Practical Application of the Dublin Core Standard for Enterprise Metadata Management

by Camille Mathieu

EDITOR'S SUMMARY

Large organizations relying heavily on knowledge work require effective capture and reuse of information, enabled through consistent use of standardized enterprise content metadata. The Jet Propulsion Laboratory (JPL) has undertaken a standardization effort, building an internal content schema based on established metadata field standards that are content- and application-agnostic but locally customizable for application to a broad variety of repositories. The JPL adopted the Dublin Core standard, with its Simple and Qualified properties as well as further refined Custom sub-properties. The JPL Resource Schema serves as an enterprise-wide metadata standard, while specific application profiles state the available fields and field labels for each repository or content management system. The schema's terms are drawn from two distinct but semantically related vocabularies and linked by an intermediary registry tying granular listings for specific applications to enterprise-level terms. The registry mappings permit the use of both local metadata and higher level or external systems. The effort has demonstrated the importance of consistent application of both granular and general metadata for information capture and revealed important lessons about adopting the Dublin Core standard in a large enterprise setting.

KEYWORDS

metadata
Dublin Core
document schemas
electronic document management systems
information reuse

Camille Mathieu is an information science specialist at the NASA Jet Propulsion Laboratory, California Institute of Technology. She can be reached at camille.e.mathieu<at>jpl.nasa.gov.

When information is generated by knowledge workers in an organization, the organization has vested practical and financial interests in maintaining the usability of that information. Once created, such information can be reused many times in an enterprise setting, minimizing redundancies in knowledge work products and increasing the overall efficiency of an organization. However, despite these benefits, knowledge-producing organizations have long struggled to provide workers with a straightforward means of reusing business information and this gap in enterprise knowledge management that can exact significant cost on an organization [1].

Problems associated with information capture and reuse are amplified in large organizations, where greater numbers of employees, departments and content repositories allow for the nearly unlimited siloization of enterprise knowledge. The Jet Propulsion Laboratory (JPL), a NASA research institute managed by Caltech and a leader in robotic planetary exploration, is one such large enterprise with vested interest in effective knowledge management. Most of JPL's 5,000 employees regularly perform knowledge-work tasks – and utilize hundreds of content repositories and other systems in managing information products. As JPL works to increase the sophistication of its knowledge capture and retrieval environment, one primary area of concern is in the effective management and practical standardization of enterprise content metadata.

The Value of Enterprise Metadata

Though the business value of well-managed enterprise metadata is evident in theory, in practice many organizations neglect to consistently apply metadata to their work content because they do not see any

immediate benefit in doing so. Further, if the application of metadata to a document or dataset is a laborious process, many knowledge workers may feel that applying metadata to their work products is outside the scope of their current task. Concerns such as these may help explain why recent surveys of enterprise information management practices suggest that about 50% of organizations have no metadata standards in place [2, 3].

This resistance or hesitation in adopting standards for organizational metadata is ill-founded, however, since metadata is vital for a number of business and information technology operations. The digital preservation of documents, intranet search and retrieval, the aggregation of like content across systems and repositories, document rights management, information validation, and records management and disposition are all key tasks facilitated by controlled content metadata [4]. The value of quality metadata for enterprise content is not evident immediately, but cumulatively. Over time, consistently applied metadata will yield greater and greater returns, while a lack of such metadata will progressively compound retrieval issues and further stress organizational efficacy. In order to better leverage the information generated by knowledge workers, organizations should seek to develop enterprise-level standards for metadata application and management.

Standardizing JPL Metadata

JPL presently has no formal metadata standard for internal content, although attempts to develop enterprise vocabularies or to define core metadata attributes have been made at intervals over the last decade. Recently, members of the JPL Library and other stakeholders have undertaken a new standardization effort. The aim is to create a standard schema that can be used to describe JPL's internal content, regardless of where that content is housed. Stakeholders for this effort have included not only information specialists from the JPL Library, but also repository and application managers. The content they write or manage is in need of standardized metadata to adequately describe it.

These stakeholders defined a series of parameters based on JPL's specific information management requirements to guide in the selection and adaptation of a metadata standard. The first parameter determined that standardization efforts would center on metadata *field* standards, rather than on metadata *value* or *format* standards. That is, stakeholders determined at the outset that the present standardization effort would focus primarily on the creation of an enterprise-level schema instead of on the specification of *value* standards (such as thesauri or controlled vocabularies) or *format* standards (relating to metadata encoding, such as XML, RDF, etc.). By prioritizing *field* standardization, the present effort finds a middle ground between the detailed *value* standards approach and the broad strokes *format* standards approach, since the *field* standardized metadata schema can be adopted at the enterprise-level and implemented at the application-level while neither oversimplifying the content metadata nor being stalled by too much repository-specific detail.

