IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5019

Designing a Metadata Model for Unstructured Document Management in Organizations

Federica Paganelli
Electronics and Telecommunications Department, University of Florence, Italy, via S. Marta 5, 50100, Firenze., Tel: +39 55 47 96 382, Fax: +39 55 48 88 83 , e-mail: paganelli@achille.det.unifi.it

Omar Abou Khaled
Department of Computer Science, University of Applied Sciences of Fribourg, Switzerland, Bd de Pérolles 80 - CP 32 CH-1705 Fribourg, Tel: +41 26 429 65 89, Fax: +41 26 429 66 00, e-mail: omar.aboukhaled@eif.ch

Maria Chiara Pettenati
Electronics and Telecommunications Department, University of Florence, Italy, via S. Marta 5, 50100, Firenze., Tel: +39 55 47 96 382, Fax: +39 55 48 88 83 , pettenati@achille.det.unifi.it

Franco Pirri
Electronics and Telecommunications Department, University of Florence, Italy, via S. Marta 5, 50100, Firenze., Tel: +39 55 47 96 382, Fax: +39 55 48 88 83 , e-mail: fpirri@ing.unifi.it

Dino Giuli
Electronics and Telecommunications Department, University of Florence, Italy, via S. Marta 5, 50100, Firenze., Tel: +39 55 47 96 382, Fax: +39 55 48 88 83 , e-mail: giuli@det.unifi.it

ABSTRACT

At present Document Management represents a critical problem for organizations, where a growing amount of information is produced in unstructured format. Metadata sets for Document Management addressing organizational needs are lacking, as most research efforts for Metadata specifications are focused on digital libraries and web communities requirements, rather than organizational ones. In this paper we propose a Metadata Model for Unstructured Document Management, called DMSML (Document Management and Sharing Markup Language), which aims at modeling intrinsic properties of documents (e.g. title, author and keywords) as well as the relationships with the organizational context. Furthermore, while most metadata specifications don't rely on a well-developed data modeling methodology, our proposal is based on a three-layered data modeling approach, distinguishing a conceptual, logical and physical layer. This approach encourages understanding and easy adoption of metadata specifications among end users. At present we are developing a DMSML framework prototype for validation purposes.

INTRODUCTION

At present information management is a critical requirement for all organizations. Organizational information is expected to be used in several business activities and for different purposes, accessed by users with different roles and managed by means of heterogeneous applications. One of the main obstacles towards effective information management is due to the fact that almost 80% of documents in an organization is available in unstructured format (the451, 2002).

Examples of documents usually available in unstructured form are e-mail, reports, presentations, word processing files, marketing materials and, multimedia files, encoded in different digital formats (e.g. .doc, .pdf, .wav). Unstructured formats are those, which don't provide a clear representation of the content organization, i.e. how the document is structured in information elements and the relationships among such elements. Usually information about content structure is mixed to rendering information. As a result, applications don't have enough information in order to effectively process neither content nor document rendering. Applications cannot point to specific content elements (such as clauses or customer identifier number in an electronic contract) in order to extract and process the required information. At present, information indexing and retrieval is based on a limited set of meta-information (e.g. document type, size, date of creation; etc.) or on full-text search, thus resulting into a difficult and time consuming task. Furthermore, information related to the document lifecycle management is usually hard-coded in specific applications and dependent on specific implementation details (for instance database schema) and it is hardly reusable by other applications. As a result, the limitations in automatic processing and retrieval of the information elements contained in unstructured documents compromise information reuse and sharing in organizations. Furthermore, existing Document Management Systems are proprietary and designed according to a tool-oriented rather then general and standard methodological approach. Related disadvantages are thus vendor dependence, difficult maintenance and poor interoperability with other information systems (Stickler, 2001).

Digital metadata are at present considered as the key mechanisms in order to face the requirements posed by management of unstructured information, as they enable to represent document properties in a human- and machine-understandable way (Gilliland, 2000). Several metadata sets have been proposed by research communities or industrial consortiums in order to provide a standard way to describe properties of information resources, but most efforts have been focused on digital libraries and Web Communities, rather than organizational needs (Murphy, 1998). To the state of our knowledge, metadata specifications, which aim at covering the requirements of organizations, are lacking. Furthermore, existing metadata sets do not rely on a well-developed data modeling approach (Lagoze et al., 2001). As a result, dissemination, human understanding and automatic processing of the metadata specifications may be compromised.

