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/DIII-D

Alex Romosan, Arie Shoshani - LBNL

Martin Greenwald, Josh Stillerman, John Wright – MIT/C-MOD

NGNS PI Meeting
September 17, 2014
Rockville, MD

*Supported by US Department of Energy
Insuring the Usefulness of Extreme-Scale Data: System for Fusion Science Operational

Objectives:
- Create a data model, infrastructure, and a set of tools to automatically document workflow & data provenance from any tool that processes data
- For each data element: who, what, when, how, why including tracking connections & dependencies between data elements
- Deploy interactive tools for efficient browsing and searching of workflows & associated data
- Production deployment on real-world fusion data
- Extend solution to other sciences within SC portfolio

Approach:
- Integrated metadata, provenance & ontology system
- Primitive and languages for annotation
- Database schema design for general solution
- Research on user interfaces: graphical navigation
- Demonstrate on real-world fusion applications with early deployment & agile development
- Extend to other sciences to validate generality

Progress:
- Data being loaded, system being refined based on feedback

Impact:
- Increase the value of experimental & computational data across the diverse domains of the DOE/SC
- Paper presented in China at an IAEA meeting solicited substantial interest from attending scientists
- Operational system deployed for alpha/beta users that supports data provenance, metadata, and ontology with interactive tools for browsing and searching

3-year project started October 2012
Contact: D. Schissel (schissel@fusion.gat.com)
Tracking of the Data Lineage has not kept Pace with the Explosive Growth in the Amount of Data

- **Provenance**: from the French *provenir* “to come from”
  - Where did a piece of data come from & where was it used

- **Questions that data provenance can answer**
  - *Diagnostic X* calibration changed, what about my results?
  - What data/publications does this code bug effect?

- **Associated questions that can be answered**
  - Who does “*Analysis Y*” so I can ask for advice?
  - Who else is analyzing this shot in detail?

- **DOE Digital Data Management Statement (July 2014)**
  - All stages of the digital data lifecycle: Capture, analysis, sharing, and preservation
  - Data Management Plans now required
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 (e.g. shell, Python, IDL, MATLAB)
  – Vision: an API that can be applied to any tools used to process or manipulate data (experiments & HPC)
  – Not tied to a specific workflow engine
Approach: Focused Research to Build Tools for Real-World Science

• Integrated metadata, provenance & ontology research
  – General data model and conceptual framework
  – Directed Acyclic Graph: Logic of tasks performed

• 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
The MPO System has 5 Basic Elements

- **Data Objects**
  - Structured data inside/outside the MPO with pointers

- **Activities**
  - Create, move, or transmute data from one form to another

- **Connections**
  - Data objects & activities linked to represent a workflow

- **Comments**
  - Unstructured text with other attributes (e.g. who)

- **Collections**
  - List of related objects, activities, or workflows
Project Divided into 4 Distinct Elements

- **API: primitives and languages for annotation**
  - Useful/realistic for workflow steps data & metadata entry

- **Database: metadata, provenance, ontology, workflow documentation**
  - Capture all elements needed when a workflow executes

- **UI: graphical navigation including real-time updates**
  - Display, navigate, interact, browse the metadata catalog
  - Graphical display to explore workflow and provenance

- **Users: Continual deployment/testing**
  - Starting with EFIT, SWIM, and GYRO from fusion science
Functional MPO Infrastructure is based on Model-View-Controller (MVC) Concept

- Separates data representation & user’s interaction with it

Model = Database
Views = UI & JSON from RESTful API
Controller = Logic on client response or DB interaction
Present Schema Implementation

```
Data Object
  ↓
Data Object version
  ↓
Data Object

Activity template
  ↓
Activity version
  ↓
Activity

Workflow Template
  ↓
Workflow

Ontology terms

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
```
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
  - Search & navigation directed by domain specific ontology

- **Graphical navigation**
  - Flow chart, flow map, Timeline, Radial Tree map, news-map, tag-cloud maps

- **Dynamic visualizations created from MPO data**
  - Real-time feedback
MPO Software Stack Combines Open Source Solutions

- Both API server and Web UI server uses Flask - a lightweight web application framework
  - Core components simple but extensible
  - Supports templates or HTML placeholders
  - Clean separation of components

- Twitter bootstrap to create standardized Web front-end
  - Hides Javascript complexity for easier development
  - Built-in responsive web page creation capability

- DAGs created by Graphviz software package
  - Dynamically created and embedded in HTML webpage

- MDSplus event services to create simple event server
  - Provides real-time update capabilities
MPO Web Site Operating with Ontology-based Search, Automatic Real-Time Graphics, Live Data Loading
Project’s Final Year Goal is to Expand System’s Depth and Expand the Reach of our Tools into other Sciences

- **Alpha Users evaluating, beta users by end of CY14**
  - Presentation at APS/DPP Nov. 2014 (attracting beta users)

- **1st Quarter CY15, push to a different science domain**
  - Which domain depends on who can give us the time

- **Hardening for Production**
  - Formalize schema updates, separate development/production/user sandbox, develop/guarantee our persistent store

- **Continue to evolve MPO UI and data schema**
  - For example: UI evolving to handle large quantity of workflows, adding collections
Summary

- **Substantial progress since the last PI meeting**
  - API, data store/Ontology, & UI all evolved

- **Production workflows have been MPO instrumented**
  - DIII-D experimental analysis & SWIM simulations

- **Our results validate our approach**
  - Simple API to instrument basically any existing workflows
  - General data store and UI to store and navigate

- **Include a new science domain moving forward**
  - Yield feedback to allow iteration on the MPO framework
From Rich: “What Question Does Your Research Motivate You To Now Ask?”

- **How to expand the reach of our MPO framework?**
  - Across many science domains (ease of adoption, robust)
  - Federated system within a science (fast at large scales)

- **Compatibility with W3C Standards (e.g. PROV)**
  - How to import/export to MPO?
  - Can draw in this ecosystem (e.g. Annotation WG)?

- **Efficient UI operation at large-scale**
  - How to do better/faster Graphical Navigation?

- **Provide rich data centric tools**
  - Are there different UIs to the MPO data?