ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2005, Vol. 34, No. 4

ISSUES ON FORMING METADATA OF EDITORIAL SYSTEM'S DOCUMENT MANAGEMENT

Marijus Bernotas, Remigijus Laurutis, Asta Slotkienė

Information Technology Department, Šiauliai University
Vilniaus str. 141, LT-76353 Šiauliai, Lithuania

Abstract. Documents play a very significant role in modern organizations. This role is especially important in periodical publishing editorial offices, which manipulate with huge amounts of documents and whose activity is marked by often reuse of documents of their parts. Due to this fact the need for document management emerges by using document management systems. As universal document management systems are developed for a wide use in various areas of activities, so the main obstacle for integrating them into the activity area of the editorial office is a closed and inflexible model of metadata description. Metadata for the description and efficient management of documents, which circulate in the periodical publishing editorial system are presented in this paper, their processing methods are analyzed.

Keywords: metadata, document management systems, editorial system, management document.

1. Indroduction

Being one of the most rapidly growing branches of industry information technologies have reached such a level when each organization, which seeks for a position of a leader, can gain not a little time and resources if it selects appropriate information systems. Documents have always played an important role in the activity of organizations. The storage of documents in paper form only enlarges the organization's disbursement and does not allow ensuring the dynamism of the information spread. So currently no one is surprised by their storage in the electronic form, which enables for larger possibilities of manipulation. The growing amount of circulating documents burdens the reachability, selection and use of the information contained in them. Consequently, several copies of the same document are being emerged and the information about the environment it has been created in is lost. These problems can be solved by using e-document management systems, where the additional information about the document is stored and supplemented; such information is document's metadata. This paper seeks to analyze metadata that can be applied to document management in an editorial system of a periodical publishing for document versifying, classification, search and content extraction. Moreover, methods for describing document's metadata are being analyzed.

2. The Topicality of Document Management in the Editorial System

Many software producers offer document management systems, however their integration and adjustment to enterprise's needs cause quite a few problems, and the solutions do not provide with the expected efficiency in most cases. The mentioned possibilities of document management systems are especially relevant in editorial systems whose main activity is publishing of print media (newspapers and journals). Since the process of publishing involves various documents and paper handouts, the latter systems have all the characteristic features of a document management system. However the work flows and collaboration features are based on the features of print media publishing.

Referring to the specialty of periodical publishing activity one can observe the forces which influence the need for document management in editorial systems:

- The experience of editorial offices reveals that the publications afooted often use the information that is prepared beforehand, i.e. the major part of papers are just the variations of the older versions, supplemented with new info. [4,8]
- Editorial offices of periodical edition manipulate huge amounts of documents.
- Document development, storage, modification are iterative processes. As a multiplex use of the document or its part in the preparation of an article

becomes an integral part of the process, the need for a fast selection of the required information emerges.

- Recently the Internet becomes one of the main sources of fast varying information; consequently the presentation of the periodical edition in the Internet becomes a desirable issue as it provides the reader with the interactivity and feedback.
- The archives of editions are especially valuable, where the Internet navigation and search shorten the selection of the required information in several tens of times. Such archives are useful for the organization's staff, which spends a lot of time for the search for the required documents.

The application of the document management system automated the mentioned processes, reduced the production resources and improved the efficiency of work.

3. Publishing Process's Life-Cycle and Design

Several groups of people – designers, journalists, editors, etc. – prepare the document in editorial offices of periodical editions, and the prepared information is disseminated through several means of information spread. Designing a document management system it is important to define the document's life-cycle, which determines the order of documents' processing, storing, retrieving and editing. The edition's publishing life-cycle begins with the process of obtaining information from various sources and terminates in the edition's release and publication on the Internet. Referring to the concept of document management, the operation processes of editorial system can be excluded into the following three stages. (Figure 1):

1. Paper preparation – preparing and processing the document or its parts.
2. Edition creation – documents' confirmation, conflation and edition's formation.
3. Edition publication – webpage generation, archiving and access management.

