Data Modeling and Harmonization with OWL: Opportunities and Lessons Learned

John L. McCarthy[1], Denise Warzel[2], Elisa Kendall[3], Bruce Bargmeyer[1], Harold Solbrig[4], Kevin Keck[1] and Fred Gey[5]
[1]Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA
[2]National Cancer Institute, National Institutes of Health, Bethesda, MD
[3]Sandpiper Software, Los Altos, CA
[1]Mayo Clinic, Rochester, MN
[5]School of Information Management, University of California, Berkeley

Abstract. Experience from recent projects helps illuminate the promises and limitations of OWL to specify, review, refine, harmonize and integrate diverse data and concept models. One of the attractive features of OWL is that it can be used by inference engines to help augment queries through inferred semantic relationships. But OWL, like SQL, is only a computer programming language. Using it to review and refine representations of data, metadata, and concept systems, including terminologies, thesauri, and ontologies, requires a well-defined abstraction layer – which itself can be specified in terms of OWL. In order to optimize, harmonize and integrate such information effectively for large scale projects, OWL definitions and relationships should be specified in terms of a standard metamodel, such as ISO/IEC 11179-3, Edition 3.

Keywords: OWL, metadata, UML, ISO 11179, metadata registration, data modeling

1 Introduction

Information technology experts are beginning to recognize the need to combine multi-disciplinary data, metadata, and concepts from a variety of related fields to address complex and/or large scale problems. Doing so requires integrating content from diverse communities with long established but different terminologies, concepts, and ways of naming and organizing their data. This paper explores how OWL (Web Ontology Language [1]) can be used to help advance decades-long efforts to represent, manage, harmonize, and integrate metadata and semantics for concept systems (including taxonomies, thesauri, and ontologies), databases, data elements, and value domains (*i.e.*, data types, and sets of valid values). For the purpose of this paper, the word "ontology" refers to a domain specific conceptualization, for a specific purpose [2].

As elaborated further below, OWL, like other computer languages, is intended for a specific purpose, it is a declarative representation language for representing knowledge and used for authoring ontologies. OWL does not provide built-in, standard modeling constructs to harmonize between ontologies, let alone describe their interrelationships or external relations to data, metadata about databases or application systems, nor to manage the evolution of such relationships over time. Several recent projects illustrate how OWL provides some very useful constructs that complement capabilities of other software engineering technologies and paradigms and can be used in conjunction with them to support new ways to model concept and data semantics. One such paradigm for concept and data management is the metadata registry, (MDR) particularly those based on the ISO/IEC 11179 Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes Edition 3 (E3), [3] and related ISO/IEC 19763 Metamodel Framework for Interoperability (MFI) [4] specifications which are being extended to represent relationships between and across ontologies, as well as relationships between ontologies, terminologies, data models and web services that implement or reuse them.

Our discussion is based in part on results from three recent data and semantic modeling projects that all employed OWL, each for different purposes. The first of these projects used the Ontology Definition Metamodel (ODM)[5] to represent the BRIDG[1] model [6, 7], and transform it to OWL to help analyze and identify potential shortcomings in BRIDG [8]. The second project attempted to use an automatic ODM-based conversion tool to transform the LexGrid [9] terminology model from XML Schema into OWL. The eXtended Metadata Registry (XMDR) project [10] has used OWL in conjunction with other tools to develop a prototype system that implements extensions and enhancements included in the current CD of ISO/IEC 11179 E3, Standard. Lessons learned from these efforts will be highlighted in the sections that follow.

2 Background and Motivation

Descriptions of data, how it was collected, and what it means are an essential component of modern information systems. These descriptions, called metadata (*i.e*., data about data), help ensure that data is interpretable by both humans and computers over time. Large organizations like the U.S. Environmental Protection Agency (EPA), National Cancer Institute (NCI), and Department of Defense (DOD) perform research that utilizes large amounts of data drawn from a variety of disparate systems. They have long recognized the need for *standardized metadata registry systems* to help manage and harmonize data elements from different databases and application systems [10]. At the same time, such organizations, along with others in Europe and Asia, were among the first proponents of national and international standards for terminologies, thesauri, concept systems and ontologies[2].

