Date of Submission ___________
Derived from ________________
Declassify on ________________

METADATA REQUIREMENTS IN A MULTI-LAYERED NETWORKED ENVIRONMENT

Anthony W. Isenor
DRDC Atlantic
9 Grove Street, Dartmouth, Nova Scotia
Canada B2Y 3Z7

ABSTRACT

In future coalition operations, data assets originating from coalition partners will be made available to the broader networked community. It is expected that the ability to access networked data assets will improve the speed and accuracy of Common Operating Picture (COP) formation. However, in such a networked environment there are issues related to the discovery and utilization of data assets when a multitude of similar assets are available. Metadata will play a key role in both the discovery and utilization processes. However, the metadata requirements are not consistent through all levels of the process. At the local level, metadata exists to support application level connection and interaction with the asset. This type of metadata may be particular to the application or in-house system. However, when the asset is exposed to the broader community, the metadata requirements change to being related to the discovery of the asset. After discovery, metadata requirements become related to the utilization of the data asset. This includes the necessary information to establish connection to the asset but also the semantic information required to understand the data being provided by the asset.

This paper describes the metadata requirements for the connection of an asset to an information grid. This work focuses on Canadian efforts addressing a Technology Demonstration Project (TDP) in Networked Underwater Warfare (NUW). The NUW data model utilizes metadata efforts from the international military and marine communities.

INTRODUCTION

The networking of systems is receiving considerable attention as national militaries begin to research the net centric paradigm. Although the exact nature or structure of the paradigm is open to debate, many countries are moving forward by first attempting to define their own interpretation of net-centric operations.

In Canada, the term used to describe the networked paradigm is Network Enabled Operations (NEOps) [1]. Here, NEOps is considered an approach to military operations, where the sharing of data is enabled by the culture, technology and practices of the community. This suggested definition notably includes both systems and people, working in a slightly different fashion from the traditional military operation.

The DRDC Networked Underwater Warfare (NUW) Technology Demonstration Project (TDP) [2] is researching potential changes in effectiveness of operations using one possible NEOps model. In particular the NUW project is interested in the speed and accuracy of the formation of a Common Operating Picture (COP) given the networked data assets. In the NUW project, networked platforms include two surface ships (one with a towed array), a maritime patrol aircraft deploying sonobuoys, a reach-back cell and possibly a submarine. Key systems existing on the platforms will be linked over an Internet Protocol (IP) network to provide a common understanding of the available data from the platforms.

One key requirement of the NUW project relates to the mutual understanding of data items between the platform systems. These systems were developed independently and as such do not necessarily utilize the same data space or terminology. Describing the data requirements and assets in such a way to provide semantic compatibility is one challenge to this demonstration project.

CONCEPTUAL MODEL

Conceptually, the environment established for NUW consists of numerous activity cells (Figure 1), where each cell corresponds to a coherent entity such as a ship, aircraft or reach-back cell. In this model, each cell makes available the resources within the cell. The resources are made available using services that provide access to key data or information from the cell.

Services at the cell can programmatically be implemented in various forms. An individual form, or agent [3], provides the network of users with access to local resources. In this model, local resources include computing resources and data assets.

Individual data services and descriptions of these services are maintained at each of the individual cells. This allows management at the individual cell to manipulate the agent providing the service while maintaining consistent description and operation of the service.

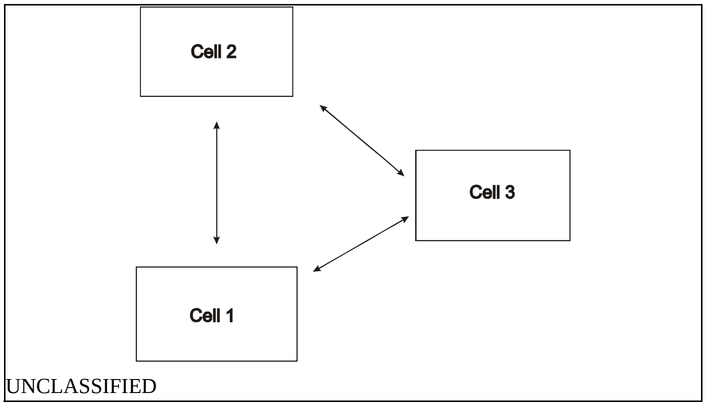

Figure 1. Cells are used to describe disparate entities within the network.

