University of Tennessee, Knoxville
Trace: Tennessee Research and Creative Exchange

DataONE Sociocultural and Usability & Assessment Working Groups

Communication and Information

10-20-2011

Architecture Documentation for NonTechies

SCWG

Follow this and additional works at: https://trace.tennessee.edu/utk_dataone

 Part of the Library and Information Science Commons

Recommended Citation
SCWG, "Architecture Documentation for NonTechies" (2011). *DataONE Sociocultural and Usability & Assessment Working Groups.*
https://trace.tennessee.edu/utk_dataone/166

This Creative Written Work is brought to you for free and open access by the Communication and Information at Trace: Tennessee Research and Creative Exchange. It has been accepted for inclusion in DataONE Sociocultural and Usability & Assessment Working Groups by an authorized administrator of Trace: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

NOTES: Rob P mentioned looking at range of datasets, also science vignette success stories. Rob P also identified what I've labeled as Purpose 2.

PURPOSE 1(Institutional orientation): to provide a quick overview of DataONE including the benefits/costs of being in DataONE, technical specifications and architecture.

PURPOSE 2 (Individual user orientation): Pull in people quickly, provide quick "users guide"

INSTITUTIONAL ORIENTATION – libraries, data centers
Audience: administratively focused people who mayor may not be technically literate but at the least have technically proficient advisors
Format: 1-2 page executive summary with bullet points. White paper 10-15 pages
Issues to address:

- Why & how to be a MN
- Cost
- Benefits
- Who needs to be involved
- Level of commitment
- Architecture: Resource focus

INDIVIDUAL ORIENTATION – users
Audience: scientists
Format: 1 page executive summary, 5 page user guide
Issues to address (Per Rob P)

- ½ page of what D1 can do for them- 2 quick science cases
- simple way to get started
- basic intuitive instructions to get started (find, search, mount dataset)
- Tree with types of data and how they come together
- Timeline on rollouts at a high level

INSTITUTIONAL ORIENTATON (Written for Academic Library Director)
THE EXECUTIVE SUMMARY

What DataONE means for your Library

DataONE provides your library with a framework to support open, persistent, robust and secure access to well described and easily discovered data about life on earth and the environment that sustains it. This includes biological data from genome to ecosystem including data from atmospheric, ecological, hydrological and other earth sciences. Becoming a member node of the DataONE network provides your library with the framework to store and provide access to your library's digital scientific holdings and also to make these data available to a much broader community. Your library may host its own data, or may partner with other data holders. Your library may also use a variety of standards or formats to describe and store your data.

As a DataONE member node your library would report its metadata to a DataONE coordinating node**,** which choreographs services that help users store, discover, and retrieve data. The coordinating node provides a host of other services including facilitating replication and preservation of data objects, authorization and authentication, and data discovery. In essence DataONE is a partner in helping to facilitate your library's implementation of a data management strategy, and to showcase your institution's intellectual capital.

DataONE is also developing a suite of tools for your scientists, researchers and students. The Investigator's Toolkit (ITK provides your users with powerful tools for finding, using and storing data on the DataONE network. The DataONE ITK already includes tools for data searching, desktop data file management, and data integration and analysis.

Potential Benefits to your library

- Improved methods of access to your data, as well as wider access to and exposure of those data
- Cost-effective preservation of your data and metadata
- Access to a larger community with expertise and best practices in data life cycle management
- A toolkit for analysis, visualization and modeling your data
- A response to increasing demands by funding agencies for long-term data management planning
- Visibility as a community leader in supporting open and rapid access to scientific data
- Recognition and credit for data and services provided through data citations in published literature
- Opportunities for collaborative research
- Potential funding opportunities resulting from data discovery
- Showcase for your institution's intellectual capital

Roles and Responsibilities for your library

- Provide access to actively managed data
- Allow replication of data for preservation and increased access
- Provide descriptive metadata to Coordinating Nodes to facilitate data discovery, access and usability
- Ensure availability and reliability of physical data access
- Engage with the larger DataONE community to advance DataONE services, support emerging Member Nodes, and promote best practices in data management.
- Utilize a unique identifier for each dataset by participating in DataONE user identity federation
- Deploy an implementation of DataONE APIs (Application Programming Interfaces
- Have an identifier system to uniquely identify data items within the Member Node.

