

Project: JDEMSOC for FY10

Purpose of the project:

This document provides a high-level description of the tasks we are working on at Fermilab for JDEM in FY10, and describes the reasons for doing this work.

1. Together with the DOE JDEM Project Office we are attempting to establish Fermilab as the future location of the JDEM Science Operations Center and to position Fermilab for a significant role in JDEM science operations. Several tasks that the Project Office believes will strengthen our case for hosting the SOC at Fermilab have been identified, including the following:

- **Develop a design for slitless spectroscopy data processing for JDEM**

In the current mission design, slitless spectroscopy is the most challenging aspect of JDEM science operations. By developing a design for slitless spectroscopy data processing, we will have solved the most challenging aspect of JDEM science operations. This should be a powerful argument for locating the SOC at Fermilab and for playing a significant role in JDEM science operations.

- **Develop slitless spectroscopy simulations**

Simulations are needed to develop a design for slitless spectroscopy data processing. The deliverable for this task consists of generating simulated images for slitless spectroscopy.

- **Develop a prototype workflow for slitless spectroscopy and NIR processing**

Slitless spectroscopy and NIR data processing have workflow participants that are common to both types of data processing. The goal of this task is to develop a prototype workflow with an emphasis on software infrastructure. By working on the prototype we will identify algorithms and applications that are needed for the workflow and develop some of these applications. We will substitute “dummy” applications for workflow participants that we are unable to develop this fiscal year due to time constraints. The workflow will need access to a database, which will be developed as part of this task.

- **Develop prototype workflows for calibrations generated by EGSE at LBNL**

LBNL and SLAC are developing the EGSE demonstrator. One of our roles for this fiscal year is to develop prototype workflows for the generation of calibrations based on darks and flats generated by the EGSE demonstrator. A second role of ours is to demonstrate a science image calibration workflow based on examples of EGSE point source images (simulated star observations provided as FITS files.)

- **Develop requirements management capabilities**

We will work with the DOORS requirements management software to develop expertise in the use of this software for JDEM. Requirements will be developed for the QuIDS quality control software, and we will use DOORS to track these requirements.

2. Based on requirements developed for the SNAP Science Operations Center, we have identified several aspects of science operations that need to be addressed to reduce cost and risk of hosting the JDEM SOC at Fermilab. These aspects will be addressed by the following R&D efforts.

- **Investigate DDS for quality control**

We are working with Tech-X to develop QuIDS for quality control. Tech-X received a Phase 1 SBIR grant from DOE

to develop a system based on DDS. The goals of this effort are described in the QuIDS ConOps document. We are considering collaborating with Tech-X on a Phase 2 SBIR proposal. There are alternatives to working with Tech-X. For example, there is DDS expertise at Vanderbilt University, and we have developed DDS expertise at Fermilab so that we might be able to develop QuIDS without having to rely on Tech-X.

- **Investigate workflow management and provenance tracking**

For JDEM we are interested in a workflow management system that will simplify specification and execution of workflows. Ideally, the system will also record provenance and other metadata involved in the execution of the workflow.

- **Investigate large-scale databases**

At this time we do not have a good understanding of the database needs for JDEM. R&D on large-scale databases is one area that could turn out to be needed for JDEM. New database products for high-performance large-scale databases are available. Our goal is to support the database needs for slitless spectroscopy while investigating some of these new products.

3. Based on requirements developed for the SNAP SOC, we believe that we can provide better support for JDEM scientists' data analyses compared to NASA. By following a model used in HEP for science data analyses we expect that we can provide a better data-analysis environment for JDEM scientists than NASA. This could be an important argument for locating the JDEM SOC at Fermilab.
4. We are collaborating with members of the JDEM GDS team at LBNL. It is important that we establish an effective collaboration with LBNL and other institutions that receive DOE funding to demonstrate to DOE and NASA that we are an important part of the DOE team working on JDEM.

5. We are collaborating with members of the JWST Science Operations team at STScI. We benefit from this collaboration by learning more about working with NASA. JDEM may benefit from this collaboration if it turns out that STScI is selected to host the JDEM science archive, since we will have established a working relationship.
6. By developing the WBS for FY10 and executing the tasks outlined in the WBS, we are establishing a baseline budget for JDEM at Fermilab. The work that we do this year and the cost of doing this work will be used to develop future budget requests to support JDEM work at Fermilab.
7. We are working with members of other groups on developing capabilities that are of interest to JDEM and others. We are currently collaborating with STScI, LSST, NOAO, ESO (European Southern Observatory), the Kepler Project for scientific workflows, and projects at Fermilab such as LQCD and NOvA. This approach reduces our development costs now, and is likely to reduce future costs of developing software for JDEM.

Client:

The deliverables for FY10 are being developed for the DOE JDEM Project Office.

Stakeholders:

The stakeholders for our work on JDEM this fiscal year are the following:

- DOE's JDEM GDS team (FNAL and LBNL),
- DOE's JDEM Project Office,
- DOE's JDEM Scientists,
- Fermilab Directorate,
- Fermilab Computing Division Management,

- Fermilab Center for Particle Astrophysics.

While there are individuals from other projects who are interested in our software development efforts, we include these individuals (see next section) as “potential users” of the software.

Users:

Software that we will be developing this year will be used by members of the following groups:

- DOE’s JDEM GDS team (FNAL and LBNL),
- DOE’s JDEM Engineers,
- DOE’s JDEM Scientists.

Potential users of the software we are developing include projects at Fermilab as well as other projects such as JWST and LSST.

