A Metadata-Based Approach for Unstructured Document Management in Organizations

Federica Paganelli, University of Florence, Italy

Maria Chiara Pettenati, University of Florence, Italy

Dino Giuli, University of Florence, Italy

ABSTRACT

Effectively managing documents is a strategic requirement for every organization. Available document management systems (DMSs) often lack effective functions for automatic document management. One reason is that relevant information frequently is conveyed by unstructured documents, whose content cannot be easily accessed and processed by applications. This article proposes a metadata model, the DMSML (Document Management and Sharing Markup Language) to enable and to ease unstructured document management by supporting the design of DMSs. We argue that the extensive use of this metadata language will render organizational information explicit, promoting information reuse and interoperability in a more profitable way than what is guaranteed by proprietary DMSs. We also briefly depict the design and deployment phases of a Web-based DMS prototype based on DMSML. Our overall intent is to increase the awareness of what managers should account for when considering the possibility of adopting a DMS.

Keywords: design methodologies; document management systems; information system design; metadata; software development tools; Web-based applications

INTRODUCTION

Document Management (DM) is the scientific domain dealing with the use of ICTs for the effective "storage, organization, transmission, retrieval, manipulation, update, and eventual disposition of documents to fulfill an organizational purpose" (Sprague, 1995, p. 32). Existing ICT-based DM solutions, hereafter called document management systems (DMSs), do not completely fulfill the expectations of providing enough effective tools for information creation, sharing, and retrieval inside an organization, often causing user frustration, dissatisfaction, and inefficiencies (Ginsburg, 2001).

A typical situation that creates problems in many organizations is the management of unstructured documents that often convey important information and knowledge (the451,

2002); due to their lack of structure, these documents cannot be easily and effectively accessed and processed by applications, thus limiting effective document management. As a consequence, members of organizations have difficulty retrieving the information contained in these documents. Moreover, existing DMSs seldom are designed according to a general and/or standard methodological approach and are built around open data and process models. Thus, related disadvantages are vendor dependence, difficult maintenance, and poor interoperability with other information systems (Stickler, 2001).

In order to deal with these issues, we propose in this article a metadata model, enabling the design of DMSs and aiming at combining the benefits of metadata for document description with the use of Web standards. The metadata language described in this work has been named DMSML (Document Management and Sharing Markup Language). DMSML offers a solution to representing a set of document properties that are relevant to document management and to rendering business and organizational information explicit in a way that promotes reuse, user-driven extensibility, and interoperability with heterogeneous systems. A Web-based prototype developed according to DMSML specifications will help make the theoretical arguments presented throughout this article concrete.

This article is organized as follows: initially, some considerations on the management of unstructured documents are made in order to point out which relevant characteristics of an unstructured document are worth being described in order to improve its efficient management (its *content* and *context of use*). Then, an example of a typical document frequently managed in many types of organizations, the *project proposal*, will sustain what is said from a general perspective. The following paragraphs will be devoted to the analytical description of the requirements for high-quality DMSs. The fulfillment of these requirements will be taken as the basis for the design of the metadata language (DMSML) as well as for the develop-

ment of the Web-based DMS prototype presented in this work. Examples and comparative evaluation of available products — both commercial and open source — will show that meeting all of the aforementioned requirements is a characteristic satisfied by none of the presented products, to the best of our knowledge.

The central part of this article is aimed at describing (from a general point of view) the use of metadata for document management, illustrating the benefits of their usage in this domain. Existing metadata languages related to document management at the state of the art also will be recalled in this part of the article.

In the sequel, the DMSML metadata specification will be described, highlighting its fulfillment of high-level requirements. DMSML will be proposed as a declarative language for the specification of DMSs, based on existing standards and on a rigorous modeling approach. Metadata modeling, using proper formal representation techniques, then will be illustrated. The advantage of using DMSML for DMS design and development will be described in the related paragraph.

Finally, the Web-based prototype using DMSML metadata specifications will be described in its functional components and implementative details. Using these arguments, we demonstrate that the extensive use of this metadata language in document management systems will help to exploit business and organizational information in a more profitable way than what is guaranteed by proprietary document management applications, because the knowledge (properly codified through a metadata language) will allow both human and machine readability and, consequently, more effective reusability.

MANAGEMENT OF UNSTRUCTURED DOCUMENTS

Unstructured documents are text and multimedia documents such as e-mails, reports, and so forth, stored in various formats that do not provide an explicit, formal, and separate representation of either content structure (also

Related Content

A Novel Architecture for Learner-Centric Curriculum Sequencing in Adaptive Intelligent Tutoring System
Ninni Singh, Neelu Jyothi Ahujaand Amit Kumar (2018). *Journal of Cases on Information Technology (pp. 1-20).*
www.irma-international.org/article/a-novel-architecture-for-learner-centric-curriculum-sequencing-in-adaptive-intelligent-tutoring-system/207363

Salary Differences Between Male and Female Software Developers
Ronald Dattero, Stuart D. Galupand Jing "Jim" Quan (2007). *Emerging Information Resources Management and Technologies (pp. 24-42).*
www.irma-international.org/chapter/salary-differences-between-male-female/10093

Improving Virtual Teams through Creativity
Teresa Torres-Coronasand Mila Gasco-Hernandez (2005). *Encyclopedia of Information Science and Technology, First Edition (pp. 1419-1424).*
www.irma-international.org/chapter/improving-virtual-teams-through-creativity/14449

Information and Its Conceptual Perspectives
José Poças Rascão (2019). *Advanced Methodologies and Technologies in Library Science, Information Management, and Scholarly Inquiry (pp. 168-184).*
www.irma-international.org/chapter/information-and-its-conceptual-perspectives/215922

Developing a Learning Organization Model for Problem-Based Learning: The Emergent Lesson of Education from the IT Trenches
Kam Hou Vat (2006). *Journal of Cases on Information Technology (pp. 82-109).*
www.irma-international.org/article/developing-learning-organization-model-problem/3177