Registry Framework: Front-to-Back

TODD KING[1], JAN MERKA[3,4], THOMAS NAROCK[3,4], RAYMOND WALKER[1,2], LEE BARGATZE[1]

[1] INSTITUTE OF GEOPHYSICS AND PLANETARY PHYSICS, UNIVERSITY OF CALIFORNIA LOS ANGELES, CA 90095,
[2] EARTH AND SPACE SCIENCE DEPARTMENT, UNIVERSITY OF CALIFORNIA LOS ANGELES, CA 90095,
[3] HELIOSPHERIC PHYSICS LABORATORY, CODE 672 NASA GODDARD SPACE FLIGHT CENTER,
[4] UNIVERSITY OF MARYLAND, BALTIMORE COUNTY, BALTIMORE, MD 21250

ESSI WORKSHOP, AUGUST 3-5, 2009

Overview

- A "Registry" is a location in an organization where definitions are stored and maintained.
- A "Metadata" registry stores information about data models.
- A "Resource" registry stores structured, descriptive information about resources.
  - includes the scientific context related to the resource; its temporal, spatial or spectral range; expert contacts; and where the resource can be found and accessed.
- A Resource registry plays a central role in connecting a user to the data.
- Resource information is used to support search engines, retrieval services and resource exploration.

Registry Functional Aspects

ISO-11179 identifies functional aspects of a Metadata Registry which are also applicable (with a little twist) to Resource Registries.

Functional Aspects:

- Administration and Identification
  - Management of information related to a resource and its provenance.
  - Handled by the registry framework.
- Naming and Definition
  - Each managed item has a name and definition which conforms to established policies.
- Classification
  - The language (data model) used to describe the resource.

Domain Realities

(True in most science domains)

- Data and Metadata are not co-located.
- The Data and Metadata environments must be symbiotic.
  - Co-exist
  - Complementary
- Disciplines require different value-added services (views of the data and environment)
- Many concurrent efforts.
  - It's a large domain with multiple agencies involved.

Data Environment

- Data is provided by
  - Missions
  - Research Groups
  - Archives
  - International peer systems
- Multiple disciplines

  Heliophysics example:
    - Magnetospheres
    - Waves
    - Ionosphere-Thermosphere-Mesosphere
    - Radiation Belts
    - Energetic Particles
    - Solar Physics
    - Models and Simulations

Metadata Environment

- Metadata is needed to describe resources:
  - Orignation: Observatories, Instruments, Persons
  - Infrastructure: Registry, Repository, Service
  - Data: Numerical, Display, Catalog
- Metadata comes from many sources
  - A Virtual Observatory, data provider, researcher, resident archive and more.
- Metadata is utilized in services
  - Examples: registries, search engines, downloaders, visualization tools (autoplot)

Framework Components

- Metadata Management
  - Well defined workflow
  - Reliable and trusted information
- Registry Services
  - Update-to-date content
  - Comprehensive scope
  - Reliable and trusted access

