Provenance Metadata and Extensibility of Metadata Describing Measurement Data

Project: NetKarma
Milestone: MDOD

Scott Jensen and Beth Plale
Indiana University
scjensen@indiana.edu, plale@indiana.edu

In this whitepaper we analyze the preliminary metadata proposals for the Measurement Data Object Descriptor (MDOD) [6] whose role it is to describe a Measurement Data Object (MDO). As a starting point we use version 1.0 of the *GENI Instrumentation and Measurement Architecture* documentation [1] and the draft MDOD discussed in the Instrumentation and Measurement (I&M) group meeting at GEC11. We make recommendations on the organization and content of the MDOD such that it is more extensible for describing the data generated by future measurement tools, and that it is able to track the provenance of an MDO within the MDOD. This latter objective arose from the I&M discussion at GEC11. Our recommendations are based on careful analysis of the I&M use cases and on our recommendation for an MDOD that is extensible to MDOs generated by future measurement tools or enhancements to existing tools.

1. Instrumentation and Measurements Use Cases

The GENI I&M group has identified a number of use cases for measurement data from the perspective of different GENI user groups (e.g., experimenters versus operators) [3]. Some of these use cases involve the MDO's directly, and some of the use cases involve using the MDOD's metadata to find the relevant MDO's. It is this second scenario - data discovery, where the MDOD is relevant from the standpoint of discovery metadata. In addition to discovery metadata, some use cases identify the potential need for "use" metadata which can assist users in determining whether an MDO they have discovered meets their needs.

1.1 Use Cases

Following are the use cases identified by the I&M group that are relevant to metadata and the MDOD. Each use case is prefaced with the user group that the case relates to.

Experimenter: My experiment data shows inconsistencies, let me query the status of user slice resources so that I can trace my non-intuitive results to a problem in the environment and subsequently notify GMOC about any perceived performance problems.

Experimenter: Provide me with an archive of some or all of the slice resource performance measurements so that I can reference them during offline analysis of the data collected in my experiment after the slice expires.

Experimenter: Provide access to my opt-in users who want to query measurement data within my experiment slice using web-service clients based on GIMA compliant data sharing schemas.

Experimenter: Provide me with an archive of some or all of the slice resource performance measurements that I requested as part of my experiment.

Experimenter: Provide me with mechanisms to share my slice measurements archive with researchers and opt-in users at different levels of permissions sharing (i.e., whitelist/blacklist, sign-in, public).

Operators and Aggregate Providers: We would like to keep metadata of all the experiments, send us experiment metadata after each slice expires. *(Note: mentioned here because it is capturing metadata of the experiment, but in this case, the metadata is the MDO, not MDOD.)*

Operators and Aggregate Providers: Provide me with an archive of some or all of the slice resource performance measurements of users X and Y so that I can analyze infrastructure problems spanning multiple aggregates that may have corrupted the users experiment environments.

Archive Providers: I would like NOC staff, aggregate providers, and experiment researchers to provide me policies relating to the measurement archive sharing permissions (i.e., whitelist/blacklist, sign-in, public).

Archive Providers: I would like measurement archives corresponding to GENI experiments to be published in the repositories PQR by the experiment researchers, aggregate providers and GMOC with suitable keywords that allow me to catalog indexes for future search and retrieval purposes.

Researchers using Archived Measurement Data: I would like get search results and access to measurement archives corresponding to GENI experiments published by the experiment researchers, aggregate providers and GMOC when I use different search keywords.

Researchers using Archived Measurement Data: I would like to be able to share (e.g., email, post on Twitter), annotate, search and cite the measurement datasets in repositories of several Archive Providers.

1.2 Use Case Analysis

From an analysis of the above use cases, the key metadata that should be in the MDOD to support data discovery (identifying which MDO's may meet a user's needs) include the following:

- ID of the slice
- ID of experiments within a slice
- ID(s) of the owner or user of a slice
- Keywords describing an experiment

These four criteria relate to the above listed I&M use case as follows: experimenters will want to discover MDO's based on the slice the measurements relate to. For operators and aggregate providers, the uses cases identify a need to identify MDO's by slice, but also by user of the slice and experiment within a slice. The relationships between slices and aggregates are many-to-many, so from the

experimenter's perspective their slice may span multiple aggregates, and from the aggregate provider's perspective, multiple experiments will be running on their resources. Archive providers need the ability to search based on keywords as well as access to policy information regarding how an MDO can be shared. Researchers also need the ability to search on keywords, but additionally require the ability to cite an MDO and annotate MDOs.

The policy information on sharing that is needed by archive providers, as well as the need by researchers to be able to cite an MDO can both be classified as "use" metadata, which is needed to be able to use (or in the first case access) the MDO, but is not needed for data discovery. Additionally, the policy metadata regarding the access or sharing constraints on the MDO may be considered "registry" level metadata. In the GENI NetKarma project [9], provenance metadata is segregated into registry and instance level provenance. The instance level relates to a particular data object or workflow (experiment), but the registry level provenance relates to services or types of workflows (e.g., workflow templates). The policy metadata regarding MDO access shares similarities with the registry level provenance in that particular aggregate providers or operators may have standard policies regarding access and use constraints that would apply to broad sets or classes of MDOs.

