Federation of Metadata Registries

Giridhar Manepalli

Corporation for National Research Initiatives
Virginia, United States of America
gmanepalli@cnri.reston.va.us

Abstract. Federating metadata registries introduces a range of problems to address, from identifying the commonality in data models across registries, to agreeing on a well-defined data dictionary. As more and more registries participate in the federation, the number of problems to solve can potentially rise to unmanageable levels.

While solutions to these issues are researched, generic federation systems must be developed to address several common issues: cross-walking of registries; identifying the metadata entities; and providing a place holder to capture domain specific solutions. This paper presents a conceptual model for such a federation system that will address these common issues. The experience gained during the design and implementation of two metadata registries, ADL-R and FeDCOR, is adequately leveraged in the design.

This paper empirically defines "Federation System" as a quintuple, F = (R, C, K, I, D) where F is a Federation System; R is a set of registries participating in the federation; C is a set of common features in the federates; K is a set of domain independent algorithms to cross-walk registries (every algorithm in the set is a function over C); I is an identification system; and D is a set of domain specific algorithms to cross-walk registries (every algorithm in the set is a function over C).

Keywords: federation, metadata registry, Handle System, ADL-R, FeDCOR

1 Introduction

A Metadata Registry provides several services including discovery, browsing, and possible accessing of the content objects using the metadata items put[1] into the registry. Although the definition of a metadata registry varies widely, this paper adopts the above definition based on the experience of the author from the research efforts in the

[1] *Put* includes both push and pull techniques to register the metadata in the metadata registry.

design and development of Advanced Distributed Learning Registry (ADL-R) (Jerez, Manepalli, Blanchi & Lannom, 2006) and Federation of DSpace using CORDRA Registry (FeDCOR) (Manepalli, Jerez & Nelson, 2006).

The presence of multiple metadata registries each addressing corresponding community interests leads to the prospect of unifying such registries to provide unified service access to heterogeneous communities[2]. The unification, a.k.a. federation[3], of metadata registries raises serious problems that must be addressed that span several conceptual domains (data semantics, domain ontology etc). Among the many challenges the process of federation requires facing are the identification of commonality in data models across registries; encapsulation of common community interests; construction of community meta-models; categorization of conceptual models in registry federates; and conformation to a well-defined data dictionary by communities. All of these domain-centric problems may be thought of as a registry interoperability problem. While solutions to these problems are evolving, designing a federation system involves mitigating a variety of other issues as well. The following section investigates the scope of a federation system and highlights a few popular metadata registries.

2 Scope

Any metadata describing a content object by itself is neither infrastructural, nor self-qualifying, nor self-identifying. Infrastructural, in the metadata context, may be defined as a quality by which the elements within the metadata present scope and context for other elements consistently, thereby establishing a well-defined structure to its whole. A metadata registry may impart this quality upon the metadata by using a well defined meta-model[4]. Self-qualifying quality of a metadata makes it context/community/environment free and therefore represents the content object distinctly in heterogeneous

[2] Related research may be found from ADL-R.

[3] The set of use cases driving a federation process is specific to the group of communities participating in such federation.

[4] A meta-model, in the present context, is a model used to define the metadata model used by the corresponding registry. In its definition of a metadata model, the attributes of the structure, form, and usage are emphasized.

communities. A metadata registry may qualify the metadata by establishing the context of content and community by defining inter-registry cross-walking definitions. Self-identifying metadata establishes an identity to itself that is globally valid, and therefore allows its reference in other contexts. A metadata registry may associate an identifier with each metadata item when registered. It is important to note that not all metadata registries impart these missing qualities into the registered metadata items.

2.1 Case Studies

2.1.1 DSpace

The schema used for the metadata describing content objects ingested into a DSpace instance conforms to Dublin Core (DC) metadata schema (Powell, Nilsson, Naeve & Johnston, 2005). However the DC elements themselves have no defined meta-model associated[5] with them, and therefore DSpace does not impart the infrastructural quality onto the metadata. The metadata ingested into DSpace about the content object is only defined within the context of the content object, if at all. Also, the metadata and the content object in DSpace are both assigned a single identifier. Although an identifier is associated with every metadata item, the identifier does not help identify the different metadata items that may have existed for a given content object. In effect, the metadata is not self-identifying.

