The MPO System for Automatic Workflow Documentation

Gheni Abla
for The MPO Team

10th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research

20-24th April 2015
Ahmedabad, Gujarat, India

DIII-D NATIONAL FUSION FACILITY SAN DIEGO

BERKELEY LAB

PSFC

GENERAL ATOMICS

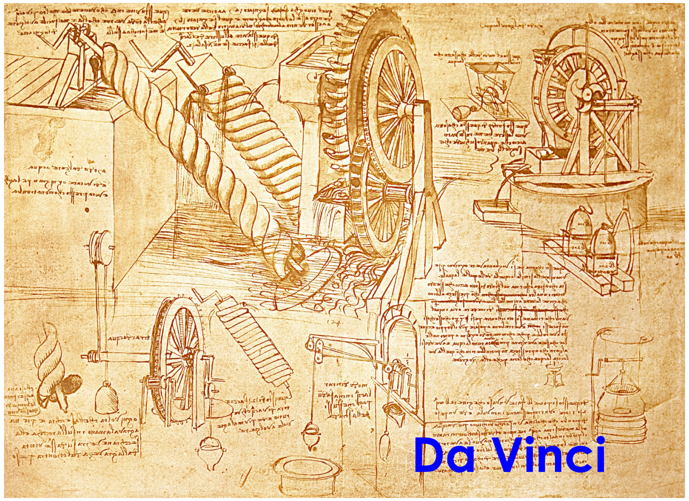
This Presentation Builds on Previous Presentations

- M. Greenwald, et al., “A Metadata Catalogue for Organization and Systemization of Fusion Simulation Data”, 8th IAEA-TM, San Francisco, CA, June 2011
- D.P. Schissel, et al., “Automated Metadata, Provenance Cataloging, Navigable Interfaces: Ensuring the Usefulness of Extreme Scale Data”, 9th IAEA-TM, Hefei, China, May 2013
- J.C. Wright, et al., “The MPO API: A Tool for Recording Scientific Workflows”, 9th IAEA-TM, Hefei, China, May 2013

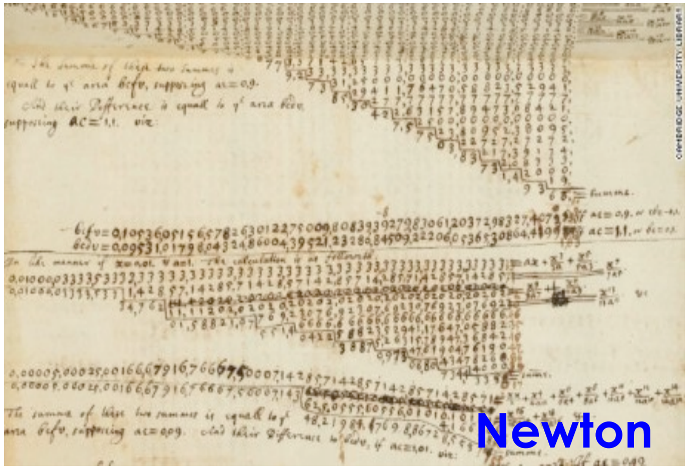
Documenting Data and Processes is Important Aspect of Scientific Research Activities

- **Data from research activities is expensive to produce and may be critical for follow-on research**
- **It is not the mere existence of data that is important, but our ability to make use of it**
- **The context and metadata makes the data more usable**
  - Hypotheses
  - Pre-process activities
  - Experiments
  - Computational process
  - Reflections
- **Documenting the process is not an easy task**

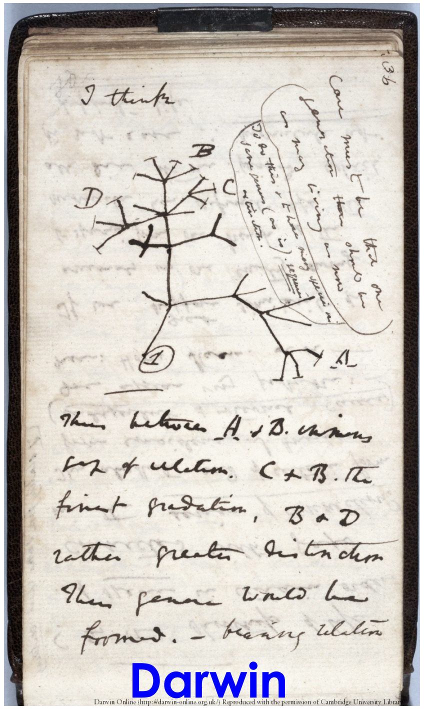
Throughout History, Scientists Generated Handwritten Logbooks to Keep Track of Data



Da Vinci



Newton



136

I think

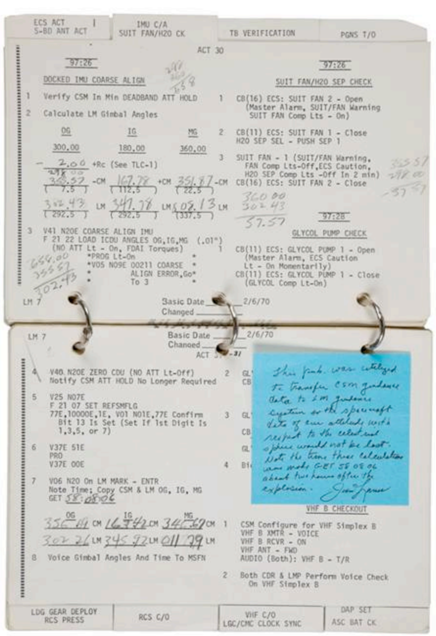
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.

