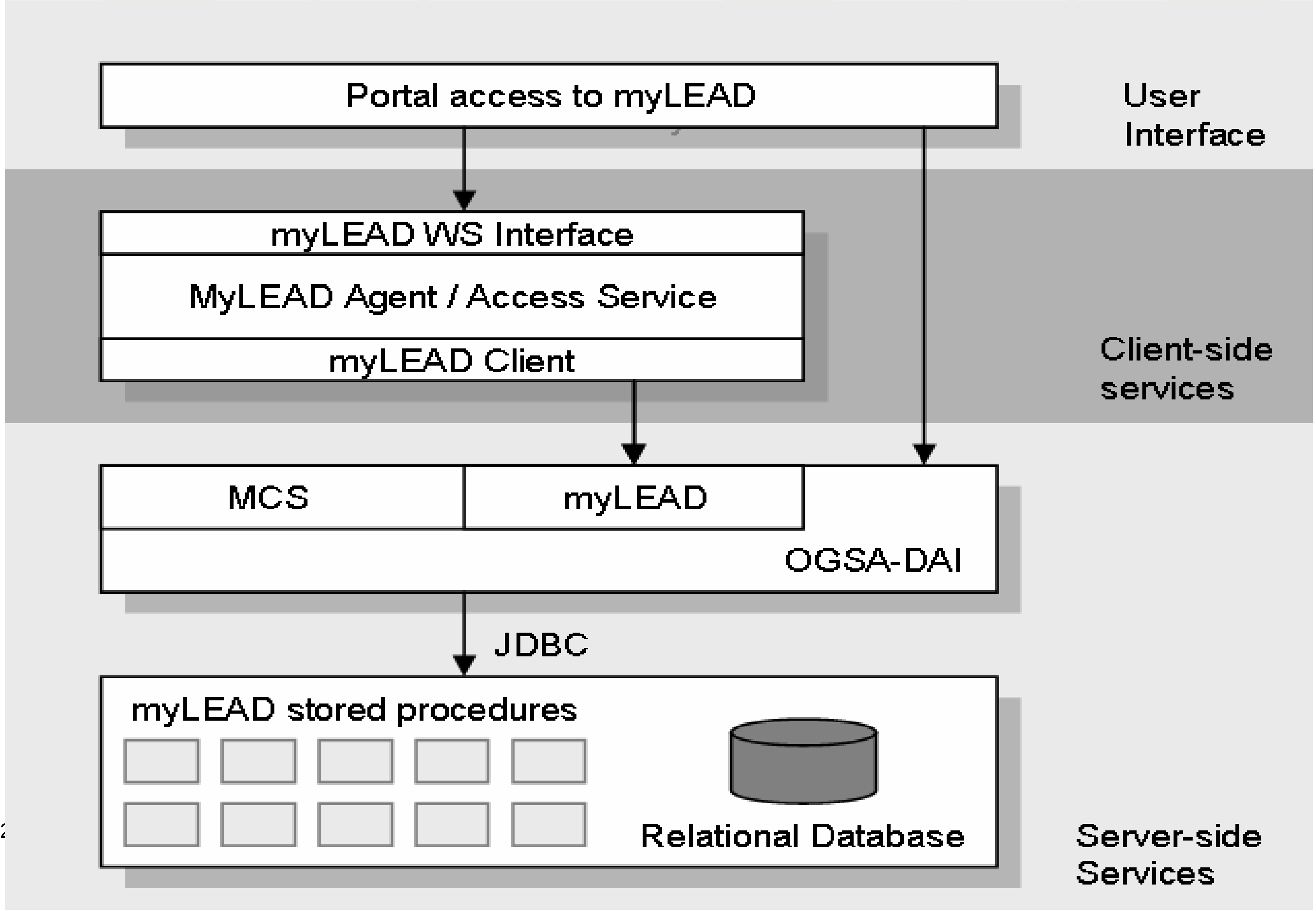
Architecture Part 1: Distribution scheme of metadata catalogues