The second parameter required that the JPL metadata standard make significant use of some established external metadata standard while still providing a level of customization that makes the metadata useful in local organizational applications. Satisfying internal information sharing requirements is of foremost importance, since it is these enterprise-specific requirements which ensure the efficient local retrieval and management of content. Adherence to external standards is similarly important, however, for sharing information outside of the organization and for making content application-neutral so that it will not be locked in to any specific application or content management system. This second parameter seeks to ensure the long-term usability of JPL content, both within and outside of the organization.

Finally, the third parameter, taking into account all of the different content types, applications, repositories and departments that comprise the JPL information environment, specified that the JPL metadata standard be both content-neutral and application-neutral and be able to be consistently applied in each of JPL's hundreds of active repositories. Such a parameter again emphasizes the importance of internal cohesion between enterprise contents in various repositories while also looking ahead to ensure that enterprise search and aggregation systems can benefit from the consistency and neutrality of the standard organizational metadata.

Assessing Dublin Core

The Dublin Core standard, as described in ISO 15836 [5] and more extensively on the Dublin Core Metadata Initiative (DCMI) website [6], was selected as the basis of the JPL Resource Schema after a review of several established metadata standards. The Dublin Core standard falls within all predefined parameters established by JPL stakeholders, as it is general enough to describe a variety of content types in a variety of contexts, but also refined enough and customizable enough to be useful in specific application instances. For clarity, metadata properties defined by or allowed by the Dublin Core standard can be broken out into three groups:

- **Simple Dublin Core** properties are those original 15 elements first defined by the Dublin Core Metadata Workshop Series in the mid-1990s. Though this categorization is conceptually useful, Simple Dublin Core is a somewhat deprecated notion now subsumed into the *dc/terms/* namespace as high-level properties.
- **Qualified Dublin Core** properties are refinements of the original 15 elements (with some elemental additions defined more recently by the DCMI), which are presently managed in the *dc/terms/* namespace.
- **Custom Dublin Core** properties are custom refinements of the controlled *dc/terms/* elements made by local schema developers. While the Dublin Core standard does not allow for custom elements to be asserted, it does allow for the custom refinement of existing elements through the Dumb-Down Principle. This principle states that local refinements on the Dublin Core elements are supported as long as external applications can "ignore any qualifier and use the description as if it were unqualified" [7]. Adherence to this principle ensures that all enterprise-specific metadata elements can be dumbed down and ingested by external systems, even with some loss of specificity, since all custom elements are sub-properties of controlled Dublin Core elements.

In early stages of JPL schema development, existing content metadata from several repositories was mapped to either a Simple, Qualified or Custom standard element field to determine how many JPL-specific metadata properties would be supported by the established Dublin Core and how many would require custom refinements within the standard schema to remain useful in enterprise operations.

Using Namespaces, Building Profiles

Knowledge workers at JPL utilize hundreds of content repositories and applications to create, manage and store their digital work products. Each of these repositories describes content with a set of metadata fields that are distinct and disambiguated within that individual repository. While certain of these metadata fields are understandable regardless of their repository context (generic fields like "author" or "title," for example), more specialized repositories and applications may make use of highly specific fields that are usable only in the context of a single repository. This situation leads to an unavoidable ambiguity as to how metadata fields are understood at the enterprise level. For example, if a meeting-notes repository makes use of only one "document identifier" field to uniquely identify content, but a specialized engineering database makes use of not only a "document identifier" field, but "parts identifier" and "revision identifier" fields as well, then the concept of an "identifier" is not universally understandable at the enterprise level, but only at the individual repository level. In order to make an "identifier" field understandable at the enterprise level, the field has to be refined enough to allow for the different permutations of identification that each repository will necessarily need to employ.