In parallel, NCI and small communities of data modelers and software engineers have been using ontologies to extend the capabilities of their metadata registry and software systems to enhance the semantics, identify potentially duplicate metadata, and increase the potential for reuse [11, 12]. Ontologies and ontology tools can facilitate automation supporting categorization and reasoning about increasingly massive amounts of data and metadata that many large organizations have to cope with. Despite apparent potential benefits from cross pollination between data management, data governance, and related disciplines that use semantic technologies, little progress has been made in marrying the two communities outside of a few isolated United States and European government agency activities. The overlap between the enterprise data and semantic web communities is very small at present, as evidenced by discussions at recent workshops at the En-

---

[1] The Biomedical Research Integrated Domain Group (BRIDG) Model is a domain analysis model that describes biomedical/clinical research data.

terprise Data World[3] and Semantic Technology Conferences[4] earlier this year.

The National Cancer Institute (NCI)'s data management infrastructure, comprised of the Cancer Data Standards Registry (caDSR) and Enterprise Vocabulary System (EVS), is a notable exception to the low level of synergy between data management and ontology communities. caDSR and EVS have enabled NCI to collect, harmonize and integrate detailed metadata and concepts describing some 5,500 data elements and case report forms in hundreds of clinical studies from over 90 different projects. The semantics of these elements are tied directly to over 10,000 concepts drawn from the NCI's Enterprise Vocabulary System (EVS) [13]. The U.S. Environmental Protection Agency's Environmental Data Registry (EDR), which has information covering water, air, and soil in databases and application systems, with many data elements, value domains, and terms from different terminology systems (though not directly linked) also illustrates the synergy between utilization of terminology systems and data management (*e.g*., GEMET, Chemical Substances Taxonomy, etc.) [14].

Use of formal languages such as OWL to represent data semantics linked to terminologies can provide a tremendous opportunity for novel research and discovery, in particular if the expression of the data semantics is based on a well-defined, shared metadata model and the terminologies are well formed. World Wide Web Consortium founder Tim Berners-Lee, speaking about the "semantic web," stated that "The concept of machine-understandable documents… indicates a machine's ability to solve a *well-defined* problem by performing *well-defined* operations on existing *well-defined* data. Instead of asking machines to understand natural language, it involves asking people to make the extra effort" [15]. (Italics are ours). The implication is that people will have to take additional steps to create machine-understandable documents. Utilizing ISO/IEC11179-3 Ed 3 integrated metamodel for data and concept systems provides well-defined data descriptions for use with the OWL representation , making data comparable within and across communities because the common structure of the metadata allows programmers to develop well-defined operations for machine interpretation. Recent projects have demonstrated the power of such integrated infrastructures, but achievement of this level of integration does not come automatically simply by adopting a new language such as OWL [16].

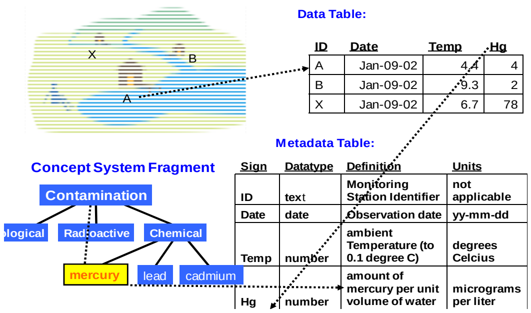
3 Levels of Abstraction for Data, Metadata and Ontologies

To elucidate these ideas more concretely, consider a hypothetical example of something in the real world about which we want to capture data, representation of the data in a database, information about this data (metadata), and a concept system fragment (which could be part of an ontology). Figure 1 shows this example with three "levels" of information – concepts, data and metadata, and some of the relationships between them. The picture at the top left represents two streams and a lake, each with its own monitoring station (A, B, and X). A formalized concept system could describe the relationships between these bodies of water and monitoring stations. The table at the top right contains three rows of data, one collected from each of

[3] See http://edw2009.wilshireconferences.com/.
[4] See http://www.semtech2009.com/.

those stations on a particular date. Each column in the table contains data for a particular observation of the variables (ID, Date, Temperature, and Mercury contamination), while each row represents values observed at a particular monitoring station location (A represents the row of values for the Lake monitoring station, B the values for the second monitoring station, and X the values for the monitoring station further downstream