1.3 Provenance Metadata Needs in Use Cases

Provenance is a type of metadata, and some of the above use cases directly address the need for provenance as part of the metadata, such as operators and aggregators searching for experiment results relating to specific experimenters. Other use cases listed above will also require provenance even if not directly stated in the requirements. For example, research users searching archived measurement data based on keywords will want to know the provenance of the measurement data that they find - were the measurements instrumented and taken by operators, aggregate providers, or experimenters? The provenance will also capture the stewardship of the measurement data - who collected it, and which services may have been stewards of the MD in its lifecycle, such as Measurement Collection (MC) services, Measurement Information (MI) services, Measurement Analysis and Presentation (MAP) services, User Workspace (UW) services, or Digital Object Archives (DOAs).

The second use case related to researchers using archived data refers to sharing measurement data. As data is shared or forwarded between users, or relied on in research papers, the provenance as to who collected the measurements, who shared the measurement data, and who calculated statistics using the measurement data will be critical to establishing the weight and value researchers and reviewers are willing to put on specific measurement data.

2. I&M Architecture Document Measurement Data Object Descriptor (MDOD) Description

Preliminary requirements for the MDOD are identified in version 1.0 of the *GENI Instrumentation and Measurement Architecture* documentation [1]. Measurement Data Object Descriptors (MDODs) provide metadata regarding Measurement Data Objects (MDOs). MDOs can be files, directories, or a temporal slice of a stream that is instantiated as a file (such as from perfSONAR). perfSONAR, OML and IPFIX are identified as instances of stream generating measurement tools.

According to[1], MDODs are most often created either when:

a) a slice experimenter/owner/or operator takes measurements, or
b) an MDO is transferred from one slice to another (the MDOD may be created either by the transferor or recipient).

Identifiers
Locators
Object type
Measurement data schema
Subject
Annotations
Owner/creator
Previous holder(s)
Current holder
Rules for sharing / disposal

Table 1. Metadata components for MDOD from architecture documentation

Metadata schemata used in the sciences to describe resources (often data and services), have converged on some general principles. Among these is the grouping of metadata into categories or sections; what these categories are is emerging as a common set and are becoming widely accepted. Drawing from this, we recommend reorganizing and aggregating the components of the existing MDOD, shown in Table 1, into the following sections, each of which is often represented by a separate schema document:

- *Identification*: Identifiers, owner/creator, subject, locators, spatial and temporal metadata, and object type.
- *Lineage/provenance*: current holder, previous holder(s).
- *Constraints/security*: rules for sharing, disposal, anonymization, encryption, and access methods.
- *Measurement data descriptions*: types of measurements (tools used?), interpretation methods, flow rates, object size and format.

The measurement data description schema is expected to be machine readable. The MDOD also contains an annotation element that can relate to specific aspects of the data and metadata or relate to the MDOD as a whole. These annotations, by their free-text nature, would be more loosely defined and not machine readable.

3. Draft MDOD Design

At GEC11, a draft version [6] of the MDOD was discussed which expands and elaborates on some of the components from the I&M Architecture document. The former identifies the following top level components:

- Identifiers
- Descriptors
- Holders

This preliminary grouping of the components from the architecture document into three categories is a start on a logical grouping of components similar to other established metadata standards, but there is overlap between the three categories in the draft MDOD. *To address this, we advocate reclassifying some of the elements and also breaking out constraints and security metadata as a separate fourth section.*

Specifically, in the current MDOD design specification, much of the identification metadata is included within the "identifiers" section, including the primary ID (introduced at GEC11), title, abstract, subject, and keywords. The descriptors section includes spatial and temporal metadata; additional identification metadata including the project, slice, experiment, and run IDs; and holder IDs and data location (e.g., path or URL). We suggest all of the identification attributes be aggregated under a single identification section that includes the IDs (identifiers) and all metadata related to the identification of the MD described by the MDOD.

The "descriptor" section contains information about the measurements made in the MDO described by the MDOD, including the target of the measurements, the types of tests, parameters, data flow rates, MDO object size, data format and interpretation method, and nested descriptors. However, the descriptor also contains a significant degree of identification metadata that either logically fits in the identifiers section or overlaps with metadata already captured in the identifiers section. The descriptor also includes security metadata such as whether the data is encrypted and access methods.

The "holders" section of the MDOD is a combination of provenance, security/access constraints, and some identification metadata. The holders section includes provenance of the original collector of the measurement data, subsequent holders of the data, and transactions related to how the measurement data was shared or disseminated (who was it shared with, assigned to, or modified by). It further contains transactions related to lineage, (who collected or received the measurement data). The primary holder contains security information set by the creator of the measurement data. This includes sharing, anonymization, and disposal policies. The holders section also contains identification metadata such as slice and project IDs that are also included in the descriptors section.

