Scientific Workflows and Provenance Working Group Charter (Draft)

January 31, 2010

1 Background

The Observation Network for Earth (DataONE) is poised to be the foundation of new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data. Supported by the U.S. National Science Foundation, DataONE will ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data. DataONE will transcend domain boundaries and make biological data available from the genome to the ecosystem; make environmental data available from atmospheric, ecological, hydrological, and oceanographic sources; provide secure and long-term preservation and access; and engage scientists, land-managers, policy makers, students, educators, and the public through logical access and intuitive visualizations. Most importantly, DataONE is not an end but a means to serve a broader range of science domains both directly and through the interoperability with the DataONE distributed network.

Working Groups

Working groups are central to DataONE in conducting research, specifying cyberinfrastructure, and engaging the community. The Working Group model allows DataONE to conduct targeted research and education activities with a broad group of scientists and users. Working Groups are also designed to enable research and education activites to evolve over time. Each Working Group will have two co-leaders who organize the activity and propose solutions to particular research, education, and cyberinfrastructure problems.

2 Purpose, Scope, Mission

The data lifecycle that informs the DataONE architecture is aimed primarily at the collection, curation, validation, and long-term preservation and accessibility of high-volume datasets. The mission of the Working Group on Scientific Workflows and Provenance (SWAP) stems from the observation that a wealth of machine-processable provenance metadata for describing data management processes and their detailed execution, can be collected and later used to provide DataONE users with additional value at each phase of the lifecycle. Delivering such value requires that process descriptions and provenance metadata become first-class citizens in the DataONE space. This group will work towards this goal, by investigating models, techniques, and tools for preserving process specifications and data provenance together with primary and derived datasets. This will allow scientists and other DataONE users to leverage this information throughout the data lifecycle.

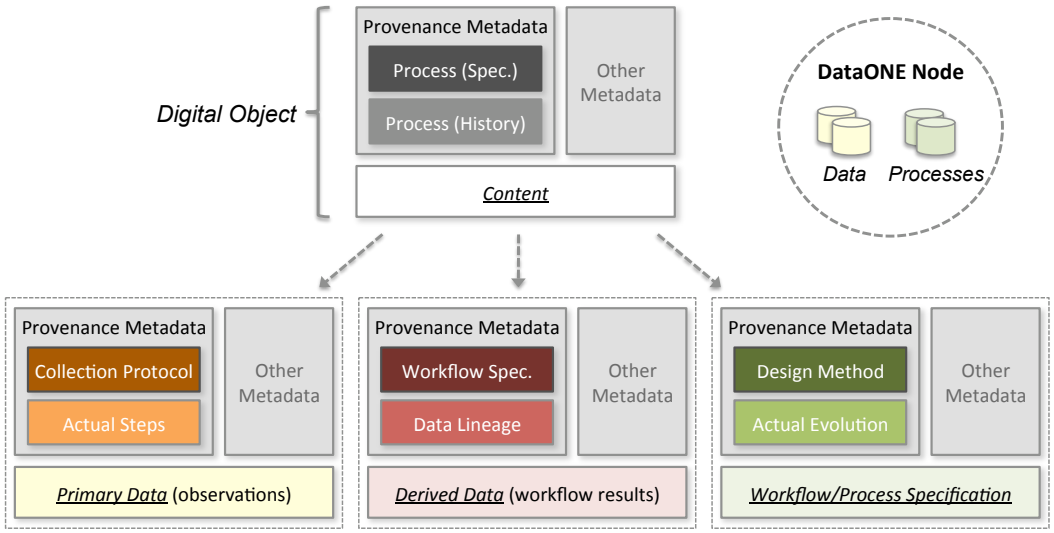

Figure 1: Abstract structure of *digital objects* considered by the SWAP working group (top) and three specific types of digital objects for managing *primary data*, *derived data*, and *workflow specifications* (bottom).

3 Motivation and Objectives

By enhancing the primary and derived data stored in DataONE repositories with provenance metadata and workflow support, DataONE will be poised to provide scientists, information experts, and other DataONE users with the ability to more effectively automate data ingestion, curation, and analysis tasks; discover and reuse data and process specifications; and assess and ensure data quality. Thus, a primary objective of the working group is to interact with members of the workflow and provenance communities to evaluate and extend existing technologies for incorporation into the DataONE infrastructure and investigator toolkit.