Thus between A & B immense gap of relation. C & B the finest gradation, B & D rather greater distinction. Thus genera would be formed. — bearing relation

Darwin Online (http://darwin-online.org.uk/) Reproduced with the permission of Cambridge University Library

Darwin

Lovell



ECS ACT
S-BD ANT ACT | IMU C/A
SUIT FAN/H2O CK | TB VERIFICATION | PGNS T/O

ACT 30

97:26

DOCKED IMU COARSE ALIGN

1 Verify CSM In Min DEADBAND ATT HOLD

2 Calculate LM Gimbal Angles

OG	IG	MG
300.00	180.00	360.00
− 2.06 +Rc (See TLC-1)		
355.57 -CM (7.5)	167.78 +CM (112.5)	351.87 -CM (22.5)
302.43 LM (292.5)	347.78 LM (292.5)	008.13 LM (337.5)

3 V41 N20E COARSE ALIGN IMU
F 21 22 LOAD ICDU ANGLES OG,IG,MG (.01°)
(NO ATT Lt - On, FDAI Torques)
*PROG Lt-On *
*V05 N09E 00211 COARSE *
* ALIGN ERROR,Go*
* To 3 *

658.00
355.57
302.43

97:26

SUIT FAN/H2O SEP CHECK

1 CB(16) ECS: SUIT FAN 2 - Open
(Master Alarm, SUIT/FAN Warning
SUIT FAN Comp Lts - On)

2 CB(11) ECS: SUIT FAN 1 - Close
H2O SEP SEL - PUSH SEP 1

3 SUIT FAN - 1 (SUIT/FAN Warning,
FAN Comp Lts-Off,ECS Caution,
H2O SEP Comp Lts -Off In 2 min)
CB(16) ECS: SUIT FAN 2 - Close

360.00
302.43
57.57

97:28

GLYCOL PUMP CHECK

1 CB(11) ECS: GLYCOL PUMP 1 - Open
(Master Alarm, ECS Caution
Lt - On Momentarily)
CB(11) ECS: GLYCOL PUMP 1 - Close
(GLYCOL Comp Lt-On)

LM 7 Basic Date 2/6/70
Changed

LM 7 Basic Date 2/6/70
Changed
ACT 31

4 V40 N20E ZERO CDU (NO ATT Lt-Off)
Notify CSM ATT HOLD No Longer Required

5 V25 N07E
F 21 07 SET REFSMFLG
77E,10000E,1E, V01 N01E,77E Confirm
Bit 13 Is Set (Set If 1st Digit Is
1,3,5, or 7)

6 V37E 51E
PRO
V37E 00E

7 V06 N20 On LM MARK - ENTR
Note Time: Copy CSM & LM OG, IG, MG
GET 58:08:06

OG	IG	MG
356.84 CM	167.42 CM	340.69 CM
302.26 LM	345.92 LM	011.79 LM

8 Voice Gimbal Angles And Time To MSFN

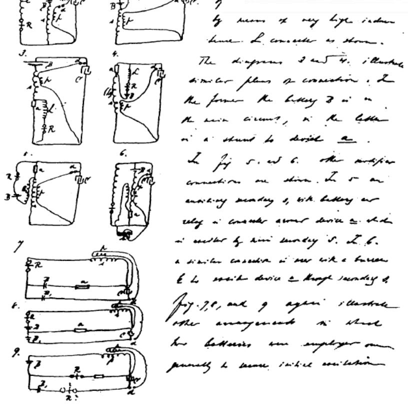
This page was utilized to transfer CSM guidance data to LM guidance system so the spacecraft data of our attitude with respect to the celestial sphere would not be lost. Note the time these calculations were made GET 58 08 06 about two hours after the explosion. Jim Lovell

VHF B CHECKOUT

1 CSM Configure for VHF Simplex B
VHF B XMTR - VOICE
VHF B RCVR - ON
VHF ANT - FWD
AUDIO (Both): VHF B - T/R

2 Both CDR & LMP Perform Voice Check
On VHF Simplex B

LDG GEAR DEPLOY
RCS PRESS | RCS C/O | VHF C/O
LGC/CMC CLOCK SYNC | DAP SET
ASC BAT CK



Tesla

In the Modern Era Documenting Research Process Made Progress and Met New Challenges

- **Personal computers and mobile devices helped electronic logbooks replace handwritten ones – brought conveniences**
  - Multi-media and hypertext support
  - Store, share and search

- **However, the content creation and log entry remained as a manual activity in the electronic logbooks**

- **As the pace of scientific research accelerated, documenting the process & data became more challenging & time consuming**
  - Increased precision of scientific instruments
  - Rise of exascale computing and arrival of Big Data
  - Result: fragmentation of data, processing, and documentation

Metadata, Provenance and Ontology (MPO) System is for Documenting Scientific Data & Workflow

- **Provenance: Preserve meaning of data by documenting all of the steps taken to produce the data**
  - Automate metadata generation as much as possible
  - Support more systematic management of data used and resulted by analysis and simulation

- **Provide and preserve answers to two key questions:**
  - ***Where did a particular piece of data come from?***
    - Assumptions, inputs, parameters used for calculation
    - The origin of inputs; reasons for assumptions & parameters
  - ***Where was this data used?***
    - Other calculations
    - Publication and presentation
    - Contributions to databases

Potential Use Cases

- **How did I get the data plotted in Fig.6 of my 2014 Phys. Plasmas article?**
- **A calibration error was found in Thomson Scattering data taken during 2011 - the data has now been recalculated.**
  - Where was the old data used?
  - What publications used that data? Were they critical for the published conclusions?
  - Did we contribute that data to a database shared by others?
- **A recently graduated PhD student left behind output from thousands of gyrokinetic simulations**
  - Which of these were used in her thesis?
  - Which might be useful in the future?
  - What were the inputs and parameters used in the interesting runs?