2.1.2 NSDL

Some of the metadata schemas, namely IEEE Learning Object Metadata (LOM) (Institute of Electrical and Electronic Engineers [IEEE], 2002) etc, supported by NSDL ("NSDL Library Architecture") have a defined structure. For example, LOM has a well defined meta-model for the elements. For instance, the meaning associated with any element may be deduced from the level of the element in the hierarchy. Moreover, the metadata describing the content objects requires qualification outside the scope of its hosting NSDL. Finally, the metadata items inside the registry have a unique identifier associated with them.

[5] The Dublin Core metadata schema reflects the principles of the initiative – the elements are optional, the elements are extensible etc. Additionally, the Dublin Core metadata specification does not provide any guidance for architects using the model.

The following table summarizes the observations of the author made in the metadata registries[6].

MDR	Infrastructural	Self-qualifying	**Self-identifying**
DSpace	No	No	**No**
NSDL	Yes	No	**Yes**
ADL-R	Yes	No	**Yes**
FeDCOR	**No**	**No**	**Yes**

Table 1 Summary of qualities imparted by various metadata registries

Based on the observations made above, it is clear that different metadata registries have different qualities missing in their metadata items. While the objective of the metadata registries may not include imparting any of these qualities onto the metadata, the federation of such metadata registries demands the composition of these qualities in the federation system.

In a federation, the above described qualities[7] play an important role in defining a unifying standard to the set of services contributed by the participating registries. The infrastructural quality – the quality of having well-defined structure in the metadata – helps in understanding the semantic associations between the metadata of participating registries. The self-qualifying quality – the quality of describing attributes in a global context – aids in estimating the information loss when a cross-walk between the participating metadata registries is made. The self-identifying quality – the quality of allowing identification to the metadata – is necessary to reference the metadata in a larger scope and context.

In addition to the described qualities, a federation system also relies heavily on the technology and standards to close the gap between problems posed and solutions offered. It is important to understand that a federation system only achieves the goal when the solutions offered are efficient, optimal, and scalable.

[6] The observations of ADL-R and FeDCOR are not described here. The details about metadata models in these registries may be found from ADL-R and FeDCOR.

[7] Although only three qualities are identified in this paper, the federation model is open to other qualities as well. However, the stated three form the lowest common denominator for any federation system.

3 Federation

Based on the identified scope and requirements of a federation, a federation system should provide both domain-independent solutions to address the performance, scalability and participation constraints of cross-walking registries, and domain-dependent solutions to mitigate the interoperability problem.

Past research efforts to solve interoperability problems emphasized using three approaches: a mapping-based approach, an intermediary-based approach, and a query-based approach (Park & Ram, 2006). The mapping-based approach attempts a concept-only solution requiring the use of a unified data model and mechanisms to translate from individual data models to this unified data model. The intermediary-based approach is partly conceptual and partly implementation oriented. The use of intermediaries (agents) with domain-specific knowledge (processing model, domain ontology) provides a pathway to achieve interoperability. The query-based approach formulates queries specific to the domain the interaction is being handled for.

3.1 Federation Model

To better understand the importance of components in a federation system, it is useful to realize that the problem at hand is a conceptual space; this conceptual space consists of subsets of space (registries). These subsets may or may not overlap with any of the other subsets. The points within each subset are the metadata items registered within the registry. In order for this conceptual space to be coherent, each point in the space satisfies certain spatial properties – the properties that emphasize the commonality in spite of the variability. Apart from the properties, the conceptual space provides path and direction to each of these points by defining rules over the existing spatial properties. The defined rules may be universal rules that do not take into consideration the geometrics of the subset, or may be local rules that consider the subset-space geometrics. These rules, which provide direction toward each point, are the algorithms that allow cross-walks between registries.

