
# Principles of Metadata Registries

## A White Paper of the DELOS Working Group on Registries

**Thomas Baker, Fraunhofer-Gesellschaft, Germany (Chair/Editor)**
**Christophe Blanchi, CNRI, USA**
**Dan Brickley, World Wide Web Consortium, UK**
**Erik Duval, Katholieke Universiteit Leuven, Belgium**
**Rachel Heery, UKOLN, UK**
**Pete Johnston, UKOLN, UK**
**Leonid Kalinichenko, Russian Academy of Sciences, Russia**
**Heike Neuroth, State and University Library Goettingen, Germany**
**Shigeo Sugimoto, University of Tsukuba, Japan**

# Table of Contents




This paper is the collective product of a Working Group on Registries sponsored by the DELOS Network of Excellence on Digital Libraries, an initiative funded by the European Commission to promote research and international cooperation in the field of Digital Libraries. The working group was convened and chaired by Thomas Baker, who integrated contributions and feedback from group members and edited the collective draft. As the product of a process that included two face-to-face meetings and email discussion in 2001 and 2002, this paper reflects a rough consensus among its participants, though its signatories may not individually agree with every point or definition therein.

# 1 Introduction

"Metadata" is typically defined as "data about other data" [1]. One classic example is a library catalog record listing the author, title, and subject of a book. A more sophisticated example might include links to related information such as book reviews. The concepts on which such structured descriptions are based – Author, Title, Subject, ReviewedBy – constitute small vocabularies. Metadata is in this sense a form of language, and the problem of interoperability among diverse metadata languages is partly linguistic in nature.

This paper is about "metadata registries" which, to stretch the analogy, are like dictionaries of metadata language. Just as there are many kinds of natural-language dictionaries – from lexica descriptive of historical usage to dictionaries prescriptive of good usage, from glossaries of professional jargon to dual-language dictionaries for translators – a similarly diverse set of registry services may evolve for presenting, processing, and understanding various types of metadata. Usage scenarios for registries, existing or planned, include the following:

- A cataloger needs to know the best practice for describing a particular type of resources. (A query to a registry might return a list of metadata element sets classified by use.)
- A federation of information providers wants to harmonize metadata usage among its members. (A registry might present descriptions of how metadata element sets have been applied so that a reader can compare areas of similarity and difference.)
- An information provider needs to translate its metadata into the shared format of a digital library federation. (A registry might link to crosswalk services that can batch-convert records from one format into another.)
- An implementer wants to construct a schema, re-using existing elements as far as possible. (A registry allows searching and browsing of data elements grouped into sets and profiles.)
- A software developer wants metadata tools to update their configurations automatically. (A registry might point to or provide machine-processable schemas.)
- Ten years from now, an archive needs to interpret and convert metadata records from 2002. (A registry might hold historical versioning information on standards or on particular applications.)
- Chinese speakers want to view or process metadata prepared in Germany. (A registry might specialize in providing translations or annotations in multiple languages.)

As of 2002, the principles underlying the design of such dictionaries are not yet well understood – a reflection in part of the relative newness of metadata languages in the context of global networks. Precisely because metadata language communities now find themselves on the common ground of a global network – an Internet Commons – attention is shifting from refining the metadata languages of particular communities towards defining design principles, data structures, and algorithms that will facilitate interoperability among them.

This paper is an attempt by researchers from different backgrounds and perspectives on the metadata world to articulate a shared set of principles underlying the construction of metadata registries. Most of the participants were themselves involved in registry-building activities and approached the task with practical aims. However, the immediate challenge for the group was to

[1] In the computer-science field of database design there is an older tradition, dating from before the "Web era", which uses the term "metadata" to designate information about the database schema, i.e., about the structure of data instances in a database. This clash of terminology caused some confusion in the mid-1990s but no longer seems to be a problem. This paper uses the term "metadata" in its Web sense rather than in the earlier database sense.

agree on a common vocabulary for talking about the object of registries. The most basic terms in current use in the metadata field – words like "schema" and "vocabulary" – can have confusingly different connotations between communities of practice. This paper, then, begins by presenting a common frame of reference – a shared vocabulary of basic concepts defined with respect to a simple model. This is followed by a brief survey of current registry activities. The paper then closes with a discussion of selected issues for research and further development in registry services.

# 2 Terms of reference

## 2.1 The elements of metadata

The "words" of metadata – concepts used to describe data, such as Author, Title, and Subject – are in this paper called "metadata elements" (or "elements")[2]. When seen in relation to the resources being described, they are called "attributes" or "properties" – e.g., a "title" is an attribute (or property) of a book.

Metadata elements are typically defined not in isolation, but as part of a group of elements that are useful for describing resources of a particular type or for a particular purpose. These functionally related groups of elements are referred to generically as "metadata element sets"[3]. In December 2002, the maintainers of several key metadata standards (GILS, ONIX, MARC 21, CERIF, DOI, IEEE/LOM, and Dublin Core) achieved consensus on a statement acknowledging that the various standards all had elements – "units of meaning comparable and mappable to elements of other standards" [CORES-RESOLUTION].