Costs

Estimating the cost for implementing and deploying as a DataONE member node involves many variables including the existing technology of the underlying data repository, services it may already expose, the structure of data and metadata being used, and the technical expertise available. A rule of thumb would be that two months of development time should be sufficient for most libraries that already have a repository of some kind to support member node services. In many cases the effort would be much smaller, and as more examples become available the cost for implementation should show a corresponding decrease.

Technical Requirements

Bandwidth: Your library's bandwidth level should be at least on a local 10base-T connection (10 Mbps) to a local backbone operating at T1 or higher rates (about 1.5 Mbps). This level of bandwidth should allow a gigabyte file to be transferred in approximately 2-4 minutes. DataONE recognizes that network capabilities around the world vary (e.g., Internet backbone bandwidth and network reliability vary between countries), and these expectations are guidelines not fixed requirements.

Storage capacity: Ideally, your library will participate in the cooperative replication scheme supported by DataONE which means you will store copies of data from other member nodes. A good rule of thumb is that your library should expect to provide additional storage capacity equivalent to the capacity dictated by replication expectation for your own data. For example, if you bring 100GB of data to the DataONE network and its replication policy is there should be two copies of each dataset somewhere in the DataONE network, that Member Node should in all fairness expect to provide 200G of storage capacity over and above the original 100GB. DataONE recognizes that some libraries may have extreme resource limitations and so this is not an absolute requirement for participation in DataONE.

Connectivity: Reliable, high-capacity connectivity contributes to good performance, and DataONE expects that your library will be available via the network 99.9% of the time. This level of availability provides for about 8.5 hours of unscheduled downtime per year.

Status checks: Part of the core DataONE operations is a testing service that runs a tests to check conformance with the API specifications as well as stress tests to gauge performance. The service generates a report detailing the overall functionality. The same test will be performed on a regular basis to ensure that your library's node is operating as expected. This service performs routine status checks to detect if your node becomes inaccessible, and automatically updates routing information to avoid directing clients to attempt connecting to your library's node when it may not be accessible.

Who Needs to Be Involved

[This section will note who the library needs to involve. What level of expertise does the tech person need? Et.]

THE FULL REPORT

DataONE has a core infrastructure that supports content replication, identifier resolution, content discovery and retrieval, and a federated identity infrastructure. The Investigator Toolkit also contains several components widely used by the community. This paper provides a high level overview of the DataONE architecture to inform those who are making the decision to implement a DataONE member node at their library. Full documentation of the services and tools is available at http://mule1.dataone.org/ArchitectureDocs-current/index.html

Data and Metadata: DataONE defines data as a discrete unit of digital content that represents information obtained from an experiment or scientific study. The data is accompanied by science metadata, which is a separate unit of digital content that describes properties of the data. Each unit of data or science metadata is also accompanied by a system metadata document that contains attributes that describe the digital object it accompanies (e.g. hash, time stamps, ownership, relationships). Science metadata may be in a variety formats which are outlined in the referenced document at http://mule1.dataone.org/ArchitectureDocs-current/design/WhatIsData.html Operationally, science data and metadata are stored on a member node and a copy of the science metadata is held by the coordinating node in order to facilitate the discovery process as people search for content. The data package contains the data, science metadata and system metadata. [ADD FIGURE HERE]

System metadata: System metadata is used by DataONE to track and manage objects across the distributed network of coordinating and member nodes. It is considered operational information and is needed to run DataONE. It is critical to ensuring that science data, science metadata, and other digital objects stored in DataONE are discoverable, accessible, auditable, verifiable, and are associated with meaningful related digital objects. Digital objects in DataONE must also be viable for the long term - for many decades - and so the system metadata must also include provenance information.