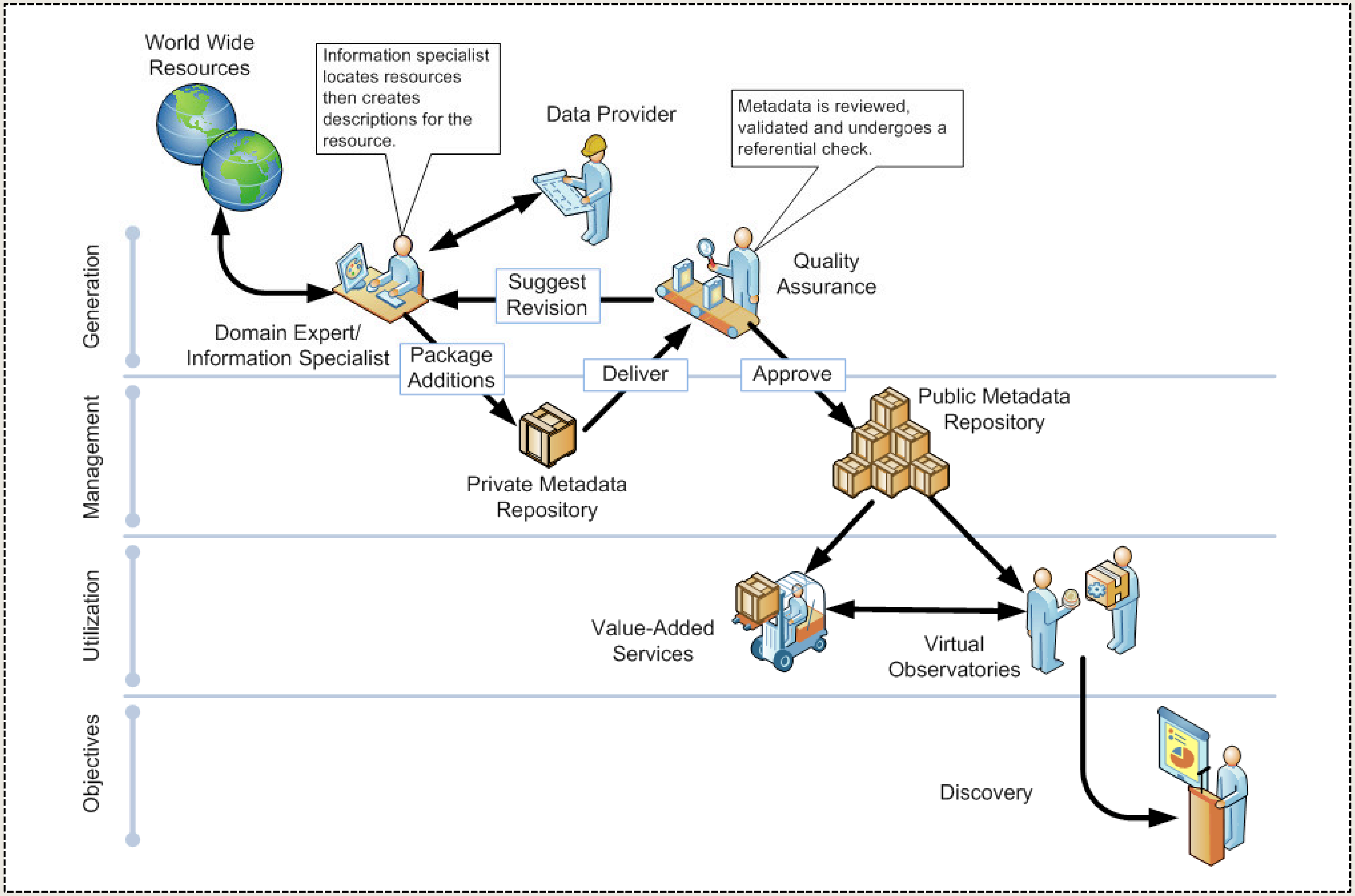
Metadata Management

- The key to a reliable and trusted data environment is well managed metadata.
- Distributed editing needs a moderator.
- There is a need for revision control.
- After exploring many alternatives we settled on "git" is for metadata development/management
  - Keeps track of history,
  - Allows a review process before “release”
  - Changes are properly attributed
    - ('we know who to blame'),
  - Simple to go back,
  - Off the shelf ready
    - no need to develop it.
    - robust and tested.

Protocols and Procedures

- Update Protocols
  - Repositories are cloned, modified locally and patches sent to moderator.
- Review Protocols
  - Domain experts review content before inclusion in the shared repository.
- Harvesting and synchronizing
  - supported by git commands.

Metadata Workflow


World Wide Resources
Information specialist locates resources then creates descriptions for the resource.
Data Provider
Metadata is reviewed, validated and undergoes a referential check.
Generation
Management
Utilization
Objectives
Suggest Revision
Quality Assurance
Domain Expert/ Information Specialist
Package Additions
Deliver
Approve
Public Metadata Repository
Private Metadata Repository
Value-Added Services
Virtual Observatories
Discovery

Registry Service
Scope

- A Registry supports initial discovery.
  - Unstructured search (keywords)
  - Core science criteria
    - Temporal Range
    - Spatial Range
- Information needed for initial discovery is very general.
  - Applicable over a broad range of disciplines.
- Allows drill-down into detailed information.
  - Detailed information may be maintain in domain specific data models.

Principles Regarding a Resource

- Has a unique identifier.
- Has a time range of observation.
- Has descriptive information (name and narrative)
- Can have multiple measurement types.
- Can observe multiple regions.
- Can be associated with any number of other resources.
- Can have any number of indexed words.
- May have a spatial extent.

