Automated Metadata, Provenance Cataloging and Navigable Interfaces:

Ensuring the Usefulness of Extreme-Scale Data

David Schissel, Gheni Abla, Bobby Chanthavong,
Sean Flanagan, Xia Lee – General Atomics

Alex Romosan, Arie Shoshani - LBNL

Martin Greenwald, Josh Stillerman, John Wright – MIT

Next-Generation Networks for
Science Program PI Meeting
March 18-20, 2013
Berkeley, CA

Goal: Support Data Tracking, Cataloging and Integration Across a Large Scientific Domain

- **Create a data model, infrastructure, and set of tools**
  - Automatically document workflow and data provenance from user scripts or any tools that process data
- **For each data element: who, what, when, how, why**
  - Connections & dependencies between data elements
  - Human or automated annotation
- **Realistic applications starting with Fusion research**
  - What scientists do today (Python scripting & MDSplus)
  - Vision: an API that can be applied to any tools used to process or manipulate data (experiments & HPC)

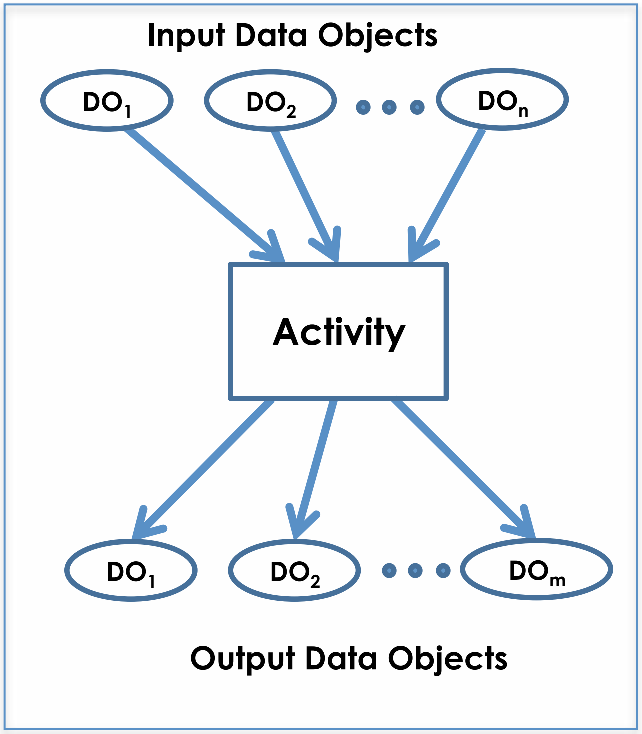
Approach: Focused Research to Build Tools for Real-World Science

- **Integrated metadata, provenance & ontology research**
  - General data model and conceptual framework
- **Research on User Interfaces: Graphical Navigation**
  - Efficiently browse and search for discovery of workflows, their components, and associated metadata
- **Demonstrate on real-world fusion applications**
  - Early deployment & agile development approach
  - Feedback and improve the design
- **Extend to other sciences to validate our generality**
  - Climate modeling and space sciences