Digital Objects. Figure 1 shows three broad kinds of digital objects that encapsulate primary data, derived data, and workflows, respectively. These different object types share a common structure (see top in Fig. 1):

(D1) **Content:** one or more datasets, or information artifacts of interest to DataONE users;

(D2) **Process specification:** a description of the process to be followed for producing a dataset;

(D3) **Process history:** a description of the actual steps followed during the production process; and

(D4) **Other metadata:** any annotation or additional information that is either manually or automatically added to complement the data such as attribution information (who created or executed the process) or the structural description of the dataset.

The bottom part of Figure 1 depicts three specific realizations of digital objects that are within the scope of the DataONE preservation effort. The leftmost realization illustrates the case of *primary data*, i.e., observational data that is not the product of a computational workflow. Here, the *actual steps* (D3) followed for data production are recorded along with the prescribed *collection protocol* (D2). For instance, consider the case of a sensor network designed to periodically measure environmental data (D1), e.g., temperature, humidity, windspeed, etc. The protocol may specify the layout of the sensors, their calibration, the query frequency, and other details that fully specify the data acquisition process. In turn, the process history for primary data

corresponds to a record of the actual steps that were taken during the acquisition. The main purpose of such a record is to help data analysts better understand the quality of the collected data, e.g., by comparing the actual steps used with the prescribed protocol.

The second type of digital object illustrates the case of *derived data* obtained using computational methods, which may range from the ad hoc invocation of tasks, typically scripts or services, to the execution of full-fledged workflows that specify an automated composition of those tasks. When the processes are specified using a workflow system (e.g., Kepler, Taverna, VisTrails, and others), the workflow specification (D2) itself becomes part of the digital object, together with (D3), the detailed trace of data lineage dependencies and other events that occurred during the computation (e.g. the invocation of scripts, together with inputs and outputs).[1] Many workflow management systems (including those mentioned above) provide facilities to record such provenance information automatically. For derived data, the adoption of a common, formal, and machine-processable model for provenance, such as the Open Provenance Model[2], facilitates the automated interpretation, analysis, and reproduction [Mes10] of the processing history of data.

Finally, the third kind of digital object in Figure 1 figure extends the notion of digital objects to the workflows themselves, now viewed as knowledge assets and artifcats that may evolve in time and should be preserved and made available for sharing, and whose evolution history should be documented. In particular, in this case the content is the specification of a process or a workflow. Correspondingly, the process specification (D2) of the digital object (here: the design method) describes a specific methodology for process design and evolution, whereas the process history (D3) captures the actual workflow evolution as a record of the changes to the process specification from one version to the next.

Data Lifecycle. Fig. 2 shows how the different digital objects can be incorporated into the general data lifecycle of DataONE. The figure illustrates the main phases of the data lifecycle (the boxes) as actions that operate upon the digital objects (shown as icons next to the arrowed lines) managed through the DataONE nodes. As shown, in this view of the data lifecycle, workflow specifications, primary data, and derived data can be deposited into, and subsequently discovered via DataONE nodes. Primary data, derived data, and workflow specifications can also be reused and repurposed via tools that will be made available through the investigator toolkit. For instance, workflow specifications could be discovered, modified, and then used over primary and derived data to produce new derived datasets. Both the derived datasets and the modified workflow specifications are deposited to DataONE nodes, and their associated provenance metdata is also stored and subsequently made available (e.g., to help others determine their suitability for reuse in other contexts).

Working Group Focus Areas. The primary focus areas of the working group are shown in Fig. 3. A summary of the **outcomes** by each area (columns) and phase (rows) of the data lifecycle is also shown. The main focus areas include processes and workflows, provenance capture, provenance storage, and provenance analytics. Underlying the outcomes is a classification of workflow types according to two orthogonal dimensions, namely their purpose and their specification model. In terms of purpose, a distinction is made between ***scientific workflows***, which contribute to data creation, analysis, and visualization, and ***data preservation workflows***, which describe actions directed at recording datasets as part of the DataONE data preservation and curation practices. In terms of the specification model, a distinction is made between workflows that are formally specified, using for example a specific workflow language (underlying Kepler, Taverna, VisTrail, etc.), and those that reflect manually constructed sequences of steps. This classification provides an initial framework to understand the role of workflows, and workflow repositories, in the different phases of the

[1] For techniques to efficiently store and query lineage data see e.g. [ABML09] and [ABML10].