Figure 1: Data, Metadata, and Concept Systems



The table on the lower right contains "metadata" – a description of the meaning and purpose of each column in the Data table -- (ID, Date, Temp, and Hg). Each column in the metadata table (*i.e*. Sign, Datatype, Description, Units) contains a piece of information about a column of the Data table, while each row represents a particular column in the Data table. Depicted at the lower left is an excerpt from the General European Multilingual Environmental Thesaurus (GEMET) concept system that shows Contamination of a body of water in three forms (Biological, Radioactive, and Chemical), along with three kinds of Chemical Contaminants, namely mercury, lead, and cadmium. Dotted lines in Figure 1 indicate some of the important relationships between different components of the three types of information. For example, the ID with the value "A" in the first cell of the first row of the Data table is the identifier for the Lake monitoring station in the picture. All the values in that row refer to the monitoring station with the ID "A". Likewise the first cell in the bottom row of the Metadata table refers to the label of the fourth column in the Data table and all the values in that metadata row pertain to the values in that column e.g. the cell of the second column in the bottom row refers to the Datatype of the fourth column in the Data table. Another dotted line shows a relationship between the Definition cell for the bottom row of the Metadata table (which relates to the 4th column of the Data table) to a particular item in the hierarchical diagram of terms from GEMET.

Ideally, we would like to be able to answer queries that span all three levels of information, Data, Metadata and Concept Systems, such as *find water bodies downstream from Fletcher Lake where the level of chemical contamination for any of a specified set of substances was greater than the allowable tolerance between December 2001 and March 2003.* Constructing and answering these kinds of queries, which was difficult at best using traditional database technology, is now possible through the open world reasoning facilities supported by OWL in conjunction with the metadata registry capabilities specified in ISO/IEC 11179-3 Ed 3.

3.1 Information Models, Concept Systems, and Ontologies

From a data engineering perspective, a conceptual or information model typically defines a set of properties and relationships that describe real world entities. Frequently, in order to define information models that will ultimately result in business applications or services, multiple information models are needed, each of which may define various aspects of the same set of entities focusing on different perspectives, context and/or processes. Each such *conceptual* information model may, in turn, correspond to one or more *logical* data models that refine various aspects of the conceptual model, which may then be realized in a number of *physical* models, or schemas that correspond to platform-specific implementations (*e.g.*, XML schema, relational databases, etc.). Depending on the level of formality imposed by the organization responsible for designing the business services, the conceptual modeling part of this process may be short circuited, or even skipped. In some cases where conceptual models *are* developed, they may not be well documented, especially if the models are only shared among a small group of developers where assumptions are implicitly understood. Consider the data values in Figure 1. if the organization had only the headings for the data table, but no other metadata, concept systems, or asserted relationships between them and the data to aid with human or machine interpretation.

From the Semantic Web perspective, an ontology provides the semantic grounding for an information model. They can be one and the same; or additional vocabularies or ontologies can be used to provide terminological support for the core business information model. Consider the often-used example of students, classes and teachers that occurs in literally hundreds of Database textbooks. These examples work because there is a relatively consistent and shared understanding of schools, teachers, students, classes, subjects, etc. within 20th century western school systems. Schools hire teachers, teachers teach classes, students attend classes, classes have schedules, etc. There are literally thousands of different information models scattered throughout these textbooks that reference this set of topics. While there is a component of each of these specific conceptualizations of these topics that is invariant, it is highly unlikely that any two of the independently developed information models or ontologies, are identical.

Typically, the purpose of developing an ontology is to define a particular conceptualization for use in a particular application or context (*e.g.*, if we are describing teachers and students, the fact that students are composed of cells, require a certain amount of nutrition each day to survive, may vote in elections and may have a preferred medical doctor, etc. are probably not relevant and are probably not included in a description of an ontology focused on schools.). Information models, particularly at the logical or physical level, may add features that are not required at the conceptual level, for example, details regarding primary or foreign keys, unique identifiers that are application specific GUIDs, and so forth. Obviously an information model attempts to maintain some sort of correlation between these identifiers and the things being identified, but these are still artifacts of the information model, not of the conceptual level.

3.2 Harmonizing Data across Multiple Systems and Ontologies

Metadata registries provide an abstraction "layer" to systematically describe, manage, and query metadata for databases, applications, and concept systems, and are particularly useful for large-scale, distributed environments. The ISO/IEC 11179 Metadata Registry family of standards also provides guidance for managing the evolution of such information over time. Figure 2 shows conceptually NCI's caDSR metadata model and Enterprise Vocabulary System's supported mappings – using concept systems to defining semantics from high-level, conceptual definitions of ISO/IEC 11179 Object Class and Property, to increasing refinement of meaning at the value set level.