System metadata describes managed objects such as science data, science metadata, and resource map objects by identifying low level elements (e.g. size, type, owner, access control rules) and the relationships between objects (e.g. describes and describedBy). System metadata have some elements that are provided by the client software including the low level elements. Other elements are generated by DataONE during the course of managing objects and include information about provenance and replication. System metadata is maintained by the coordinating nodes and reflects the current state of an object in the system. Full documentation about system metadata is at http://mule1.dataone.org/ArchitectureDocs-current/design/SystemMetadata.html
Details about the generation of system metadata associated with a data object and its associated science metadata is at http://mule1.dataone.org/ArchitectureDocs-current/design/SysmetaLifecycle.html

System metadata is indexed to support access control rules for search results, support efficient discovery of objects by their properties, and augment the science metadata indexes. Three indexes support this: the permission index, system index and discovery index. Complete information about the indexing process can be found at http://mule1.dataone.org/ArchitectureDocs-current/design/indexing_systemmetadata.html

More details about preservation metadata related to system metadata is at http://mule1.dataone.org/ArchitectureDocs-current/design/SystemMetadataAnalysis.html

Identifiers: Identifiers are handles utilized as a mechanism to specifically identify the individual objects (documents, data sets, data records) they manage. Identifiers refer to objects in DataONE. Initially data and science metadata documents describing data have identifiers. The definition of an individual data object may be a single record within some larger collection, or may refer to an entire set of records contained within some package. Full documentation on identifiers is at http://mule1.dataone.org/ArchitectureDocs-current/design/PIDs.html

Uniqueness of identifiers in DataONE is largely under the control of the Member Nodes, with the requirement that an existing identifier can not be reused. DataONE treats the original identifier (i.e. the first assignment of the identifier to an object that becomes known to DataONE) as the authoritative identifier for an object.

Identifiers utilized by Member Nodes can take many different forms from automatically generated sequential or random character strings to strings that conform to schemes such as the LSID and DOI specifications. DataONE does not directly utilize implied functionality and services that might be available for some of the identifier schemes. The DataONE infrastructure and services operate independently of such external services, with no functional dependency. An identifier that has been registered by DataONE will always refer to the same set of bytes in the case of data and metadata objects. Generation of other representations of the objects may be supported by services (e.g. an image may be transformed from TIFF to JPEG), but the identifier will always refer to the original form.

A fundamental goal of DataONE is to ensure that any identifier utilized in the system is resolvable, that is, DataONE provides a method that will enable retrieval of the bytes associated with the object (the actual object bytes or metadata associated with the object). Resolution is handled by the Coordinating Nodes through the CN_crud.resolve() method of the CRUD API. A guarantee of resolvability is a fundamental function of the DataONE infrastructure upon which many other services may be constructed, both within DataONE and by third party systems.

DataONE preservation strategy: DataONE's preservation goal is to protect the content, meaning, and behavior of data sets registered in its global network of member nodes. This a complex undertaking that warrants a layered, prioritized approach. To get started on a solid footing, our first objective was to build a platform that immediately provides a significant degree of preservation assurance and makes it easy to add more sophisticated preservation function over time. Initially, DataONE's focus is on preventing loss due to non-malicious causes, such as,

- Technological obsolescence (e.g., loss of support for rendering software and hardware),
- Accidental loss (human error, natural disaster, etc.), and
- Financial instability (loss of funding).

To meet the objective of "easy, secure, and persistent storage of data", DataONE adopts a simple three-tiered approach.

1. **Keep the bits safe.** Retaining the actual bits that comprise the data is paramount, as all other preservation and access questions are moot if the bits are lost. Key sub-strategies for this tier are (a) persistent identification, (b) replication of data and metadata, (c) periodic verification that stored content remains uncorrupted, and (d) reliance on member nodes to adhere to DataONE protocols and guidelines consistent with widely adopted public and private sector standards for IT infrastructure management.
2. **Protect the form, meaning, and behavior of the bits.** Assuming the bits are kept undamaged, users must also be able to make sense of them into the future, so protecting their form, meaning, and behavior is critical. In this tier we rely on collecting characterization metadata, encouraging use of standardized formats, and securing legal rights appropriate to long-term archival management, all of which supports future access and, as needed to preserve meaning and behavior, format migration and emulation.
3. **Safeguard the guardians.** The DataONE network itself provides resiliency against the occasional loss of member nodes, and this will be shored up by succession planning, ongoing investigations into preservation cost models, and open-source software tools that can sustained by external developer communities.