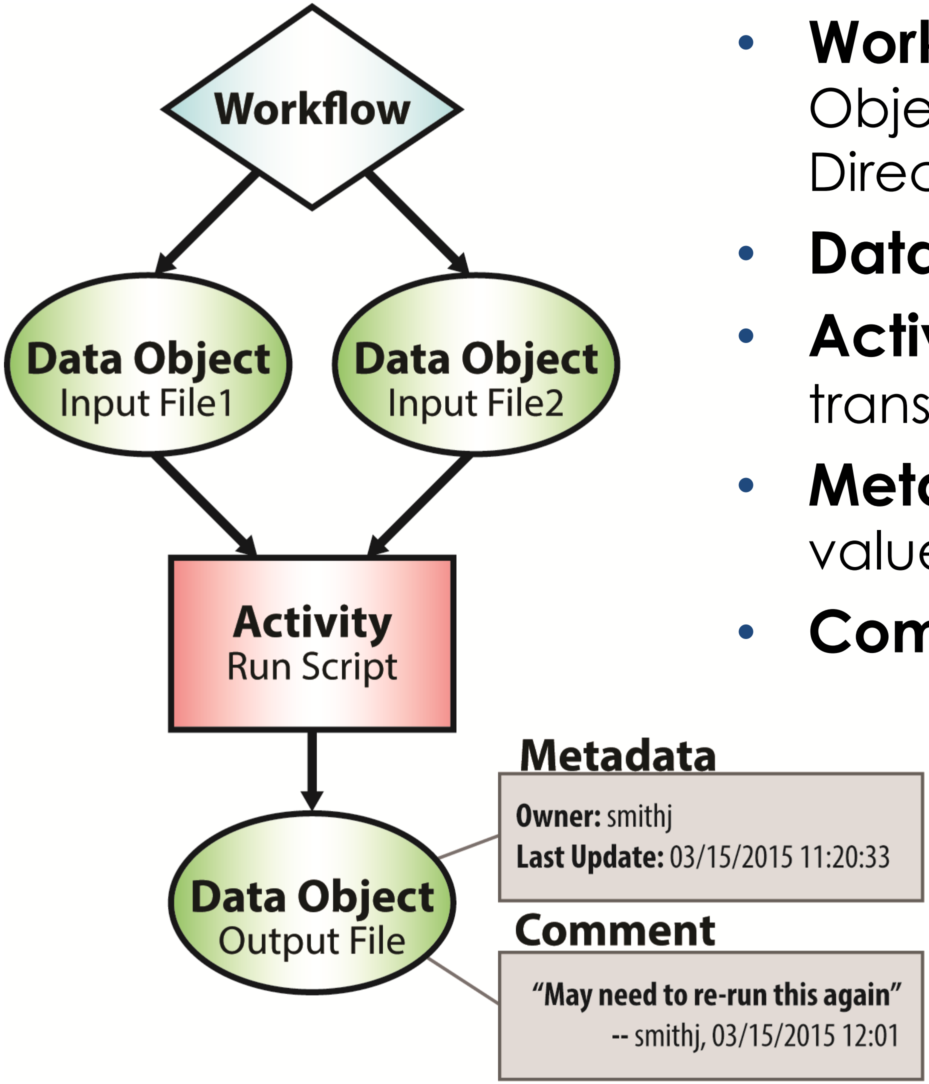
Capabilities of the MPO System

- **Support all types of the scientific workflows – both experimental and computational**

- **Function in a heterogeneous environment and interoperate with workflow tools people are already using**
  - Many different computing platforms – laptop to supercomputer
  - Researchers use many different languages (Shell scripts, Python, IDL, Matlab, etc.) and tools to get their work done
  - Data is stored in different formats (MDSplus, HDF5, NetCDF, ASCII)

- **Once set up, work as automatically as possible**
  - Best suited for scripted rather than one-time use