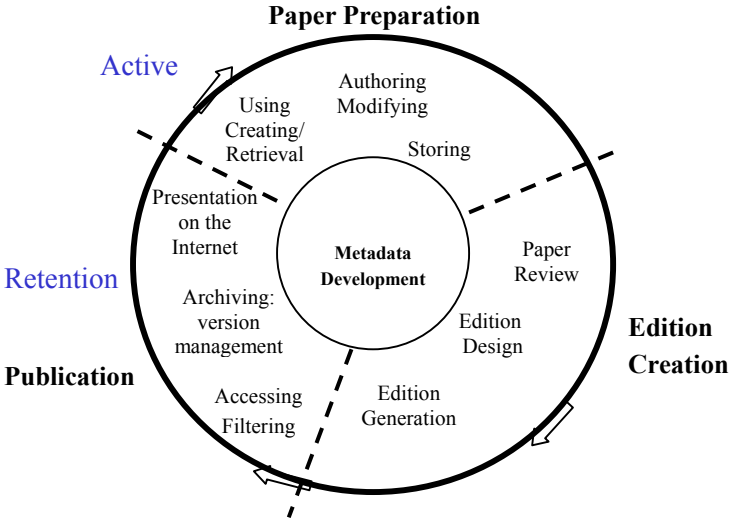

Figure 1. The Cycle of Periodical Edition Publishing

Referring to the processes executed during the publishing process life-cycle we can distinguish the editorial system's tasks:

1. The placement, search and retrieval of the document or its part,
2. The classification of documents and their elements,
3. Metadata development and management,
4. Control of the access to documents,
5. Publication of the edition content on the Internet,
6. Editions' archiving and versifying.

In the process of publishing a periodical edition one can see the topicality of preparing a quality content edition in the shortest period. This process is influenced by time consumptions for each document preparation – the search for information in documents, the retrieval of documents or their parts. The semantic quality of the search results can be guaranteed by the use of document metadata and their management.

4. The Use of Metadata for Document Management

Metadata in a document management system are defined as data about document development environment and document structure; such metadata are necessary for understanding the document, for selecting the required information or for classifying its parts. The metadata record is composed of predefined elements, which describe specific features of the

document, and each element can have one or more values.

Speaking about document's metadata no document's physical qualities, related to information fixing way, media, document form and other physical qualities are considered; metadata provide with the information about document developer, document development circumstances, place and time, as well they describe its relations to other documents. As e-documents have no elements, which define their contextual information, this function is executed by metadata. The document and its metadata compose a part of the contextual information of the e-document and have to be stored together with the document insofar the document is stored itself [1, 6, 10, 12].

Metadata in document management systems are used for [2, 7, 10]:

1. Document description – these are metadata that describe the document or its content: the creation date, content's key words, the author, etc.
2. Access control – these are metadata, which store the system of rights for users who are enabled to use the documents or the list of permissible actions of access to the document for each user;
3. Document versifying – these are metadata, where the data about the performed changes in the document are stored: date, author, the place of previous versions, the performed changes.
4. Document classification – these are metadata that define the document's place in their structure and the documents' interrelations.

Document metadata can be conditionally classified according to their development source and their application purpose (Figure 2).

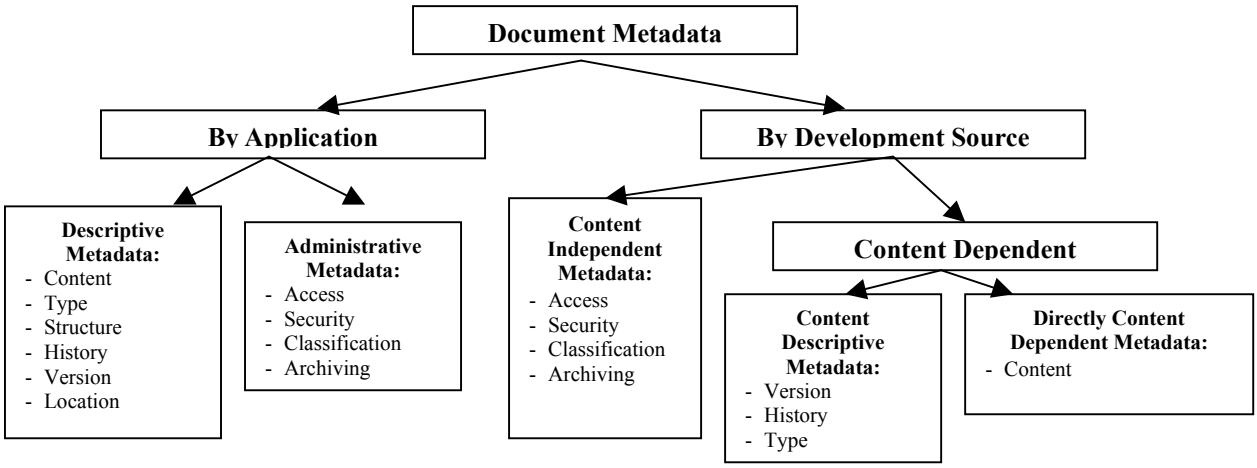

Figure 2. Document Metadata Classification

Generally, metadata elements are formed by adjusting to the application domain of the enterprise, to the activity specialty and the functioning work flows. The process of periodical publishing inherits many elements of document metadata description used in e-publishing. The facts, events, significant dates, presented in the papers, and the document's history are relevant of highlighting in publishing of newspapers and journals. The main metadata, which can be applied to document management in the editorial system, are presented in Table 1.