A catalog record using such elements to describe something is a "metadata instance" – an example of "instance metadata". This metadata instance may be seen as a particular collection of metadata elements associated with a set of "values" for those elements – e.g., "Author: William Shakespeare", "Title: Antony and Cleopatra", and "Subject: Roman history".

## 2.2 Layers of interoperability

---

[2] Like most other words that are used to talk about words, the word *element* carries different associations for different communities. In the Dublin Core community, *elements* are defined as a specific type of element in contrast to other types, such as *element refinements* and *encoding schemes*. At the other extreme, people who see no essential difference between words used as attributes and words used as *values* for attributes may see a list of subject headings (used as values for a "subject" attribute) as a set of metadata *elements*. And although *elements* can be considered the *atoms* of metadata language, some metadata communities have *composite elements* – elements that themselves contain sub-elements, perhaps more closely analogous to *molecules*. In some systems, composite data elements do not have values directly, but only through their component elements. To what extent such complexities will hamper the construction of registries that span standards communities remains to be seen.

[3] One alternative word for the concept of metadata element sets is *vocabularies* – a term preferred in contexts as diverse as W3C and indecs to refer to a wide range of metadata element and value sets [INDECS, W3C]. Such a generic use of *vocabulary* is, however, confusing for people who associate the term more narrowly with controlled lists of values, such as language codes or subject headings, or with other types of controlled listings such as thesaurus entries or subject headings. A growing community of researchers sees all of the above as *ontologies*, but this term is both too new (in its current usage) and too specifically associated with Semantic Web developments to be usable in a generic sense. When used as synonyms for "element sets", the words "scheme" and "schema" can be even more problematic, as discussed in the text and in footnote 5. Just as problematically, the word *namespace* (which has a very specific meaning, discussed in footnote 4) has sometimes been used to designate a group of *names* seen as an abstract set – i.e., as what we are here calling a *metadata element set* or *vocabulary*.

The integration of a diversity of resources into coherent digital libraries depends on an elusive quality called "interoperability". Interoperability is typically defined as the "ability of systems to provide services to and accept services from other systems" [MILLER]; more specifically as "enabling information that originates in one context to be used in another in ways that are as highly automated as possible" [RUST]; or even more concisely as “recombinant potential” [DEMPSEY].

Central to such definitions is the potential for metadata to cross boundaries between different information contexts. These boundaries may be technical, as when metadata is produced in different formats or made available with different protocols. But the boundaries can also be linguistic (as when metadata are translated), social (as when metadata for teachers are used by learners), or cultural (as with the specifically French educational term 'bac+2'). Indeed, boundaries of a cultural nature are usually harder to cross than the merely technical.

With regard to metadata, the problem of interoperability is multi-layered. In the abstract, metadata element sets can be seen as an Attribute Space of resource attributes such as "Author" and "Subject" (Layer 3a in the table below). These sets may be enshrined in standards such as the official specifications of IEEE Learning Object Metadata or the Dublin Core Metadata Element Set [DCMI, IEEE-LOM]. Alternatively, these sets may be adaptations of standards for specific purposes, such as an application profile of the IEEE LOM used by the Ariadne Foundation [ARIADNE].

The values associated with those attributes may, in turn, be constrained or defined by a variety of classifications, authority-control systems, controlled vocabularies, ontologies or taxonomies – a Value Space of values such as "Mark Twain" and "China, History" (Layer 3b). Value Space vocabularies may include intermediate constructs such as the “core ontology” developed by the DELOS Working Group on Ontology Harmonization as a bridge between the CIDOC Conceptual Reference Model (an ontology for describing cultural heritage information) and the ABC Harmony Model (an ontology for integrating information from multiple genres of multimedia information) [DELOS-ONTOLOGY].

| | | |
|---|---|---|
| **Layer 3** | **(a) Attribute Space**<br>(e.g LOM, Dublin Core MES, indecs) | **(b) Value Space**<br>(e.g. ontologies, classifications, controlled vocabularies, taxonomies) |
| **Layer 2** | **Representation**<br>(e.g. XML, RDF, DAML-OIL) | |
| **Layer 1** | **Transport and Exchange**<br>(e.g HTTP Get, OAI Protocol for Metadata Harvesting) | |

**Table 1**

In Layer 2, the attributes and values of Layer 3 are represented or instantiated using particular syntactic bindings in encoding languages such as XML or XML/RDF, which are processable by machines. In Layer 1, closest to machines and networks, metadata is transported and exchanged using protocols such as the Protocol for Metadata Harvesting of Open Archives Initiative [OAIPMH].

In terms of this model, metadata registries are applications that use metadata languages (Layer 3) in a form processable by machines (Layer 2) in order to make those languages available for use by both humans and machines. To be processable in automated ways, in other words, the conceptual structures must be bound to machine-processable formats. Problematic though these basic distinctions may seem, they are useful as a first approximation and point of reference.