Satellite catalogues at each of 5 sites

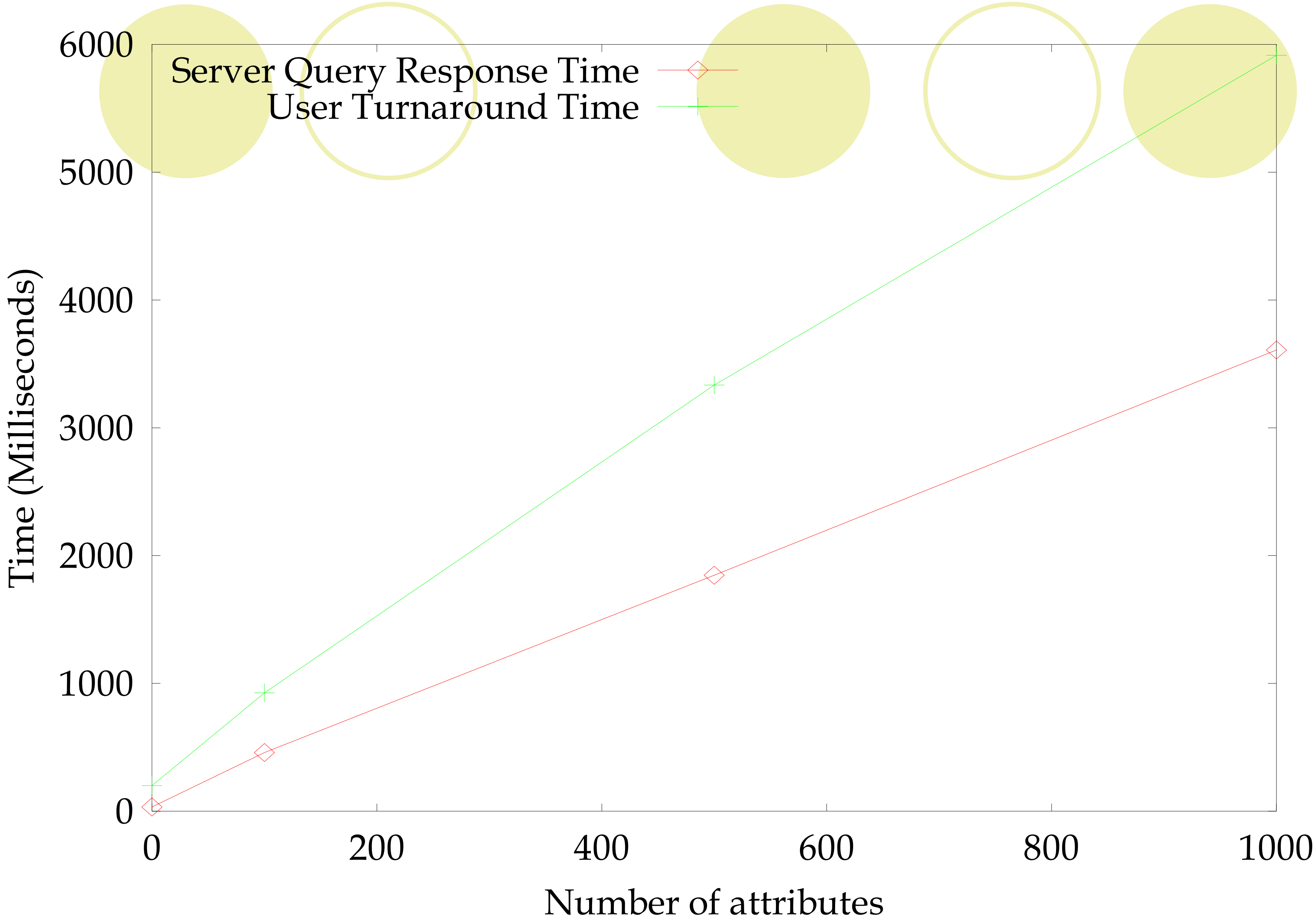
Each satellite replicates its contents to the master catalog