The work presented in this paper aims at covering the lack of metadata for Unstructured Document Management in Organizations,

proposing the DMSML (Document Management and Sharing Markup Language) metadata specification. The original value of the DMSML metadata model consists in the attempt to represent the context of use of documents in organizations (i.e. who, where, how, under which role a document is accessed), besides their intrinsic properties (like title, author, keywords, etc.). Furthermore, in order to address requirements of easy management, extensibility and adaptation to specific cases, we investigate the benefits of a rigorous formalization and data modeling approach, built on top of XML, eXtensible Markup Language (W3C, 2000), and XML Schema (W3C, 2001).

The remainder of the article is organized as follows: section 2 shows related work. In section 3 we describe the main requirements related to unstructured document management and we introduce the DMSML metadata model. In section 4 we describe the adopted three-layered data model approach. In section 5 we show the first results in the development of a DMSML framework prototype. Conclusions and future work are presented in section 6.

RELATED WORK

Several metadata sets have been proposed by research communities and/or industrial consortiums in order provide a standard way to describe information resource properties. However, several metadata standards (Dörr, 2003) cover specific application domains (e.g. archiving, eLearning, geographical information, medical informatics, etc.), and are not general enough to cope with requirements of information management in organizations. Other metadata sets provide descriptive models of generic document properties, but they are focused on requirements of digital libraries and WWW communities, rather than organizational needs. Readers are likely to be familiar with the Dublin Core (DC) metadata set (DCMI, 2003). DC aims at cross-domain information resource description, thus facilitating information indexing, search and retrieval. However, the DC metadata set cannot cope with organizational requirements (Murphy 1998), as its elements cannot conveniently represent organization concepts (e.g. department units and organizational roles).

In the research field of organizational metadata it is worth citing the following contributions:

- The Metia Framework (Stickler, 2001), which provides a metadata vocabulary, for management, storage and retrieval of electronic media. The contribution provided by METIA differs from our approach in that the proposed vocabulary aims at representing intrinsic characteristics and attributes of data, without taking into account issues related to the context-of-use. Furthermore, interoperability and integration with other official or de-facto metadata standards is not envisaged, and at the state of our knowledge, encoding solutions have to be provided.
- The work of Karjalainen (Karjalainen et al., 2000), proposing a method for collecting and analyzing organizational metadata for Document Management Systems. The method distinguishes metadata classes describing users, technological implementations and processing (e.g. workflow, archiving procedures, etc.). However, a rigorous data model is lacking, as well as policies and mechanisms for metadata encoding and management.
- The IEC 82045, which is a standard for Document Management/ Management data (metadata) for technical documents (IEC 2001). The objective is to propose a metadata model to describe the lifecycle of technical documents. Unfortunately, as the standard has not been published yet, an evaluation cannot be provided.

This overview shows that a lot of work in the field of organizational metadata for Unstructured Document Management is required. The lack of organizational metadata specifications motivates our DMSML metadata proposal, described in the next section.

DOCUMENT MANAGEMENT AND SHARING MARKUP LANGUAGE

The intrinsic properties of information resources (e.g. title, author, date of creation, etc.) do not effectively represent the value that electronic documents have inside organizations. Electronic documents are the artifacts that enable users with different roles to share and transmit information and that can be accessed for different purposes and in different business activities. Considering a document like a standalone resource poses several limitations to its effective exploitation. The added value that metadata can bring consists in the capability of modeling the context of use, as well as the intrinsic properties of documents. With the term context of use we conceive the set of factors, which identify who, when, for which purposes a document is accessed. Examples of context of use factors are: the organizational unit, which is responsible for the document creation or publication, the document lifecycle and the related access policy.

In order to cope with such issues the DMSML metadata model covers the following areas:

- Description: the set of intrinsic properties that describe the information resource. Such properties are usually logically related to the content and remain almost unchanged during the lifecycle of the document (e.g. title, author and subject). They usually serve for information indexing and retrieval.
- Collaboration: the document is conceived as the artifact enabling users with different roles and authority level to share information. Metadata should thus represent the relationships of the document with the organizational entities, such as the organizational model and the access policies.
- Lifecycle: the document is conceived as an information object, which evolves during time and the execution of business activities, across different phases (e.g. creation, review, publishing, etc.).