### 2.3 Data models

Data models are the "grammars" of metadata language – formalized world views that provide a context for metadata by defining the structural relationships between different types of elements (analogously to parts of speech in language) and sometimes by characterizing the *things* to which the elements refer. In our model, data models underpin the element and value structures of Layer 3 and provide a context for their use. In this sense, data models cross-cut the Attribute and Value Spaces. To take four examples at different points on a continuum of complexity and formalization:

- Dublin Core is based on a simple grammar of Elements and Element Refinements (resource attributes in a generic sense, such as Date Published) and Encoding Schemes (for giving context to element values) [DC].
- The IEEE Learning Object Metadata group is based on a more elaborate and hierarchical model which groups data elements under categories such as General, Lifecycle, Metametadata, Technical, Education, Rights, Relation, Annotation, and Classification [LOM].
- The Indecs model uses an extensive framework for describing Entities and their Attributes (Elements) with a focus on Events that relate Parties to Creations by way of Transactions [RUST].
- Pre-Web standards such as Machine Readable Catalogue in the library world (MARC), finally, may be based less on formal data models per se than on detailed, formalized rules that have evolved in a pragmatic manner over many decades of cataloging practice [MARC].

Moving down to Layer 2, closer to specific application environments, one finds a diversity of adaptations and extensions to the standard attribute and value sets, such as XML schemas that mix and match elements from multiple sources. In the translation to Layer 2, the conceptual systems of Layer 3 may be adapted or even distorted in subtle ways by the constraints of particular encoding languages.

The problem of distortion in the translation between layers may hold in the opposite direction as well: Metadata designed entirely for use within a specific application environment may be wholly pragmatic and ad-hoc in nature, as is the case with hard-coded templates designed without reference to particular data models, standard attributes, or controlled values of Layer 3. Where interoperability requires conceptually clean mappings, retrospectively deriving such mappings from metadata that was not designed with this need in mind can involve its selective interpretation in terms of Layer 3 models and element sets.

### 2.4 Identifying elements

The World Wide Web Consortium (W3C) has articulated a vision of global networks in which resources are identified with names that are globally unique. The form of identifier promoted for this purpose is known generically as the Uniform Resource Identifier [URI]. URIs are not limited to identifying network-retrievable resources: a URI can refer to anything one might want to point to or talk about. Since metadata elements are, in Web terms, "resources" along with everything else, it follows that metadata elements can be uniquely identified using URIs. The Dublin Core Metadata Initiative, for example, assigns URIs to the elements it has defined: "Extent" and

"Medium" can be referenced unambiguously using the URIs "http://purl.org/dc/terms/extent" and "http://purl.org/dc/terms/medium"[4].

In the "CORES Resolution on Metadata Element Identifiers" of December 2002, maintainers of the GILS, ONIX, MARC21, CERIF, DOI, IEEE/LOM, and Dublin Core standards have agreed to assign URIs as identifiers for their metadata elements. Analogously to ISBN numbers for books, URIs will allow specific metadata elements to be used or cited with precision, which is aseen as a useful first step towards the development of mapping infrastructures and interoperability services. Just as importantly, the signers of the CORES Resolution have committed their organizations to the formulation of official policies regarding the stability, persistence, and maintenance of these URIs over the long term [CORES-RESOLUTION].

Although designed to be identifiers, URIs that begin with http: (the most common prefix) also look alot like URLs – i.e., addresses of specific files on specific servers somewhere on the Web. The W3C specifications do not themselves require the URI for an element to resolve to a particular document. Not unreasonably, however, many people expect that "clicking on" such a URI in a browser will call up a representation of that element, for example in a Web page with authoritative definitions. Some people have long argued that the URIs of metadata elements should be considered "just identifiers" (character strings), but the notion that they should resolve to "something" has gained some acceptance. Exactly what they should resolve to – e.g., RDF schemas, XML schemas, Web pages in XHTML, or RDF embedded in XHTML – is still unclear. Emerging technologies for content negotiation may let users choose between several such options.

### 2.5 Describing metadata elements in context

Alongside a large community of users oriented primarily to the World Wide Web, a community of practice around the ISO/IEC 11179 family of standards focuses on the problem of anchoring metadata elements in a generically defined context. Considering "data elements" as the smallest, most irreducible units of fact within a given system, ISO/IEC 11179 Part 3 defines each such element in terms of attributes of five types: identifying, definitional, relational, representational, and administrative. Building on Part 3, ISO/IEC 11179 Part 6 describes a hierarchy of central and domain-specific registration authorities for associating data elements with maintenance agencies [ISO11179-6].