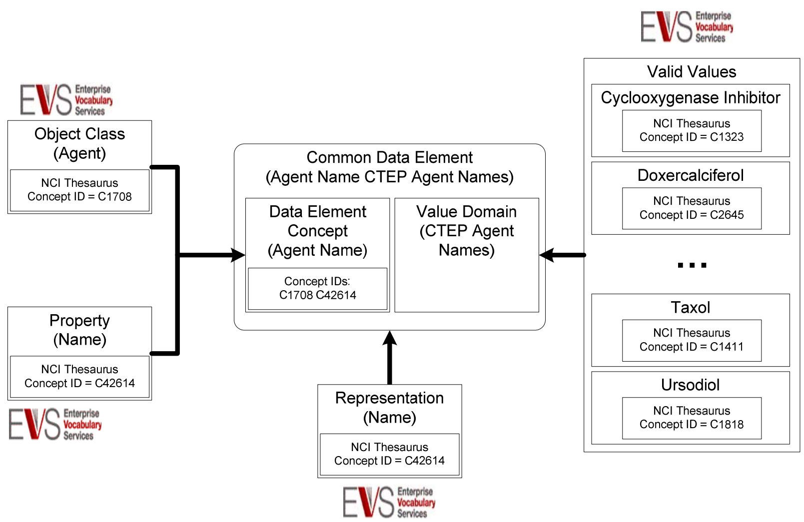


Figure 2. Refinement from Conceptual to Logical Definitions Using ISO/IEC 11179 MDR (left to right)

Just as SQL can be used to describe tables, columns, and relationships between columns (such as foreign keys), OWL is well suited for describing terminologies, concept systems or ontologies and various different types of relationships between their components (*e.g.*, subsets, inverses, aggregations, etc.). SQL in and of itself does not provide built-in constructs to describe, manage and query metadata registries. It can be used to construct standard tables and columns containing metadata registry information so it can be stored, managed and queried using SQL, however. As noted earlier, OWL does not provide built-in constructs to create or harmonize ontologies. Absent standard constructs or techniques for anchoring the semantics of the elements represented in OWL to an external reference that could render them comparable across models, even if those elements are grounded in the same higher-level ontologies, reasoning over multiple OWL representations may be limited to evaluation of potentially related content solely on the basis of text labels for the elements and relations. But like SQL, OWL is a powerful language that can be used to create standard constructs for registries of multiple ontologies, metadata, and their inter-relationships, either through an additional external ontology or via annotations, and through a combination of open and closed world reasoning, can enable new capabilities that make such registries invaluable to their users, particularly for question answering over large distributed repositories.

One recent effort employed OWL to evaluate the BRIDG Domain Analysis Model (DAM) – a multi-agency effort to develop a shared view of the data, relationships, and processes which collectively define “protocol-driven research and its associated regulatory artifacts.” The first formal release of BRIDG, using the Unified Modeling Language (UML) [14] was published in June 2007 [5]. In late 2008, NCI commissioned a study to use the Ontology Definition Metamodel (ODM) [8] to translate the BRIDG DAM from UML to OWL in order to help identify potential problems with BRIDG as a data model. The absence of built-in constructs for comparative purposes resulted in a lot of manual effort to ensure that comparisons between versions of the model were apples-to-apples. The assignment of immutable identifiers and versions to the elements of the owl model emulate an ISO/IEC 11179 MDR approach.

The resulting analysis concluded that there were several flaws and ambiguities in the BRIDG model, including problems with relationship names and types, and what was termed an "*explosion of attributes in the model...due to the creation of a new Common Data Element (CDE) when the concept is the same but the context of use is different."* The report cites postalAddress as an example: *"This attribute occurs 8 times in the model, all with the same AD datatype expression, and all referencing a physical postal address of an entity but all with a slightly different definition."* In this case, however, showing different variants could be considered a feature or strength of the BRIDG model rather than a "bug," depending on what purposes the model is intended to serve. A more general solution might be to have a single generic postalAddress class, with subclasses for different variants that are used by different groups and agencies. The new draft ISO/IEC 11179 (E3) metadata registry standard and XMDR prototype support just this kind of capability to document and manage the evolution of application and database-specific variants while at the same time showing their commonalities and translation requirements. If the semantics of each component of a particular variant are identical then they should be modeled as one object, but if they differ even slightly, it often is helpful to be able to distinguish and identify those subtle differences, as well as to note how one may be transformed to the other, with or without loss of information in one or both directions.