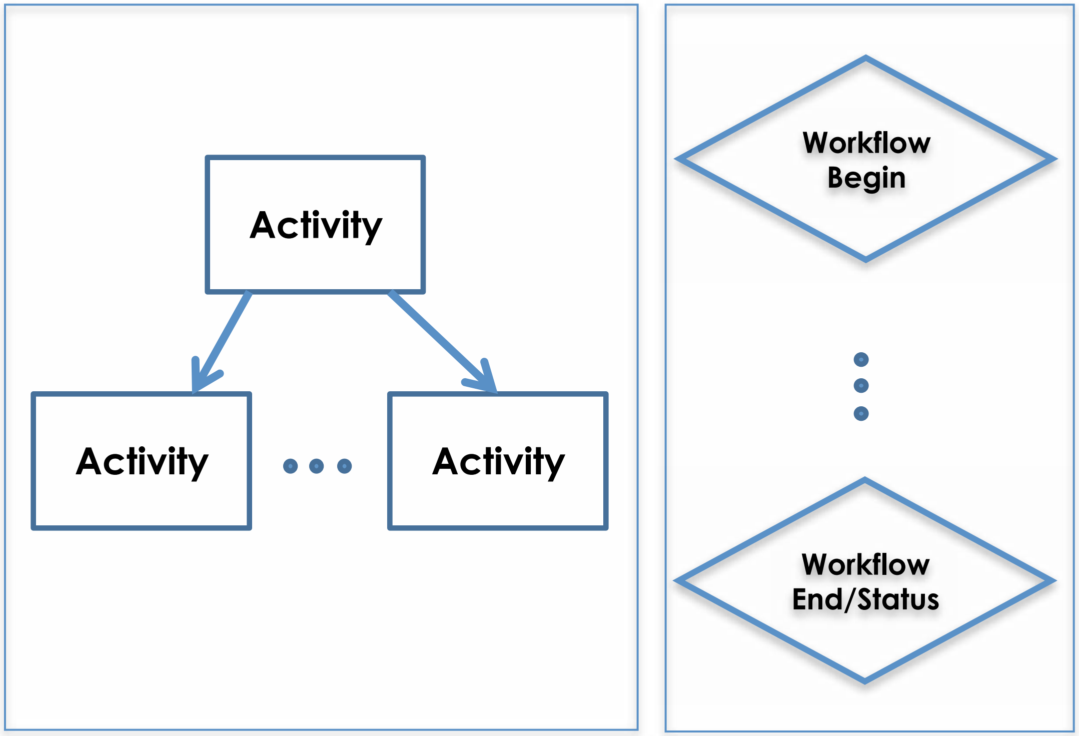
Relationship between Workflow, Provenance, Metadata, and Ontology

- **Workflow: specification of actions as DAG structure**
  - Directed Acyclic Graph: Logic of tasks performed
- **Provenance: automatically generated by the workflow**
  - Input/output for each step & relationship between steps
- **Metadata: information about each process step**
  - Process step can be a code & include documentation
- **Ontology: a structure that captures the common terms used to describe object properties in a specific domain**
  - Necessary for information search such as provenance

Workflow Primitives



Basic workflow structure



Special workflow structures

Project Divided into 4 Distinct Elements

- **Primitives and languages for annotation**
  - Useful/realistic for workflow steps data & metadata entry

- **Integrating, provenance and workflow documentation**
  - Investigate best approaches and technologies

- **User interfaces including graphical navigation**
  - Display, navigate, interact, browse the metadata catalog
  - Interactively explore data relationships
  - Graphical display to explore workflow and provenance

- **Software Suite MPO: Continual deployment/testing**
  - Starting with EFIT and Gyro from fusion science