MPO System Entities and Data Model



Workflow

Data Object
Input File1

Data Object
Input File2

Activity
Run Script

Data Object
Output File

Metadata

Owner: smithj
Last Update: 03/15/2015 11:20:33

Comment

"May need to re-run this again"
-- smithj, 03/15/2015 12:01

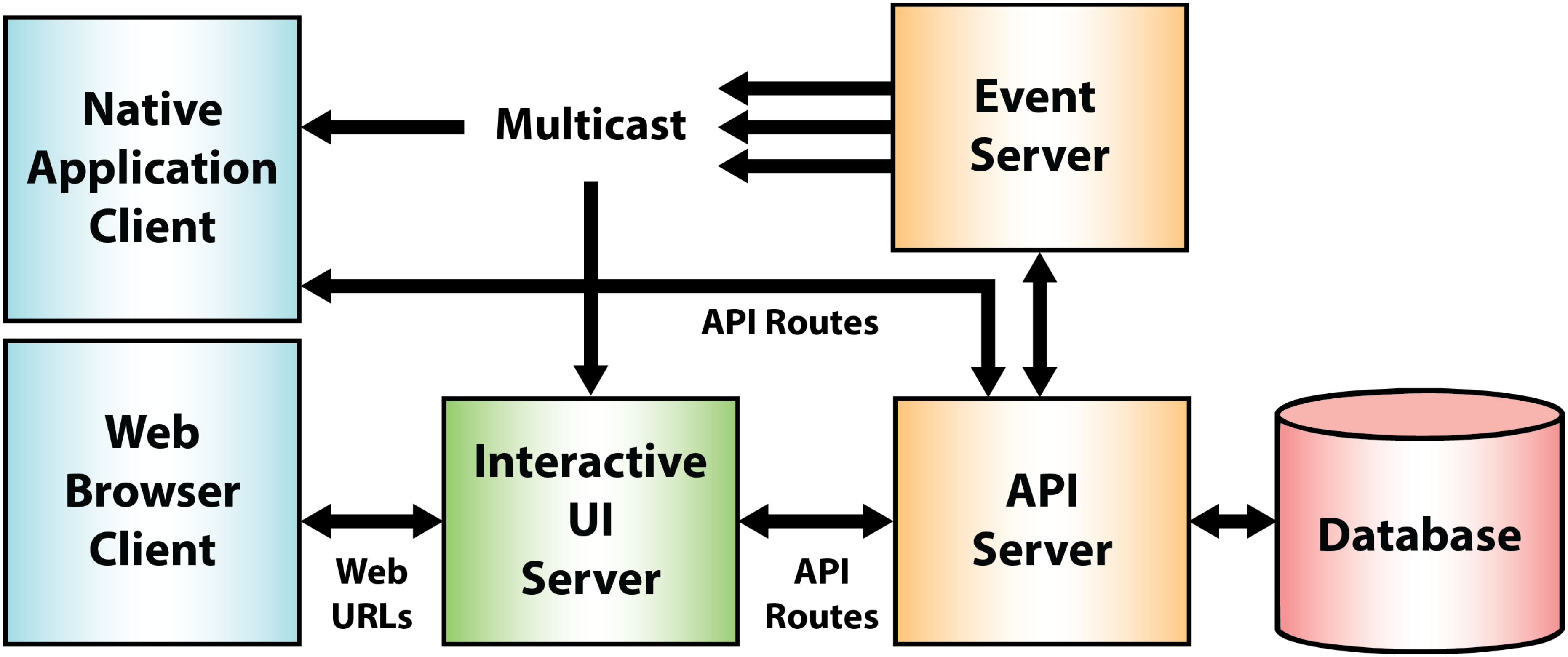
- **Workflow:** A series of connected Data Objects and Activities, organized as a Directed Acyclic Graph (DAG)
- **Data Object:** Structured data
- **Activity :** Process that creates, moves or transmutes data to a new form
- **Metadata:** Text-based, arbitrary name-value pairs
- **Comment:** User annotation

- **Connection:** The links between inputs, actions and results
- **Collection:** Group of Data Object, Activity, Workflow, and Collection

MPO Entities Are Uniquely Identified

- **Each MPO entity is given a global unique numerical identifier**
  - UID – Unique ID
  - 128 bit, pseudo random numbers
- **Data object is also identified by a URI (Uniform Resource Identifier)**
  - The URI is the pointer to the data object
  - URI includes the data protocol name and the path to the data
  - Examples: MDSplus server+node path, HDF5 file +data location
- **Workflows also can be identified by composite ID**
  - Examples: doej/EFIT/52, smitha/OMFIT/1002
- **Searching is enhanced by defining a “controlled vocabulary”**
  - User-defined, hierarchical ontology