Thus, because of the nuance required to adequately describe work products at JPL, the standard (Simple/Qualified) Dublin Core schema had to be customized and further refined to meet JPL content management needs for use as a standard within the organization. JPL schema developers created a series of custom refinements of the established Dublin Core terms and organized these refinements in a separate */jpldc/* namespace, meaning that, in an XML record, custom properties are prefixed by the "jpldc:" prefix instead of the standard "dcterms:" prefix. All properties described in the (Custom) */jpldc/* namespace are refinements of properties in the established (Simple/Qualified) *dc/terms/* namespace and are expressible in RDF as sub-properties of these *dc/terms/* terms. Because of this property/sub-property relationship between terms in the two namespaces,

all the terms in the */jpldc/* namespace inherit effectively the same definition and usability as that of their parent *dc/terms/* property. Rather than being assertions of new properties, the custom terms in the */jpldc/* namespace serve to refine the standard Dublin Core terms so as to make them usable within the specific contexts of the enterprise – while, at the same time, also ensuring that these specific fields could be understood more generically by external systems because of their sub-property relationship with terms in the *dc/terms/* namespace (Figure 1).

FIGURE 1. A sample of the terms in both the *dc/terms/* namespace and the */jpldc/* namespace, with sub-property relationships identified. Terms which are not sub-properties of any other properties are akin to Simple Dublin Core aspects, which cannot themselves be altered but merely refined.

dc/terms/		/jpldc/	
term	**subproperty of**	**term**	**subproperty of**
Date	none	dateArchived	Date
Created	Date	Keyword	Subject
Modified	Date	Phase	Subject
Valid	Date	Instrument	Subject
dateSubmitted	Date	Project	Subject
dateCopyrighted	Date	Approver	Contributor
Subject	none	Reviewer	Contributor
Contributor	none	Modifier	Contributor
Creator	Contributor		

Although the creation of custom properties in the */jpldc/* namespace, used in combination with the controlled *dc/terms/* properties, makes the Dublin Core standard useful and usable as an enterprise-level generic schema, its design is too capacious for implementation at the application level. The JPL Resource Schema defines a large series of controlled properties, and while most applications-specific fields can be mapped to this schema, it is not practical to encode the entire schema in broad strokes for documents at the individual application level. There will be many fields in the generic, application-neutral JPL Resource Schema that are not useful for certain applications, but that are vital for adequately describing the content in other repositories. Thus, while the JPL Resource Schema may represent the enterprise standard for metadata, the implementation of this schema at the application level requires the creation of application profiles for each repository or content management system where the standard will be adopted. These application profiles will specify which standard fields are utilized by individual applications, as well as the local repository-specific label for each controlled field. By mapping between application-specific fields and those controlled properties in the JPL Resource Schema, and by selecting for inclusion in each application profile only those fields which are core to describing an application's content, the JPL Resource Schema can be implemented at the application level with limited impact on the existing workflows of the repository owners or users.

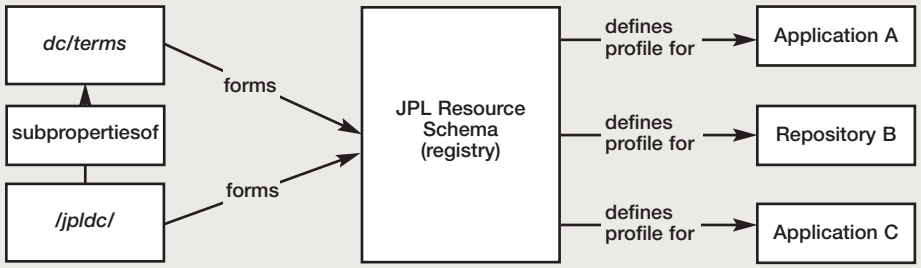
JPL Resource Schema

The JPL Resource Schema has been in development since May 2016. Presently, the schema comprises 54 properties, 34 of which are controlled *dc/terms/* terms and 20 of which are custom (sub-)properties in the */jpldc/* namespace (Figure 1). Because it draws terms from two namespaces, the JPL Resource Schema can be considered a composite schema, built from two separate yet semantically interrelated vocabularies. This composite schema represents all of the resource metadata properties that JPL tracks at the enterprise-level and the relationship of these properties to one another.

In order to bridge the gap between this enterprise-level composite schema and the granular field listings of application profiles, an intermediary registry has been proposed which will relate generic, controlled schema terms to the existing fields used by individual applications. This registry, which has at its core the composite standard schema, will associate controlled properties from the composite schema with uncontrolled application-specific fields (Figure 2). This approach of mapping individual application metadata profiles to an enterprise-level schema interferes minimally with the application's existing metadata usage, while also allowing the application-specific content to better interoperate between applications. Further, because the JPL Resource Schema is constructed in compliance with the Dublin Core metadata standard, the

architecture of mapping an application-specific field to a JPL schema property means that the application-specific field is ultimately readable as a standard term in the *dc/terms/* namespace. By applying the JPL Resource Schema in this way, applications are able to continue operating with specific, local metadata while, at the same time, registry mappings allow enterprise-level and even external systems to understand application-level metadata (though certainly with less specificity than would be captured at the application level). Thus, the application loses none of its required metadata granularity, and enterprise-level information search and aggregation systems are able to, through the metadata registry mapping, treat specific fields more generally and relate generic property fields across repositories.