DMSML SPECIFICATION

In order to support interoperability we adopted the XML schema syntax for metadata encoding. XML schema syntax enables rich description of the metadata set, as it provides a rich set of mechanisms for expressing data typing, value, uniqueness constraints and reuse of existing schemas. Furthermore, the adoption of XML Schema syntax facilitates integration with widely adopted standards, already available in XML schema format: Dublin Core and Petri net Markup Language (PNML) (Weber et al., 2002), which have been adapted in order to respectively address the Description and Lifecycle issues and the eXtensible Access Control Markup Language (XACML) (OASIS, 2003), which is a standard proposition for access control policy description.

However, effective exploitation of organizational metadata poses other needs, which cannot be addressed by XML Schema features, but require a richer modeling approach:

- Genericity: the model aims at modeling generic properties of unstructured documents and it is not tied to a specific application domain (e.g. technical documents, health records, etc.)
- Easy understanding: few organizations are likely to employ professional catalogers (Murphy, 1998) to create and organize descriptive information. Murphy (1998) suggests that metadata in organizations will tend to be more and more author-created, contextual and less formal. As a result a framework facilitating conceptual understanding and sharing of metadata is needed.
- Extensibility: in order to cope with the specific and contextual needs of organization, the properties expressed in DMSML are expected to be modified and extended for specific purposes. Some questions arise: who should be responsible for metadata adaptation? Which competences should be required: professional cataloging expertise, knowledge of the organizational context and business processes, or technical know-how?

XML Schema is a rich description language, but it doesn't encourage easy understanding of data model, as it requires a strong technical expertise. In order to facilitate understanding and extensibility of the

DMSML specifications, by hiding implementation and XML Schema syntax details, we applied the layered data modeling approach, which is traditionally used in database design, to the formalization of the DMSML metadata set. The DMSML model is described in the next section.

DATA MODELING

The DMSML three-layered data model consists of a conceptual, logical and physical layer. As we apply such approach in order to formalize the specification of metadata, which are encoded in an XML-based format and not for the purpose of database design, our needs are not the same of traditional database design. In order to avoid any ambiguity, we define our objectives for data modeling in the followings.

At the conceptual layer we aim at modeling the real world-view and understanding of data. Conceptual models enable people with low technical expertise to understand meaning of data and manipulate the data model.

The logical layer provides an abstraction, based on rigorous and standard data modeling language. In order to maintain the rich modeling constructs of XML Schema, but at the same time abstracting from implementation and XML syntax details, we adopt a UML notation (Booch et al., 1998), extended by means of a UML profile for XML Schema, as proposed in (Routledge, 2002).

At the physical layer data are encoded according to the XML Schema syntax.

Conceptual Layer

At conceptual layer we define the main concepts of the problem domain. We establish the meaning, content and context of metadata at a business level, i.e. an organization/real world perspective of the data. The main entities, which populate the model, are:

- Actor - The actor is the entity, which performs actions on some information resources. An actor may be a user, a group of users or an organization unit.
- Information resource - This entity represents the digital information resources. An information resource may be a workspace, a folder or a document.
- Document - It is the atomic unit of information that we take into account. The document is seen as a black box.
- Folder - It is a collection of documents.
- Workspace - This entity represents the working environment: it contains documents and folders.
- Document Lifecycle - it is defined as a sequence of tasks in (creation, revision, review, annotation, sharing, publication, searching, etc.), triggered by actions performed by users and groups (actors) within the organization.
- Rights- it is the set of access policies, which regulate dissemination and use of information within an organization.

