INTEGRATING HUBZERO AND IRODS

GEOSPATIAL DATA MANAGEMENT FOR COLLABORATIVE SCIENTIFIC RESEARCH

Rajesh Kalyanam, Robert Campbell, Samuel Wilson, Pascal Meunier, Lan Zhao, Elizabett Hillery, Carol Song

Purdue University
HISTORY – GEOSSHARE, DRINET, U2U

- Time-series, geospatial data and regular files
- Researchers
  - Manage, query, access, share scientific data
  - Research collaboration
  - Quick preview
  - Run shared post-processing tools
  - Compare data from different sources
• Cyberinfrastructure platform
• User collaboration
  ➢ Groups, projects, blogs, message boards
• Instruction
  ➢ Courses, tutorials, lectures, seminar series
• Data sharing, simple preview, curation
  ➢ Publications with file bundles, supporting documents, DOI generation
HUB TOOLS

• Web-enable scientific tools
• Rappture Tool Kit
  ➢ Common GUI elements
  ➢ Support for various programming languages
  ➢ Output visualization
• Containerized
  ➢ OpenVZ containers with VNC support
• Data transfer to/from local desktop
• Reusable building blocks for geospatial data
  ➢ Processing
  ➢ Metadata extraction
  ➢ Map visualization
  ➢ Search

• Part of the NSF DIBBS initiative
  ➢ Data sharing for collaborative research
  ➢ Diverse domains
**GABBS ARCHITECTURE**

- **End User**
  - Computation
  - Visualization
  - Data Sharing

- **New Capabilities**
  - Maps
  - Control widgets
  - Data presentation
  - Remote servers
  - Overlays
  - Data processing
  - Tool builder
  - Geo-processing
  - Standard protocols
  - Data management
  - Data formats
  - Data sharing
  - Data-Tool connectors

**HUBzero Platform for Scientific Collaboration**

- Computation tools and online databases
- Content publishing
- Collaboration (group, project)
- Learning (courses, self-help)
- Support (tickets, Q&A)
- Community (forum, review, calendar)
HUBZERO AND IRODS INTEGRATION

• Require central storage mechanism uniformly accessible throughout data lifecycle
• Needs to support easy extensibility to handle large file quantities
• Support for processing co-located with data
• iRODS storage underlies Hub Projects Filespace
  ➢ iRODS FUSE mount onto hub webserver
  ➢ PHP Flysystem adapter for CMS access, future expansion
Hubzero and IRODS Integration

- Hub tools have local access to Hub Project files
  - Bind mount users’ accessible collections on webserver into tool OpenVZ container
  - Can serve as tool input source and output destination, simplifying development
- Supports pre, post-processing of files
  - Automatic metadata extraction, ingestion into Apache Solr on file creation
  - On-demand bulk metadata update
  - On-demand visualization of geospatial files
HUBZERO AND IRODS INTEGRATION

Hub Tool

Bind mount

Hub Project Files

Fuse mount

Update File Metadata

Run microservice

Run microservice

Geospatial Preview

Serve layer

Update index

Register Layer

MyGeoHub.org
GEOSPATIAL METADATA EXTRACTION

• Implemented as iRODS microservice
  ➢ Runs on file creation, attached to `acPostProcForPut`
  ➢ Uses GDAL C++ APIs to process vector, raster geospatial files
  ➢ Abstracts extracted information into 15 common Dublin Core Metadata Initiative (DCMI) fields
  ➢ Also extracts geospatial bounds for subsequent geo-search

• Metadata storage
  ➢ Extracted metadata stored as iRODS AVU triples
  ➢ Ingested into Apache Solr for subsequent search
METADATA UPDATE

• Implemented as iRODS microservice
  ➢ Runs on-demand from Hub Project Files UI
  ➢ iRODS PHP APIs used to execute iRODS rule
  ➢ Metadata to be updated provided as key-value pair array input
  ➢ Supports arbitrary additional non-DCMI key-value pairs

• Index update
  ➢ Solr index updated with changes to DCMI fields only
GEOSPATIAL PREVIEW

• Implemented as iRODS microservice
  ➢ Runs on-demand from Hub Project Files UI
  ➢ Enabled for supported file extensions

• Preview Implementation
  ➢ Files registered as GeoServer layers after appropriate processing
  ➢ GDAL APIs used for reprojection, format conversion and subdataset extraction
  ➢ Layer name, projection information returned as rule output
  ➢ OpenLayers Javascript library used for map display
GOING FORWARD

• iRODS Federation to link distinct hubs for data and tool sharing
  ➢ Potentially enable tool workflows across hubs
• Integrate other storage mechanisms into hub projects
  ➢ Support offline data replication between iRODS storage and these other storage providers (Globus, Dropbox, Google Drive)
• Integrate data access protocols (OpenDAP)
  ➢ Allow data subsetting for chunked access to larger files
This work was supported by the NSF Award ACI - 1261727 CIF21 DIBBs : Integrating Geospatial Capabilities into HUBzero