Architecture Part II. Single myLEAD

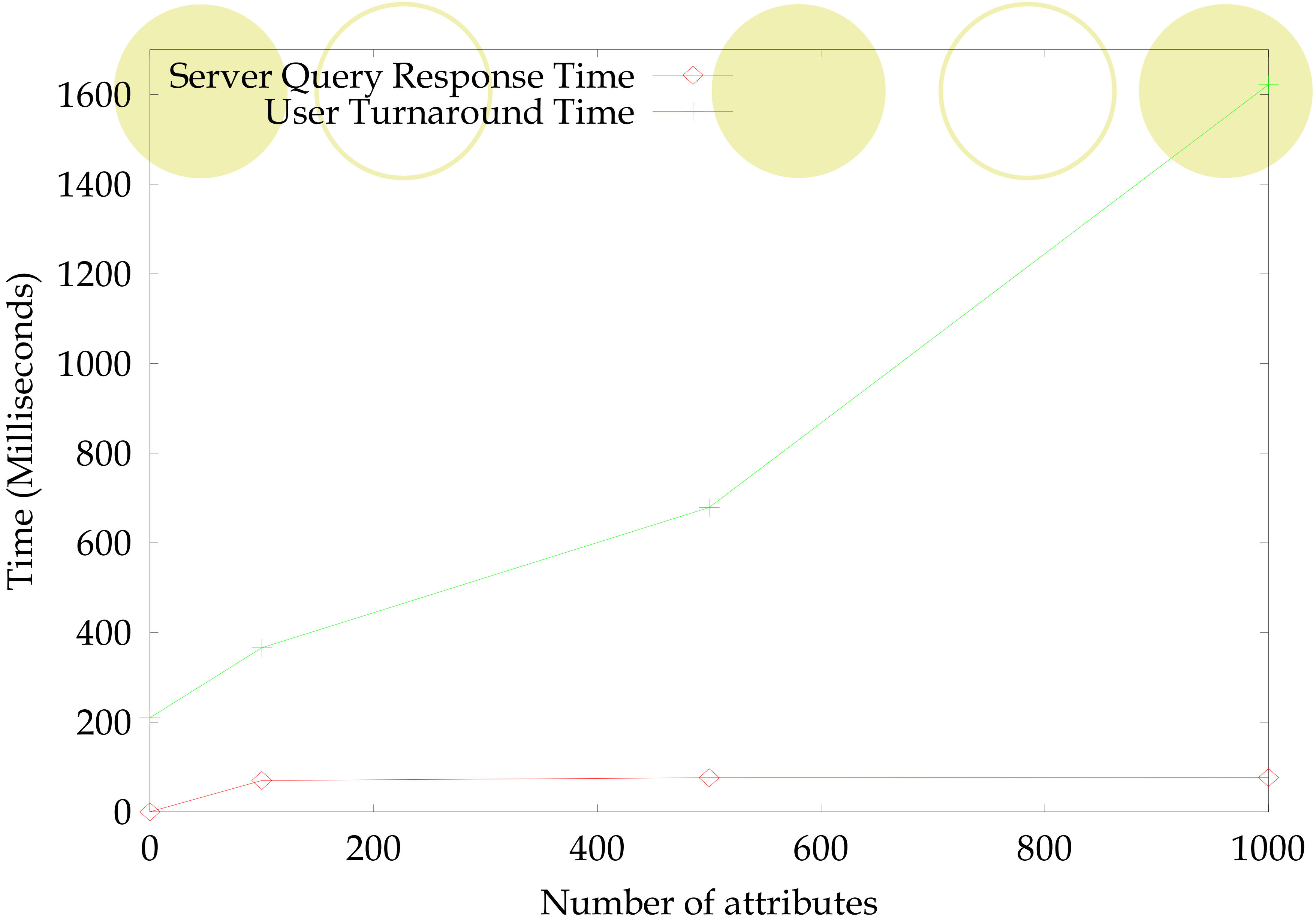


Performance Evaluation

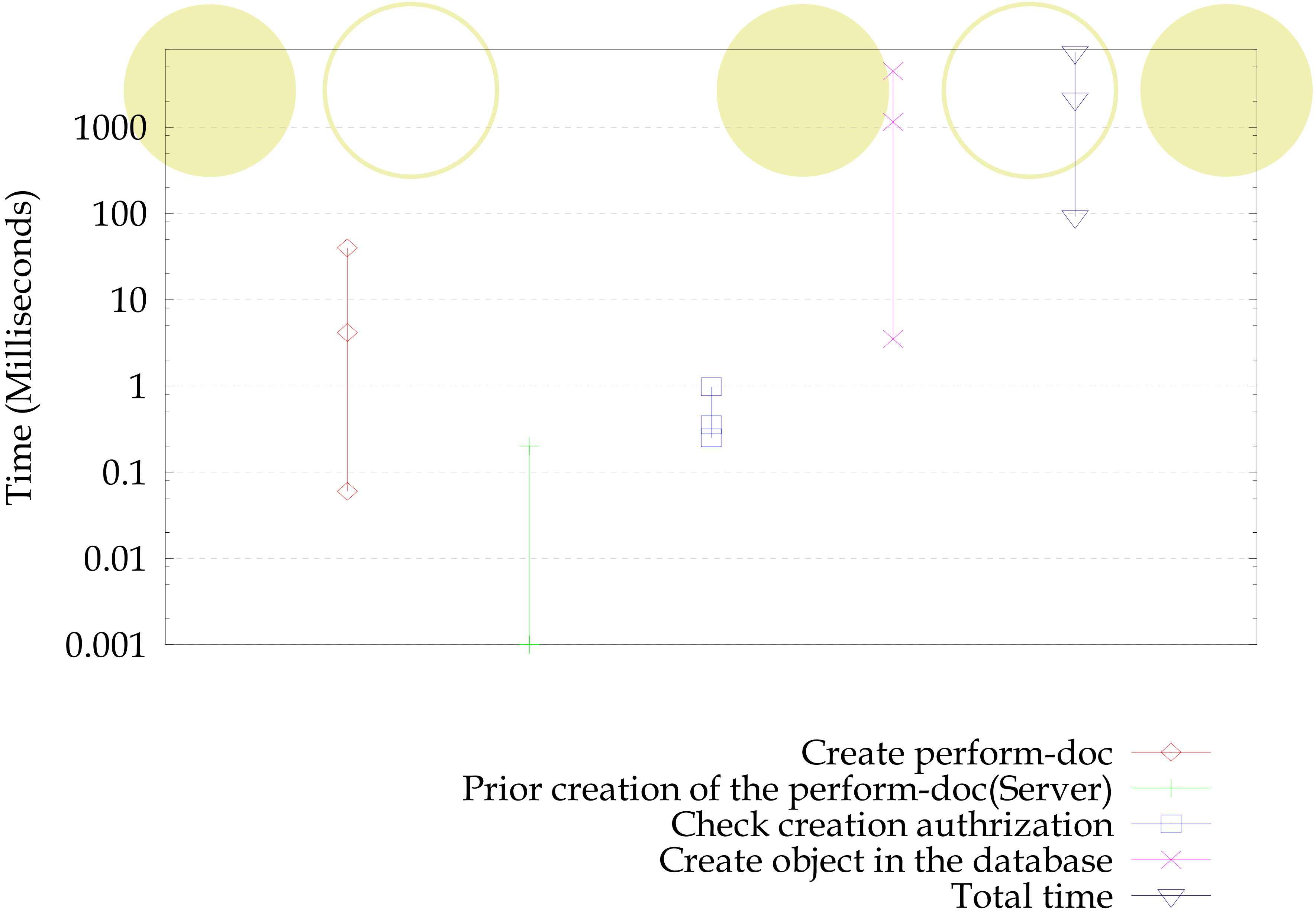
- MyLEAD extends Globus MCS
  - Extending the schema by including support for spatial and temporal attributes
- Client
  - A dual processor Dell PowerEdge 6400 Xeon server (700MHz PentiumIII), 2GB RAM, 100GB Raid 5, RedHat 7.2, JDK1.4.2.
- The myLEAD server
  - A dual processor 2.0 MHz Opterons, 16GB RAM, GENTOO Linux.
  - The OGSA-DAI version 3.0, Globus MCS version 3.1 and provides access to the database platform, mySQL-version 5.0.0.
- The myLEAD client and the myLEAD server are interconnected through a 1 Gbps switched Ethernet LAN.
- Single user