In the NUW project, the emphasis is on providing data assets to the coalition via specialized services. These services provide the ability to find, understand, extract and utilize the data assets from the cell. For example, the common subscribe/publish mechanism may be one agent implementation of a service.

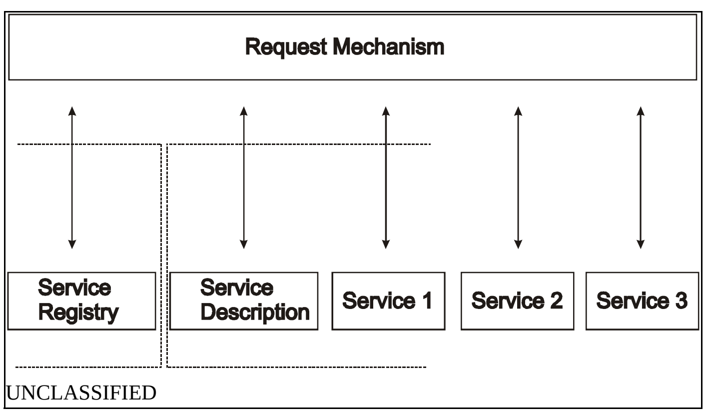

Figure 2. Requests to the service registry identify individual services. Service descriptions provide details on how to interact with a service.

One or more cells are also charged with providing a centralized repository for all available services within the network. Cells entering the coalition may make requests to the central repository, thereby identifying services of interest to the requesting agent (Figure 2). The central repository may be located on one of the cells providing data, or alternately a platform that provides the function of a central repository.

METADATA DEFINITION

Metadata is at a sufficient level of abstraction to be somewhat difficult to understand. Many groups and organizations describe metadata as ‘data about data’. However, such a definition typically confuses the issue making it difficult to distinguish the data from the metadata.

Perhaps a better definition is that metadata are *the values of characteristics that qualitatively or quantitatively describe or support a resource*. This definition provides several advantages over the more traditional definition.

The central point of the definition is the resource. A resource can be any data or service asset that is available to the local or networked environment. The resource is described using characteristics. These characteristics may be either qualitative or quantitative. The value of the descriptive characteristic is the metadata.

As an example, consider a dictionary of data terms. These terms may be used to describe elements or items within a data structure. In turn, these structures are filled with data to form data records and ultimately data sets. Suppose the dictionary contains a term ‘latitude’. The dictionary would likely contain a descriptive characteristic called ‘definition’. As an example, for the term ‘latitude’, the definition characteristic may contain ‘the angular distance of a point from the equator of the earth’. The value of the descriptive characteristic ‘definition’ is the metadata that supports the term ‘latitude’.

Metadata may also include quantitative descriptions of the resource. For example, a quantitative characteristic that supports latitude may be the range of acceptable values. If latitude were being used to describe the position of an object on the earth, then a quantitative limit on the range may be -90 degrees to +90 degrees or similarly, limits defined in terms of North and South.

This content or description is the metadata that describes the single term ‘latitude’. Given this content, we see one role of metadata is to provide the semantic understanding of the terms used within a particular resource. In the case of the example, the metadata provides the semantics of the data item 'latitude'.

Metadata may also support a complete data set. In this case, the differences between describe and support are important. Describe implies the citing of details to provide a more realistic view of the data. For example, the latitude range defines values that directly describe the allowed content of the latitude data. Support implies that the metadata provides a level of assistance to the data, but does not directly define or limit the data. Support also includes the support of processes applied to the data asset. For example, a supporting characteristic may be the (IP) address of the computer where the data asset may be obtained. This type of metadata supports the discovery of the data asset, but does not describe the data asset.

