CEN/ISSS MII-DC (WI3)

Guidance for the Deployment of Dublin Core Metadata in Corporate Environments

DRAFT

Contents

Guidance for the deployment of Dublin Core metadata in corporate environments

Background

Work Item 3 builds on DC-Corporate, a Dublin Core Metadata Initiative (DCMI) group formed at the DC 2002 meeting to identify and address corporate metadata needs to support modern business organizational functions like internal knowledge management. Since January 2004, the project team has held conference calls approximately every two weeks and begun work on tasks in the workplan. The team has:

- Prepared an overview of issues extracted from a review and summarization of the DC-Biz (DC-Corporate precursor from 2000-2002) and DC-Corporate listserv archives;
- Identified and prioritized a list of people who work with metadata in corporations in Europe and North America;
- Developed a semi-structured interview questionnaire regarding current metadata practices in large companies;
- Conducted and summarized 10 pilot telephone interviews of people from the list using the questionnaire (please see list at end of this document); and
- Prepared this annotated table of contents document.

Review Process

This document in various iterations has undergone and is undergoing continual community discussion and review. This process has included review in the following ways:

- Individual review and sign-off from all those interviewed.
- Validation of guidance areas by survey.
- Postings of drafts to the DC-Corporate and MMI-DC lists to gather feedback.
- Public workshop at DCMI October 2004 conference in Shanghai, China.
- Public workshop in December 2004 in Brussells, Belgium.

Guidance Areas

This section contains five areas of guidance on the deployment of Dublin Core in corporate environments. These areas were identified and are being validated in consultation with representatives of large companies.

1. How is Dublin Core used in corporate environments?

Most organizations we contacted are using Dublin Core as the de facto standard for descriptive metadata to uniquely identify document-like (unstructured) resources so that these resources can be found and re-used later on shared storage drives, corporate intranets and portals, or formal repositories to support specific business purposes. In most cases this is a selection of Dublin Core elements deemed appropriate, sometimes relabeled for corporate- or business area-specific use, and implemented as “simple Dublin Core”. Because there is a large amount of unstructured information in corporations, and most of it has little or no metadata, the bar is very low. Corporate metadata stewards are desperate to keep the metadata scheme as simple and transparent as possible.

In many organizations, the interest in defining coherent ways of applying metadata has intensified as a result of attempts to implement enterprise portals as a single point of access to company-wide information resources and applications. Another key driver of the focus on metadata in the

U.S. has been HIPAA[1] privacy issues and Sarbanes-Oxley[2] regulation, and in Europe by comparable compliance requirements.

Dublin Core is often seen and used as "integration metadata", to enable the federation of information in multiple, heterogeneous repositories such as those containing structured (databased), and those containing unstructured (document-like) resources as well as semi-structured resources described by specific metadata elements and encoding schemes. Repositories frequently include legacy (that is, pre-existing) resources. Applications and procedures (typically facilitated by manual workflow) are developed to add metadata to new content.

Among DC elements, ***Type***, ***Coverage*** (usually geographic), ***Publisher*** (or ***Source***) and ***Date*** are frequently used.

2. How is Dublin Core extended in corporate environments?

Corporate metadata applications are based on a combination of standards. While Dublin Core is the de facto standard for descriptive metadata, sometimes industry standards are combined with local schemas to meet specific functional requirements. These depend largely on the applications that are being supported. For example, in order to manage a document's lifecycle, various types of dates and access restrictions may be specified. Two document lifecycle themes are common—facilitating routine purging and/or archiving of electronic resources including both formal document types like project specifications or clinical study protocols, and informal communications especially emails.

Facilitating research and development, intellectual property management, and regulatory compliance are common business processes. Defining and declaring the purpose of a type of information resource in the context of a business process is a common Dublin Core extension in the business environment. An example is a list of document types relevant for different business processes that becomes a controlled vocabulary to be used for populating the ***Type*** metadata element. In many cases the metadata enables people to find documents describing an instance of a business process, such as a project or a clinical study, as a path to the information resources themselves, such as the clinical data or specific images.

Every business has particular products and services. Managing this controlled vocabulary and tagging every information resource related to a product and service is a common Dublin Core extension in the business environment. This is done either by adding an extra metadata data element (e.g. ***Product***) or by using a products and services controlled vocabulary encoding scheme to populate the ***Subject*** metadata element.

Corporations need to have lots of metadata about people, for example, in the human resources management and in research and development areas. It is sometimes more important to find the people who are doing the research and development than finding a particular report. Metadata about people such as information about competencies, roles, and access needs is usually developed and maintained separately from Dublin Core. This often means that the people metadata is not integrated with the objects and documents metadata.

At an even more basic level, there are problems with inconsistencies in the actual encoding of simple Dublin Core in the various tools used to create and access DC metadata. It is not at all clear that most people are following similar conventions partly because the implementation

---

[1] HIPAA is the abbreviation for The Health Insurance Portability and Accountability Act of 1996.