A mapping between the components from the I&M Architecture documentation [1], the draft version 0.2.1 MDOD design discussed at GEC11 [6], and the four independent sections we suggest in Section 2 are related as shown in Figure 1.

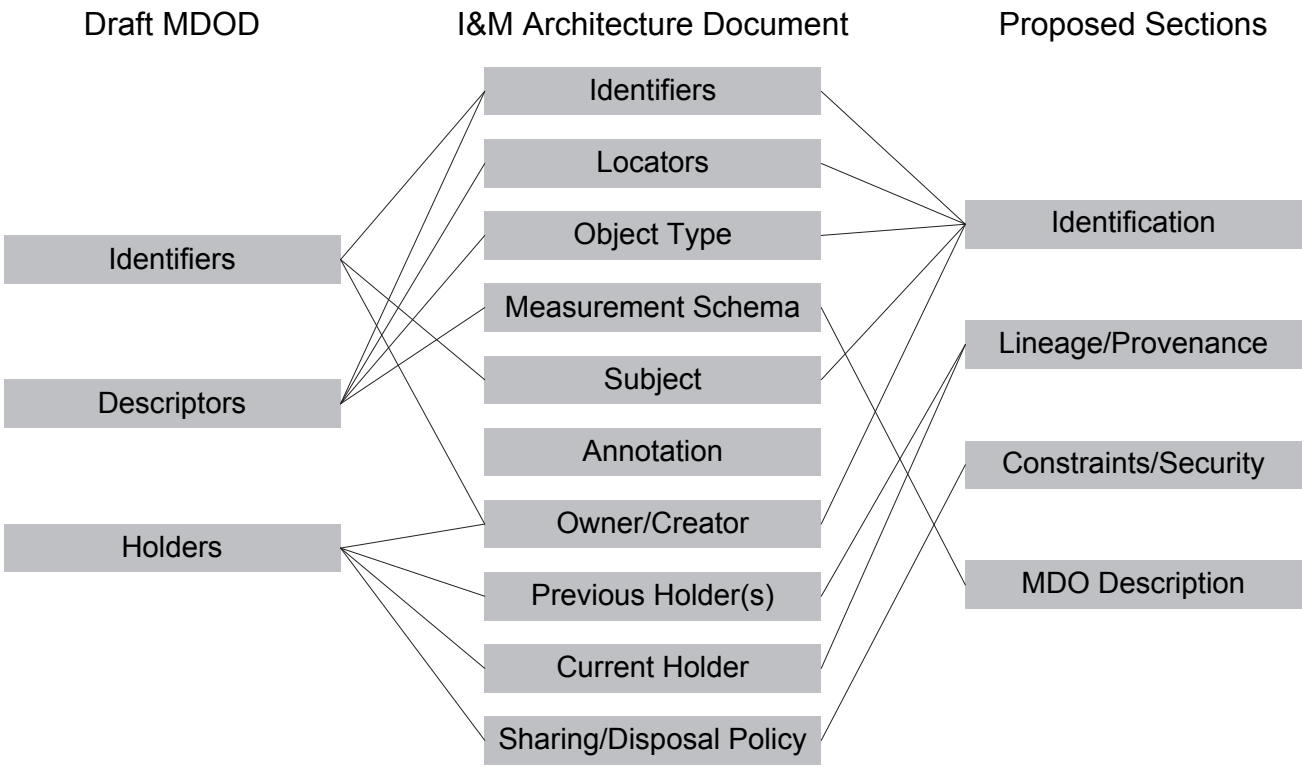

Figure 1. The draft MDOD, v0.2.1 on the left is related to the I&M Architecture document components in the center and the proposed independent metadata sections on the right. The annotations apply to all categories, so those lines were omitted. The MDO Description in the proposed sections could also represent a derived product such as an analysis or graphical presentation of the measurement data.

Figure 2 depicts at a high level the draft schema we present as one possible representation of the MDOD. This schema uses three of the four independent sections shown in Figure 1 (Identification, Security, and Description) to describe each MDO or derived product contained in the collection described by the MDOD (both MDOs and derived products would be described by a "dataDescriptor"). In addition, there would be an identification section at the MDOD level that identifies the MDOD itself as a collection as well as a provenance section that describes the provenance of each dataDescriptor in the MDOD and the MDOD itself (including any MDOD(s) it was derived from).