The MPO System is Based on a Multi-Tier Software Architecture

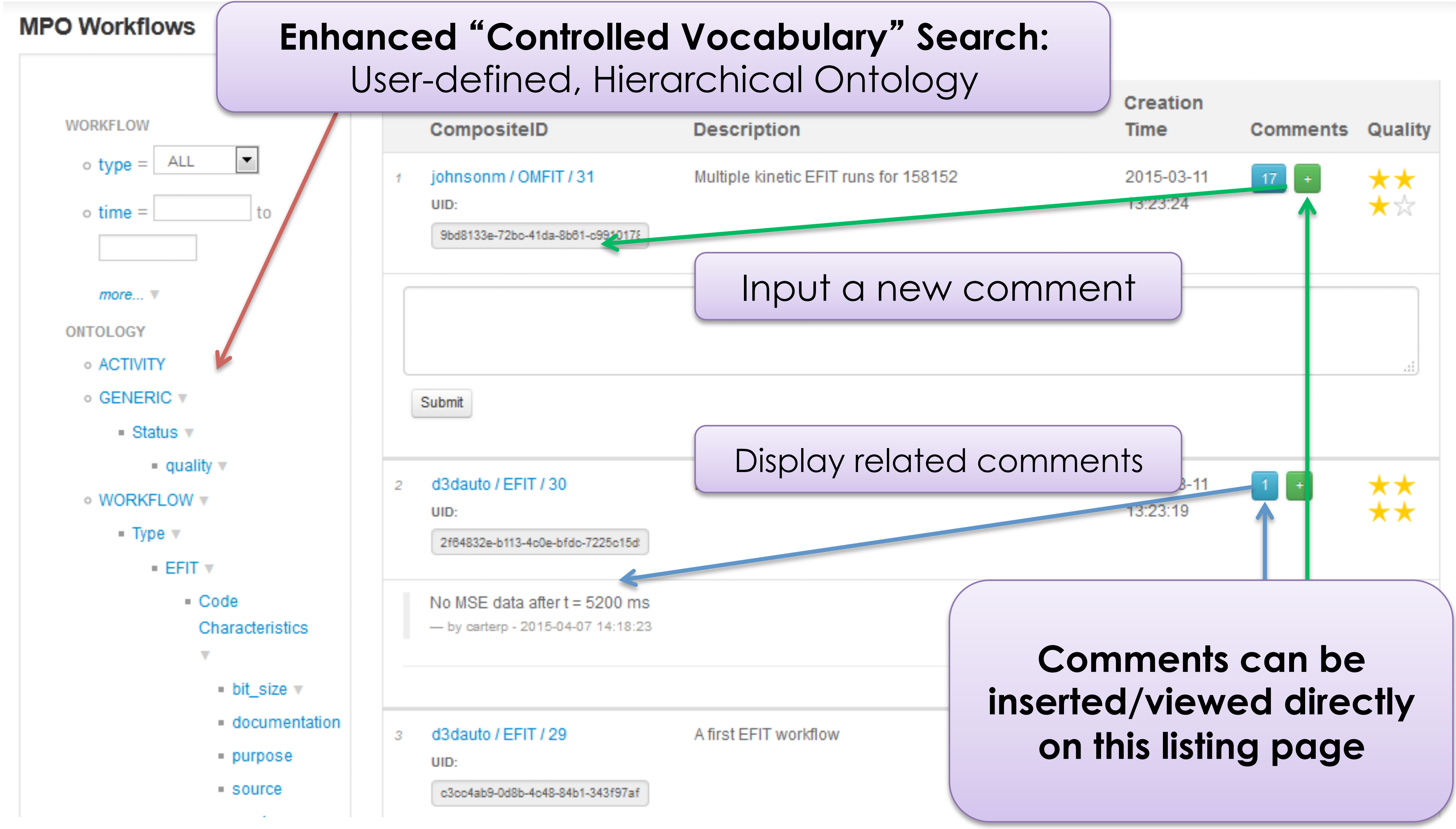


MPO System Is Based on Open-Source Software

- **MPO Technology Stack**
  - "PostgreSQL" database used for current implementation
  - Both API server and Interactive UI server use "Flask", a lightweight Python web application framework
  - Twitter "Bootstrap" to create standardized Web front-end
  - DAGs rendered by "Graphviz" software
  - Authentication via x.509 certificates (currently support OSG, MIT & MPO certs)
  - MDSplus event services
  - SQLAlchemy for Object Relational Mapping

- **API is based on RESTful abstraction**
  - Services are exposed via RESTful methods (GET, POST) and URIs

Interactive UI Page Example: Workflow List



MPO Workflows

WORKFLOW

- type = ALL
- time = to

more... ▼

ONTOLOGY

- ACTIVITY
- GENERIC ▼
  - Status ▼
    - quality ▼
- WORKFLOW ▼
  - Type ▼
    - EFIT ▼
      - Code Characteristics ▼
        - bit_size ▼
        - documentation
        - purpose
        - source

	CompositeID	Description	Creation Time	Comments	Quality
1	johnsonm / OMFIT / 31 UID: 9bd8133e-72bc-41da-8b61-c991017f	Multiple kinetic EFIT runs for 158152	2015-03-11 13:23:24	17 +	★★★☆
2	d3dauto / EFIT / 30 UID: 2f64832e-b113-4c0e-bfdc-7225c15d		3-11 13:23:19	1 +	★★★★
3	d3dauto / EFIT / 29 UID: c3cc4ab9-0d8b-4c48-84b1-343f97af	A first EFIT workflow			

Submit

No MSE data after t = 5200 ms
— by carterp - 2015-04-07 14:18:23

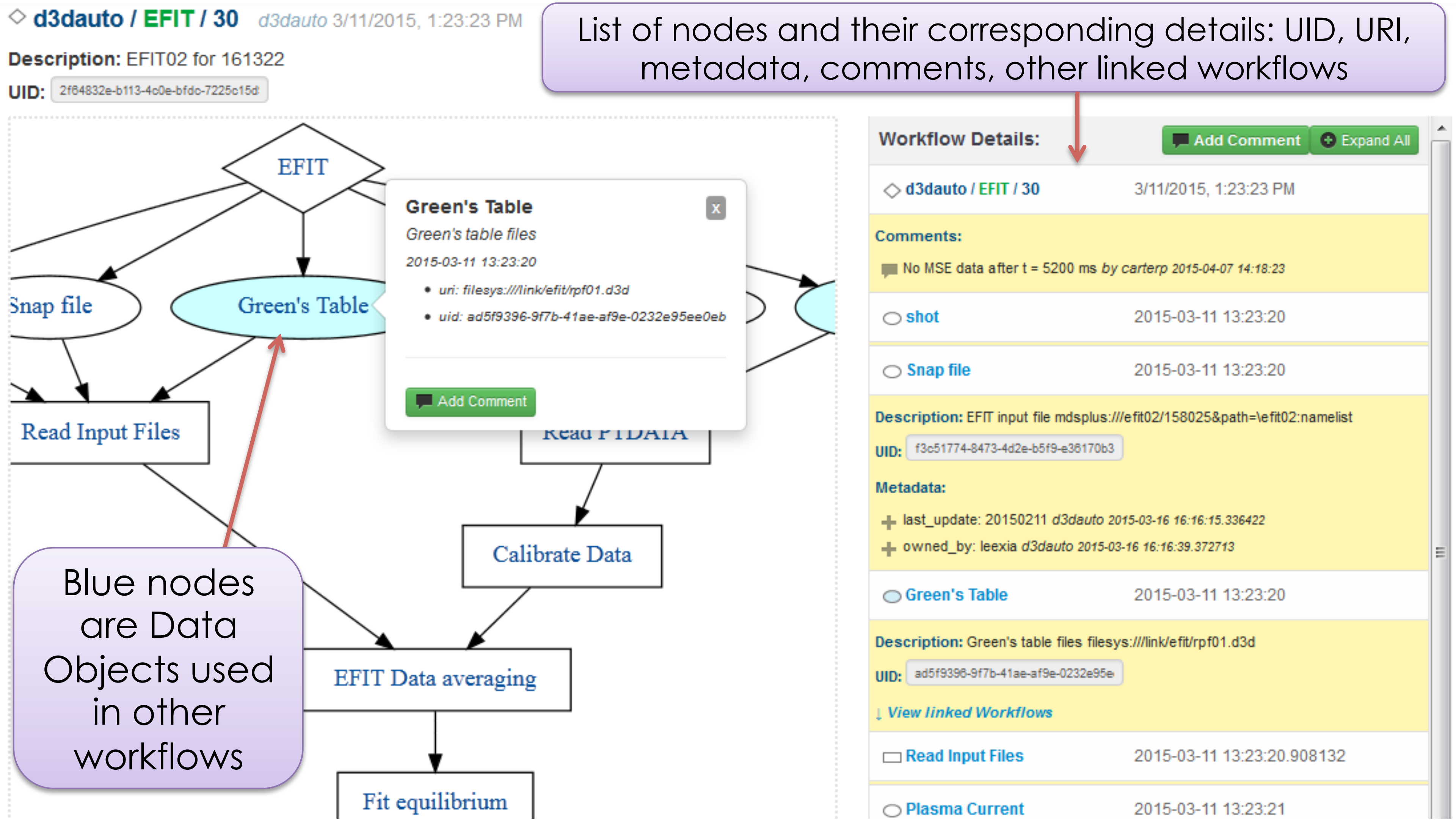
Enhanced "Controlled Vocabulary" Search:
User-defined, Hierarchical Ontology