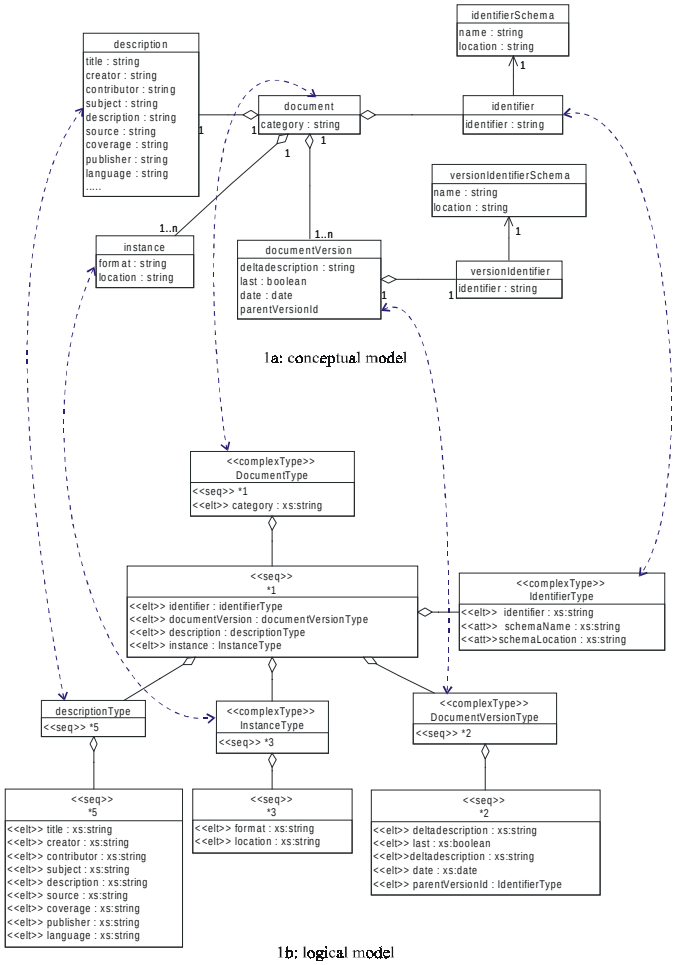
Figure 1a shows an extract of the conceptual model, in UML class diagram notation. The *document* is described in terms of a *category* attribute (for instance administrative, technical, etc. according to the categorization scheme that is adopted in the organization) and a *description* class (which contains static descriptive properties, such as title, creator and subject). Each document has a unique *identifier*, for instance ISBN or an internal identifier schema. A document may have one or more *versions*. A document may have one more *instances*, specified by the physical location and format.

Logical Layer

The logical layer aims at refining the conceptual layer through logical and rigorous modeling constructs, i.e. provides description of data structure in an abstract way, usually by means of graphical notation. We adopted the UML profile for XML schema, as proposed in (Routledge, 2002), which provides a 1-to-1 mapping between UML and XML schema model. At the logical layer, class diagrams use the stereotypes defined in the UML profile in order to represent the XML Schema concepts.

Figure 1b shows the logical layer UML class diagram. This extract shows some of the stereotypes that are defined in the UML profile for XML Schema: <<elt>>, <<att>>, <<complexType>>, <<seq>>, which represent respectively the "element", "attribute", "complexType" and

Figure 1: Extract of DMSML model



"sequence" XML Schema data constructs. The dotted arrows show the relations between concepts at the conceptual layer and XML Schema modeling constructs represented at the logical one.

Physical Layer

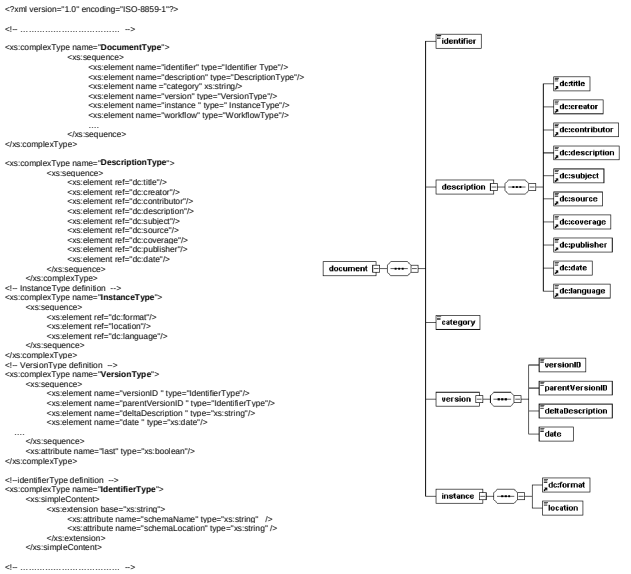
XML Schema syntax is used in order to represent the metadata model in an XML-based language, called DMSML (Document Management and Sharing Markup Language). Figure 2 shows an extract of DMSML specifications. XML Schema is a language enabling to define the grammar, i.e. describing the valid content and structure, of an XML document. For such a purpose, an XML Schema defines the elements, attributes and related data types that can appear in a document. XML Schema features include:

- Rich data typing capability: XML Schema supports a broad range of built-in datatypes, such as numeric date/time, Boolean, URI, string, etc. and it provides mechanism for user-defined types specification
- OO model: it supports a single inheritance model that allows reuse of type definitions. .
- Constraints: a rich set of constraint types is supported. For instance it provides format-based constraints (pattern facets), uniqueness (key and unique labels), relationship specifications (key references) and cardinality constructs.

DMSML FRAMEWORK

At present we are developing a DMSML framework prototype, which provides the user with automated support to adaptation and usage of the metadata set and the deployment and maintenance of a Document

Figure 2: extract of XML Schema DMSML specifications



Management System. The DMSML framework consists of three parts: a Conceptual Graphical User Interface, a DMS Engine and a Web-based DMS (see figure 3).

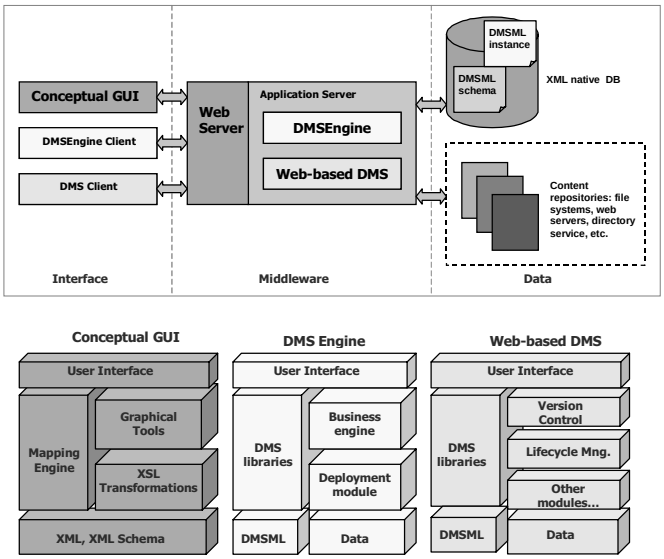
Conceptual GUI

The Conceptual GUI provides the user with a conceptual view of the DMSML metadata set, abstracting from the complexity and implementation details of the XML schema syntax. XSLT transformations are the mechanisms, which enable the mapping from the conceptual to the physical layer and vice versa.

DMS Engine

The DMS Engine is an application capable of dynamically generating a Document Management System (DMS), according to the high-level specifications expressed in the DMSML instance document. The DMS Engine is a web-based application deployed on the J2EE platform (Sun). It includes templates of both presentation and business logic software components of a DMS with basic functionalities (version control, lifecycle management, etc.). The parameters, which are required to customize the template and deploy a web-based DMS, are contained in a set of XML documents. The technical configuration parameters (such as connection to databases, authentication and directory service, web servers, etc.) are encoded in a *Technical Configuration* XML Document. The DMS Engine uses these data in order to actualize the J2EE *deployment descriptors* for the new web-based DMS. The parameters, which describe the organization of the workspace, the categorization of information resources, the description of document lifecycle and the relation with the organizational entities, are expressed in a DMSML instance document. Leveraging on this declarative approach, the DMSEngine can deploy an ad-hoc DMS, according to the high-level organizational specifications encoded in the DMSML syntax.

Figure 3: DMSML Framework Architecture



Web-based DMS

Thanks to the automation support of the DMSEngine, it is possible to deploy a web-based DMS. The DMSML metadata describe the workspaces and the documents managed by the DMS. At this phase of the implementation the metadata can be stored in an XML-native database or a relational one. At the time of configuration the database used for metadata storage is specified in the Technical Configuration XML document.