[2] `http://openprovenance.org/`

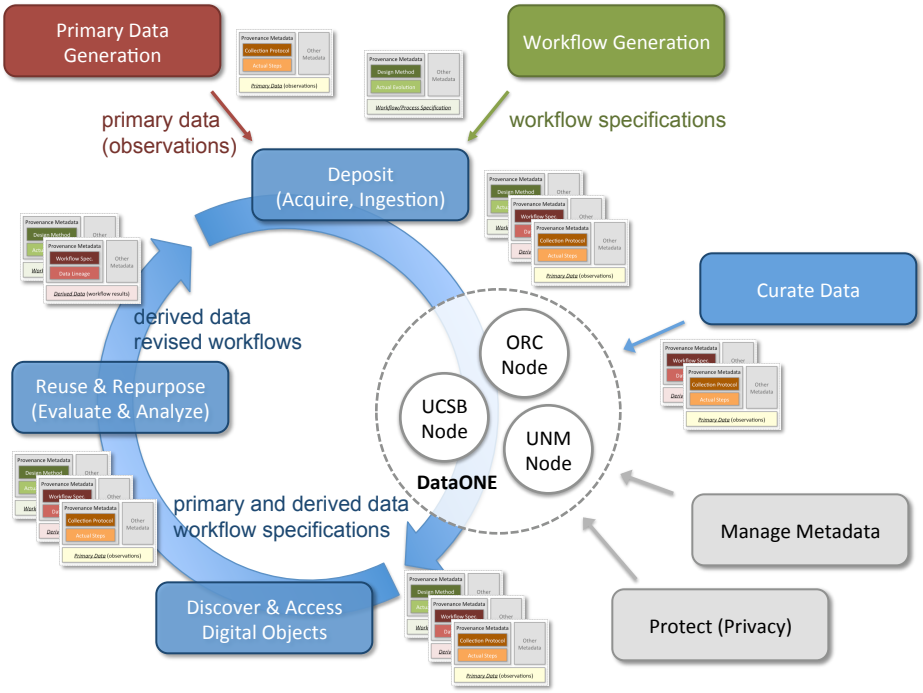

Figure 2: The DataONE lifecycle of data incorporating primary data, derived data, and workflow artifacts with corresponding provenance metadata for quality assessment.

lifecycle. Similarly, it is useful to distinguish between capturing provenance, which includes capturing both the description and the execution trace of workflows, and provenance mining, which involves a variety of query and analysis capabilities on large-scale provenance databases. To fully account for the importance of workflow and provenance in the DataONE context, it is useful to combine the data lifecycle with a simple specification of a workflow lifecycle, whereby workflows are designed, executed, assessed and improved upon as part of an iterative incremental refinement process (Figure 2).

In addition to the high-level objectives summarized in Fig. 3, the working group will also focus on the following **technical objectives**:

- Analysis of existing provenance models suitable to address the specific user and technical needs of DataONE. Although community efforts are under way on a common model for representing provenance (the Open Provenance Model amongst others), a gap analysis with respect to specific DataONE use cases may reveal additional requirements that are not met by such specification. Some of the working-group members are expected to be involved in these community efforts and provide input to it, as part of the activity associated with this topic.

- Architectures and systems for access to and storage of large-scale provenance metadata. As with provenance models, a number of community efforts are under way on efficient approaches for storing provenance information, and members of these groups are expected to be involved in the working group.

- Languages and systems for querying and mining provenance information to support data discovery, access, and use. A number of efforts are also underway and have been in place for querying prove-