Input a new comment

Display related comments

Comments can be inserted/viewed directly on this listing page

Interactive UI Page Example: Workflow Details



◇ d3dauto / EFIT / 30 *d3dauto* 3/11/2015, 1:23:23 PM

Description: EFIT02 for 161322

UID: 2f64832e-b113-4c0e-bfdc-7225c15d

Green's Table

Green's table files

2015-03-11 13:23:20

- *uri: filesys:///link/efit/rpf01.d3d*
- *uid: ad5f9396-9f7b-41ae-af9e-0232e95ee0eb*

Workflow Details:

◇ d3dauto / EFIT / 30 3/11/2015, 1:23:23 PM

Comments:

No MSE data after t = 5200 ms *by carterp 2015-04-07 14:18:23*

○ shot 2015-03-11 13:23:20

○ Snap file 2015-03-11 13:23:20

Description: EFIT input file mdsplus:///efit02/158025&path=\efit02:namelist

UID: f3c51774-8473-4d2e-b5f9-e36170b3

Metadata:

+ last_update: 20150211 *d3dauto 2015-03-16 16:16:15.336422*

+ owned_by: leexia *d3dauto 2015-03-16 16:16:39.372713*

○ Green's Table 2015-03-11 13:23:20

Description: Green's table files filesys:///link/efit/rpf01.d3d

UID: ad5f9396-9f7b-41ae-af9e-0232e95e

↓ *View linked Workflows*

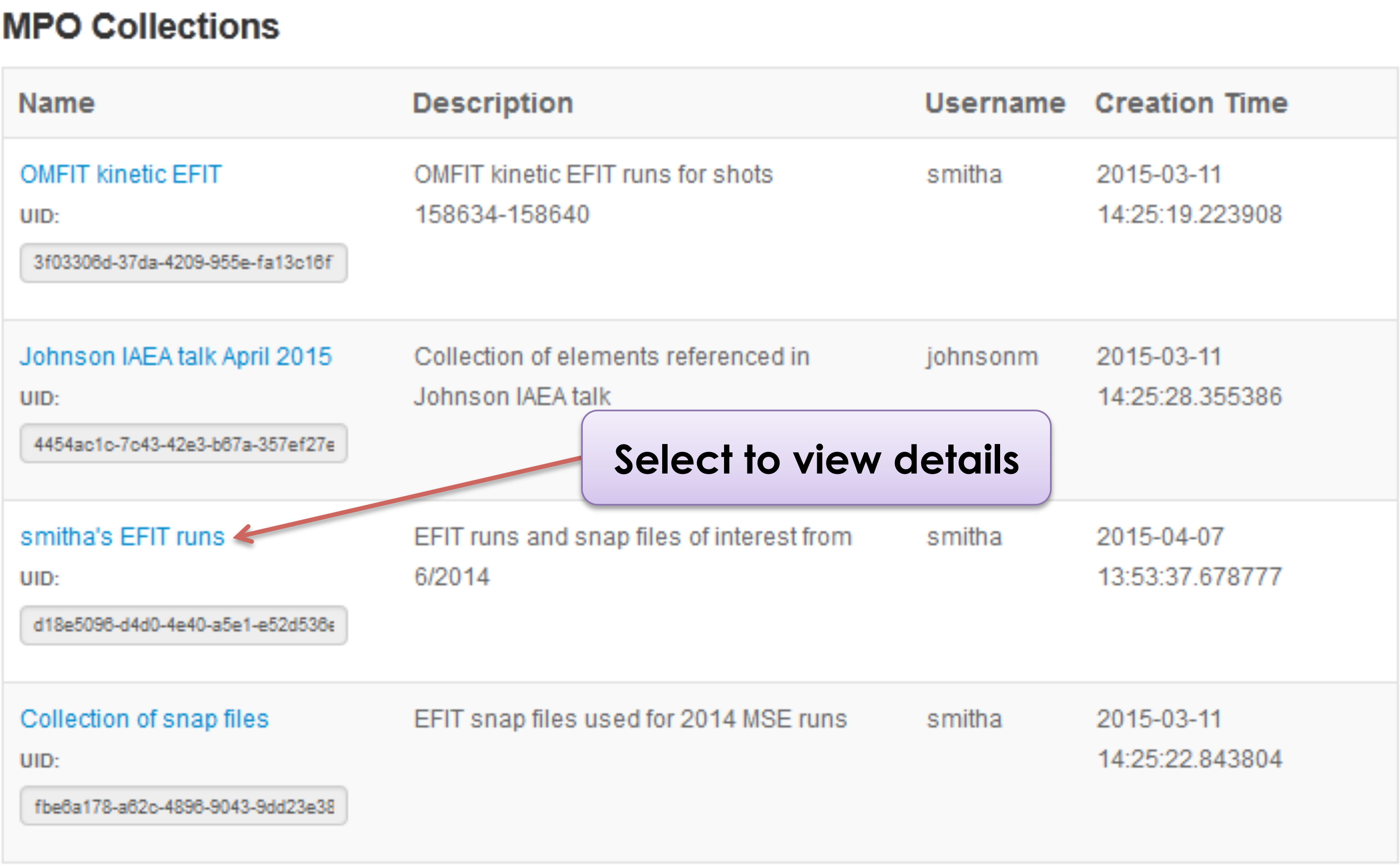
▭ Read Input Files 2015-03-11 13:23:20.908132