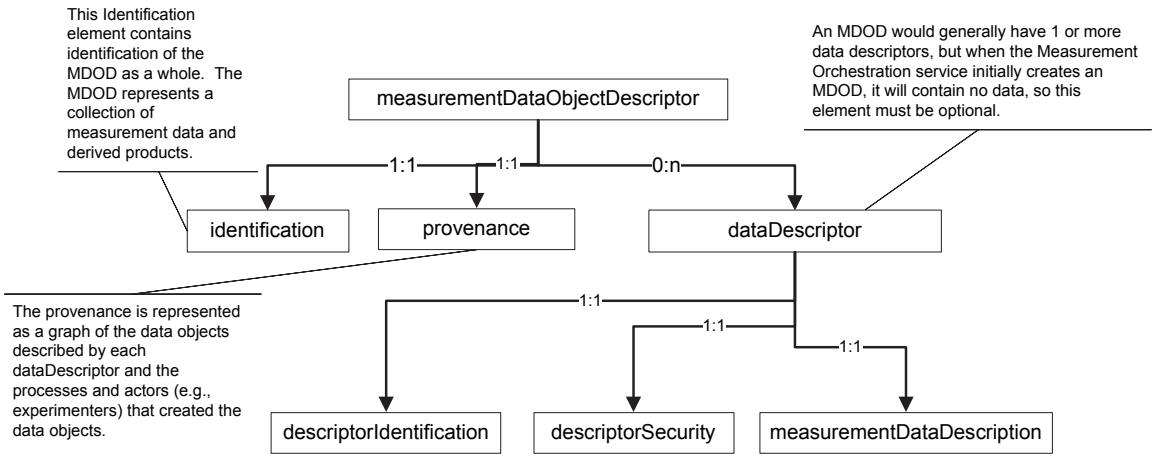

Figure 2. High-level view of the proposed MDOD schema changes. Each MDO or other data object is described by a dataDescriptor containing 3 of the 4 independent sections in Figure 1. At the MDOD collection level there is an identification section for the MDOD itself and provenance as an OPM graph for all of the dataDescriptors and the MDOD.

4. Schema Issues

We observed a number of issues with the schema that we list here.

a.) What does the primary identifier identify? At GEC11 it was decided that a primary identifier was needed. However, it was unclear from the discussion or documents whether this is an identifier for an MDO, a collection of MDOs, or the MDOD itself. In [5], the MDOD is described as the metadata for all entities involving measurements for an experiment. From that perspective, an MDOD is the metadata for a collection of MDOs that could be collected by a number of different sensors or MPs. The Measurement Orchestration (MO) service is the initial creator of the MDOD and assigns the primary ID [5], which is the mandatory identification for the MDOD.

Further, from this perspective, the primary ID is the unique identifier of the MDOD, which in turn is the metadata for a collection of MDOs. However, is it possible to assign a unique ID at this level? Based on [5], the MDOD is initially created by a Measurement Orchestration (MO) service and assigned a primary ID, but when published by the MO, the MDOD is initialized with only minimal metadata and is not immutable (it does not describe any measurement data at that point). The MDOD is published by the MO to a number of different users/destinations such as the User Workspace (UW) service of the experimenter, a measurement data archive (MDA), measurement collection (MC) services, measurement analysis and presentation (MAP) services, as well as other subscribers.

Under this scenario, users could end up with different subscriptions to the measurement data subsequently generated by MPs in that slice, different policies as to updating the MDOD, or perform different transformations on the measurement data (e.g., summarization or statistics for analysis and presentation). Although the MO initially creates the MDOD, it is the MPs that collect the measurement data, and when the measurement flow stops, the experimenter adds policies through the MP service [5]

and then publishes it to all subscribers. Depending on whether all subscribers are consistent throughout the lifecycle of the MDOD, and whether all subscribers consistently receive all of the published measurements, ***multiple versions of the MDOD (all with the same primary ID) could be available from different services.***

Further, in [5], each experiment initiated in GENI is identified by a name assigned by the GPO that includes the managing party and an integer ID. Since the experiment ID uniquely identifies the experiment the MDOD relates to, what additional information does the MDOD primary ID provide? Different subscribers could end up populating MDODs that use the same primary ID but contain different metadata content and represent different MD. The primary ID of the MDOD would identify the different versions of the MDOD (and the collections of measurement data they represent) as having started at the same MO, but is that of significance to a researcher later searching for the measurement data? ***To address this, we suggest:***

- ***(a) That the Measurement Orchestration (MO) service subscribe to the MPs generating the measurement data, populate the MDOD as a collection prior to publishing it, and assign the MDOD a unique ID at that point.***
- (b) ***The MP, MAP or any other service creating an MDO assign a unique ID to the MDO before making it available for inclusion in any MDOD. This will identify identical MDOs contained within different MDODs (e.g., two experimenters creating MDODs that include measurements from an aggregate provider or operator).***
- (c) ***Any subsequent additions to an existing MDOD collection be treated as a new MDOD and assigned a new ID, with the provenance tracing back to the prior MDOD. This would avoid the implication that MDODs are identical when they are not and also trace the provenance of who modified an MDOD.***

A related issue is that the same measurement data will be contained within different MDODs that have different primary IDs. In the example given in the Measurement Data Flows/Transfers slides from GEC11, an operator has set up measurements and the related MO would have created a primary ID. The measurement data represented by that MDOD is also shared with "experimenter B" who combines the MD from the operator with MD collected from their experiment. The MO for the experimenter will have created a separate MDOD with a different primary ID. The operator and the experimenter both put the MD through a MAP service that does some analysis, but in the case of the experimenter they will also have included MD generated in their experiment. From the viewpoint of a researcher later searching for data in a Digital Object Archive (DOA), does it matter that different digital objects, described by separate MDODs with different IDs contain some of the same initial MD collected from the same MP services, but then mixed it with different measurements? This issue is addressed by item (b) above and discussed below in the provenance issue as to the granularity of the provenance that should be tracked for measurement data.