Focus Areas of the DataONE Scientific Workflow and Provenance Working Group

Digital Object Lifecycle	Processes & Workflows	Provenance Capture	Provenance Storage	Provenance Analytics
Deposit (Acquire, Ingest)	• Employ workflows for **automatically ingesting** primary data into DataONE repositories (*Ingestion & Preservation Workflows*) • Specify **data collection protocols** and the **actual steps** used in obtaining primary data (*Collection Protocol* and *Actual Steps*)	• Automatically **track which Ingestion Workflows** were used for depositing primary data • Automatically **record detailed provenance** produced during execution of Ingestion Workflows • Manually **link Data Collection Protocols and Actual Steps** to associated primary data (as metadata)	• **Archive Ingestion and Preservation Workflows** with associated **detailed provenance** for primary data deposited into DataONE • **Archive Data Collection Protocols and the Actual Steps** used in obtaining primary data deposited into DataONE	• Provide access to Ingestion Workflows and detailed lineage information to **assess quality of primary data based on ingestion process** • Provide access to Collection Protocols and Actual Steps to **assess quality of primary data based on collection protocol and approaches used**
Curate	• Employ workflows for **automatic data curation activities** (*Curation Workflows*)	• Automatically **track Curation Workflows** and **record detailed provenance** produced during execution	• **Archive Curation Workflows** with associated **detailed provenance** for curation activities	• Provide relevant views of Curation Worklows and detailed lineage information to **assess quality of curation activities**
Discover & Access			• Allow DataONE users to search for and retrieve **relevant workflows** as well as associated **primary and derived datasets** based on provenance metadata	• Allow DataONE users to **query and analyze provenance metadata** to support decision making regarding the fitness of purpose of digital objects, including **quality filtering and relevance determination**.
Reuse & Repurpose (Evaluate, Analyze)	• Employ workflows for **automating evaluation and analysis tasks** for generating derived data from DataONE holdings (*Analysis Workflows*) • Modify, extend, and combine workflows for reuse (*Workflow Reuse & Evolution*)	• Automatically **track which Analysis Workflows** were used for generating derived data • Automatically **record detailed provenance** produced during execution of Analysis Workflows • Automatically **record the evolution history** of workflows	• **Archive Analysis Workflows** with associated **detailed provenance** for derived data deposited into DataOne • **Archive the evolution history** of workflow modifications and extensions	• Provide access to Analysis Workflows and detailed lineage information to **assess quality of derived data and use of primary data** • Provide access to evolution history of workflows for **determining attribution and supporting workflow design**

Figure 3: Focus areas of the working group and their objectives and potential benefits to users of DataONE.

nance information and leveraging provenance for quality assessment. Similarly, the working group will include representatives of this work as members of the working group.

- Provenance mining in support of assistive workflow design, workflow evolution, and enhancement of data retrieval from DataONE node;
- Workflow and provenance requirements to enable *reproducible research*;
- Benefits of, requirements, and techniques for managing semantic-enhanced provenance

The above list roughly corresponds to the order in which the WG plans to address the various issues raised. As new use requirements are being developed within DataONE, an ongoing task of the working group will be to analyze the role and potential contribution of workflow and provenance technology in the context of such use cases.

4 Duration of the working group

This Working Group intends to be active throughout the duration of the DataONE project. s

5 Expected Deliverables, Outcome & Schedule

The WG will play a number of roles with respect to the core DataONE CCIT. Here we list expected deliverables associated to each of these roles.

1. In its **technology advisory role**, the group will provide periodic recommendations on evolving specifications and technology to the DataONE CCIT, as well as promote a common understanding of the different directions in which the technology is moving. Specific deliverables include:

   - A document that synthesizes a common understanding of provenance in the context of DataONE workflows. Producing such document is the objective of the **first WG meeting**, which aims at bringing together experts with different perspectives on a DataONE-specific case study.
   - One or more white papers on the progress at the cutting edge of the area of scientific workflows and provenance management, and on open challenges.
   - Contingent upon available resources, **case-studies** and **demonstrations** of promising technology, and/or **prototypes** can be developed, that then can be further developed and hardened by the CI-Core team.

2. In its **technology transfer facilitator role**, the group will provide ongoing support towards the adoption of sufficiently mature specifications and software into the DataONE technology suite. Resources permitting, the group will interface with DataONE architects and developers to integrate software that is within its remit, into the DataONE nodes functionality and the investigators' toolkit.

3. In its role as **link between DataONE and the workflow and provenance community**, the group will work to establish and maintain working relationships and collaborations with relevant communities and standardization bodies, in order to:

   (a) promote DataONE **requirements** to the community;

   (b) promote DataONE as a valuable **case study** to be used for large-scale deployment and experimentation of cutting edge provenance technology;

   (c) periodically explore the state of the art in academia and industry.

   Specific deliverables in this capacity include the promotion of discussion, dissemination and other fact-finding activities, through participation to, and organization of workshops, preferably co-located with popular venues (conferences, experts' groups meetings in various domains of science).

6 Potential Risks

This group's activities are based on the expectation of a sustained momentum around workflow and data provenance models and technology, and of their growing uptake by the community across different domains of science. The group's impact on DataONE will be reduced if this expectation goes unfulfilled in the long term, or conversely, if insufficient resources will be available to take the best advantage of the momentum. Specific risks include the following:

- The resources assigned to facilitating effective technology transfer are insufficient;
- The technology explored by this group is not sufficiently mature to meet DataONE's needs;

- The group fails to stir enough interest within the provenance and workshop community in taking up DataONE as an interesting case study.