CONCLUSIONS AND FUTURE WORK

Management of unstructured document management is an unresolved issue, which is posing several obstacles towards effective information reuse in organizations. Lack of theoretical contribution for metadata specification in the field of unstructured document management in organization motivates our work. The original value of the DMSML metadata model consists in the representation of the document in relation to its context of use, rather than as a stand-alone resource. Furthermore our proposition is based on a rigorous and rich modeling approach, based on a conceptual and logical layer built on top of XML schema syntax. These abstraction mechanisms encourage easy understanding and extension of the metadata set. Further research is needed to optimize the integration of the Dublin Core, XACML and PNML schemas in the DMSML framework in order to avoid possible semantic collisions and inconsistency among the elements coming from heterogeneous schemas. At present we are working on the development of a DMSML Framework prototype, consisting in a conceptual GUI, a DMS Engine and a web-based DMS.

REFERENCES

Booch, G., Rumbaugh, J. et Jacobson, I. (1998). *The Unified Modeling Language User Guide* Addison-Wesley Pub Co.

the451 (2002). *Unstructured Data Management: the elephant in the corner*. Retrieved April 9, 2003, from www.the451.org.

DCMI (2003), *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. Retrieved April 2003 from http://www.dublincore.org.

Dörr, M., Guarino, N., Lopez, M.F., Schulten, E., Stefanova, M., & Tate., A. (2001). *State of the Art in Content Standards*. Retrieved January 2003, from: http://www.ontoweb.org/download/deliverables/D3.1.pdf.

Gilliland-Swetland, A. J. (2000). *Setting the Stage*. Retrieved October 20, 2002, from http://www.getty.edu/research/institute/standards/intrometadata.

IEC (2001). *IEC 82045 - Specifications for Document Management Systems*. Retrieved January 2003 from: http://tc3.iec.ch/txt/3osl11.pdf

Karjalainen, A., Päivärinta, T., Tyrväinen,P., Rajala, J. (2000). Genre-Based Metadata for Enterprise Document Management. In *Proceedings of HICSS2000*. Retrieved March 2003 from http://www.jyu.fi/~ankarjal/HICSS2000.pdf

Lagoze, L. & Hunter, J. (2001). The ABC Ontology and Model. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*. Retrieved 24 January 2003 from http://www.nii.ac.jp/dc2001/proceedings/Contents.html

Murphy, L. D. (1998). Digital Document Metadata in Organizations: Roles, Analytical Approaches, and Future Research Directions. *Proceedings of the 31st Hawaii International Conference on System*

Sciences: Digital Documents. IEEE Computer Society Press, Los Alamitos CA, 1998.

OASIS (2003). *Extensible Access Control Markup Language (XACML), V. 1.0*. Retrieved April 2003 from http://www.oasis.org

Päivärinta, T., & Ylimäki T. (2002). Defining Organizational Document Metadata: A Case beyond Standards. In *Proceedings of European Conference on Information systems*. ECIS 2002, Poland. Retrieved March 10, 2003,

Routledge, N., Bird, L., & Goodchild, A., (2002). UML and XML Schema. In *Proceedings of Thirteenth Australasian Database Conference* (ADC2002). Retrieved from http://titanium.dstc.edu.au/papers/adc2002.pdf

Sun. Java 2 Enterprise Edition (J2EE) Documentation. Retrieved October 2002 from: http://java.sun.com/docs/index.html

Stickler, P. (2001). Metia- A Metadata Driven Framework for the Management and Distribution of Electronic Media. *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*. Retrieved January 24, 2003 from http://www.nii.ac.jp/dc2001/proceedings/Contents.html

Weber, M. et Kindler, E. (2002). *The Petri Net Markup Language*. Retrieved April 2003 from www.informatik.hu-berlin.de/top/pnm

W3C (2000). *Extensible Markup language (XML) 1.0. W3C XML Working Group*. Retrieved from http://www.w3.org/TR/REC-xml

W3C (2001). *XML Schema Part 0-2: [Primer, Structures, Datatypes]. W3C XML Working Group*. Retrieved from: http://www.w3.org/TR/xmlschema-0/, http://www.w3.org/TR/xmlschema-1, http://www.w3.org/TR/xmlschema-2d