Rosetta Attributes

The "Dublin Core" for data.

Attribute names and occurrence. All "type" are enumerations.

ResourceID [1] ResourceName [1] ResourceType [0..1] Description [1] MeasurementType [0..*] PhenomenonType [0..1] ObservedRegion [0..*] ObservatoryID [1] ObservatoryName [1] ObservatoryType [0..*] ObservatoryGroup [0..*]	InstrumentID [1] InstrumentName [1] InstrumentType [0..1] ReleaseDate [1] StartDate [1] StopDate [1] Cadence [0..1] Latitude [0..1] LatitudeExtent [0..1] Longitude [0..1] LongitudeExtent [0..1] Association [0..*] Word [0..*]

Registry Service
Required Functions

Essential registry functions (reformulation of IVOA, OAI-PMH and new requirements)

- Search
  - Keyword
  - Facet (Constraint in IVOA, Sets in OAI-PMH)
  - Constrained (XQuery search in IVOA)

  Note: A time range can be used to set the scope of any search.
- Retrieval
  - Resource Description
  - Resource references (URL and Resource ID)
- Existence
  - ID Stemming
  - ID Resolution

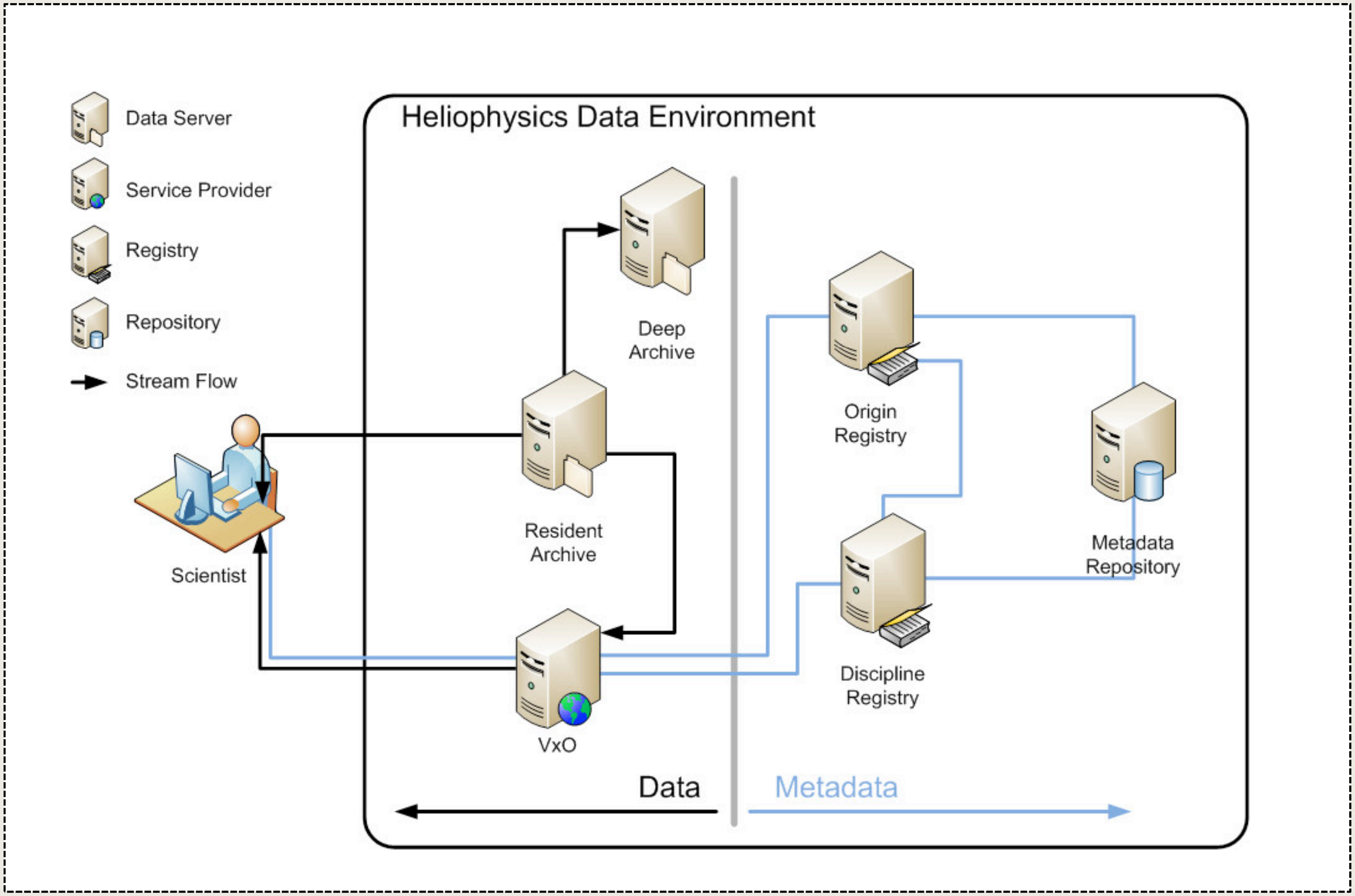
Will this work?