7 Membership

This group consists of a core set of members, and will invite selected experts from the community to join as appropriate to meet its goals. The initial members include:[3]

- Bertram Ludäscher, UC Davis (Kepler and provenance research)
- Paolo Missier, University of Manchester, UK (Taverna and provenance research)
- Shawn Bowers, Gonzaga University (Workflow and provenance research)
- Steve Kelling, Cornell (EVA WG and EVA workflow liaison)
- Bob Cook, ORNL (DataONE and Data Preservation)
- Carole Goble, University of Manchester, UK (Taverna and myExperiment project lead)
- David De Roure, University of Southampton, UK (myExperiment project lead)
- Juliana Freire, University of Utah (Vistrails and provenance research)
- Matt Jones (Kepler liasion)

8 Roles and Responsibilities

Members of this working group will have an expert understanding, and an active interest in any of the following areas:

- workflow modelling for scientific applications, in the areas of interest to DataONE;
- workflow management technology and its theoretical underpinnings;
- data and workflow provenance and its management (storage, querying, etc.);
- large-scale data management architectures;
- social models and technology for collaborative science.

Decisions within the group are aimed to be based on consensus to the extent possible. When dissent is observed, the WG Chair(s) will promote a discussion, and then record a decision based on a majority of votes, along with any objections. New members will be nominated by the WG Chair(s).

[3]There is now a growing community of researchers from the workflow and databases communities actively working on data and workflow provenance. Thus, this list can be easily expanded. The challenge is to recruit members that are available and committed to supporting the DataONE effort.

9 Resources

Meeting organization and travel resources will be required for active members who volunteer to undertake significant and recognized responsibilities within the group. These will include both the organization of planned WG meetings, as well as additional meetings amongst WG members in the context of occasional events (e.g. conferences), and sponsorship of workshops that are within the scope of the group. In addition, to support transfer of technology and know-how from the data and workflow provenance community into DataONE via the development of case-studies and prototypes, some R&D support will be needed. These resources will be used to support student, postdoc, or developer time to work on a focused project (further defined as part of the the initial WG meetings).

10 Relationship to other WGs

The SWAP WG will collaborate primarily with the following DataONE WGs:

Scientific Exploration, Visualization, and Analysis: This WG will collaborate with the Scientific EVA group, which is expected to provide valuable use cases for this group to analyze and work on.

Preservation and Metadata: This group's emphasis on the preservation lifecycle and on the progression of datasets into digital objects make the Preservation and Metadata group a natural first choice for close collaboration.

Integration and Semantics: Interaction with the Integration and Semantics group is also expected to lead to useful collaboration, based on the emerging role of semantics in metadata.

Core CI team: Finally, and not less important, the nature of this group suggests a natural liasion role with the Core CI team, as articulated in Sec. 5 and elsewhere in this document.

11 Communication Plan, Reporting Requirements

- It is envisioned that the group will hold periodic meetings through teleconferences and other online mechanisms, as well as ongoing e-mail communication and periodic face-to-face meetings, at venues which a significant number of WG participants are likely to attend, and will typically be co-located with events such as conferences, workshops, etc.

  The frequency of such meetings is to be determined and may change in time, depending on the types of ongoing activities.

- The group will use a dedicated area of the DataONE site for exchanging documents and other archival purposes.

- In addition to its normal deliverables, the group will prepare an annual report describing past activities and their outcome.

- In addition, during periods where the WG includes a specific R&D project, the supported FTE will communicate and periodically report to the DataONE Core-CI team.

References

[ABML09] Manish Kumar Anand, Shawn Bowers, Timothy M. McPhillips, and Bertram Ludäscher. Efficient provenance storage over nested data collections. In *Intl. Conf. on Extending Database Technology (EDBT)*, pages 958–969, St. Petersburg, Russia, 2009.

[ABML10] Manish Kumar Anand, Shawn Bowers, Timothy M. McPhillips, and Bertram Ludäscher. Techniques for efficiently querying scientific workflow provenance graphs. In *Intl. Conf. on Extending Database Technology (EDBT)*, Lausanne, Switzerland, 2010. to appear.

[Mes10] P. Mesirov , Jill. Accessible Reproducible Research. *Science*, 327, 2010.