A RESTful API Provides a Robust Interface

- **REST : Representational State Transfer**

  Provides database operations through http verbs

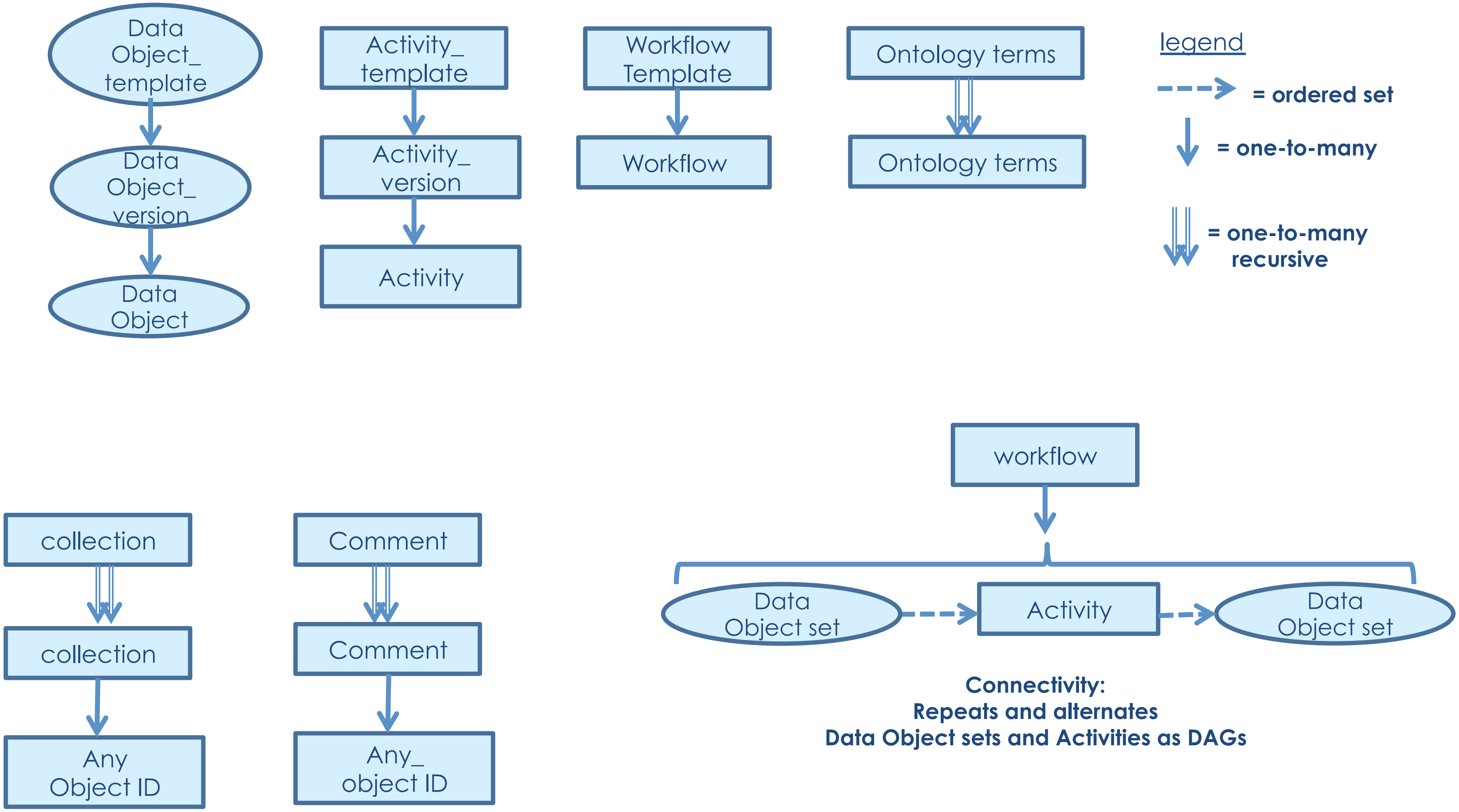
  - Create=PUT with a new URI
    POST to a base URI returning a newly created URI
  - Read = GET
  - Update = PUT with an existing URI
  - Delete = DELETE

- **Leverages existing web infrastructure**
  - URIs are nouns (http:://host/workflow, http:://host/comment) defining resources to be created or accessed
  - Data server is accessed with standard http queries supported in nearly all languages
  - Simple implementation and use (but design is hard)

Clients Manipulate Resources through the RESTful API

- **POST /resource, GET resource/:uid**
  - /workflow
  - /dataobject
  - /activity
  - /comment
  - /metadata
  - /ontology
- **Support for facets of resources and queries**
  - GET /workflow/:uid/graph
  - GET /workflow/:uid/alias
  - GET /activity?name=EFIT&user=schissel

Abstract Schema Design



Connectivity:
Repeats and alternates
Data Object sets and Activities as DAGs

UI Vision: Integrated Interface for Accessing all Types of Data in a Scientific Environment

- **One intuitive interface to accelerate scientific discovery**
  - Data, data analysis methods, interactive vis, collaboration
  - Hypertext based and graphical
- **Context enable navigation**
  - Search, navigate, interactive access to MPO data
- **Graphical navigation**
  - Flow chart, flow map, Timeline, Radial Tree map, news-map, tag-cloud maps
- **Dynamic visualizations created from MPO data**
  - Real-time feedback