○ Plasma Current 2015-03-11 13:23:21

List of nodes and their corresponding details: UID, URI, metadata, comments, other linked workflows

Blue nodes are Data Objects used in other workflows

Interactive UI Page Example: Collections List

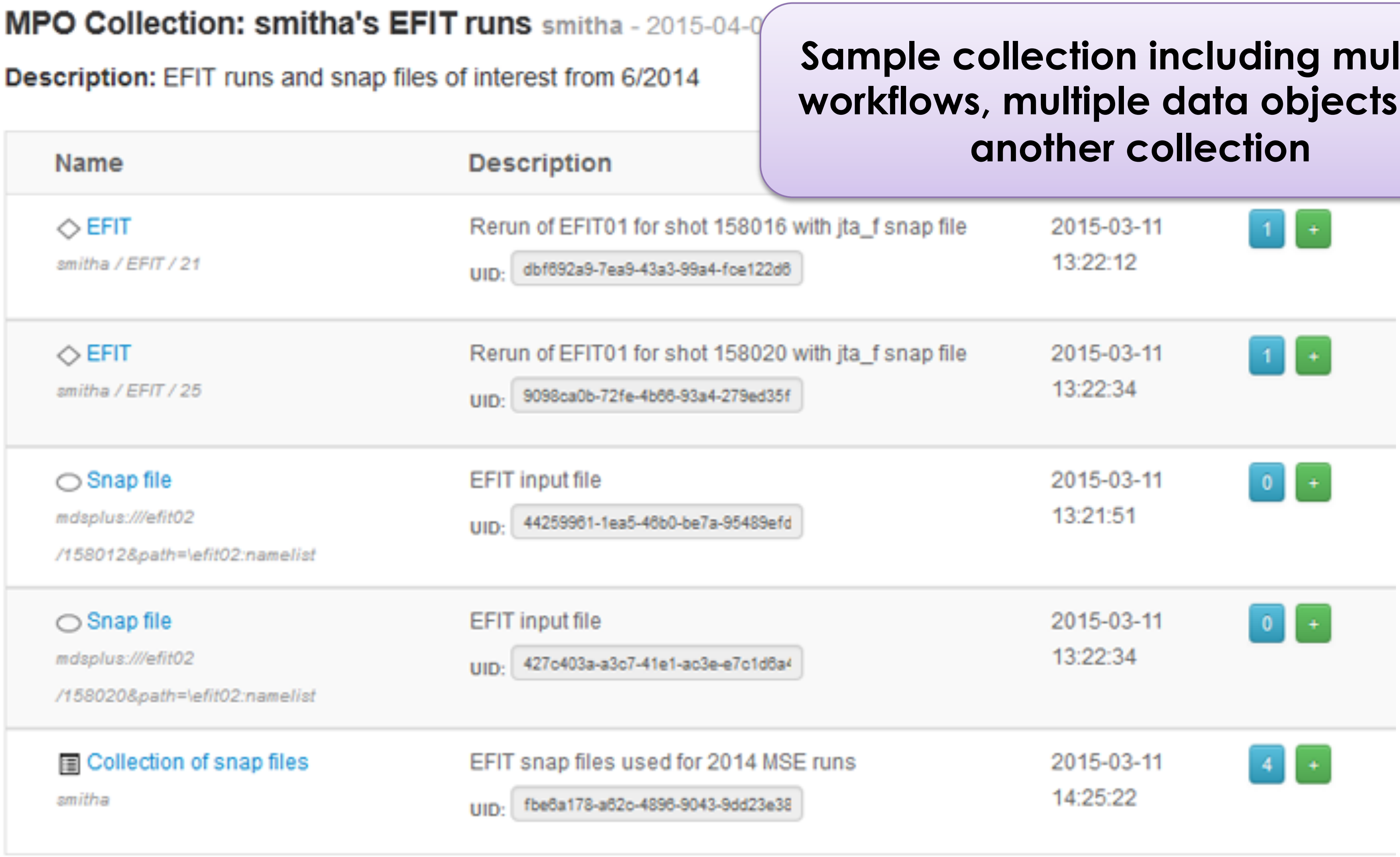


MPO Collections

Name	Description	Username	Creation Time
OMFIT kinetic EFIT UID: 3f03306d-37da-4209-955e-fa13c16f	OMFIT kinetic EFIT runs for shots 158634-158640	smitha	2015-03-11 14:25:19.223908
Johnson IAEA talk April 2015 UID: 4454ac1c-7c43-42e3-b67a-357ef27e	Collection of elements referenced in Johnson IAEA talk	johnsonm	2015-03-11 14:25:28.355386
smitha's EFIT runs UID: d18e5096-d4d0-4e40-a5e1-e52d536e	EFIT runs and snap files of interest from 6/2014	smitha	2015-04-07 13:53:37.678777
Collection of snap files UID: fbe6a178-a62c-4896-9043-9dd23e38	EFIT snap files used for 2014 MSE runs	smitha	2015-03-11 14:25:22.843804