---

[4] Within the content of a document encoded in XML, a metadata element is identified in relation to a construct called XML Namespace. The XML Namespace is defined in a W3C Recommendation as "a collection of names, identified by a URI reference, which are used in XML documents as element types and attribute names" [XML-NAMESPACE]. On the example of the Dublin Core term "Extent", the Dublin Core "namespace" (designated with the URI "http://purl.org/dc/terms/") provides a context for an element named "extent". In an XML document, then, the Dublin Core term "extent" is represented with an XML Qualified Name, or "Qname", which is formed by prefixing the element name with a placeholder for the namespace, as with the Qname "dcterms:extent" (where the prefix "dcterms:" stands for a namespace which, in turn, is identified as "http://purl.org/dc/terms/"). The pair of namespace URI ("http://purl.org/dc/terms/") plus element name ("extent") – in XML terms, a *universal name* or *expanded name* – may in practice be used to derive a URI (e.g., "http://purl.org/dc/terms/extent"). However, the syntactic equivalence between a URI (seen as a string) and an XML expanded name (seen as a string) is in some sense superficial, and the potential dangers of relying on this practical equivalence for the purposes of interoperability is currently a topic of much debate.

The goal of the ISO 11179 standards is to provide a precise and unambiguous description of the nature, conditions of use, and maintenance context of data elements such that independent parties can understand, find, and reuse them in other systems. Although promoted by an active user community, however, ISO 11179 has hitherto not been widely used by standards-developing organizations for declaring their metadata elements [CORES-SURVEY].

### 2.6 Schemas

In current usage, the term "schema" refers to a variety of constructs ranging from the very abstract to the very specific[5]. At the most abstract, "schema" can be used to designate a set of terms – e.g., metadata elements or subject headings – along with their attributes, such as name, identifier, definition, or relationship to other concepts (in our model the Attribute and Value Spaces on Layer 3). As discussed above, however, we prefer to call these "metadata element sets" or "value sets" and to reserve the term "schema" for representations of the same on Layer 2.

On Layer 2, "schema" can refer to any one of several quite different constructs. A file describing a set of XML elements and their interrelationships might loosely be considered a "document schema". XML Document Type Definitions (DTDs) and XML schemas, for example, are used to parse and validate the element structure of a specific document or metadata record [XML-SCHEMA]. An RDF schema, in contrast, is designed to describe the semantic relationships between terms identified with URIs in the global Web space and might in this sense be considered a "semantic schema" [RDF-SCHEMA].

Document schemas have gained rapid acceptance among industrial users because it is straightforward to write software to validate and process documents in XML. As DTDs and XML schemas proliferate by the thousands, however, it becomes proportionally more difficult to merge metadata from multiple sources into integrated wholes. The problem is most evident when faced with the uncontrolled diversity of resources on the open Web.

### 2.7 Semantic Web

The problem of integrating information from a diversity of sources is a key motivator of "Semantic Web", a vision articulated by Tim Berners-Lee that is being pursued by the World Wide Web Consortium [SEMANTIC-WEB]. The Semantic Web idea rests on a few core architectural principles: a simple, linked data model for creating webs of information about related things using metadata statements of a common pattern; the use of Uniform Resource Identifiers (URIs) and XML namespaces to give unique identity to the metadata elements used to describe resources; and the use of XML as a universal file and data exchange format. The vision is based largely on the notion that a shared grammar for metadata "statements", such as that provided by RDF, is needed to ensure that humans and software will interpret metadata consistently. According to this approach, the URIs used in RDF statements serve as "anchor points" for merging statements drawn or extracted from multiple sources. URIs can likewise associate a statement's terms with

[5] Use of the Greek plural for schema, "schemata", seems to be waning. Confusingly, the word "scheme" is sometimes used as a synonym for "schema" in all of the senses discussed in this section. However, "scheme" also has quite different meanings in particular contexts. In the jargon of Dublin Core, for example, an "encoding scheme" is a type of metadata term that provides context for interpreting an element value; and the Dewey Decimal System is sometimes called a "classification scheme". This paper generally avoids the term "scheme" because of this ambiguity.

appropriately identified element sets, controlled vocabularies, or ontologies to anchor the statement in a specific semantic context.

It is recognised that the process of normalizing the diversity of metadata constructs of the world to a simple, uniform, almost pidgin-like statement grammar may involve a certain loss of specificity, and that exporting statements to unintended contexts may not always make sense, but these problems are accepted as an inevitable aspect of imperfect communication in an imperfect world. Rather, the more modest goal is "partial understanding" – the inevitably imperfect, lossy, selective merging of data from underlying models that are semantically and structurally richer and more diverse.

### 2.8 Application Profiles

Profiles, or application profiles, have emerged over the past few years as a vaguely defined but recognizable construct for adapting standard terms for specific purposes. Analogous notions of "profile" have been formulated in parallel for standards as diverse as Z39.50 (a protocol for accessing distributed databases), IEEE Learning Object Metadata (for describing educational materials), Digital Object Identifiers (for describing intellectual property), Dublin Core (for simple resource description), and the National Spatial Data Infrastructure in the US [DC], [Z3950], [DOI], [IEEE1484]. All of these notions of profile aim at providing a way to extend or constrain the use of a standard in order to optimize it for a particular application, function, organization, or user community.