Continual Deployment/Testing Critical to Project's Success

- **Early deployment of software for user engagement**
  - Provide useful feedback & shorten development lifecycle
- **Working prototypes (database/interfaces) to users early**
  - Evaluate, revise, & release based on their experience
- **Near-term: two fusion codes**
  - EFIT (plasma shape) during operations via MDSpus
  - Gyro (large sim code) with results in large file repository
- **Longer-term: Additional fusion applications and other sciences**

Current RESTful API Supports Workflow Instrumentation

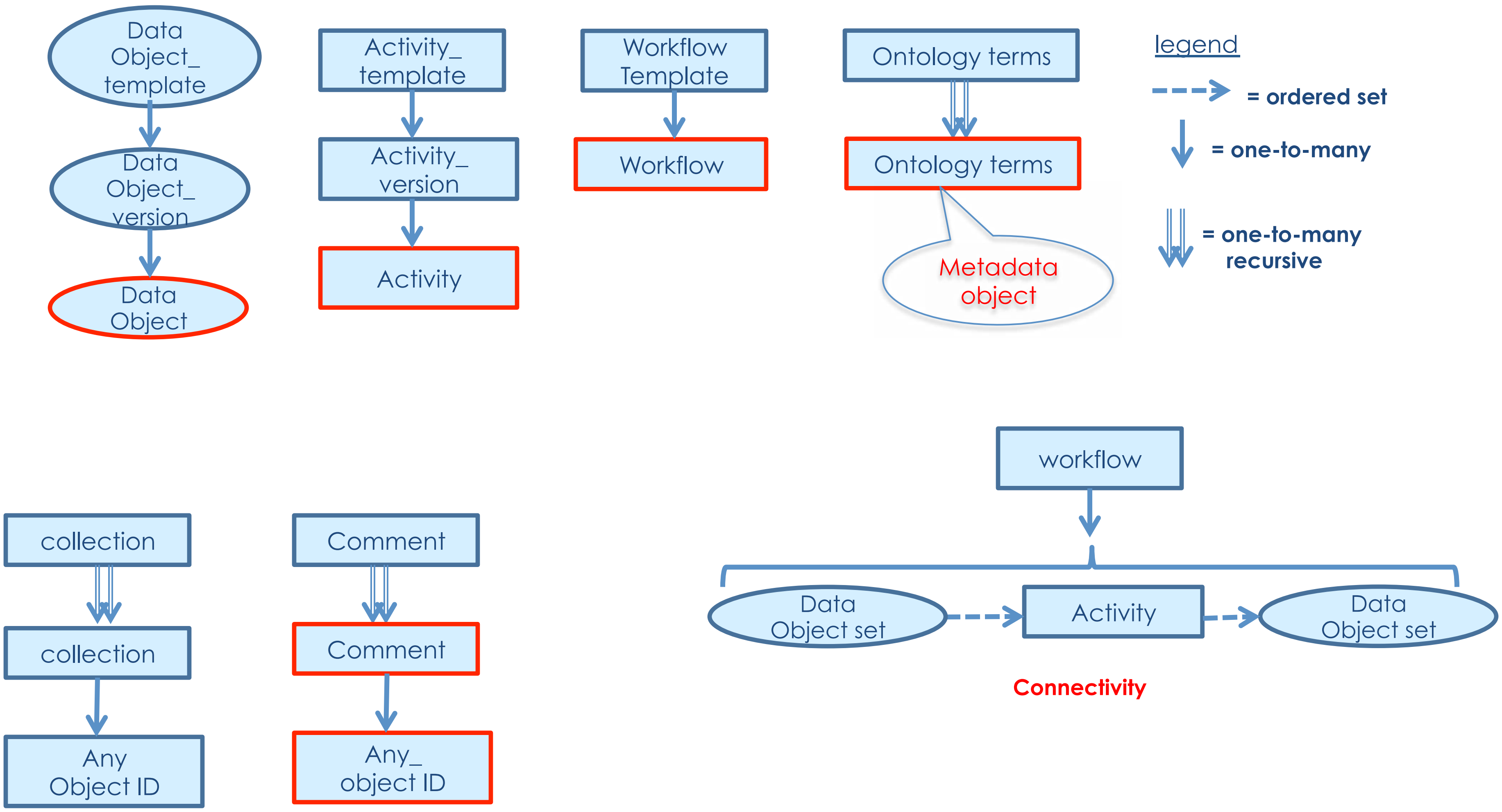
- **Routes for workflow creation and annotation**
  - /workflow, /activity, /dataobject, /comment, /metadata
  - Each route supports POST for object creation and GET:uid for object retrieval
  - Objects are encoded in JSON for POSTing and GETting
    - POST /workflow
      BODY: { "name":”GYRO”,
      "description":”Important ITER run”}
    - GET /metadata?work_uid=f20b23ec-aefb-481c-8c08-6443f
      Returns: {"target_uid":” f20b23ec-aefb-481c-8c08-6443f”,
      "key": “Te(kev)”,
      "value": 3,
      "uid": “e1b13f63-97ca-490d-9218-15c8f5cae1d5”,
      "time": 2013-03-14 19:44:34.235565,
      "uri": http://mpohost/metadata/e1b13f63-97ca-490d-9218-15c8f5cae1d5)}