Using metadata as a support for the data discovery function is one metadata usage that is easily understood. However, other views of metadata may be used to elucidate the meaning of the term metadata. For example, a unit of metadata may be considered to consist of a descriptive characteristic (e.g., termed a property), a value for this characteristic (e.g., termed a value), and the subject that the metadata refers to (e.g., termed a resource) (see [4] for further description).

This model is also the basis of the Resource Description Framework (RDF) [5]. RDF was developed by the World Wide Web Consortium (W3C) to represent metadata for web resources, where the term web resource can include anything identifiable on the web as well as things retrievable from the web. The RDF model uses the resource, property, value combination with a slightly different terminology, namely a subject, predicate and object, respectively.

In terms of functional uses, metadata contributes to the process of distributing, advertising, using and combining data assets. Internationally, these functions are being explored in community-based efforts focused on marine data. Experts in the marine metadata initiative [6] are helping to explain many of the metadata issues by providing definitions, guides and examples to clarify the use of metadata in these functions.

CLIENT CATEGORIES

In the process of defining a system, designers must keep in mind the overall objectives for the system. In one respect, the objectives may focus on meeting the requirements of the users or systems requesting data. This effectively means that the focus is directed to satisfying user or automated computer demands on the system. Within this document, both the user and system level demands will be categorized together as client level demands.

To address the needs of the client it is important to understand the level of knowledge possessed by the client. In this regard, three levels of clients may be defined. At the highest level, the clients possess considerable a priori knowledge with regard to the data assets of the cell. At this level, users or other system designers have access to the original designers of the data structures within the asset.

At level two are those clients with an intricate knowledge of the functional aspect of the data asset, but no knowledge of the detailed structure of the asset. In this case, the asset and functions are known to exist, however, the details of the data and data structures contained within the asset are not known. At level two, the client recognizes the existence of the asset but does not possess knowledge on the details of the internal structures.

The lowest level of knowledge for a client is level one. At this level the client has no a priori knowledge regarding the structure or the data asset. The client is not aware of the existence of the data asset or of the internal structure of the data asset. This level of knowledge is characterized by a client entering a network with no knowledge of the assets available within the network.

MODEL OF DATA UTILIZATION

The most general client category is level one and this is the level that will be addressed by the NUW architecture. At this level, the client requires an assortment of metadata, primarily oriented to aid the client in discovery, understanding and utilization of the data asset.

Before proceeding we must be clear about the descriptions being provided. The metadata content must assist the client by providing the information to aid in the discovery, understanding and utilization of the asset. Equally important is the metadata structure used to house the metadata content. The metadata structure obviously must support the content, but also must be understandable and utilizable by the client. The metadata content and metadata structure are two different yet connected concepts.

Consider the metadata content. The content must support the discovery process, where we define discovery as the searching and locating of data that meets a particular client requirement. The understanding is obtained by providing content with sufficient definition to allow clients the ability to independently judge the data assets applicability to the client requirement. Finally, utilization is the process of obtaining or extracting the asset and then using the asset in a proper fashion (some metadata groups separate extraction from utilization [7] but here for simplicity, the two are combined).

Building a structure that supports data discovery is not a sufficient condition for discovery; however, it is a necessary condition. The discovery process relies on the metadata to the point that the metadata must exist, it must be accessible, and must be syntactically and semantically understandable by the client. Here, accessible implies that the metadata exist in a common and known location, or be registered through common procedures. Syntactically understandable means the metadata structure is readable by the client while semantically understandable implies that the metadata content is in a form that the client can understand and interpret.

VOCABULARIES - METADATA CONTENT

One form of metadata content are the terms that are used within a particular subject area. Collectively, these terms represent a vocabulary for the subject area. Unique vocabularies are common is specialized fields of study or professions.

Previously, we noted that the term ‘latitude’ could be defined within a dictionary. It was also noted that the latitude could be part of a larger collection of data, called ‘position’. This example provides an opportunity to distinguish between two important types of vocabularies: a data vocabulary and a discovery vocabulary.