The federation model defined above is an imaginary universe with various elements, R. The behavior of these elements is guided by the exhibited properties, C and I. The boundaries of the universe are defined by certain rules, K and D. The model is therefore formally defined as a quintuple F = (R, C, K, I, D). The common properties along with algorithms allow the federation system to compose the metadata qualities missing in the participating registries. In addition, the model provides a place holder for mitigating the interoperability problem. For instance, in a federation system, if the participating registries are non-infrastructural, the federation may adopt mapping-based interoperability solutions; and the domain dependent algorithms may then provide cross-walk into a unified data model.

4 Practical Approaches

This section describes the value domain for each of the quintuple elements, and concludes with some practical scenarios.

4.1 Common Properties – C

The definition of the common properties present in different federates implies that the domain semantics are not considered when choosing such properties. Recall that the federation model underscores the importance of metadata items satisfying these common properties. These properties form the lowest common denominator set. The following table lists the properties along with a filtering use case.

Property	Definition	**Filter**
Identifier	The identifier of the metadata item unique to the entire federation system	**Identifies metadata item within registry**
Location	The location of the metadata item as understood by the corresponding MDR	**Retrieves metadata item from registry**
Timestamp	The timestamp at which the metadata item is created, modified, possibly deleted in the MDR	**Reduces noise using provenance**
Substrate	**The representation of the metadata item useful for its discovery. Possible values are high dimensional keywords, classification scheme values, index segments**	**Filters metadata items from multiple items**

Table 2 Non-normative set of common properties

The table above lists a non-normative set of common properties that the metadata items demonstrate across registries that do not necessarily demand any domain knowledge. From a use case perspective, the common properties identified may not directly aid in discovering the metadata item from the registry. Nonetheless, they help in filtering the process of discovery.

4.2 Identification System – I

The Identifier System acts as an underlying technology in order to identify several entities that participate in the realization of the federation system.

The CNRI Handle System® (Sun, Lannom & Boesch, 2003) is an advanced technology used to manage and resolve unique, persistent identifiers. The Handle System defines a flexible data model that each persistent identifier – a handle – adopts. ADL-R and FeDCOR harnessed the Handle System technology to efficiently solve multiple problems related to security, application profiles, repository objects dissemination etc., the details of which are beyond the scope of this paper. The use of the Handle System, therefore, would bring in a distinguished dimension to a federation system.

4.3 Algorithms – K & D

The purpose of a federation system is to provide unified services to its clients. Since the services provided by the federation system are similar to the services provided by any registry at the community level, it is appropriate to consider such a registry to be a ‘service provider registry’ (SPR). Assuming the existence of SPR there are two possible scenarios: (1) Every registry in the system acts as a SPR; and (2) There exists a single SPR for the entire system (without considering the option of having mirrors for load balancing and fault tolerance). Depending on the approach being considered for building a SPR the topology of the system changes, and therefore the set of algorithms that defines the interaction between the registries changes.

A few system scenarios to emphasize the adequacy of the federation model are illustrated below with a (non-normative) solution.

System Scenario: Every registry in the system acts as a SPR

Regardless of the approach adopted to solve the interoperability problem, the fact that every registry is a SPR demands certain geometric sanctions. (1) Since every registry is a SPR, it has to be grouped closely (in a federation network) with every other registry to enable high communication bandwidth. (2) Every registry should be able to reach other registries in a small number of hops. This is to ensure that the registry is readily accessible to all service clients.

The first requirement is to have a cliquish network which is measured by the degree of its cluster coefficient. The second requirement is met by keeping the "characteristic path length" minimum. The characteristic path length (in the federation scenario) may be defined as the average number of hops each registry takes to reach another registry. The fact that the cluster coefficient needs to be high, while the characteristic path length needs to be low, makes the registry network behave like a true social network (Hong, 2001). Without considering the domain dependent features of the registries, an algorithm based on Freenet linking algorithm (Hong, 2001), is stated below:

Assume the total number of registries in the federation system to be 'n'. Let k be any number between '1' and 'n'. Each registry r_i, where 'i' is ranging between 1 and n, holds a hash table with k entries. These entries point to 'k' other registries in the system, thereby forming the links. The links are created in such a way that the network results in a social network[8]. The diagram below is a typical structure of the network thus formed.