In some of these variants, an application profile is a fully conforming instantiation of an element set for a particular community. It may involve making some elements mandatory; constraining value spaces; or imposing specific relationships between elements. The purpose of such an application profile is to adapt an element set into a package tailored to the functional requirements of a particular application while retaining interoperability with the base standard.

Other, more documentary styles of profile describe how information providers "mix and match" terms from multiple standards in order to meet the descriptive needs of a particular project or service [HEERY]. In the Dublin Core Metadata Initiative and in several European projects, application profiles stand by definition in contrast to element sets: elements are declared in element set schemas and reused in profiles [DCMI, SCHEMAS, DESIRE, CORES]. Such profiles are seen as a loosely defined construct with which any number of usage notes and annotations may be associated.

## 3 Registry services

### 3.1 Pioneering registry activities

There are historic precedents in the database world for the registries that are now emerging in the context of the World Wide Web. A “data dictionary” describing the information structure of a traditional database – its data elements and their interrelationships, attributes, data types, and uniqueness constraints – is roughly analogous to a registry description of metadata on the Web. Directories of data elements arose from a recognition of the benefits of shared data dictionaries leading to the specification of a formal registration process in the standard ISO/IEC 11179. In this tradition, some registries aim at providing a reference tool for interpreting or reusing a wide range of complex data sets. The Environmental Data Registry (EDR) hosted by the US Environmental Protection Agency, for example, provides a tool for interpreting environmental data. The EDR documents over 9,000 data elements used by 54 submitting organizations with the aim of

improving the sharing of data among environmental programs [EDR]. The National Health Information Knowledgebase hosted by the Australian Institute of Health provides access to data definitions and standards related to health, housing, and community services using a data model and registration process based closely on ISO 11179 [NHIK].

More recent activities in the context of the Web share similar goals of enabling interoperability between systems, promoting the re-use of existing data elements, and ensuring the authoritativeness of data definitions. Many of the recent activities, however, put more emphasis on the "set" of elements, reflecting the significance within the bibliographic and information management tradition of relatively small-scale, coherent "vocabularies" as compared to large-scale data dictionaries. Some registries aim at controlling terminologies in use within particular domains. The LEXML initiative, for example, is developing a multi-lingual and multi-jurisdictional dictionary for the legal world. Using Resource Description Frame (RDF) as its modeling basis, the intention is for the LEXML prototype to act as a catalyst for a network with other RDF dictionaries in Europe [LEXML]. The UN's Food and Agriculture Organization is developing an Agricultural Ontology Server as a reference tool for standardizing terminology for use by builders of information resources in the agricultural domain [AOS].

Other types of registry aim at providing specific services or operational components of services. The xml.org directory hosted by the Organization for the Advancement of Structured Information Standards (OASIS) indexes descriptions of XML document specifications such as Document Type Definitions (DTDs) for sharing and re-use among applications [XML-REGISTRY]. The Distributed Metadata Services under development at the Corporation for National Research Initiatives are designed to associate various types of metadata with services capable of converting instance metadata, on demand, from one format to another [BLANCHI].

One cluster of related registry efforts focuses to some extent on Dublin Core, its adaptations and extensions. A prototype registry maintained by the Dublin Core Metadata Initiative provides an interface for exploring DCMI term sets, relationships between terms, and translations of their labels and definitions into various languages. The intention is to provide both users and software applications with reliable information updated in various forms upon demand [DC-REGISTRY]. The MetaForm database at the State and University Library in Goettingen describes adaptations and crosswalks for Dublin Core. MetaForm is in part an attempt to track "dialects" in the practical use of one particular standard in various "manifestations", especially as it is used in Germany [METAFORM]. A Metadata Observatory maintained by the CEN/ISSS Workshop on Multimedia Information looks at the relationship between Dublin Core and emerging standards for multimedia [CEN-OBSERVATORY]. The ULIS Open Metadata Registry, now maintained at the University of Tsukuba in Japan, has focused on linking reference descriptions of Dublin Core metadata terms in several different languages to related materials, such as descriptive elements of Nippon Cataloging Rules [NAGAMORI].

Other registry activities are designed to serve specific user communities. The UK's Metadata for Education Group indexes standard element sets and application profiles used within the UK educational community [MEG]. A registry established by the EU-funded SCHEMAS Project and currently being advanced in a successor project, CORES, targets the universe of European projects, indexing 'standard' metadata element sets along with application profiles that use those standards and activity reports describing metadata-related activities and initiatives [SCHEMAS, CORES].

### 3.2 Registry service models

Registry models differ in several dimensions. A database of pointers to elements sets and application profiles, each perhaps with a simple description, might be called a "shallow" registry. Although such a registry is valuable as a resource locator, there is significantly more potential in "deep" registries that provide machine readable access to various sorts of schemas, indexing them for structured search and browsing. Ideally, such registries would provide search and browse access across the boundaries of many different element sets.