b.) Is collection the right level of granularity for the MDOD? At GEC11, the I&M team discussed a set of use cases that expanded on [7] and illustrated different possible flows and transfers of measurement data. Some of the cases illustrated MPs providing measurement data to different Measurement

Collection (MC) services or Measurement Analysis and Presentation (MAP) services. Some of the MC and MAP services were subscribed to all of the MPs, while others only received measurement data from a subset of the MPs. In the cases illustrated, the different MAP services have different users (an operator vs. an experimenter) and the MDODs would be created by different Measurement Orchestrators (MOs). In this case, some of the same measurement data would be contained in collections described by different MDODs that were created by independent MOs. Should the measurement dataset generated by each MP be described by its own MDOD (or a subset of the MDOD) and be assigned its own ID? The experimenter or operator that collects the MD is considered the "owner" and sets the storage and disposal policy (and presumably sharing policy). In a case where multiple users subscribe to an MP, who is the owner? For example, in [7] where an experimenter collects MD from their experiment and also gets MD provided by an operator, the experimenter and operator could have different policies as to storage, disposal, and sharing. What policies would be allowed? Could an operator specify that their MD is freely available, but if combined with other data in a MAP service, any derived product is also freely available? ***We advocate for keeping the MDOD as a collection of the metadata descriptors describing a set of MDOs and possibly derived products, but there is an open issue as to how rights propagate. The MDOD (which is strictly metadata) could be public within GENI while the MDO itself or the derived analysis is private based on the rights assigned by the owner of that MDO. An MDOD could thus describe a set of MDOs that have different access rights.***

c.) Hierarchy of descriptors: In the draft MDOD design [6], the MDOD contains a single top-level descriptor and can then contain any number of nested descriptors. The goal of this design is that "one MDOD can be used to describe all of the data associated with one slice over a period of time." However, this seems restrictive in that if a descriptor were to represent the measurement data from a specific MP for some temporal window, the MD represented by these descriptors could be used to generate different derived products (such as a statistical analysis or graph by a MAP service). In such a case, a strict hierarchy does not cleanly represent the relationship between the descriptor for the MD and the different derived products it is used in. In feedback on the initial MDOD design, Jason Zurawski also raised some issues with having a single top-level descriptor and suggested that there could be multiple descriptors within an MDOD.

d.) Identification of definitive sources or enumerations: Multiple parts of the MDOD need to be populated based on enumerated lists of values. In the draft MDOD discussed at GEC11, this is acknowledged for many elements by comments as to where the definitive list is defined. Since this is a fast evolving field, it would be difficult to define an enumeration that would not need to evolve fairly frequently, so these fields could instead be populated based on controlled vocabularies that evolve over time. This would prevent the need for frequent modifications to the MDOD schema as the list of enumerated values evolve. To accommodate the use of multiple controlled vocabularies, additional elements should be included in the schema to identify the sources (vocabularies) for these definitions. Particularly for the measurement metadata (captured in the descriptor in the current design) the definitions could come from different sources as new tools are developed. The Trusted Computing Group (TCG), in its specification for endpoint metadata [2], allows for vendor-specific metadata, and

archives may be able to interpret such metadata but are not required to interpret it and are expected to ignore vendor-specific metadata they do not understand. A similar approach could be used to provide extensibility for new measurement tools in the MDOD, but being able to identify the definitive source for the definitions used in such extensions would make it more interpretable. The ability to define new attributes of a dataset and the source of those definitions is the approach that was taken in the widely used spatial metadata standard defined by the Federal Geographic Data Committee (FGDC) [11].

d.) Partitioning MDOD into sections that serve distinct purposes: In the draft from GEC11 there are three major sections to the MDOD, but the fields contained within them overlap as to their purpose. We recommend partitioning these elements into distinct sections as is common in other metadata standards. This was discussed in Section 2. As a starting point, we suggest identification, security, provenance, and measurement (descriptor). Since MD aggregated in the collection represented by the MDOD is likely to come from different MD owners as illustrated by the examples in [7], there would be an identification section for the MDOD itself and provenance could be captured as a graph at the MDOD level describing the lineage of the MDOD itself and the MDOs it describes. The security metadata, additional identification metadata, and metadata about the measurements themselves would be within the descriptors for each set of measurements (MDO).