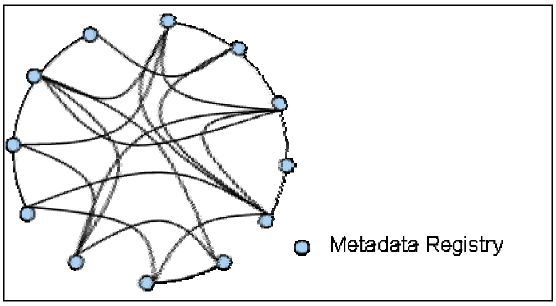

Figure 1 Social network based federation system (Hong, 2001)

[8] For further details refer Freenet linking algorithm.

When the communication request for r_j is issued by r_i (a typical scenario when every registry is a SPR), the request is traversed from r_i through the registries using the links, until it reaches the destination. The property of this federation topology, a social network, ensures that it takes very few hops to reach the destination (on an average).

Scenario Update: There are communities of common interest formed and communities may expand or collapse dynamically

The above specification requires links between registries to be dynamically adjustable. A domain specific algorithm, based on the SETS (Bawa, Manku & Raghavan, 2003), may be used to group the communities of common interest into segments. The following diagram illustrates snap-shot of a registry network forming segments of common interest.

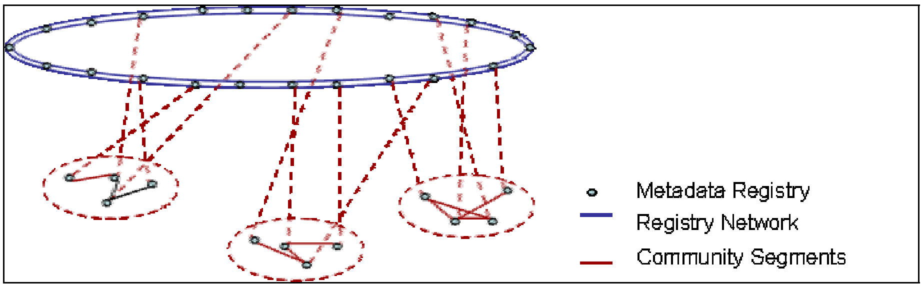

Figure 2 Community based federation system ("FreeLib")

It can be proved that if there are 'k' links between registries, the query can be routed between any two registries in O((log-square n)/k) hops using topic-based routing technique[9].

System Scenario: There is a single SPR for the entire system. There are communities of common interest formed that have a hierarchy of domains, sorted from abstract registries to domain specific registries

The following geometric sanctions are required. (1) The network is a tree with SPR as the root node. (2) The level of (domain) abstractness decides the level of the registry in the tree. The more abstract the registry, the closer it is to the SPR. The following algorithm, based on

[9] Further details about the routing technique may be found in SETS.

Adaptive Tree Walk protocol (Tanenbaum, 2003), assumes certain domain specific requirements.

The root node of any sub tree is assumed to have the domain knowledge of the underlying nodes. Specifically, in a query-based solution, the root node arbitrates, if the query belongs to the domain of one of the child nodes or not. Consequently, the root node forwards the query down the decided sub-tree nodes. This process continues until the query reaches the leaf node – if there is one – that belongs to same domain. The following graphic, a tree diagram, illustrates the network structure.

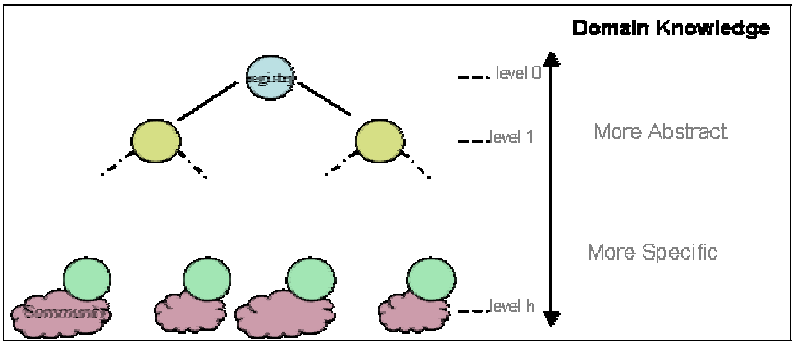

Figure 3 Registry Hierarchy

4.4 A Complete Example

The following section explains the system dynamics for a given process scenario using the federation model laid out in the paper, assuming the query-based approach is adopted by the system. The particular process scenario is the following query sent to the SPR by the client: “Find metadata items that have as the content object identifier 100.50.10.1/1234 and that are registered in the system after March-2005 but not later than September-2006”.