Registry services can be organized around a number of alternative foci:

- An individual standard – providing authoritative current and historical information on a particular metadata standard, perhaps linked to user guidelines;
- A core standard or ontology – serving as a target for mappings from a diversity of other standards or ontologies for the purpose of information integration;
- Extensions of a standard – providing information on how a particular standard has been extended and localized by communities of use;
- Data warehouses – storing definitions of data elements and data types for the purpose of integrating a large number of structured databases into a central repository;
- Usage within domains – providing access to schemas of interest to a particular domain such as education, cultural heritage, or commerce;
- Metadata functions – providing access to schemas of use for particular tasks, such as resource discovery, digital rights management, or user profiling;
- Corporations or communities – providing access to knowledge frameworks or taxonomies used in enterprise portals or corporate intranets;
- Application-based – providing schemas available in a particular syntax or format for use in specific software applications;
- Mappings and conversions – providing services for translating metadata between different metadata systems.

These various usage scenarios imply different business and operational models. One insight of registry-building activities regards the non-scalability of centralized solutions. A number of early prototypes achieved proof of concept from the standpoint of content by manually entering their information into a database. While this method may achieve reliable results rather quickly, the overhead involved in keeping track of information maintained by others quickly becomes overwhelming.

Harvesting methods, such as the protocol-based method of the Open Archive Initiative now used for harvesting a variety of types and formats of metadata located on dozens of repository servers, provide a compelling model for how registries might be organized in the future [OAIPMH]. In the context of metadata languages, however, a harvesting strategy presupposes widely-understood guidelines that allow maintainers to publish their element sets, application profiles, or controlled vocabularies in a form that can be harvested.

Some schema languages, such as Resource Description Framework, lend themselves especially well to harvesting from multiple sources for merging into central databases. Indeed, making an RDF schema available on the Web with a URL to be harvested with an HTTP "get" command may itself be seen as a form of registration. The need to fulfil different requirements has led to the development of diverse and sometimes competing standards such as XML Schema and RDF Schema. This has hampered convergence on technical solutions, while the data models for declaring metadata terms remain the object of discussion and experimentation. Registry prototypes in RDF and related Semantic Web technologies will continue to be an active topic of research for the foreseeable future.

# 4 Open issues

## 4.1 Good practice for declaring element sets

At present, metadata element sets are declared in a variety of publication formats, from paper documents and Web pages intended for human consumption to XML and RDF schemas. In the recent "CORES Resolution", many of the major metadata standards are now undertaking to identify their metadata terms with URIs [CORES-RESOLUTION]. The resolution itself specifies that URIs are used as identifiers with no requirement or expectation that the URIs will reference any particular content on the Web, such as documentation pages or machine-processible representations of metadata elements. Clearly, though, there is a potential here for a broader consensus on the form of machine-understandable schemas. Agreement on such issues would move the Web metadata community one step closer to the integrated environment envisioned in the idea of Semantic Web.

## 4.2 Simple methods for declaring application profiles

As illustrated by some of the registry activities described above, there is a widespread need for implementers and user communities to declare how they are using or adapting standards in their metadata. Similar "profile" constructs have been invented within most standardization communities, but the differences between these constructs are still poorly understood. Convergence on simple principles for declaring profiles would help meet a growing demand on the part of registries and software vendors to use and incorporate profiles in their services, which would in turn promote the harmonization and stabilization of good-practice profiles within user communities. Registries that make visible the broad landscape of metadata practice could facilitate the identification of empirical usage trends and feed back into more "bottom-up" processes for standardization.

## 4.3 Machine-processible crosswalks

The process of mapping between diverse element sets cannot reliably be left entirely to algorithms and heuristics; manual intervention by experts is usually needed to resolve unusual constructs or idiosyncrasies. At present, this work is usually done on paper, then hard-coded into software modules. Ideally, however, such mappings would be expressed in a form reusable in automated environments for converting metadata on the fly. Such "mapping profiles" might be used for declaring equivalencies or near-equivalencies between different element sets retrospectively – for example, to declare that the non-standard Title element of a local application is equivalent to the Title in Dublin Core, which is in turn equivalent to the General.Title element in LOM. The use of URIs provides a common method of citing such elements across many standards communities.

## 4.4 Generalized metadata "types"

There is at present no common understanding of the types, formats, and genres of metadata in use on the Web. Some level of agreement on basic types and on methods for identifying those types are a precondition for some of the more sophisticated approaches for metadata interoperability currently under development. CNRI's Distributed Metadata Services, for example, is a metadata registry infrastructure capable of describing, managing and accessing heterogeneous metadata to be dynamically rendered interoperable using a registered metadata type mechanism. This metadata registry leverages the functionality of the CNRI digital object architecture which seeks to manage, describe and provide distributed access to information in terms of its intended use. In this metadata registry, metadata records are contained within one or more digital objects and can be associated with a set of uniquely identified distributed services. Each identified metadata service provides

functionality specific to the metadata it represents such as describing the specific characteristics of the metadata's schema or providing services to convert any of its associated metadata record from one schema into another. Standardization of such high-level "content types" could bear some resemblance to methods for enabling browsers to acquire new plugins or to Multipurpose Internet Mail Extension (MIME) registries for types of Internet email attachments [BLANCHI].