[2] The Sarbanes-Oxley Act of 2002 refers to the Public Company Accounting Reform and Investor Protection Act.

guidelines have evolved over the years. In this context, there are issues around conversion, mapping, dumbing-down (or normalizing) DC metadata for export and index integration, and how to mix-and-match Dublin Core and other XML namespaces. We are seeking some examples of practice to allow us to derive a small set of criteria that can be used in selecting the right level of complexity to achieve maximum result.

3. What metadata tools are being used to create and maintain metadata?

There are two methods to create and maintain metadata—by requiring resource creators to add metadata, or by using a centralized staff. Most organizations use a combination where metadata is collected during resource origination and development, and providing a central staff to clean-up and/or add metadata so that it meets the corporate standard. All organizations that value and use metadata require a high level of completeness and consistency, and typically have a metadata staff to provide quality assurance.

Very few organizations we spoke with are using any tools to automatically generate metadata using business rules or more exotic statistical algorithms available in information management applications, or as stand-alone tools. Paper or web-based forms with in some cases pull down lists of controlled values have sometimes been developed locally and are available to content originators and metadata staff. In most cases, these metadata forms provide little or no customization in specific contexts. For example, web-based forms usually include all values instead of only the applicable set of values for a particular document type. In a well-designed form fill-in application, the pull down list should include only the relevant parts of a more extensive list of values depending on the context.

We will seek contributions from vendors of these types of tools to get further input identifying the most pressing issues and potential solutions. Our preliminary typology of the types of tools is:

1. Integrated metadata tagging tools (modules of enterprise information management applications like Documentum, SAP, etc.)
2. Auto-categorization tools (stand-alone like InXight or part of search engine like Autonomy)
3. Vocabulary/taxonomy editing tools (stand-alone like MultiTes or part of an application like SchemaLogic)
4. Guided navigation applications (applications that can use metadata to create a user experience like Endeca)
5. Federated search and repository “wrappers” (such as AskOnce, an information integration layer, develop by Xerox Multilingual & Knowledge Management Solution, on top of information sources and their own search capabilities. The key functionality is the way they make it relatively easy to leverage existing metadata through metadata harvesting and harmonization)

4. What controlled vocabularies are being used?

Controlled vocabularies play a major role in business metadata applications. As mentioned above, lists of products and services, as well as business processes are frequently important and need to be created and maintained as part of the corporate- or business area-specific namespace. This area is not generally well-developed in organizations unless the corporate- or business area-specific metadata group has linkages to structured corporate applications like SAP, a data warehouse reference data database, or some other enterprise-wide application.

Most organizations use industry standard vocabularies such as ISO 3166 for geographic ***coverage*** or use local standard vocabularies such as LDAP for ***creator*** to populate common Dublin Core elements. Documenting best practices in the area would be highly beneficial, especially to instruct on how to take advantage of such industry standard and internal organizational authorities, rather than for metadata groups to take on the added responsibility of

developing and maintaining vocabularies. For example, how should a company handle different versions of external reference data such as country codes.

There are a few areas where a broader community could undertake vocabulary development, especially for ***content types***. Most organizations report a need for a Type vocabulary, but an adequate list of electronic resource types does not yet exist. In regulated industries such as pharmaceuticals, or in industries with a rapid product life cycle documents types are being standardized as part of standards initiatives such as eCTD[3] in the pharmaceutical industry and RosettaNet[4] in the electronic components industry. Where corporate- or business area-specific vocabularies are developed such as a products and services controlled vocabulary, there is a need for version handling, and a common way of refereeing a version.

We are asking the corporations that we speak with whether they would be willing to share their experiences on how they address these issues from an organizational perspective.

5. What specific guidelines are needed for the use of specific elements?

An inventory of which DC elements corporate users need specific guidance on will be developed. For example, while all DC elements are optional and repeatable, what is the current practice at corporations? How are people using them specifically?

An example mentioned above is how the ***Subject*** metadata element could be used for Product/Service classifications, as an alternative to adding a new data element to hold these values. Another example is the area of actors, such as the ***Publisher*** and ***Contributor***, which may need to be adapted and interpreted to better capture the different types of roles involved in production and publishing of information resources in corporations.

In different initiatives and standard organizations there seem to be a lot of interest in the concept of "Metadata Registries"[5] and the updated ISO standard for data elements (in general as well as the elements being used as metadata) (ISO11179). Will we see similar efforts internally to register external standardized schemas and internal specifications of data elements?

---

[3] eCTD, Electronic Common Technical Document (www.ectd.com), is a standard from ICH-International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use.

[4] RosettaNet (www.rosettanet.org) is a consortium of IT, electronic components, semiconductor manufacturing, telecommunications and logistics companies working to create and implement a common e-business language.