5. Provenance, Lineage, and Dissemination of the MDOD

The holders section of the current MDOD design includes identification information as to the creator and owner of the MDO, provenance metadata as to where the MD came from, and forward-looking information as to who the MDOD or related MD was shared with. Other metadata standards, such as the ISO 19115 metadata standard for spatial metadata [4], document the provenance or lineage of the data. In the ISO standard, the Lineage section tracks the history of a data object based on the sources that were used in creating the object as well as the process steps that were involved in transforming the sources into the final data product. In the ISO model, as depicted by the illustration from Ted Habermann at NOAA in Figure 3, the output of one process step can be the input to one or more subsequent steps, and the relationships between sources and process steps are many-to-many:

Ted Habermann, NOAA January 31, 2011.
Available at: https://geo-ide.noaa.gov/wiki/index.php?title=File:ISOLineageModel.png

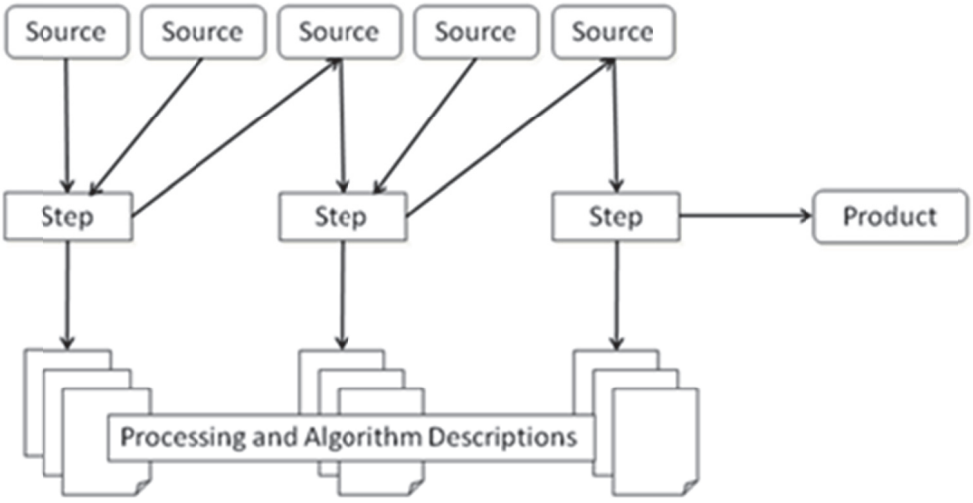

Figure 3. ISO 19115 lineage model.

The original ISO 19115 lineage model was extended in ISO 19115-2. The same source and process step framework is used in the extension but additional properties are captured that describe both sources and process steps.

The Open Provenance Model (OPM) [8], which is similar to 19115 Lineage and underlies the data model used in the GENI NetKarma project [9], has actors, processes and artifacts as its main components, where OPM processes and artifacts map to ISO process steps and sources. Both the ISO and OPM models are focused on provenance - essentially the trail of bread crumbs as to how a data product got to its current state. The OPM representation of provenance is as XML that forms a provenance graph, which in NetKarma can be visualized and played as a movie to see the sequence of processes taken by actors (holders) to use, create, or transform artifacts (MDOs). The draft MDOD design [6] addresses provenance in part through multiple "holder" elements that can contain "transaction" elements. The transaction indicates the action taken and there can be multiple transactions for a holder. In the draft MDOD, the holder maps partially to actors in the OPM. However, the process steps and sources found in the ISO model (or the processes and artifacts in OPM) do not map to the MDOD model. This may be in part due to the characteristics of measurement data and methods of collecting such data, but that is unclear. Some issues that need to be addressed in the MDOD holder and capturing provenance in the MDOD are:

- The transactions in the MDOD relate both to provenance and dissemination of the MDO. The provenance transactions include such actions as collection of the measurements and modifications to the measurement data. The holder also tracks the dissemination of the data through transactions related to the sharing or assignment of the data. In the case of measurement data available through a graphical web interface, providing access to the web

interface would result in a "sharing" transaction, so the holder may actually track potential dissemination in addition to the actual sharing of data. The metadata in the MDOD should capture the provenance of the measurement data (how it came to be in its current state), but the dissemination of the MD or the related metadata should be tracked by the source that is distributing or publishing the data and not in the MDOD.