Command Line Client For Use in Scripts

- **Client uses ‘meta’ command and method names**
- **Shell scripts and batch scripts can be instrumented**
- **User can make queries & comments via command line**
- **Example script or command line session:**

```
wid = mpo init --name=EFIT --desc=test`
oid = mpo add  $wid --parent=$wid --name=shot --desc="Plasma shot
    number" --uri=150335`
oid2 = mpo add  $wid --parent=$wid --name="Snap file" --desc="EFIT input
    file" --uri="\\efit01:namelist"`
aid = mpo step $wid --input=$oid --input=$oid2 --name="EFIT exec"
    --desc="Fit equilibrium and compute plasma parameters" --uri=EFIT`
cid = mpo comment $aid "This program is the only one in this workflow"
```

Initial Schema Implementation



Data Object_ template

Data Object_ version

Data Object

Activity_ template

Activity_ version

Activity

Workflow Template

Workflow

Ontology terms

Ontology terms

Metadata object

legend

= ordered set

= one-to-many

= one-to-many recursive

collection

collection

Any Object ID

Comment

Comment

Any_ object ID

workflow

Data Object set

Activity

Data Object set

Connectivity

Preliminary Database Schema as Operating Today: Implemented in PostGreSQL but any DB will be Sufficient

Workflow: W_GUID, name, WS_GUID, description, U_GUID (owner), start_time, end_time, completion_status, status_explanation