A data vocabulary is a collection of terms that identify or name the individual data items. For example, the term ‘latitude’ would be contained in the data vocabulary as this term applies to a data item.

A discovery vocabulary typically names a group of data labeled to assist in the discovery of data items that are in some way related. Discovery vocabularies are typically hierarchical, containing labels that often contain other labels but ultimately relating to the available data. This results in potentially broad discovery terms such as ‘atmospheric’ to contain all atmospheric data at the asset. In the previous example, ‘position’ would be in the discovery vocabulary.

For both data and discovery vocabularies, the labels must be known and defined. Latitude is a somewhat common term and one may consider it to be obvious. However, other terms (data or discovery) may not be obvious. For example, ‘waveform type’ may be well known in one specialized subject area but unknown in another. Alternately stated, both vocabularies are often domain specific.

DICTIONARY STRUCTURE

The set of terms that constitute a vocabulary may be grouped in an assortment of ways. For example, the set may be grouped in a glossary or taxonomy. The terms may also be grouped in a dictionary, similar to the common language dictionary but containing only the specialized set of terms in the particular vocabulary.

Merriam-Webster provides several definitions of the term dictionary. A slightly modify version of the Merriam-Webster definition [8] is:

a list of terms or names important to a particular subject or activity along with discussion of their meanings and applications

For clients to be successful in discovery and utilization of the data asset, the topic vocabulary must be defined, accessible and understood by the clients. One method to accomplish this is to create a dictionary to support the terms used in both the discovery and data vocabularies.

Such a dictionary may be modelled after a common language dictionary. This would provide a structure that is familiar to users. This was the tactic taken by an international study group formed under the International Council for the Exploration of the Seas (ICES) and the Intergovernmental Oceanographic Commission (IOC). The Study Group was tasked to examine marine data exchange systems using XML and became known as the Study Group on XML (SGXML) [9].

The SGXML developed a dictionary structure (partially shown in Figure 3) intended to aid in the discovery and mapping of dictionary terms. The intent was to populate the dictionary with terms used in the marine data community, thereby allowing this particular community of interest the ability to query and identify existing dictionary terms. The SGXML hoped that by providing such a dictionary of terms, users would reuse existing terms rather than develop new terms.

The SGXML dictionary structure is useful for both the data and discovery vocabularies. For the data vocabularies, the dictionary structure allows a multiple set of parameter codes for any term definition. A single term may be defined and described in such a way that it is common across many systems. However, internal to the local system, individual codes or abbreviations may be used to name the specific term.

A simple example of this may be developed. Consider the bearing of a target from a platform. The bearing may be defined as the angle between a reference direction (e.g., the ship heading) and a line towards the target. The angle may be defined in terms of the System International derived unit of radians or perhaps in degrees.

- dictionary
  - dictionary_owner
  - dictionary_citation
  - dictionary_entry
    - dictionary_term
    - role
    - definition
      - instance
      - definition_owner
      - short_name
      - creation_date
      - methollodgy
      - unit_of_measure
      - min_value
      - max_value
      - codeset
        - codeset_name
        - code
        - codeset_owner

UNCLASSIFIED

Figure 3. An abbreviated illustration of the SGXML dictionary structure. This shows the hierarchical nature of the structure.

It is likely that such a common definition would be applicable to many systems. However, the systems may be storing or manipulating bearing data using an assortment of codes that identify the data. For example, one system may refer to the bearing data as 'brn' while another system refers to the same data as 'bearing'. The SGXML structure allows the term bearing to be defined and also allows the term to be connected to multiple codes. In this example the codes are represented by 'brn' and 'bearing'.

The explicit definition of these terms then opens the possibility of automated systems manipulating the data content to address the unique needs of the local system. For example, an automated system may recognize unit discrepancies between codes used for the same definition. The system could then apply appropriate conversions based on standard conversion algorithms and coefficients.