Solution: Here, I is the Handle system. C, the set of common properties, is inherited from the Table-2 above, namely identifier, location, and timestamp. D, the domain dependent algorithm, is as defined above.

Process: Based on the problem request, the query from root node is passed on to its child nodes at level ‘1’. Specifically, the left child at level ‘1’ is queried. Based on the query, the registry arbitrates that its child registries are not responsible for that query. Next, the right child at level ‘1’ is queried. The registry passes the request to its child nodes, as it finds out that this side of tree has the corresponding metadata items. The subsequent steps in the process follow the same approach, before it finally reaches the leaf node r_n. The r_n processes the request and finds that there are three metadata items for the content object ‘100.50.10.1/1234’. The following diagram illustrates the described process.

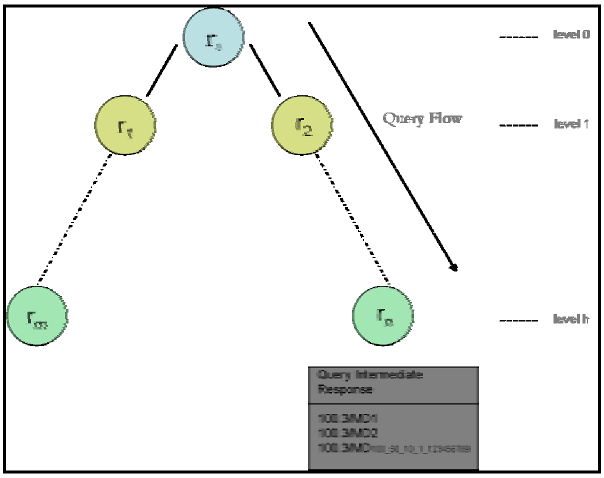

Figure 4 Query Flow

The common property, timestamp, is used to filter the results to discover one metadata item: 100.3/MD100_50_10_1_1234123456789. This result – a handle - is passed to the client. The handle records within the handle are displayed in the table below. The URL when resolved using any HTTP client uses the r_n registry services to retrieve the metadata item from the registry.

Handle Records of **100.3/MD100_50_10_1_123456789**	
URL	**http://handle.cordra/?m=100.3/MD100_50_10_1_123456789**
100.TYPES/CONTENT_OBJECT	**100.50.10.1/1234**

Table 3 Handle Resolution

5 Other Federation Efforts

The value–domain of the elements defined in the federation system allows many combinations. Each combination of values for the elements may be suitable for a specific scenario. The federation system instances, as illustrated in the above section, emphasize how solutions differ with varying federation scenarios. However, the different instances demonstrate the need for the defined elements.

5.1 The China Digital Museum Project (Tansley, 2006)

The China Digital Museum effort was to create a large-scale, federated deployment of DSpace. The federation results in two data centers that harvest the digital content from the participating federates - DSpace instances in this case. The data centers use OAI-PMH ("Open Archives Initiative Protocol for Metadata Harvesting") to harvest METS ("Metadata Encoding and Transmission Standard") Dissemination Information Packages from the federates. The metadata provided by the participating federates conform to a single metadata schema. Consequently, the domain dependent metadata cross-walk algorithms are virtually eliminated from the federation system. However, the federation system needs an identification system to identify the resources (digital objects) uniquely, and also locate the copies that may exist in other DSpace instances. This project uses CNRI Handle System as an identification system to accomplish the above requirements.

5.2 NSDL[10]

NSDL is a federation system harvesting metadata from multiple providers. The set of digital resources from each provider – a collection – is harvested using protocols including OAI-PMH and NSDL API. NSDL, as existing at the time of this writing, supports Dublin Core metadata schema. Accordingly, each collection provider uses recommended cross-walking algorithms to convert from collection specific metadata schema to Dublin Core metadata schema prior to providing the metadata for harvesting. The harvested metadata is ingested into corresponding component of the digital library after assigning an unique identifier.

