

A Metadata Model for the Design and Deployment of Document Management Systems

Federica Paganelli¹, Omar Abou Khaled², Maria Chiara Pettenati¹, and
Dino Giuli¹

¹ University of Florence, Department of Electronics and Telecommunications
via S. Marta 3, Florence, 50100, Italy
{paganelli, pettenati}@achille.det.unifi.it
giuli@det.unifi.it

² University of Applied Sciences of Western Switzerland,
Department of Computer Sciences
Boulevard de Perolles 80, Fribourg, 1700, Switzerland
omar.aboukhaled@eif.ch

Abstract. This work aims to address the issues of unstructured document management, by proposing a set of metadata: the DMSML (Document Management and Sharing Markup Language). DMSML represents a set of document properties, which are relevant to document management and render business and organizational information explicit, in a way which promotes reuse, interoperability and integration with heterogeneous systems.

A huge amount of organizational information is formalized in unstructured documents. Due to their intrinsic characteristics, management of unstructured documents presents critical issues: difficult information search and retrieval, poor interoperability among information systems, poor reuse of content, as well as of business information, related to the context of use of documents in organizations (i.e. business processes and organizational schema).

The objective of this work is to address these issues, by proposing a Metadata Model, called DMSML (Document Management and Sharing Markup Language). DMSML represents a set of document properties, which are relevant to document management and render business and organizational information explicit, in a way which promotes reuse, user-driven extensibility and interoperability with heterogeneous systems.

In order to cope with the issues of document indexing, search and retrieval, and reuse of documented and business information, the process of metadata specification has been focused on the selection a set of labels representing content as well as context-related properties of documents. Content properties relate to what the document contains or is about, thus providing to users and applications useful hints to help document search and retrieval and improve the reuse of documented information. Context-related metadata express the "by whom,

where, how, under which constraints and for which purpose” a document is being accessed, transmitted and modified. In this way, business information related to the practices of use of documents is made explicit, promoting formalization, exchange and reuse of this valuable information.

In DMSML we represent these two dimensions of unstructured document properties, by distinguishing three main parts, or modules:

- The *Descriptive Information Model*, i.e. the set of properties, which describes and identifies the document (e.g. title, author, date and subject).
- The *Collaboration Model*, which formalizes how the organizational resources are structured (the organizational model) and how access to information resources is regulated (the access right policy), on the basis of the organizational roles or responsibilities of individuals.
- The *Process Model*, which specifies the lifecycle of the document. The document lifecycle usually consists of the following stages: creation, review, publication, access, archive and deletion. A specific lifecycle may not implement all these stages, or implement others.

The DMSML specification is based on a three-layered model, promoting human understanding of the metadata specification (*Conceptual Layer*), logical data modeling (*Logical Layer*), and machine-understanding and interoperability with other applications (XML Schema-based *Physical Layer*).

DMSML supports the conception of a completely declarative approach for the design and automatic deployment of Document Management Systems. The use and proper adaptation of DMSML enable to configure a DMS according to the specific requirements of an organization, providing specific methods and mechanisms to exploit the business knowledge detained by the end users, and leveraging on the compliance with technical and business metadata standards.

A DMSML Framework prototype has been developed, which provides the user with automated support for the adaptation and use of the metadata set and the deployment and operation of a Document Management System. It consists of three parts: a DMS Configurator, which is a graphical interface for the creation of DMSML instance documents, representing the requirements of an organization, a DMS Engine, which automatically deploys a Document Management System, properly configured according to the specifications encoded in the DMSML instance, and a DMS Web Application, providing basic Document Management features. The description of the information resources and the policies for the proper management of these resources (i.e. document lifecycles, access policies, etc.) are encoded in the DMSML instance. The operation of the DMS Web Application (e.g. upload of new documents and creation of new folders) is mapped into proper updating of the DMSML instance.

Some testing activities in real application scenarios are planned in the research progress. Possible scenarios to test the approach are envisaged within two application domains: the management of legal documents (with the support of the Institute for Legal Documentation, <http://www.ittig.cnr.it/>) and the lifecycle of scientific grey literature in a research institution (with the CERN Library, <http://library.cern.ch/>).