Whether we use OWL, UML, or other representation paradigms such as Entity-Relationship modeling, each of which can serve multiple purposes in the context we’ve described, we still need an additional level of abstraction for management of both data model and ontology information, along with the relationships between them in order to document, manage, and harmonize data and semantics from diverse systems – particularly as they evolve over time.

4 ISO/IEC 11179 Metadata Registry Standard

The first edition of ISO/IEC 11179, which became a formal standard in 1994, established a discipline for standardization of data elements such as those found in databases and data interchange. Organizations leading the development of this standard, including DOD, EPA, and NCI, recognized the need for better ways to represent relationships, and to express relational semantics sufficient to enable machine-processing and elementary logical operations. Subsequent editions add tighter linkage between the semantics found in data, metadata and concept systems. Edition 3 of ISO/IEC 11179 Part 3 provides explicit specifications for registering ontologies, thesauri, taxonomies and other semantic artifacts useful in managing the semantics of data. Work on additions and extensions to ISO/IEC 11179 Edition 2 began in 2004. At the same time, staff at the Lawrence Berkeley National Laboratory began to develop a prototype system to demonstrate the feasibility of implementing software that could be used to load, manage, and query an eXtended Metadata Registry, as specified by the evolving third edition working drafts.

ISO/IEC 11179-3 (E3) provides a means by which multiple definitions of particular terms (be they defined via ontologies, taxonomies, other kinds of models, or other metadata), can be articulated and reused across systems and modeling paradigms. It includes a rich model for provenance about the models managed in a compliant registry, providing a means by which one can say "what sources were used to develop this model/concept", "where is this model/concept used", "who depends on it", "what domain was it intended for", etc. -- which the modeling paradigms themselves do not do (nor do they claim to do). It also supports data stewardship, change management/change control, management of technical status and management of the status of organizational acceptance. Figure 3 shows a summary of the top level classes and a few of their major sub-classes from the current draft of ISO/IEC 11179-3 (E3).

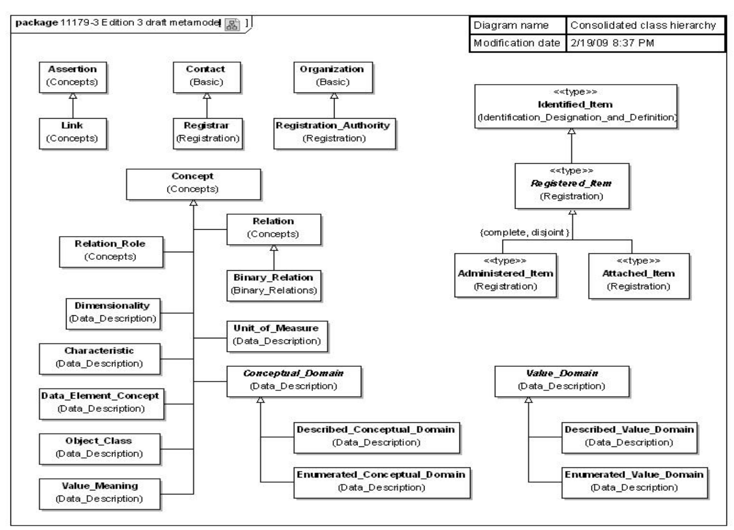


Figure 3. Current Draft ISO/IEC 11179-3 (E3) Consolidated Class Hierarchy

Leading the progression from ISO/IEC 11179 (E2) to (E3), NCI added linkage from the semantic metadata content of the metadata registry to controlled terminology, thus providing a more granular way to compare Object Classes and Properties, and providing a way to link other metadata to controlled terminology so researchers can explore/discover what might be similar or related, to the data item of interest. An overview of the architecture, current implementation, including the use of OWL and UML models, and directions this work is taking at NCI as a part of the cancer Biomedical Informatics Grid (caBIG) program is presented in [17].