DISCOVERY METADATA STRUCTURE

The US released the DOD Net-Centric Data Strategy [10] in May 2003. The strategy outlines the DOD vision of how communities-of-interest, metadata, and the Global Information Grid will be combined to form the net-centric environment. Many of the strategy goals (e.g., data visibility, data accessibility, data management [10]) are reliant on the availability and use of metadata.

The US Department of Defense Discovery Metadata Specification (DDMS) [11] outlines the intended structure for the metadata content to meet the Net-Centric Data Strategy. The DDMS identifies and describes characteristics of the data asset. This type of description describes the asset as a single unit. For example, the asset may have an associated publisher, it may have a title, a creation date, etc. These attributes pertain to the asset as a whole and do not describe the content of the asset. This level of description supports the discovery of the asset and initial assessment of the asset's applicability of use.

The DDMS has become very well aligned with the Dublin Core Metadata Initiative (DCMI) [12] specification, with extensions beyond the DCMI to address the particular business needs of the US DOD. As an example of the extensions, the DCMI element 'Coverage' is defined by the Core as specifying the extent or scope of the resource. The DDMS extends the coverage by introducing refinements that include geospatial coverage and virtual coverage. Geospatial coverage provides information on the reference frame of the coordinates used in the resource. Virtual coverage identifies the one or more addresses on a computer network where the asset is located. Note that this definition does not specify information about the content of the asset, but rather the virtual location of the asset.

Other specific elements within the DDMS assist in meeting the goals of the Net-Centric Data Strategy. For example, the DDMS 'Security' element contains 18 security information items such as the classification of the data asset, who classified the asset, the data producer, release restrictions, dates of classification, and exemptions. All of this information supports the accessibility goal of the Net-Centric Data Strategy. The accessibility is realized only when access is controlled via appropriate security metadata.

The DCMI and DDMS both represent controlled structures to be used for transfer of metadata. However, it is worth noting two other well-recognized metadata standards. Internationally, work involving digital geographic metadata has resulted in the International Organization for Standardization (ISO) metadata standard 19115. Alternately, in the US geographic metadata is mandated to comply with the Federal Geographic Data Committee (FGDC) standard. Both of these metadata standards have a large user base.

WRAPPING IT ALL TOGETHER

The various conceptual components of the model being implemented under the NUW project are combined into a single diagram shown in Figure 4. The model components now begin to form an architecture. The diagram shows a particular implementation of the model that utilizes extensible markup language (XML). The figure represents an implementation at a single cell to address client level 1 requirements.

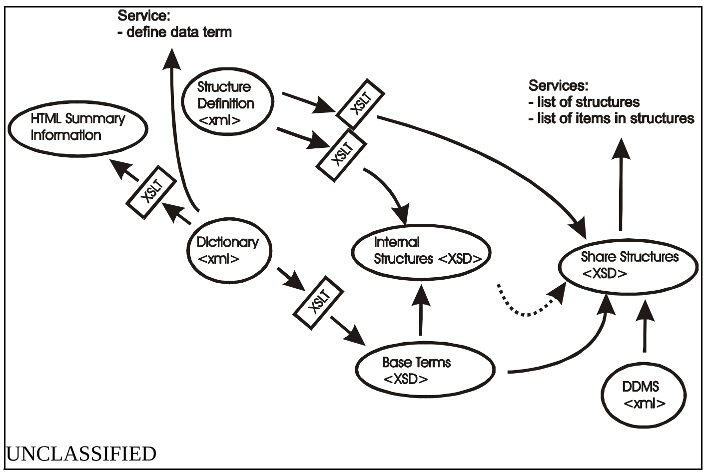

Figure 4. The metadata descriptions in the dictionary and DDMS content provide the foundation for the services within a particular cell.

The dictionary of terms is shown as an oval, with extensible stylesheet language transformations (XSLT) to produce hyper text markup language (HTML) output for quality control. The dictionary is also utilized to produce a set of base terms, expressed in XML schema language. These base terms represent all allowed parameter content within the local system. The base terms may be combined according to the structure definitions. The structure definitions represent a simple XML form that describes the internal or local data structures.