Data_object: DO_GUID, name, DOV_GUID, W_GUID, description, URI_of_data

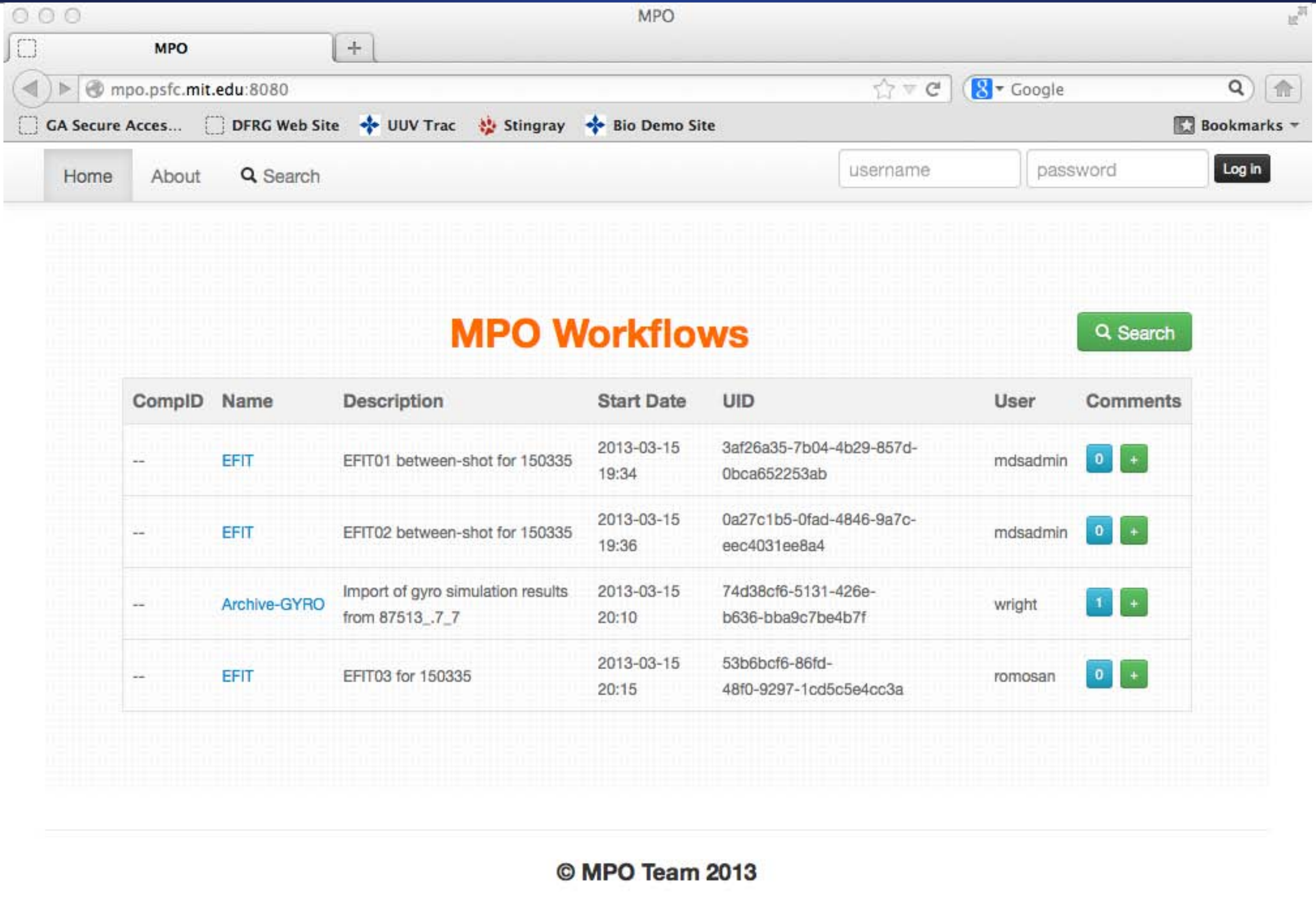
Activity: A_GUID, name, AV_GUID, W_GUID, description, URI_of_executable_file, start_time, end_time, completion_status, status_explanation

Workflow_connectivity: WC_GUID, W_GUID, child_GUID (DO_GUID or A_GUID), child_type, parent_GUID (DO_GUID or A_GUID), parent_type

Comment: CM_GUID, name, text, URI_of_comment, comment_type, parent_GUID (any object), parent_type, time_entered

Metadata: M_GUID, key, value, metadata_type, parent_GUID (any object), parent_type, time_entered