The implication for your library is that content should be stored in widely-used formats that are expected to be readable for many years into the future. Although DataONE replicates content to provide bit-T level preservation services and will provide ongoing support for some widely-used data formats, DataONE does not guard against the obsolescence of poorly-supported data formats.

Information about preservation metadata related to system metadata is at http://mule1.dataone.org/ArchitectureDocs-current/design/SystemMetadataAnalysis.html

[More information about thee DataONE preservation strategy is at http://mule1.dataone.org/ArchitectureDocs-current/design/PreservationStrategy.html]

DataONE Cyber security: Cyber security for DataONE is part of the initial system planning and design, thereby providing solid integration to all services. The design is based on the fact that DataONE is a virtual organization composed of the collaboration of researchers, data providers, institutions, coordinating nodes, member nodes, data collections and other infrastructure components. Therefore, DataONE, as an entity, spans many physical organizations and administrative domains. The primary goal of DataONE cybersecurity is to protect the data and metadata collections that those organizations and administrative domains contribute, as well as their infrastructure and the DataONE user community.

DataONE recognizes that your library must simultaneously meet local requirements and must also integrate into the DataONE cyberinfrastructure, and has designed cybersecurity management that is flexible and can support the very different needs of each member node.

The bottom line is that DataONE security is focused on protecting data in the member node at your library from intentional and non-intentional harm; for example, unauthorized viewing of private data and or alteration or deletion of another user's data. System operations are also protected from malicious activity that often occurs on the Internet.

Security in DataONE consists of two processes: 1) "authentication" which verifies the identity of users and 2) "authorization" which applies rules that govern access to system resources (e.g., data).

Authentication is based on the CILogon proect which allows for the use of user identities from academic and commercial institutions in the us that are members of the InCommon federation or through more globally accessible identity providers like Google, Facebook, and Yahoo!. The advantage of this approach for your library's users is that they will be able to access DataONE resources without yet another identity to manage. It also means your institution has full control over who may view or modify their data by defining your own local authorization rules, which will be propagated and enforced by DataONE. Authentication details can be found at: http://mule1.dataone.org/ArchitectureDocs-current/design/Authentication.html

Authorization is the process for confirming whether a user has privileges to access a resource in DataONE. It provides privacy and access control to user contributed data and metadata, as well as protects DataONE services and resources at both the coordinating and member nodes. In order to serve the many administrative domains and political boundaries represented throughout DataONE member and coordinating nodes, a "trust" relationship is crucial. Therefore access control rules that are defined by your library are honored by other DataONE members. The

language that specifies the policy for a given access control rule dictates only whether a user is allowed access to a given resource. Including the ability to explicitly deny access to a resource overly complicates managing the authorization process and is seldom used in practice. Rules will consist of the system identity of the user, the type of permission granted (e.g., read, write, and or execute), and the identifier of the resource being requested. DataONE will provide, where reasonable, a conversion of the internal access control rule to a subset of one or more industry standard policy languages to support interoperability between different organizations

Further details about authorization are at:
http://mule1.dataone.org/ArchitectureDocs-current/design/Authorization.html

The cybersecurity posture of DataONE will evolve over time both because of the continuing maturation of DataONE operational strategies and because of an ever-evolving cybersecurity landscape.