5 OWL AND THE 11179 XMDR PROTOTYPE

UML models, OWL ontologies, and XML schemas each capture important, but not identical content useful for design and implementation of information systems. The XMDR project demonstrated this at two levels for metadata registries. First, based on formative ISO/IEC 11179 (E3) specifications, the XMDR was designed to capture and interrelate selected content of UML models, XML schemas, and OWL ontologies, as well as concept systems in terms of semantic relationships as well as traditional metadata. Second, UML models, XML schemas and OWL based technologies were used to implement the XMDR prototype.

In response to user needs for extensions to the ISO/IEC 11179 (E2) metamodel such as the extensions implemented in NCI's caDSR, plus the need to test whether such extensions could be practically implemented and deployed using real data and concept systems, the XMDR project developed several types of metadata registry extensions including:

1. standardized representation of logical and other types of relationships;

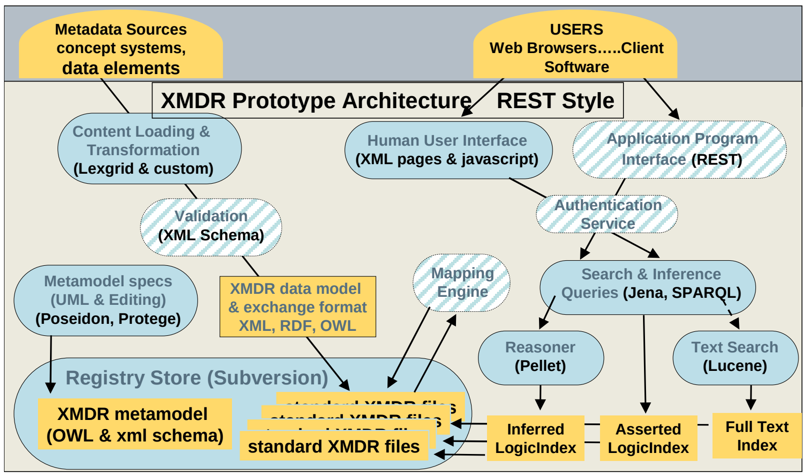
2. a metamodel framework to facilitate controlled and well-documented management and evolution of terminologies, thesauri, concept systems and ontologies in the same way that data elements, value domains, and related types of information are managed in the ISO/IEC 11179 (E2) framework;

3. use of OWL to permit and facilitate reasoning based on logical inference.

After reviewing a number of candidate languages and software packages, LBNL implemented the XMDR prototype as a modular, open architecture system, which makes it relatively easy to substitute software modules for particular components (*e.g.*, database system, reasoner).

Figure 4 shows the over-all modular component architecture of the XMDR prototype, along with particular open source software used for its current components. In this schematic diagram, ovals or rounded boxes depict major components of the XMDR system, while rectangles represent data, metadata, and indexes. Planned extensions are shown in shaded ovals. The XMDR Prototype used Poseidon and Protégé to create and edit UML and an OWL ontology that describes the XMDR metamodel. These specify how metadata is organized within the metadata registry store. The diagram also shows how content is transformed and loaded -- using LexGrid and custom XSLT scripts to create standard "XMDR files." XMDR files are indexed using Jena and Pellet to create RDF files that are stored in a Postgres database, and then further processed using Lucene to create a text index. Human users (using web browsers) can create queries and display results using a combination of JSP code and a standard REST interface. Other software (such as Exhibit[http://www.simile-widgets.org/exhibit/], from MIT's Simile Project [http://simile.mit.edu/]) can make use of the REST Application Programming Interface (API) to provide a "plug and play" Graphical User Interface (GUI).

Figure 4: XMDR Prototype Modular Architecture: with current open source software selections



The XMDR prototype demonstrates how technologies based on UML, OWL and XML Schemas can be used to create a metadata registry, which is used to register, manage, and curate selected content from UML models, XML Schemas, and concept systems (including ontologies, thesauri, taxonomies, etc.) along with traditional metadata such as descriptions of common data elements, value domains, etc. Future plans include replacing Poseidon with ODM-based UML capabilities to leverage forward and reverse engineering of OWL from UML, completion of planned modules, and revision to support the evolving ISO/IEC 11179-3 (E3) standard as it approaches formal adoption.