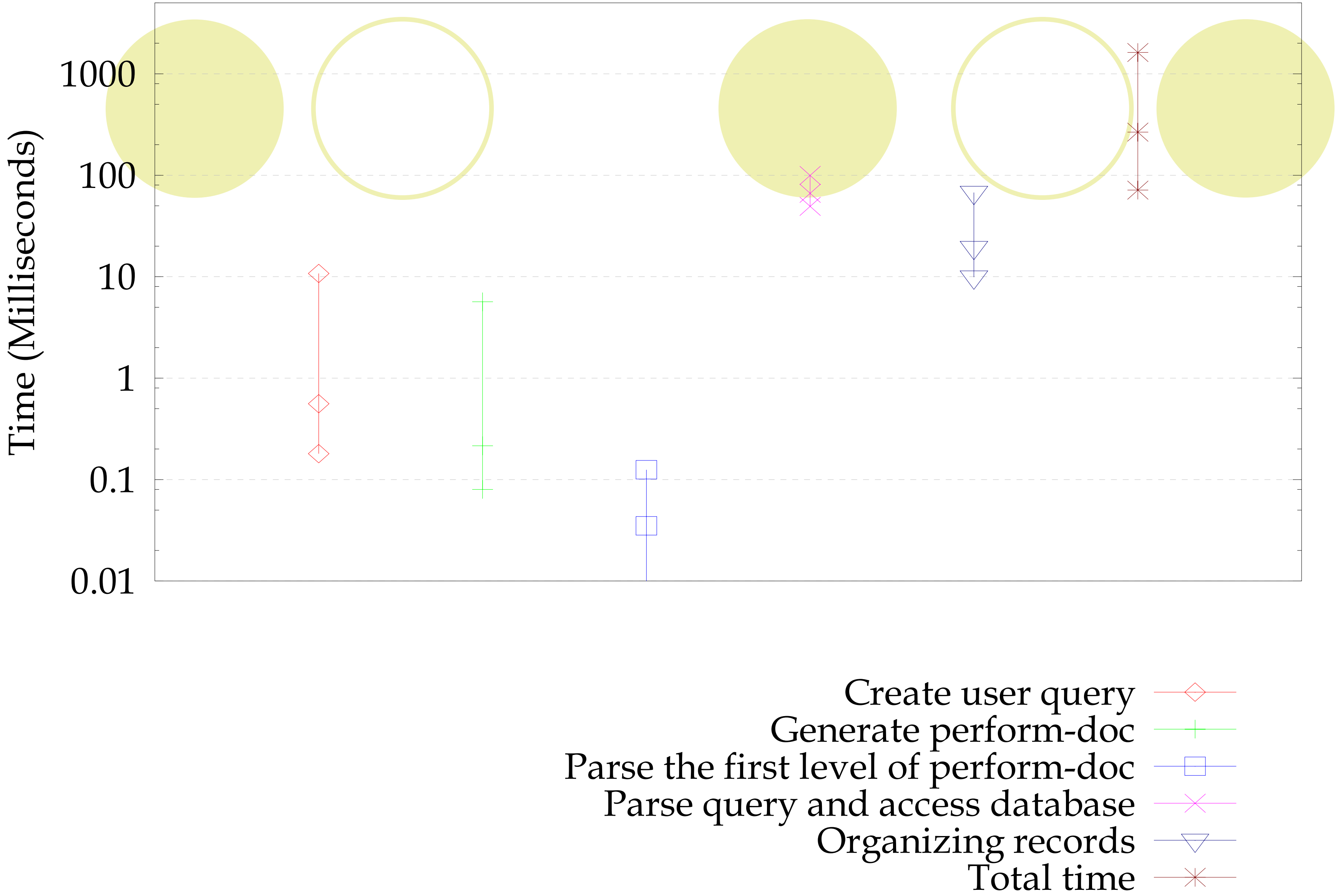
Increase of the attribute creation time by the increase in the number of attributes



Increase of the attribute query time by the increase in the number of attributes



Partial cost of creating attribute in myLEAD



Partial cost of querying attribute in myLEAD

Conclusion and summary

- MyLEAD metadata catalog provides personal workspace enabling
  - Structuring
  - Sharing
  - Preservation of the meteorological experimental data objects
- Architecture of myLEAD
- Performance

Future works

- Scalability
- Immutable experiments
- Convey visual cues of secure data access

- http://www.cs.indiana.edu/dde/projects/mylead03alpha/myLead.html