Content Discovery: The coordinating nodes have primary responsibility for supporting searches for content. This is accomplished using a web browser interface that connects to a SOLR index at the coordinating node. The SOLR index content includes system metadata, science metadata and resource maps. The basic model is that each index entry represents information about a single identifier (PID). For PIDs that refer to a science data object, the index entry will be constructed from system metadata and resource maps that reference the object. If the PID refers to a science metadata object, the index entry will have fields populated with content extracted from the science metadata document. There is also a requirement that search results only contain information that the user has permission to read. This means tat access permissions for each item min the search is examined. Technical information on how this process works is at
http://mule1.dataone.org/ArchitectureDocs-current/design/search_auth.html

Your library also benefits from this system since the content in the SOLR index can be used to created data citations that follow established best practices. This provides a means for your library to get credit for providing access to the data. A full description of the SOLR index, how it is populated, and how it can be used to support programmatic search of the DataONE holders can be found at
http://mule1.dataone.org/ArchitectureDocs-current/design/SearchMetadata.html

Event logging and reporting: The DataONE logs various interactions and operations in the system to provide operational status information about the entire system, to report on specific node operations, and to inform DataONE participants (users, contributors, administrators) about their specific domain of interest in the system. For example, your library might like to monitor use of your data and where it is being replicated to. Source: http://mule1.dataone.org/ArchitectureDocs-current/design/logging.html

Implementation: Your library can participate in DataONE as a Member Node through three different strategies:

1. Native service support: Implement the DataONE Member Node service APIs as part of the existing data repository application interfaces. Native Member Node implementations currently exist for Metacat, with Dryad close to completion, and support for Fedora Commons in the works.
2. Proxy service: Utilize a proxy service that offers a translation layer between the DataONE service interfaces and existing service interfaces exposed by a repository. A proxy service (the "Generic Member Node" or GMN) written in Python and based on the Django web framework is also available with implementation examples for Dryad and ORNL DAAC.
3. Standalone service: Deploy a standalone Member Node software stack and perform custom synchronization of content between the Member Node data store and the existing data repository. The GMN can also operate as a flexible standalone Member Node service.

DataONE has implementation examples of each that can be used directly or as examples to assist with new implementations. DataONE has a liberal open source policy, so all source code can be re-purposed as necessary to assist with new deployments.

In general there should be little to no impact on your existing operations. Existing data repositories can continue to operate in parallel with the Member Node services. Aspects that may see some impact include restriction on identifiers being used (to ensure global uniqueness), how the repository works with content that may be replicated to it from another Member Node, and there will be some additional connectivity demand to support the necessary network communications.

If your library should choose to no longer be part of the DataONE network, decommissioning is technically as simple as turning the node off. However a more controlled approach however is desirable and would ensure that the data remains in a stable state. At least a week before the shutdown, your library's member node administrator should inform DataONE of the impending change. The Coordinating Nodes will perform the synchronization and ensure that data objects are replicated as necessary. All content that was replicated up to this point will continue to be available through the coordinating node and other member nodes.

[Reference https://docs.dataone.org/member-area/planning-for-dug/dug-general-documents/member-node-documents-final-versions/MNTechnicalDoc2010Dec03.docx/view]

DataONE APIs: DataONE has application programming interfaces (APIs) available for the coordinating nodes, member nodes and investigator toolkit. A Representational State Transfer (REST) approach over HTTP is used to implement the coordinating and member node APIs. [NOTE THAT THE DOCUMENTATION DOESN"T HAVE AN INTERNAL STRUCTURE TO SUPPORT A POINTER TO KEY INFO AT THE MOST GENERAL LEVEL THAT INCLUDES POINTERS TO SUB AREAS]

Member node APIs provide mechanisms to report the level of service compliance and to specify replication policies. There are also APIs to monitor the current state of the DataONE infrastructure and to track the current operation state of associated member nodes. Member node APIs include the core API, Read API, Authorization API, Storage API, and Replication API. Complete information is at http://mule1.dataone.org/ArchitectureDocs-current/apis/MN_APIs.html

Coordinating node APIs include the the core API, Read API, Authorization API, Storage API, and Replication API. Complete information is at http://mule1.dataone.org/ArchitectureDocs-current/apis/CN_APIs.html

Investigator Toolkit APIs [NO DOCUMENTATION] Complete information is at http://mule1.dataone.org/ArchitectureDocs-current/apis/ITK_APIs.html

Time and Bandwidth Constraints: Because the DataONE architecture relies on the performance of the coordinating nodes, the transfer rate between coordinating nodes was tested to establish the rates of data acquisition, the size of data objects, and the number of simultaneous users supported.