Select to view details

Interactive UI Page Example: Collection Details



MPO Collection: smitha's EFIT runs smitha - 2015-04-0

Description: EFIT runs and snap files of interest from 6/2014

Name	Description		
◇ EFIT *smitha / EFIT / 21*	Rerun of EFIT01 for shot 158016 with jta_f snap file UID: dbf692a9-7ea9-43a3-99a4-fce122d6	2015-03-11 13:22:12	1 +
◇ EFIT *smitha / EFIT / 25*	Rerun of EFIT01 for shot 158020 with jta_f snap file UID: 9098ca0b-72fe-4b66-93a4-279ed35f	2015-03-11 13:22:34	1 +
○ Snap file *mdsplus:///efit02* */158012&path=\efit02:namelist*	EFIT input file UID: 44259961-1ea5-46b0-be7a-95489efd	2015-03-11 13:21:51	0 +
○ Snap file *mdsplus:///efit02* */158020&path=\efit02:namelist*	EFIT input file UID: 427c403a-a3c7-41e1-ac3e-e7c1d6a4	2015-03-11 13:22:34	0 +
▤ Collection of snap files *smitha*	EFIT snap files used for 2014 MSE runs UID: fbe6a178-a62c-4896-9043-9dd23e38	2015-03-11 14:25:22	4 +

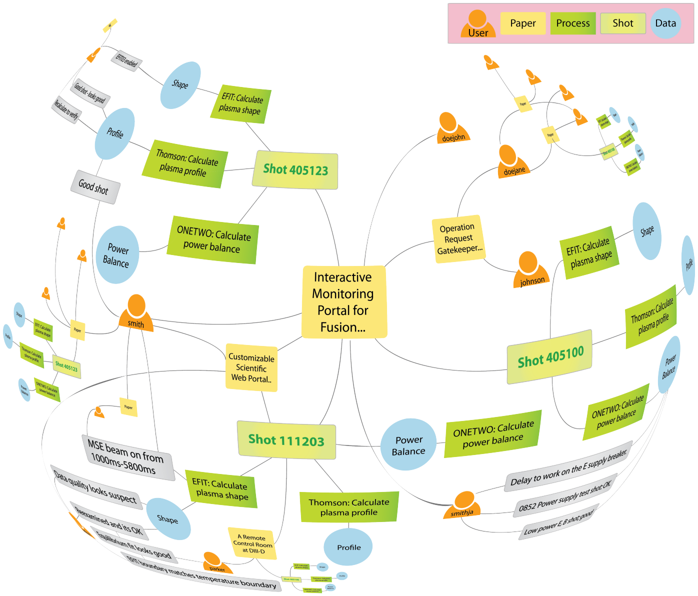
Sample collection including multiple workflows, multiple data objects and another collection

Current Status

- **The MPO System V1.0 is released**
  - http://mpo.psfc.mit.edu provides detailed information
- **Python, IDL, shell API clients are provided**
  - Used to instrument MPO calls
- **Integration with multiple workflows**
  - DIII-D between-pulse EFIT
  - SWIM (Simulation of RF Wave Interactions with Magnetohydrodynamics)
  - GYRO (Nonlinear tokamak microturbulence software package)
  - AToM (Advanced Tokamak Modeling)
- **Planned Integration with a Climate Modeling Project– Calibrated and Systematic Characterization, Attribution and Detection of Extremes (CASCADE)**

Future Work

- **Expand the reach of MPO framework**
  - Harden the system - ease of adoption, robustness, scalability
  - Reach out to more science domains – including non-fusion
- **Provide data exchange capability between MPO and W3C standard based software (e.g. PROV)**
  - W3C PROV
  - Import and export data
- **Improve user interface and analysis**
  - How to provide better/faster graphical navigation?
  - Additional visualizations and analysis



User
Paper
Process
Shot
Data
EFIT03 enabled
Good shot - looks good
Recalculate to verify
Shape
EFIT: Calculate plasma shape
Profile
Thomson: Calculate plasma profile
Shot 405123
Good shot
Power Balance
ONETWO: Calculate power balance
smith
Paper
Interactive Monitoring Portal for Fusion...
Customizable Scientific Web Portal..
Shot 111203
MSE beam on from 1000ms-5800ms
Data quality looks suspect
Reexamined and its OK
Equilibrium fit looks good
EFIT boundary matches temperature boundary
parker
A Remote Control Room at DIII-D
Profile
doejohn
doejane
Operation Request Gatekeeper...
johnson
Shot 405100
smithja
Delay to work on the E supply breaker.
0852 Power supply test shot:OK.
Low power E, B shot:good

Summary

- **MPO System is a software for documenting scientific workflows and data**
  - A new type of logbook with automation and analysis capabilities built-in
- **Production workflows have been MPO instrumented**
  - Proven useful
  - Approach is valid and general
- **MPO team seeks partners**
  - Test, deploy and feedback
  - Contribute
  - Contact email: mpo-info@fusion.gat.com

Acknowledgments

- **MPO team members**
  - Gheni Abla, Liz Coviello, Sean Flanagan, Xia Lee, David Schissel– ***GA/DIII-D***
  - Alex Romosan, Arie Shoshani, John Wu – ***LBNL***
  - Martin Greenwald, Josh Stillerman, John Wright – ***MIT/PSFC***
- **Our colleagues**
  - Members of SWIM, AToM teams
  - Staff at GA, LBNL, MIT
- **This work is supported by US Department of Energy**
  - Office of Advanced Scientific Computing Research
  - Office of Fusion Energy Sciences