### 4.5 Conventions for declaring controlled vocabularies

Analogously to methods for declaring metadata element sets and application profiles, there is a need for good-practice methods to declare controlled vocabularies such as classification schemes, thesauri, subject headings, and ontologies. Like element sets, controlled vocabularies would ideally be declared using URIs and in ways that are reusable by a broad range of registries and applications. The declaration of vocabularies such as thesauri present unique problems in modeling complex webs of related terms, but many requirements are analogous to those for element sets, such as a need for application profile constructs that package subsets of a standard for subject headings or that adapt those headings for particular uses. Standardization work in this area is currently taking place in contexts such as the CEN/ISSS Workshop on Learning Technologies and in W3C technical committees on ontology languages [CEN-ISSS-LT, OWL].

### 4.6 Registries for controlled vocabularies

In principle, controlled vocabularies could be accessed through the same registries as metadata element sets, but differences in the nature of these vocabularies – their sizes, granularity, inner structure, and expected use – imply different sorts of interfaces. Some of the initiatives to build registry environments that harvest and link controlled vocabularies in machine-processible ways are described above. Numerous experimental systems are being developed in academic and industrial research environments. Research into "subject mediators" at the Russian Academy of Sciences, for example, aims at capturing community agreements on data structures, thesauri, and ontologies for specific subject domains in a metainformation base analogous to the registries of metadata elements discussed above. Subject mediators are designed to convert and reconcile a diversity of structures relevant to subject domains with that of the mediator and to provide a uniform query interface to the data sources registered at the mediator. Registration processes use an Intermediator Protocol based on a harvesting concept similar to that of the Open Archives Initiative [KALINICHENKO].

## 5 Conclusion

Within the limited context of the DELOS Working Group on Registries, participants from a wide range of backgrounds and perspectives were able to agree on a common language for talking about metadata structures. Underlying this agreement was the broad assumption that future developments in metadata registries would play out in the global context of the World Wide Web. The layered model of interoperability allowed the group to make helpful distinctions such as those between Value Spaces and Attribute Spaces and between conceptual systems and bindings or encodings based on those systems.

Difficulties encountered in reaching this shared view, however – phenomena that straddled the neat boundaries of our model – hinted at the deeper challenge of reaching a shared language for these issues in the broader Web community. The conceptual "interoperability" we achieved around our shared model entails a measure of simplification, lossiness, and ambiguity. By analogy, metadata registry applications can do no more than present, process, use, or navigate sets of data, and any

interoperability that results will depend on the fit between the models underlying that data. "Semantic interoperability" is no more perfectly attainable in metadata than in any other domain of human understanding.

In the end, moreover, the group found itself unsure of the distinction between a registry application and the data structures underlying that application. One view holds that "the Web is the registry" – a collection of machine-understandable schemas fetched from a distributed body of Web servers with the HTTP protocol for use in various applications. A simple set of pointers to such a distributed body of schemas is what some people call a "thin registry". However, many of the application scenarios outlined above imply "thick" registries, where the integration of content from multiple sources, processes of editorial selection, the addition of descriptions or annotations, or the provision of navigation or searching add value beyond the sum of raw data. Building such applications, however, seems much easier than the task of getting providers on the Web to supply their information in a form usable by registries in the first place, which explains why the group perhaps inevitably ended up focusing on foundational technologies and data conventions.

## References


[AOS] Agricultural Ontology Service Project, http://www.fao.org/agris/aos/.

[ARIADNE] Ariadne Foundation, http://www.ariadne-eu.org/.

[BLANCHI] Blanchi C., Petrone J., "Distributed Interoperable Metadata Registry", D-Lib Magazine 7:12 (December 2001), http://www.dlib.org/dlib/december01/blanchi/12blanchi.html

[CEN-ISSS-LT] CEN/ISSS Workshop on Learning Technologies, http://www.cenorm.be/isss/Workshop/LT/Default.htm.

[CEN-OBSERVATORY] CEN/ISSS Workshop on Multimedia Information Metadata Observatory, http://www.sub.uni-goettingen.de/ssgfi/observatory/.

[CORES] CORES Project,http://www.cores-eu.net/.

[CORES-RESOLUTION] CORES resolution on metadata element identifiers, CORES Project, http://www.cores-eu.net/interoperability/.

[CORES-SURVEY] Issues in cross-standard interoperability, CORES Project, http://www.cores-eu.net/interoperability/d31/.

[DC-REGISTRY] Dublin Core Metadata Registry, http://dublincore.org/dcregistry/.

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

[DELOS-ONTOLOGY] "Building Core Ontologies: A White Paper of the DELOS Working Group on Ontology Harmonization", http://delos-noe.iei.pi.cnr.it/activities/standardizationforum/ontology/ontology.html.

[DEMPSEY] Dempsey L., "Divided by a Common Language: Digital Library Developments in the US and UK," presented at JISC/CNI Conference, Edinburgh, 2002.