5. Metadata Description Methods

Formats and standards of metadata presentation, storage, search and exchange can influence the durability of system life-cycle as well. The method of document metadata processing, as well as the technological solution of software, can influence the benefit and quality of the provided results. The methods of document metadata description may be excluded into the following categories:

1. The application of universally accepted document description metadata standards. For example, Dublin Core, Resource Description Framework (RDF) are characterized by versatility, they can be used for any e-resources and do not require additional time for creating the metadata description [3,9]. However, their application in specific document management system, as the editorial system, causes the problem of elements' flexibility. Their application may overload the system with unnecessary metadata: bibliographical (editor, co-author), document presentation type (PDF, HTML, DOC and similar), reviewers and reference.
2. Metadata description in database tables.
   2.1. If the document management system is realized using the hierarchical model of document management, then metadata are stored in the table of domain base; the document's location is indicated in this table. The main disadvantages are as follows:
   - One document – one location. Irrespective of the fact that often the document

should belong to several branches with reference to its content it is possible to hold it in one place only regarding to a hierarchical system and the documents systematization is inconvenienced very much.

- Limited search and complicated segmentation. Generally, the documents are uploaded into the system using one criterion of systematization, and they are retrieved using the other ones. Unfortunately, the hierarchical model can perform the document search only by one section. Moreover, the document systematization with reference to criteria, which are the matter-of-course for one man, may not conform to the perception of the other man at all.

2.2. If documents and their metadata are stored and processed in a database, then metadata are retrieved using queries. The main disadvantage is that if additional metadata are entered, then the database structure has to be changed.

3. A specialized use of document management metadata description using XML (Extensible Markup Language) technologies. The XML format provides with flexible possibilities to work with document metadata, consequently the developed elements of metadata correspond to the environment of organization document development and management [13]. The main criteria in selecting XML format are as follows [11,13]:
   - XML can be used for describing metadata used in one system so, that it would be possible to integrate with metadata used in another system.
   - A clear syntax and semantics of the structure enables to read XML type documents with ease.
   - XML enables to group unstructured data of the organization into the categories with an initial, search-adjusted structure of metadata.

Table 1. Elements of Document Metadata in the Editorial System

The Type of Metadata	Description
Document content	Metadata, which are formed from the document's content. The main elements: the author's name, key words, the title, date – event.
Document type	Metadata, which describe the format of document presentation (pdf, html, doc, and similar) and development (XML data description standards).
Document history	Information about the development, editing and use of the document or its parts: sources of various parts of the document, the beginning of historical facts, the editing date and the performed changes, document publishing date.
Document structure	Information about the elements, which compose the document, and their interrelations: titles, the division of sections and subsections, headers.
Document state	Information about the state of document's readiness for publication: draft for publication, draft for validation, draft for designing, a published document, and similar.
Document location	Information about the location where the document or its elements are stored, about possibilities and conditions of the access to the document or its elements.
Documents classification	Features, which determine the system of classifying the document or its parts. Edition's columns, significant repetitive events and similar can be used for descriptions.
Document access	Information about possibilities of using the document or its parts and security: passwords, access facilities, exchange protocol.
Documents archive	Information, which describes physical and logical requirements for documents or its parts. Limitations on actions with archive documents, the required hardware and software can be the examples of such information.

The use such metadata elements may improve the selection of data and the quality of search.

6. Further Works

The implementation of editorial document management system is planned using the WebDAV (*Web-based Distributed Authoring and Versioning*) protocol, which enables to solve all the tasks raised for its functioning. WebDAV protocol is an extension of HTTP1.1 and offers additional exchange, which allows users to update sources (to work with documents) safely on remote Web servers: resource management, metadata management, concurrency control, resource namespace manipulation and other (Figure 5). In order to provide effective resource authoring on the Web and to enable to exchange of documents, WebDAV supports multiple versions of resources [2, 5].

The information associated with a Web resource, such as its title, author and so on, is defined as

metadata using properties. A property is a name/value pair. The *property name* is a URI consisting of the resource URI concatenated with the tag name of the property value. This allows properties to have globally unique names. The *property value* is considered as an XML element, which allows the property value to contain a potentially complex structure. Properties are either live or dead. A *live property* is one that is maintained by the server. Client applications can access the live properties, but they cannot update them. The server controls their updates. However a *dead property* is maintained by the client and simply stored and retrieved by the server. The client application is responsible for setting and updating the value of dead properties) [2,5].