Prototype MPO Web Site Operating

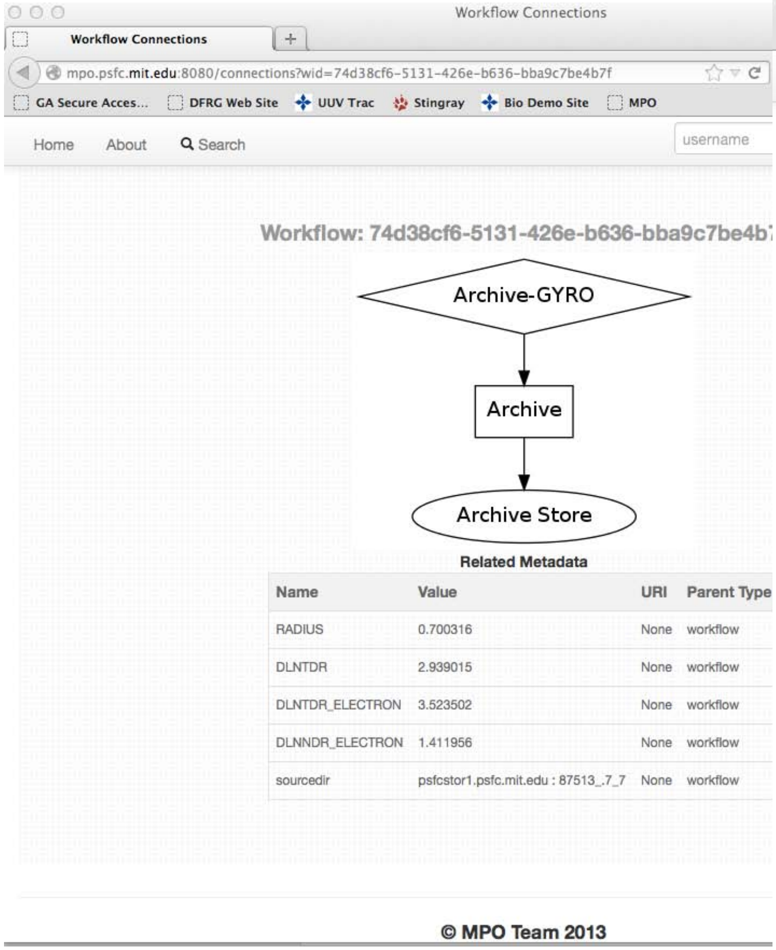


MPO Workflows

CompID	Name	Description	Start Date	UID	User	Comments
--	EFIT	EFIT01 between-shot for 150335	2013-03-15 19:34	3af26a35-7b04-4b29-857d-0bca652253ab	mdsadmin	0 +
--	EFIT	EFIT02 between-shot for 150335	2013-03-15 19:36	0a27c1b5-0fad-4846-9a7c-eec4031ee8a4	mdsadmin	0 +
--	Archive-GYRO	Import of gyro simulation results from 87513_.7_7	2013-03-15 20:10	74d38cf6-5131-426e-b636-bba9c7be4b7f	wright	1 +
--	EFIT	EFIT03 for 150335	2013-03-15 20:15	53b6bcf6-86fd-48f0-9297-1cd5c5e4cc3a	romosan	0 +