- In the draft MDOD, there is a single "first" holder which identifies the collector of the data who is considered to be the "owner" of the data object. However, in the example depicted in [7] for the flow and transfer of MD, the experimenter has combined MD that they collected in their experiment (for which they are the owner), with MD that was collected by an operator (who would own that data). In the case of the MDOD describing the MD collected from these two owners, there would need to be the ability to associate each set of measurements with their respective owners and the policies of those owners.
- The holder contains policy information regarding the collection, anonymization, sharing, and disposal of measurement data. Currently the design tracks this by holder, but these policies would be set by the original owner who collected the data and it would not seem that a subsequent holder could modify those policies.
- In the draft MDOD, there is an ordering to the holders, and this ordering is opposite the hierarchy for the descriptors in that the first holder in the ordering of holders is the creator of the data, but the hierarchy of descriptors would need to have the derived data product at the top of the hierarchy with the constituent data products used being contained within it.
- The most significant issue for the holder is whether it is tracking the provenance that is most relevant to the MDOD. Aside from the issue of whether the holder should capture only provenance (historical) metadata or also the potential dissemination of the data, the current draft MDOD tracks the sequence of individuals who have possessed the data - regardless of whether they created or modified the data. This tracking of holders based on possession more closely tracks the concept of provenance as used in the library sciences or the art world than the concept of provenance used in scientific metadata. In the library sciences, the determination as to the authenticity of a document depends on being able to track the provenance of the document from the standpoint of control or ownership. In scientific metadata, as illustrated by the Open Provenance Model and the ISO 19115 Lineage model, the focus of provenance is on the data sources that went into a data object and the processes that were used to transform the sources into the resulting data product. Both the OPM and ISO 19115 focus on how a data product was transformed into its current state instead of on the chain of ownership. Unless there is a GENI-specific need that makes the chain of possession important, the MDOD should instead focus on the data sources and processes that resulted in the current state of the measurement data represented by the MDOD.
- In the draft MDOD model, the provenance chain is determined based on an "order" element within each holder, with the creator of the MD being assigned an order of 1. The measurement data objects are separately arranged as a hierarchy of descriptor elements. Other than that holder #1 is the originator of the MD, and that both the descriptor and holder contain the slice ID, there is not a connection between descriptors and holders. An approach similar to that used

in the OPM could be adapted for use in the MDOD. Descriptors in the MDOD represent MDOs, which are similar to artifacts in the OPM and would be assigned a unique ID. Descriptors that built on other sources described within the MDOD (e.g., a MAP service's analysis based on MD from multiple MPs), would have edges in their provenance graph that link the source MDOs used and the processes that performed the analysis. Similar to processes in the OPM, actors (holders) would only be included in the MDOD to the extent that they created or transformed the MDOD itself or data described by the MDOD (they created or modified an MDO).

- When is an MDOD complete or immutable? If a service transforms the measurement data, or creates an analysis from multiple data sources, could that process use measurement data described by multiple existing MDODs? In such a case, should such a transformation be described by a new MDOD that references the existing MDODs or should the new MDOD incorporate the subset of descriptor metadata that describes the measurement data used from each of the datasets described by their existing MDODs?

We would advocate capturing the provenance of the collection of MDOs (and derived data products) described by an MDOD as an XML provenance graph of the MDOD itself and the MDOs it describes. This would allow for a flexible representation of the relationships between the MDOs, the processes that created, used, or transformed them, and the GENI experimenters, operators, or aggregate providers that controlled the various processes that used, created, or transformed the collection of MDOs described by the MDOD. Since the graph would include the actors that modified or created an object (and not only the processes and data objects themselves), it can also describe the evolution of an MDOD based on an earlier MDOD collection as additional measurement data is added or analysis is performed that may even merge measurement data from multiple existing MDODs. Since dissemination of the MDOD does not change its state, the provenance captured would be the trail of bread crumbs – similar to the provenance captured in e-Science. The record of dissemination of measurement data should be tracked by the service disseminating the data.

6. Multiple Schemata

The architecture document discusses the MDOD data model as being available as three different schemata for different purposes: (1) File (XML?) schema to be used locally within a slice, (2) A schema for registering MDOs with a Measurement Information (MI) service, and (3) archiving external to GENI (Datcat, DataCite). One caveat (section 4.1.6 of the I&M architecture documentation) is that based on prior experience, sharing will only be done if it requires no effort and can be automated. If the archive schema is different from the collection schema it will require an automated transformation and cannot require additional metadata that is not either already in the collection metadata or that cannot be retrieved programmatically from an existing archive. An archive may disseminate metadata or data in different formats, such as in the OAIS model [10], where an archive stores data in an Archival Information Package (AIP) but can then distribute that data in different formats referred to as Dissemination Information Packages (DIPs). It would be beneficial to clarify if the MDOD model itself, which could be viewed as corresponding to the AIP in the OAIS model, needs to be in different formats or whether archives just need to be able to publish using different formats.

7. Measurement Data Being Documented

The Descriptor section of the draft MDOD captures metadata describing the measurement data and how it can be interpreted and accessed. In GENI, Measurement Point (MP) services instrument the GENI infrastructure (or the infrastructure specific to a slice), using link, node, and timestamp sensors that output measurement data using different, but standardized, schemata (presumably XML). These schemata include both the measurements and related metadata.

perfSONAR wraps individual monitoring services into web services that are instantiated as Measurement Points (MPs). The measurements are retrieved using the schema developed by the Open Grid Forum's Network Measurement Working Group (NMWG). In GENI Spiral 3, the GEMINI project will build on perfSONAR (including LAMP), INSTOOLS, and other GENI spiral 1 & 2 efforts to build an instrumentation and measurement framework for GENI. Although the Descriptor section of the MDOD should be extensible to accommodate future tools through an approach such as used in IF-MAP for vendor-specific extensions, the descriptor should be able to accommodate the metadata of Spiral 3 measurement tools in an interpretable manner to the extent possible.