[10] NSDL is both referred to as a metadata registry and a federation system in this paper. It is noteworthy that the corresponding denomination of NSDL is based on the context in which the system is considered.

From the above two efforts, it may be identified that any federation system involves identification system, cross-walking algorithms, common properties across the participants, and federates themselves – the elements identified in the federation system defined in this paper. It is important to note that these elements are already described and used in the aforementioned efforts, but in an ad hoc manner. The federation system, defined in this paper, punctuates the prominence of these elements in forming a generic system – a system that addresses common federation issues.

6 Conclusion

The federation system instances illustrated in the "Practical Approaches" section are derived from the federation model defined in this paper. The model presents an open universe with mandatory elements, which are governed by the rules that define the relationships between them. The openness of the universe is still to be researched, as there are several other key factors to be considered that drive the federation model, such as interoperability approaches, ranking among registries, levels of heterogeneity, etc. Research efforts in these areas will not result in a global solution to all the described problems. Nevertheless, the efforts provide guidelines and reasons to follow a particular solution in a given context. Encouragingly, the federation model defined acts as a starting point for research in federating metadata registries.

References

1. Jerez, H., Manepalli, G., Blanchi, C. & Lannom L.W. (2006, February). ADL-R The First Instance of a CORDRA Registry. *D-Lib Magazine*, Volume 12, Number 2, ISSN 1082-9873. Retrieved August 15, 2006 from http://dlib.org/dlib/february06/jerez/02jerez.html.
2. Manepalli, G., Jerez, H. & Nelson, M.L. (2006, February). FeDCOR: An Institutional CORDRA Registry. *D-Lib Magazine*, Volume 12, Number 2, ISSN 1082-9873. Retrieved August 15, 2006 from http://dlib.org/dlib/february06/manepalli/02manepalli.html.
3. Powell, A., Nilsson, M., Naeve, A. & Johnston, P. (2005, March). DCMI Abstract Model. Retrieved August 15, 2006 from http://www.dublincore.org/documents/abstract-model/

4. (2002) IEEE Standard for Learning Object Metadata. *IEEE*.
5. NSDL Library Architecture: An Overview. Retrieved August 15, 2006 from http://nsdl.comm.nsdl.org/docs/nsdl_arch_overview.pdf.
6. Park, J. & Ram, S. (2004, October). Information Systems Interoperability: What Lies Beneath?. *ACM Transactions on Information Systems,* Volume 22, Issue 4, 595-632.
7. Sun, S., Lannom, L. & Boesch, B. (2003, November). Handle System Overview. *Internet Engineering Task Force (IETF) Request for Comments (RFC)*, RFC 3650. Retrieved August 15, 2006 from http://hdl.handle.net/4263537/4069.
8. Hong, T. (2001). Performance. In A. Oram (Ed.), *PEER-TO-PEER* (pp. 203-241). Sebastopol, CA: O'Reilly & Associates.
9. Bawa, M., Manku, G.S. & Raghavan, P. (2003, August). SETS: Search Enhanced by Topic Segmentation. *SIGIR*, 306-313.
10. FreeLib – Peer to Peer Digital Library. Retrieved August 15, 2006 from http://p2pdl.cs.odu.edu/proposal.pdf.
11. Tanenbaum, A.S. (2003). The Medium Access Control Sublayer. In *Computer Networks* (pp. 247-342). Upper Saddle River, NJ: Prentice-Hall.
12. Tansley, R. (2006, July). Building a Distributed, Standards-based Repository Federation. *D-Lib Magazine,* Volume 12, Number 7/8, ISSN 1082-9873. Retrieved August 15, 2006 from http://dlib.org/dlib/july06/tansley/07tansley.html.
13. Open Archives Initiative Protocol for Metadata Harvesting. Retrieved August 15, 2006 from http://www.openarchives.org/OAI/openarchivesprotocol.html.
14. Metadata Encoding and Transmission Standard. Retrieved August 15, 2006 from http://www.loc.gov/standards/mets/.