© MPO Team 2013

Workflow Graphics Automatically Generated from MPO Data

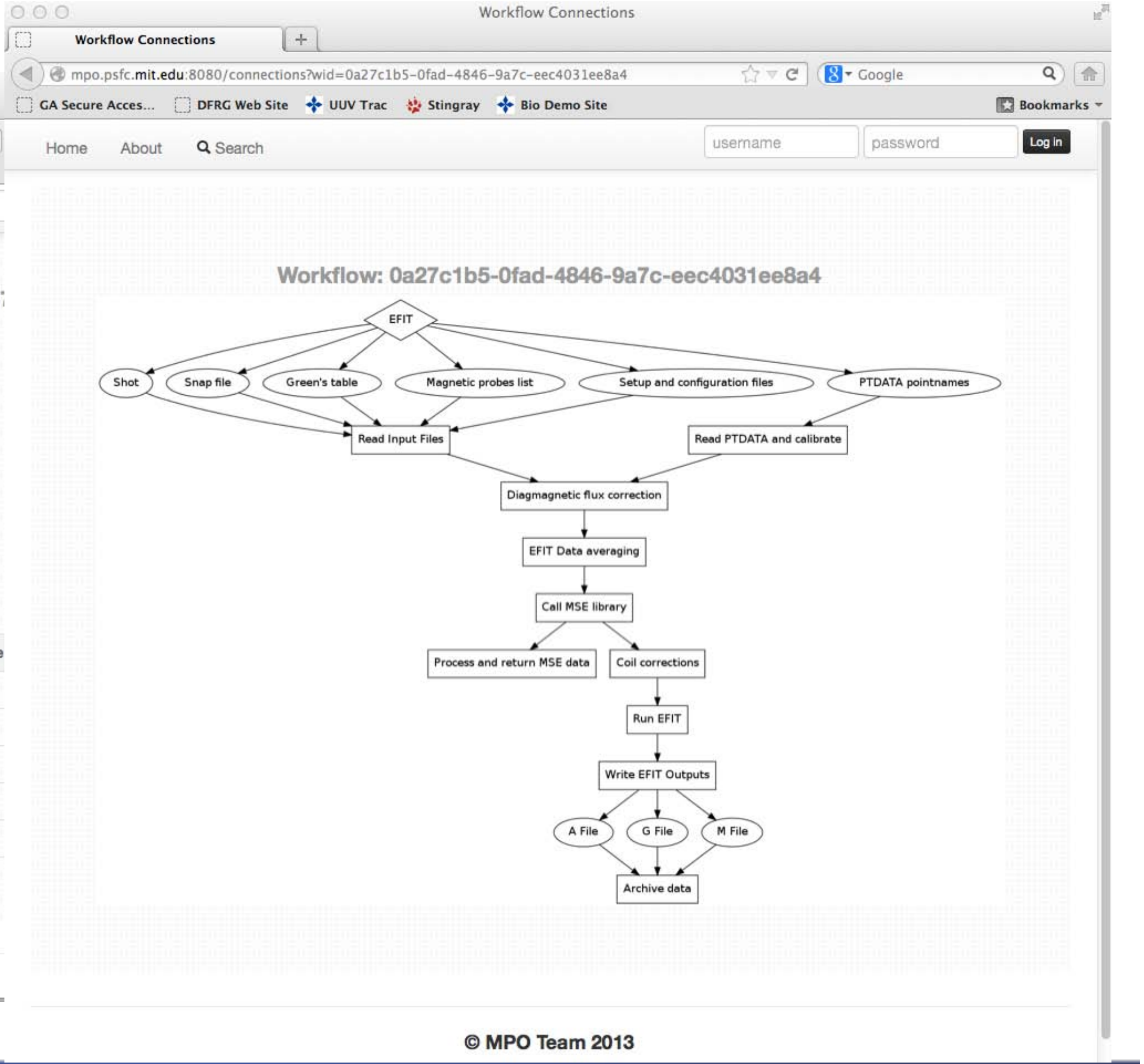


Workflow: 74d38cf6-5131-426e-b636-bba9c7be4b7

Related Metadata

Name	Value	URI	Parent Type
RADIUS	0.700316	None	workflow
DLNTDR	2.939015	None	workflow
DLNTDR_ELECTRON	3.523502	None	workflow
DLNNDR_ELECTRON	1.411956	None	workflow
sourcedir	psfcstor1.psfc.mit.edu : 87513_.7_7	None	workflow

© MPO Team 2013



Workflow: 0a27c1b5-0fad-4846-9a7c-eec4031ee8a4

© MPO Team 2013

Goal in the Next Year is to Expand System's Depth as well as Expand the Reach of our Tools

- **API**
  - Complex queries, Ontology support, fine-grain ACLs
- **Database System**
  - Add support for hierarchical ontologies using controlled vocabularies with broader and narrower terms
  - Support template structure for workflow, activities, and data-objects
- **UI workflow graphic extended to be interactive and graphical views of many workflows**
- **Extend to at least several new sciences**
- **Push deeper into fusion science**
  - Instrument data input preparation phase

Summary

- **Instrumenting existing workflows allowing automation**
  - General API and framework for general solution
- **Rapid prototyping with real-world fusion problems**
  - Quicker feedback and rich datasets to draw upon
- **General solution that will extend to other sciences**
  - Narrow early focus but with a broad long-term vision
- **Validate our approach seek other sciences for testing**
  - Are there other projects who might desire to test?