8. Relation to Other Standards

The architecture documentation suggests using Dublin Core elements where appropriate. One issue that has been raised with Dublin Core is that while it lowers the bar to entry for capturing metadata, the general free-text elements in unqualified Dublin Core can result in inconsistent representations that are then difficult to search.

SensorML is another metadata standard that may give ideas that could be utilized in the MDOD. The architecture document discusses the capture of measurements from sensors in the network. Are there similarities that these network measurement sensors share with sensors in a physical environment that may be described by SensorML?

9. Suggested Future Steps

Based on the draft MDOD and architecture document, we suggest determining which metadata should be captured at the level of a collection of MDOs in the MDOD and which metadata should be captured at the level of each MP's dataset of measurement data in a descriptor. The metadata at each level should be divided into independent sections – we propose one approach above. The provenance of the MDOs and derived data products as well as the MDOD itself should be described as an XML graph at the MDOD level. The illustrated use cases discussed at GEC11 could be used to create MDODs that would then track through the process from creation by the MOs through to the MD being stored in MI services, MAP services, and DOAs. This mapping to the use cases will highlight any issues as to whether particular metadata is being captured at the correct level of granularity and other issues that would arise as the MD is transferred or aggregated and combined with other MD.

Extensibility of the descriptor: The extensibility of the descriptor in the MDOD for future measurement sources or tools could be accomplished in a manner similar to the IF-MAP approach of allowing vendor-

specific extensions for capturing metadata for network end-points. Additionally, instead of using enumerations for elements such as the format or object type, elements could use values based on controlled vocabularies where the URI of the definitive source for the controlled vocabulary is captured as a "source" attribute of the element. To allow for measurement data being captured from multiple sources, and "owned" by different creators, the MDOD descriptor should not use a strict tree hierarchy, but should instead be described in a provenance graph. The relationships between descriptors would be represented as edges in the graph instead of a strict hierarchical relationship.

Provenance: The draft MDOD uses the holder to track the chain of ownership with less of a focus on transformations and the state of the MDOs described by the MDOD. The I&M group should determine if measurement data is different from other scientific data such that it requires a different approach to provenance or whether a focus on data products (MDOs), processes, and actors similar to the OPM could be adapted for provenance in the MDOD.

A provenance graph with edges linking the sources for data products, the processes that created the data products, and the actors that triggered the processes is more flexible than the ordering of holders in the draft MDOD. Using a similar approach in the MDOD would more easily accommodate measurement data that has been gathered from multiple sources (e.g., a slice's experiment and from an operator) such that a strict ordering of the holders is not possible (the creator of each source of measurement data could be considered the first holder).

Schema: We are working on a draft schema that follows the structure proposed in Figure 1. This schema modifies the previous MDOD draft by creating four independent sections including a new security section and reassigning elements between sections. Additionally, at the top level, we propose having identification and provenance section for the MDOD collection of data objects as a whole and an unbounded number of "dataDescriptor" elements that would each describe each data object contained in the MDOD collection. Three of the four main sections would then be child elements contained within each dataDescriptor element.

References

[1] GENI Instrumentation and Measurement Architecture Version 1.0, Document ID: GENI-SE-IM-ARCH-1.0, December, 20, 2010.

[2] TCG Trusted NetworkConnect, TNC IF-MAP Binding for SOAP, Specification version 2.0, revision 36, July 30, 2010.

[3] GENI Instrumentation and Measurement priority topics: GENIN I&M Use cases. Available at: http://groups.geni.net/geni/wiki/GeniInstMeas

[4] International Organization for Standardization, "Geographic Information – Metadata," (ISO19115:2003), 2003.

[5] D. Gurkan and A. Arun Daga, Architecture and Data Structure Mapping of GENI I&M to IF-MAP, University of Houston, September 28, 2011.

[6] Instrumentation and Measurement Group Draft MDOD Data Model, Elements, and Values, Version 0.2.1, July 25th, 2011. Last accessed 2011-10-28: http://groups.geni.net/geni/attachment/wiki/GEC11InstMeasWorkingSession/072511_ver0.2.1_MDOD_DataModel.txt

[7] MD Objects and Descriptors Presentation at GEC11. Last accessed 2011-10-29: http://groups.geni.net/geni/attachment/wiki/GEC10InstMeasWorkingSession/031711_MDObjectsDescriptors%20_I%26MArch_Overview.pptx

[8] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. Van den Bussche, The Open Provenance Model core specification (v1.1), *Future Generation Computer Systems*, vol. 27, issue 6, June 2011, pp. 743-756.

[9] NetKarma: GENI Provenance Registry. GENI wiki project page: http://groups.geni.net/geni/wiki/netKarma

[10] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," CCSDS 650.0-B-1, BLUE BOOK, January 2002. Available at: http://public.ccsds.org/publications/archive/650x0b1.PDF

[11] Federal Geographic Data Committee, Washington, D.C., "Content Standard for Digital Geospatial Metadata Workbook," Version 2.0, 2000.