The structure definitions are manipulated by XSLT to express the internal data structures in XML schema language. The internal structures must be checked against the base terms to ensure that only allowed terms are used within the structures. For example, if an internal structure contains the data item ‘latitude’, then the ‘latitude’ term must be present in the base terms schema. Since the base terms are constructed from the dictionary, this also ensures that ‘latitude’ term is defined in the dictionary.

The structure definition also describes the data structures that may be shared. The shared structures do not necessarily coincide with the internal structures. However, the data item composition of the shared structure must be checked against the base terms to ensure the data items in the shared structure are present in the local system.

The DDMS content plays a role in the shared structures. DDMS is chosen over the other standards because of the desire to align with US DOD developments. Components of the DDMS content will be used to define the characteristics of the cell and local data asset. Thus, the DDMS content provides a level of uniqueness for the data item within the asset.

The services are labeled near the top of Figure 4. This does not represent a complete list of services, but rather indicates the type of services that are possible given the metadata content. For example, services provide discovery vocabulary terms by accessing the shared structures. Services also provide the data items within the shared structure, and definitions for these data items.

CONCLUDING REMARKS

The intent of the NUW project is to demonstrate the advantages of producing a COP using networked assets. It is expected that the networking of data sources will have implications on the accuracy and compilation speed of the COP. Metadata descriptions of data assets will play a central role in both the data discovery and utilization aspect of the demonstration.

Careful consideration must be given to the metadata content and structure to ensure the understanding of the transferred data asset. It is important that clients at all levels, and the potential queries to the system that the clients will be making, are recognized during the conceptual phase. This helps to ensure that the metadata present within the local systems can support the queries. Combining the components in an XML architecture also assists in providing an open system for external use.

REFERENCES

[1] Sandy Babcock. “Canadian Network Enabled Operations Initiatives”, in the 9th ICCRTS, September 2004.

[2] Dr. Marcel Lefrançois, Dr. Bill Roger, Dr. Ian Fraser, Anthony Isenor, Dr. Garfield Mellema, Nicole Collison, LCdr Bruce MacLennan, “Data Exchange Between Air and Surface Platforms in Networked Underwater Warfare”, NATO RTO Systems Concepts and Integration Panel Symposium on Architectures for Network-Centric Operations, Athens Greece, Oct 20-22, 2003, DRDC Atlantic SL 2003-080, September 2003, 12pp.

[3] David Booth, Hugo Haas, Francis McCabe, Eric Newcomer, Michael Champion, Chris Ferris, and David Orchard (Editors), “Web Services Architecture”, W3C Working Group Note, 11 February 2004.

[4] Luis E. Bermudez, “ONTOMET: Ontology Metadata Framework”, Phd Thesis, Drexel University, December 2004.

[5] Frank Manola and Eric Miller (Editors), “RDF Primer”, W3C Recommendation, 10 February 2004.

[6] John Graybeal and Luis Bermudez, “Marine Metadata Interoperability - Adding Science Value”. Available: http://marinemetadata.org/

[7] Bryan Lawrence, David Boyd, Kerstin Kleese van Dam, Roy Lowry, Dean Williams, Mike Fiorino and Bob Drach, “The NERC DataGrid”, unpublished manuscript, January 2003.

[8] “dictionary" Merriam-Webster Online Dictionary. 2005. http://www.merriam-webster.com (1 May 2005).

[9] Anthony W. Isenor and Roy K. Lowry. 2004. Final Report of the ICES/IOC Study Group on the Development of Marine Data Exchange Systems using XML, DRDC Atlantic ECR 2005-005, Defence R&D Canada – Atlantic. Available: http://pubs.drdc-rddc.gc.ca

[10] “Department of Defense Net-Centric Data Strategy”, Prepared by Department of Defense Chief Information Officer, May 9, 2003.

[11] “Department of Defense Discovery Metadata Specification (DDMS)”, Office of the Assistant Secretary of Defense, 3 January 2005.

[12] Dublin Core Metadata Initiative, Available: http://dublincore.org/