[DESIRE] DESIRE Metadata Registry, http://desire.ukoln.ac.uk/registry/.

[DIFFUSE] Metadata Interchange Standards, DIFFUSE Project, http://www.diffuse.org/meta.html.

[DOI] Digital Object Identifier System, DOI Application Profile, http://www.doi.org/doi-ap.html.

[DUVAL] Duval E., Hodgins W., Sutton S., Weibel S., "Metadata Principles and Practicalities", D-Lib Magazine 8:4 (April 2002), http://www.dlib.org/dlib/april02/weibel/04weibel.html.

[EDR] Environmental Data Registry, http://www.epa.gov/edr.

[FGDC] Guidelines for creating a profile for the Content Standard for Digital Geospatial Metadata, Federal Geographic Data Committee, http://www.fgdc.gov/metadata/csdgm/profile.html.

[HEERY] Heery R., Patel M., Application profiles: mixing and matching metadata schemas, Ariadne 25 (September 2000), http://www.ariadne.ac.uk/issue25/app-profiles/intro.html.

[IEEE1484] IEEE 1484.18: Platform and Media Profiles, http://edutool.com/pmp/.

[IEEE-LOM] IEEE P1484.12 Learning Object Metadata Working Group, http://ltsc.ieee.org/wg12/.

[INDECS] Indecs, http://www.indecs.org/.

[ISO11179-6] ISO/IEC 11179-6, Registration of Data Elements, http://www.sdct.itl.nist.gov/~ftp/l8/other/Standards/iso11179/iso11179-6.PDF.

[KALINICHENKO] Kalinichenko L.A., Briukhov D.O., Tyurin I.N., Skvortsov N.A. "Intermediator framework protocol for information sources registration at heterogeneous mediators", *Proceedings of the DELOS Workshop "Interoperability in Digital Libraries"*, 8-9 September 2001, Darmstadt, Germany, http://synthesis.ipi.ac.ru/synthesis/publications/intframe/intframe.zip.

[LEXML] Open Source Development of an RDF Dictionary, http://home.snafu.de/mmuller/lexmlde/rdf.htm.

[MARC] MARC Standards, http://lcweb.loc.gov/marc/marc.html.

[MEG] MEG Registry Project, http://www.ukoln.ac.uk/metadata/education/regproj/.

[METAFORM] MetaForm, http://www2.sub.uni-goettingen.de/metaform/.

[MILLER] Miller P., "Interoperability What is it and Why should I want it?", 21 June 2000, Ariadne Issue 24, http://www.ariadne.ac.uk/issue24/interoperability/.

[NAGAMORI] Nagamori M., Baker T., Sakaguchi T., Sugimoto S., Tabata K., "A multilingual metadata schema registry based on RDF Schema", *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001 (DC-2001)*, pp. 209-212, Tokyo, http://www.nii.ac.jp/dc2001/proceedings/product/paper-31.pdf.

[NAMES] Namespaces in XML, World Wide Web Consortium, http://www.w3.org/TR/REC-xml-names.

[NHIK] National Health Information Knowledgebase, http://www.aihw.gov.au/knowledgebase/index.html.

[OAIPMH] The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.htm.

[OWL] Web Ontology Language, http://www.w3.org/TR/owl-ref/.

[PAEPCKE] Paepcke A., Chang C.C.K., Winograd T., Garcia-Molina H. "Interoperability for digital libraries worldwide". Communications of the ACM, 41(4):33-42, 1998, http://www.acm.org/pubs/citations/journals/cacm/1998-41-4/p33-paepcke/.

[RDF-SCHEMA] Resource Description Framework, Schema Specification, http://www.w3.org/TR/rdf-schema/.

[RUST] Rust G., Bide M., "The <indecs> metadata framework: Principles, model and data dictionary", http://www.indecs.org/pdf/framework.pdf.

[SCHEMAS] SCHEMAS Forum for Metadata Implementors Registry, http://www.schemas-forum.org/registry/.

[SEMANTIC-WEB] W3C Semantic Web Activity, http://www.w3.org/2001/sw/Activity.

[SIGMOD] Special section on semantic interoperability in global information systems, ACM SIGMOD Record 28:1 (March 1999), http://www.acm.org/sigmod/record/issues/9903.

[URL] URI Generic Syntax, http://www.ietf.org/rfc/rfc2396.txt.

[W3C] World Wide Web Consortium, http://www.w3.org/.

[XML-NAMESPACE] Namespaces in XML, World Wide Web Consortium, http://www.w3.org/TR/REC-xml-names/.

[XML-REGISTRY] XML.ORG, http://www.xml.org/xml/registry.jsp.

[XML-SCHEMA] XML Schema, World Wide Web Consortium, http://www.w3.org/TR/xmlschema-0/.

[Z3950] About Profiles, Z39.50 International Standard Maintenance Agency, http://lcweb.loc.gov/z3950/agency/profiles/about.html.