FIGURE 2. A depiction of how the properties in the two namespaces interact to form the composite JPL Resource Schema and how this schema is applied at the repository-level through the definition of controlled application profiles for each repository/application.



Implementation and Lessons Learned

As the metadata registry is developed, implementation of the standard at the application-level will occur in stages, with individualized analysis and mapping of local fields to schema properties required for each of the hundreds of JPL content repositories. As repositories at JPL progressively begin to adhere to a common enterprise-level schema, finding relevant information will become easier for JPL employees, regardless of the repository in which the information is housed. Search connectors, which currently work to index content from a variety of repositories for JPL's enterprise-wide search, can be updated to also index controlled metadata fields, improving the relevancy of free-text searches and the accuracy of cross-repository federated aggregations. Similarly, the standardization of the organization's content metadata in this way will make it more usable in event tracking systems, records management modules and any other enterprise content management systems that work by aggregating content from a variety of sources. Additional future work in the development of controlled vocabularies and thesauri as metadata value standards will only further these benefits, since enterprise systems must be able to rely on not only a known set of fields, but a defined set of values for those fields as well. Future work in modelling entities in the JPL information environment, similar to the DCMI's work on its own Abstract Model (DCAM) [8], will also factor heavily in the continuing effort to standardize JPL metadata.

The practical application of the Dublin Core standard in an enterprise environment has not been without lessons learned. Efforts to standardize JPL metadata in accordance with the ISO 15836 standard brought to light some difficulties in practically utilizing this standard, as it is very brief and does not provide many implementation guidelines. The 2016 revision ISO/NWIP 15836 is currently up for a vote by ISO, and if the revision is ratified its formalization of properties in the *dc/terms/* namespace and of aspects of the DCAM model will make the standard more practical and applicable. However, even with this expansion, implementing the standard requires frequent use of content distributed throughout the Dublin Core website. Any institution wishing to standardize metadata according to the Dublin Core standard should familiarize itself with both the ISO 15836 standard and Dublin Core website materials, should work within the stated confines of the standard when developing customized aspects and should realize that the precise method of implementation will be determined more by the needs of the organization than by any external guidelines.

Conclusion

Much work remains for organizations which, like JPL, are seeking to tackle longstanding issues surrounding information storage and reuse in the

enterprise. As applications, repositories and other information technologies are used increasingly in the workplace, it is important to remember the vital role the consistent application of metadata plays in making content more accessible to users. Enterprise-level metadata schemas, especially those which are built on established standards, are integral to increasing the interoperability of information both within and outside of the organization. Practically speaking, an organization looking to gain the most value from its information and digital content will construct a metadata standard that is granular enough to be used locally in and between an organization's repositories, yet is also general enough to effectively incorporate an established standard.

Acknowledgements

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. An additional special thanks to the JPL Ontology Working Group and to Robert Powers and Sara Tompson for their support in this work. ■

Resources Mentioned in the Article

[1] Feldman, S., & Sherman, C. (2001). The high cost of not finding information: An IDC white paper. Retrieved from www.ejitime.com/materials/IDC%20on%20The%20High%20Cost%20Of%20Not%20Finding%20Information.pdf

[2] Miles, D. (2014). *AIIM industry watch search and discovery - Exploiting knowledge, minimizing risk.* AIIM. Retrieved from http://info.aiim.org/search-and-discovery

[3] Findwise. (2015). *Enterprise search and findability survey 2015.* Retrieved from www2.findwise.com/findabilitysurvey2015

[4] Baca, M. (2016). *Introduction to metadata.* Getty Research Institute. Retrieved from www.getty.edu/publications/intrometadata

[5] International Standards Organization. (2009). *Information and documentation — The Dublin Core metadata element set.* ISO 15836.

[6] Dublin Core Metadata Initiative (DCMI). (2016). *Dublin Core Metadata Initiative. http://dublincore.org/*

[7] DCMI (2000). Dublin Core Qualifiers. *Dublin Core Metadata Initiative.* http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

[8] DCMI (2008). DCMI Abstract model. *Dublin Core Metadata Initiative.* http://dublincore.org/documents/abstract-model/