Constraints:

We are receiving funding for 2.5 FTE-years of effort for FY10, and approximately \$60K for M&S and travel. The funding consists of Fermilab carry-over from FY09, and DOE funding provided by LBNL.

We will use the following design solutions for FY10:

- Fermilab computing environment,
- OpenSplice DDS message passing software,
- Kepler workflow system,
- and DOORS for requirements management.

This does not imply that we have decided to use these solutions for JDEM in the long term, but due to funding and time constraints we cannot evaluate other alternatives at this time.

Scope:

Our priority for FY10 is to work on slitless spectroscopy and NIR calibration workflows. Work on the optical imaging workflow is a stretch goal for FY10. A specific goal is to develop a prototype computational framework that can be used to run slitless spectroscopy and NIR workflows. Substitute applications will be used when actual applications are not available for a particular workflow participant. In this case the substitute application will be implemented with characteristics representative of expected processing times in the workflow, and will have representative inputs and outputs.

Risks:

The main risks that we face for work on JDEM for FY10 include the following:

1. JDEM cancellation

The JDEM Project may get cancelled due to any number of reasons.

2. Fermilab leaves JDEM

There are a number of reasons that could prompt Fermilab Management to choose to leave the JDEM Project, such as inadequate LBNL funding for future Fermilab work on JDEM, inadequate DOE funding for FCPA scientists, or an insufficient role for Fermilab in future construction and operation of the JDEM SOC.

3. NASA rejects DOE's role as host for the SOC

We anticipate a future meeting (or meetings) in which DOE and NASA will discuss the role of each agency in JDEM science operations. We are currently working on building a strong case for hosting the SOC at Fermilab, but NASA may come to a different conclusion as to how and where the SOC will operate.

4. JDEM GDS team at FNAL is unable to produce a functioning prototype

One of our main tasks for this year is to produce a prototype workflow, which is referred to as the “Calibration Demonstration System” in the FY10 Statement of Work. If we are unable to produce such a prototype then this could jeopardize future Fermilab participation in JDEM.

5. JDEM GDS team at FNAL falls behind in the FY10 schedule

We have tasks in our WBS with start dates and durations. If we were to fall behind by several months in the schedule this will likely reduce our credibility within the JDEM Collaboration. A possible outcome may be that we receive more scrutiny in the future with less freedom to work on tasks that we consider important for JDEM. We might also receive less funding in the future, since we were not able to establish a baseline and then satisfy the baseline.

Terminology:

actor - a software component that performs a task, typically by reading input and producing output.

campaign - a *workflow* initiated by a human.

DDS - Data Distribution Service, a customizable *quality of service* publish/subscribe standard from the Object Management Group (OMG).

Fermigrid - the GLOBUS-based grid computing infrastructure installed and in use at Fermilab.

FITS - A FITS file is a sequence of *HDUs*. The first *HDU* in a FITS file (the primary *HDU*) has special requirements placed upon it.

HDU - a *header data unit* in a *FITS* file. This is a sequence of name/value pairs (the header) followed by the data described by the header.

instantiated workflow - a *workflow* in which all data sources and *participants*, as well as their configurations, are specified.

job - a submission to a batch queue. A *workflow* may contain *participants* that perform job submissions and manage interactions with the batch system. The submission itself may consist of a *workflow* (or subworkflow) together with the engine required to execute it, provided it is compatible with the batch system.

participant - an *actor* whose action is triggered by a *workflow engine*. From the point of view of the *workflow engine*, the task carried out by a participant is *atomic* in that the task completes successfully or fails to complete. Moreover, the *workflow engine* does not manipulate the internal state of a participant. A participant might, for example, be a shell script that runs a data processing application to perform a specific task.

pipeline - a *workflow* in which the *participants* are arranged according to a pipe and filter architecture. In such an architecture, the *participants* process *units of work* and can execute concurrently, each reading input from its predecessor and providing output to its successor. See, for example, the wikipedia entry on [software pipelines](#), Avgeriou & Zdun, and Clements et al.. (Often, when astronomers speak of “pipelines,” what is meant is either a [psuedopipeline](#) or some other *workflow*, or sometimes even just an isolated *participant*.)

publisher - a software entity that prepares data for transmission based on one or more *writers*.

quality of service (QoS) - the ability to provide different priorities to data flow in a network. Quality of service refers to control mechanisms that are used to reserve resources for different applications, users or data flows, or to guarantee a certain level of performance.

reader - a software entity that reconstitutes an object from the data that has been received from a *writer*.

roving laptop environment - an environment in which the user can disconnect from the network and later reconnect (possibly with a different IP address), and is able to continue the work being done before the interruption in the connection.

stream of data - a sequence of *units of work* of the same type.

subscriber - a software entity that receives data using one or more *readers*.

unit of work - the smallest data element that is processed in its entirety by an actor. Usually actors operate on a sequence of data elements.

workflow - a collection of *participants* and a defined set of rules specifying when (under what conditions) and how (with what parameters, configurations, and input data) the actions performed by the *participants* should be triggered.

workflow engine - a software application that manages and executes *workflows*.

workflow management system - a software application that triggers the actions performed by *participants* according to the rules that define that *workflow*. A workflow management system may additionally record provenance and other metadata about the execution of the *workflow*, and may provide tools for users to specify *workflows*.

workflow template - A *workflow* in which a subset of the data and/or *participants* are specified abstractly. At a minimum, a workflow template contains only the workflow rules, which refer to data and *actors* using undefined symbols.

writer - a software entity that makes instances of a user-defined data structure available over a network.