Yes.

Example: The Heliophysics Solution

- SPASE metadata
  - Designed for highly distributed data.
  - Links metadata to data by a URL reference.
  - Assigned universal identifiers to each described resource.
- Operational separation of functions
  - Resident Archives for data
  - Virtual Observatory for services/views
  - Resources (objects) passed by Resource ID
  - SPASE aware sevices.

NASA's Heliophysics System Model



NASA's Heliophysics VxOs

- VxOs (and others) create resource descriptions using SPASE terms.
  - Expressed in XML.
  - One resource description per file.
- Files are stored in a metadata repository.
  - Repositories are organized by "authority" (e.g, one for VMO, VHO etc.)
- Repositories are harvested and query services are added forming registries.
  - All repositories are harvested to form the general Inventory.
  - Discovered Resource ID collisions are resolved.
- Registries can be queried by others
  - Serve as a basis for value added services.

Example Registry Search Interface

- Keyword Search

```
/registry/resolver?w={words}
/registry/resolver?w={words}&b={time}&e={time}
```

{words} is expressed in Lucene query syntax.

```
Example: /registry/resolver?w=plasma
```

- Facet (Constraint in IVOA, Sets in OAI-PMH)

```
/registry/resolver?f={name:value}
```

{name} is pre-defined.

```
Example: /registry/resolver?f=instrumenttype:magnetometer
```

- Constrained
  - SPASE-QL protocols

Registry Retrieval

- Resource Description

  `/registry/resolver?i={id}`

  {id} is a Resource ID.

  ```
  Example:
  /registry/resolver?i=spase://SMWG/Observatory/ISEE1
  ```

- Resource References (Granules)

  ```
  /registry/resolver?g={id}
  /registry/resolver?g={id}&b={time}&e={time}
  ```

  {id} is parent Resource ID.

  ```
  Example:
  /registry/resolver?g=spase://VMO/NumericalData/DE1/MAGA/PT0.062S
  ```

Registry Existence

- ID Stemming (used to discover resource by walking trees)

  ```
  /registry/resolver?t={stem}
  ```

  {stem} is a Resource ID stem.

  ```
  Example:
  /registry/resolver?t=spase://VMO/NumericalData/DE1
  ```

- ID Resolution

  ```
  /registry/resolver?e={id}
  ```

  {id} is a Resource ID.

  ```
  Example:
  /registry/resolver?e=spase://VMO/NumericalData/DE1
  ```

Registry Tools

Metadata Management needs tools for:

- Validation (against data model schema)
- Referential checking (ID and URL)
- Collator (storage policy enforcer)
- Consolidated reports (Report card)

Registry Service is used by tools to:

- Find and access resources.
- Drill-down for details.

Tools already exist in the HPDE to do each of these.

What's Next?
Community Agreement

- Rosetta Attributes
  - Derived from Heliophysics and Planetary data models.
  - We need broad community vetting. Have we got it right?
- Service Interface
  - Is REST enough?
- Resource Identifier Conventions
  - Use URI with data model name as scheme

```
scheme://authority/path
```

Scheme name	Data Model
spase	SPASE used by Heliophysics Data Environment
pds	Planetary Data System
ipda	International Planetary Data Alliance
ivoa	International Virtual Observatory Alliance