6 Future Research

While the ISO/IEC 11179 standard itself has been in use for over a decade, its application to the management and use of ontologies and other concept systems is relatively new. Some of the organizations involved in the writing of this paper, and other colleagues within the US/ANSI DM32.8 task force and in the broader international metadata standards community have been actively involved in this evolution for at least the last five years. Implementations are nascent, and the standard itself is just now reaching final committee draft stage. We anticipate additional work will be needed to support development of reference implementations and further evolve the standard, informed by those efforts, before finalization is complete. Integration of increasing levels of automation, through ODM-based tools, through the use of reasoning to facilitate validation, search and retrieval, and other analysis activities are also planned. Evolution of the core ISO/IEC 11179 ontology, potentially reflecting the availability of new features in OWL 2, is also on our roadmap.

7 Conclusion

UML, OWL, and ODM have been instrumental in enabling better ways of characterizing and documenting information models and knowledge but they only provide the basic representation language constructs required as a starting point. Our experiences with the BRIDG 2.0 domain analysis model, development of the new draft ISO/IEC 11179 (E3) Metadata Registry Standard, and implementation of the Extended Metadata Registry (XMDR) Prototype System help illustrate the need to combine these technologies with those such as UML/ODM, OWL, and the Semantic Web with more traditional metadata registry tools and procedures to begin to address some of the really difficult information interchange issues faced by large organizations today. Utilizing a combination of ontology and knowledge representation languages, linked with standard metamodels for describing data and metadata provide an unprecedented opportunity to leverage semantic web for knowledge mining and discovery of hidden links and improve their utility in data management.

Disclaimer

This article is based in part upon work supported by the Environmental Protection Agency, Department of Defense, National Cancer Institute, and the National Science Foundation under NSF Grant No. 0637122. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the sponsors.

References

1. Mike Dean and Guus Schreiber, eds., Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language 1.0 Reference, W3C Recommendation, World Wide Web Consortium, Amsterdam, The Netherlands, 10 February 2004. Latest version is available at http://www.w3.org/TR/owl-ref/.
2. Tom Gruber (1993). "A translation approach to portable ontology specifications". In: *Knowledge Acquisition*. 5: 199
3. ISO/IEC 11179 – Metadata Registries - Part 3 (Edition 3), Available at http://metadata-stds.org/11179/index.html#A3.
4. ISO/IEC 19763, Information Technology – Metamodel Framework for Interoperability (MFI), Available at http://metadata-standards.org/19763/index.html.
5. Christopher A. Welty and Elisa F. Kendall, eds. Ontology Definition Metamodel (ODM), Version 1.0 Specification, Object Management Group, Inc., Needham, MA, May 2009. Latest version is available at: http://www.omg.org/spec/ODM/1.0/.
6. http://gforge.nci.nih.gov/projects/bridg-model and
7. http://bridgmodel.org/
8. https://gforge.nci.nih.gov/projects/bridgrev
9. http://informatics.mayo.edu/LexGrid/

10. http://XMDR.org/
11. Peter A. Covitz∗, Frank Hartel, Carl Schaefer, Sherri De Coronado, Gilberto Fragoso, Himanso Sahni, Scott Gustafson and Kenneth H. Buetow, *caCORE: A common infrastructure for cancer informatics,* Bioinformatics, Vol. 19 no. 18 2003, pages 2404–2412
12. Komatsoulis, GA, Warzel, D.B, Hartel, F.W., Shanbhag, K, Chilukuri, R, Fragoso, G., de Coronado, S, Reeves, D.M, Hadfield, J.G., Ludet, C., and Covitz, P. A. (2008) *caCORE version 3; Implementation of a model driven service-oriented architecture for semantic interoperability*. Journal of Bioinformatics, 41:1
13. caDSR http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr and EVS http://bioportal.nci.nih.gov/ncbo/faces/index.xhtml statistics citation
14. EDR http://iaspub.epa.gov/sor_internet/registry/datareg/home/overview/home.do
15. Berners-Lee, T, http://www.w3.org/DesignIssues/RDFnot.html
16. Unified Modeling Language™ (UML®) Infrastructure and Superstructure Specifications, Version 2.1.2, Object Management Group, Inc., Needham, MA, November 2007. Latest versions of this and related UML specification components is available at: http://www.omg.org/spec/UML/2.1.2/.
17. Alejandra González Beltrán, Anthony Finkelstein, J. Max Wilkinson, and Jeff Kramer *(forthcoming)*. “Domain Concept-Based Queries for Cancer Research Data Sources”.