This means your library's member node should have individual data objects that are small enough to be easily transferred across the network to other nodes. Data objects on the order of Megabytes or (low) Gigabytes are reasonable, while data objects on the order of Terabytes will not work well.

Additionally, your library's bandwidth should be at least a local 10base-T connection (10 Mbps) to a local backbone operating at T1 or higher rates (about 1.5 Mbps). This level of bandwidth should allow a gigabyte file to be transferred in approximately 2-4 minutes.

Full documentation regarding these constraints can be found at http://mule1.dataone.org/ArchitectureDocs-current/notes/time_bandwidth_constraints.html

The Investigator Toolkit: The Investigator Toolkit provides a suite of software tools that are useful for DataONE users. The tools categories are (1) web portals and tools, (2) metadata and data management tools, (3) analysis and modeling tools, and (4) DataONE libraries. Documentation about the investigator toolkit is at http://mule1.dataone.org/ArchitectureDocs-current/design/itk-overview.html

Web portals and tools: NEED TEXT TO DESCRIBE AND NOTE CURRENT STATE

Metadata and data management tools: NEED TEXT TO DESCRIBE AND NOTE CURRENT STATE

Analysis and modeling tools: Quantitiative analysis is at the center of all science, and represents the main point where scientists utilize data and produce derived data products. Thus, the analysis and modeling tools in the Investigator Toolkit will represent the most important suite of tools for accessing data from DataONE and contributing results to the DataONE Member Nodes. [NEED MORE]

DataONE libraries: NEED TEXT TO DESCRIBE AND NOTE CURRENT STATE

License and copyright policy: DataONE is developing software and distributing software that originates from multiple institutions that each have their own policies governing copyright and licensing of software. As a project, we are committed both philosophically and through our cooperative agreement with NSF to distribute our work as open source software. However, we must do so within the guidelines provided by each of our institutions.

DataONE software products have been and will continue to be jointly designed and created by individuals spread across the DataONE participating institutions. No one individual or institution claims creative ownership of any of the DataONE software products. Therefore, the general policy is that each DataONE software product contains a copyright statement that indicates that all of the participating organizations hold joint copyright in the work.

DataONE is committed to open science and open source principles, so all software developed will be licensed under an approved open source license. Because different open source licenses can have implications for the re-use and distribution of that software, we do have a general policy that guides choice of licenses. However, we also recognize that existing software that we may wish to use may already contain existing open source licenses that can not be changed easily or at all, and we do not want as a matter of policy to exclude the use of these open source products. We recognize the benefits of various open source licenses, but as a general policy DataONE will use the Apache 2.0 open source license for all newly developed code: http://www.opensource.org/licenses/apache2.0.php.

Each DataONE product includes a copyright and license statement in a file at the root of the source and binary distributions and in each source code file that reads:

This work was created by participants in the DataONE project, and is jointly copyrighted by participating institutions in DataONE. For more information on DataONE, see our web site at http://dataone.org. Copyright 2010 Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at http://www.apache.org/licenses/LICENSE-2.0 Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Each DataONE product also includes a copy of the Apache 2.0 license at the root of its source code tree in order to make the license easily accessible to people that receive the product.

Complete information regarding license and copyright policy is at http://mule1.dataone.org/ArchitectureDocs-current/license_and_copyright_policy.html

Glossary: DataONE developers have created a glossary of all terms used in the architecture documentation. This glossary also points to relevant use cases. The glossary is available at http://mule1.dataone.org/ArchitectureDocs-current/glossary.html

NOTE: Do we want to point people to the user scenarios and use cases? If so how?

TOPICS FROM ARCHITECTURE DOC THAT AREN'T IN THIS DOCUMENT

- Querying DataONE : seemed out of the scope of this document
- Supporting Online Citation Managers through COinS : there wasn't enouch content here for me to understand the outcome
- Serialization of Types for Transfer Over HTTP : seemed out of the scope of this document.
- NodeList seemed out of the scope of this document.