[5] Two examples of Metadata Registry implementations: caCORE: A common infrastructure for cancer informatics http://ncicb.nci.nih.gov/core United States Health Information Knowledgebase (USHIK) Metadata registry http://hmrha.hirs.osd.mil/registry/ and in the area of statistical metadata (see Work Session on statistical metadata, Geneva, 9-11 February 2004 http://www.unece.org/stats/documents/2004.02.metis.htm and the area of archiving "Integrating Metadata Schema Registries with Digital Preservation Systems to Support Interoperability: a Proposal http://www.siderean.com/dc2003/101_paper38.pdf

Interviews

Ten semi-structured telephone interviews were conducted to discuss current metadata practices in large companies, key issues faced, and areas where guidance is desired. Between January and April, 2004 representatives from the following copmpanies were interviewed.

- Applied Information Technique
- AstraZeneca (2)
- BBC
- BellSouth
- GlaxoSmithKline
- Intel
- John Wiley & Sons
- Rohm & Haas
- Software AG

Validation Surveys

An email survey with telephone follow-up was conducted to validate the guidance areas that were identified in the interviews. Between May and October 2004 representatives from the following companies responded to the survey.

- IBM
- SAP

Note: *Survey and response collection is still in process.*

Case Studies

A call was published to identify candidates for case studies discussing metadata practices in large companies. Between May and December 2004 case studies from the following large companies were collected.

- Unnamed high tech company
- Halliburton

Note: *Case study gathering, writing, and obtaining publication permissions are in process.*

Public Meetings

Two public meetings on metadata and knowledge management are being planned as part of this work item to be held September in Brussels and October in Shanghai. Program descriptions and participant lists are being developed.

1. Managing Interoperability Across Cultures, Oct 10, Shanghai.

The Shanghai workshop is co-sponsored with DC-Corporate Circle, the DCMI Working Group, focused on promoting adoption of the Dublin Core standard by enterprise organizations. The workshop esaddress the metadata lifecycle—creation, management, and propagation—as it applies to enterprise applications and activities, with special focus on interoperability enabling

international business. CEN/ISSS MII-DC WI3 will report on WI3 activities related to best practices around metadata creation, and seek input from workshop participants.

2. Content Management and Metadata, Dec 9, Brussells.

Interest in defining coherent ways of applying metadata has intensified as a result of attempts to implement enterprise portals as a single point of access to company-wide information resources and applications. Another key driver of the focus on metadata in the U.S. has been HIPAA[6] privacy issues and Sarbanes-Oxley[7] regulation, and in Europe by comparable compliance requirements. Today the procedures for adding metadata to content are typically facilitated by manual rather than automated workflow. The key challenge of enterprise content management applications is integrated metadata management.

This all day workshop will focus on automated metadata workflow and enterprise content management systems (ECMS). There will be presentations by experts, implementation case studies, demonstrations by search technology vendors, and lots of time for Q&A with the experts, practitioners, and vendors, and among participants.

Some important questions that will be addressed include— How do ECMS automate the process of applying metadata to content? What metadata tools are being used to create and maintain metadata in ECMS? How can metadata values be controlled using standard vocabularies in ECMS? How can the Dublin Core metadata schema be extended in ECMS?

Published Resources

"A passion for metadata – An interview with Todd Stephens of BellSouth." In: *Data Discussions*, 2003. http://www.wilshireconferences.com/interviews/Stephens.htm.

K. Forsberg and L. Dannstedt. "Extensible use of RDF in a business context." 33:1-6 *Computer Networks Issues* (June 2000) 347-364. Presented at the 9th International World Wide Web Conference, Amsterdam, Netherlands, (2000) http://www9.org/w9cdrom/323/323.html.

R. T. Stephens. "Broken windows, data quality, and the future of meta data." *Data Management Review Online* (April 2004)

R. T. Stephens. "The metadata experience." *Data Management Review Online* (March 2004)

R. T. Stephens. "The five disciplines of data." *Data Management Review Online* (February 2004)

R. T. Stephens. "The metadata brand map." *IRM UK's Strategic IT Newsletter* (January 2004)

R. T. Stephens. "Meta data is the great big book of everything." *Data Management Review Online* (January 2004)

R. T. Stephens. "One of the biggest secrets in metadata delivery." *The Data Administration Newsletter* (Jan. 2004) http://www.tdan.com/i027fe02.htm.

J. Ward. "Unqualified Dublin Core usage in OAI-PMH data providers." 20:1 *OCLC Systems & Services* (2004) 40-47.

---

[6] HIPAA is the abbreviation for The Health Insurance Portability and Accountability Act of 1996.

[7] The Sarbanes-Oxley Act of 2002 refers to the Public Company Accounting Reform and Investor Protection Act.

Note: *Gathering of published resources is still in process.*