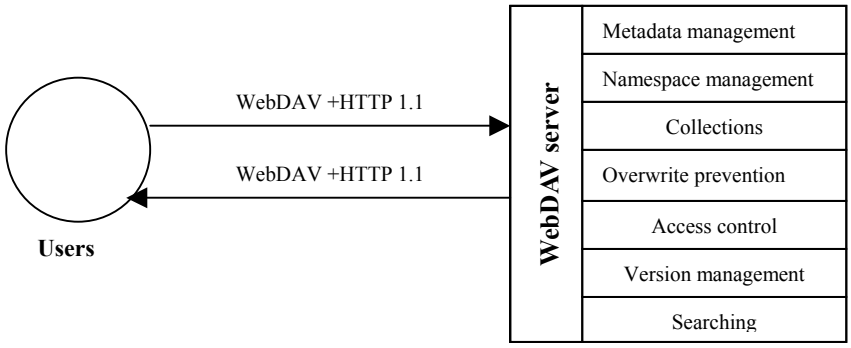

Figure 3. The Technological Solution of the Editorial System

7. Conclusions

The editorial system, which works with large amounts of information, encounters the problem of document management. The solution of applying standard document management systems does not provide with a potential efficiency as such systems are not characterized by flexibility in adjusting to document flows that circulate in the publishing process and to the information life-cycle. Whereas e-documents are used in the process of periodical publishing, so the smallest costs would be in using universal ways of describing e-documents metadata, however the main obstacle that limits their integration is a closed way of describing documents metadata. While analyzing documents, which circulate in publishing, and their description features for a fast selection of a document or its parts and for reuse, it was noticed that the overload of unnecessary information may occur. Therefore for the editorial system's document management, we suggest to develop a document metadata description format based on the XML standard, which defines the syntax and rules only. As WebDAV is based on an open standard, it is a mature and accepted protocol, currently used in many software products, so in the realization of an editorial system for the description and management of documents and their metadata we suggest to use flexible technologies – the WebDAV protocol.

Acknowledgements

This research was carried out during the project Software platform for remote workstations in editorial offices of the Eureka program with the co-financing of Lithuanian Research and Studies Fund.

References

[1] **A. Waard, J Kircz, P. Hendrikkade.** Metadata in Science Publishing. 2003. *Online version accessed on* 2005 09 10: *http://wwwis.win.tue.nl/infwet03/proceedings/8.pdf.*

[2] **B. Shadgar, I. Holyer.** WebDAV-Based Distributed Authoring of Databases. *In: The First Eurasia Conference on Advances in Information & Communication Technology. Austrian Computer Society, October* 2002. *Online version accessed on* 2005 09 25: *http:// www.cs.bris.ac.uk/Publications/ Papers/1000686.pdf.*

[3] DCMI Usage Board, DCMI Metadata Terms. 2004. *Online version accessed on* 2004 11 14: *http://dublincore.org/ documents/dcmi-terms/.*

[4] **D. Small.** A Model-Driven Architecture for Enterprise Document. *Management, Supporting Discovery and Reuse*, 1999. *Online version accessed on* 2004 11 14: *http://www.comp.leeds.ac.uk/research/pubs/theses/ small.ps.gz.*

[5] **E.J. Whitehead, Y.Y. Goland.** The WebDAV property design. *Software-Practice & Experience*, 34/2, 2004, 135-161.

[6] **J. Zwicker.** Some problems of authenticity in an electronic environment. *From ICA precongress conference in Elbląg*, 22–24 May, 2003.

[7] **K. Jeffery.** Metadata: The Future of Information Systems. *World Meteorological Organization*. 2000. *Online version accessed on* 2004-04-21: *http://www. wmo.ch/web/www/WDM/ET-IDM/Doc-2-3.pdf.*

[8] **M. Demarest.** Understanding Knowledge Management. *Long Range Planning*, 30/3, 1997, 374-384.

[9] Resource Description Framework (RDF). 2004. *Online version accessed on* 2004 11 14: *http://www. w3.org/RDF/.*

[10] **P. Dourish, W.K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D.B. Terry, J. Thornton.** Extending Document Management Systems with User-Specific Active Properties. *ACM Transactions on Information Systems*, 18/2, 2000, 140-170.

[11] **Y.H. Yao, A. Trappey, P.S. Ho.** XML-based ISO9000 electronic document management system. *Robotics and Computer Integrated Manufacturing*, 2003, *No*.19, 355 – 370.

[12] **V. Lyytikäinen**. Contextual and Structural Metadata in Enterprise Document Management, 2004. *Online version accessed on* 2004 11 14: *http://selene.lib.jyu.fi:8080/vaitos/studies/studcomp/9513917835.pdf*.

[13] World Wide Web Consortium. Extensible Markup Language (XML) 1.0 (Third Edition). *Online version accessed on* 2004-05-03: *http://www.w3.org/TR/REC-xml*.