Document Management Techniques & Technologies

Joseph P. Sathiadas[1], G.N. Wikramanayake[2]
1. Virtusa (Pvt) Ltd., Tel: 0777 313815, Fax: 074 724161, email: jpsathiadas@eureka.lk
2. University of Colombo School of Computing

Abstract

Electronic Document Management System (EDMS) is a rapidly developing technology and is considered as the solution for organizations that needs a way to manage the information efficiently. EDMS applications focus on the control of electronic documents throughout their entire life cycle, from creation to eventual archiving. Its functions include document creation, storage and retrieval, management, version control, workflow and multiple delivery formats.

Document management is not a single entity or technology, but rather a combination of elements. It is the use of information and different users in a business process, combined with the technology that permits the interaction. The technologies that make up the EDMS are categorized into distinct functional groupings. We present these and describes the techniques used to electronically manage documents. We also explores the immediate future of the EDMS and conclude that having the EDMS industry is at crossroads in its own lifecycle and is made up of a highly fragmented group of products with no single integrated vendor or framework for automating the entire cradle to grave document life cycle.

Keywords: Electronic Document Management System; EDMS; DMS; Resource Management

1. Introduction

Most of the organizations have vast amount of information that are required for their on-going projects or for their future projects in the form of knowledge of their workers or in documents. But, lack of information sharing among people and various project groups, lack of good management of information assets and lack of support from the knowledge workers make this information not available and not useful. Hence, the need for a system that could cater for this requirement and address these issues came up.

An Electronic Document Management System (EDMS) [4, 7] address most of these issues and is considered as the solution for organizations that needs a solution to manage the information efficiently. Although data management has been there for the last 30 years or so, document management came into the picture only about 10 years back. EDMS became popular with the advent of technology growth and computers.

2. Functions of EDMS

EDMS applications focus on the control of electronic documents throughout their entire life cycle, from creation to eventual archiving. Its functions [2, 6] include document creation, storage and retrieval, management, version control, workflow and multiple delivery formats.

Document creation: A document is a container, which brings together information from a variety of sources, in a number of formats, around a specific topic to meet the need of a particular individual or an organization.

Storage and retrieval: This involves storing and retrieving documents in a storage device such as, hard disk, tape etc.

Management: This covers a wide area of managing all the documents efficiently to cater to the needs of the organization and the individuals.

Version control: This is a way of tracking changes done to a document and the ability to retrieve old versions of a document.

Workflow: This is a way of tracking the state of the document and who is responsible for that step.

Multiple delivery formats: Ways of delivering the document content in various formats, such as PDF, Word, Image etc., to cater to the requirements of the end users.

3. The Document Management Space

Document management is not a single entity or technology, but rather a combination of elements. It is the use of information and different users in a business process, combined with the technology that permits the interaction. Hence, the Document Management Space can be divided into four major areas namely: documents, people, processes and technology.

Documents: The wealth of an organization is the information it has. Approximately, 20% of this information lies as data and the remaining 80% lies in documents. The 20% of data are normally well managed and maintained in databases. Lot of effort have gone into managing and utilizing this data, without giving much care about the documents that have most part of the information.

People: As like any other systems, document management system also serves a variety of different users. Users can be a Creator, Coordinator or a Consumer. A single person can also play multiple roles. The creator is the author and generates document content. Coordinators ensure that a document is properly reviewed and approved for release. They are responsible for assigning tasks for other members to perform on the document. They are responsible for the delivery of the document to the Consumer. Consumers are the real end users of the documents, who read or study them. Consumers rely on the Coordinator to get them what they want in an appropriate format.

Processes: When a document goes from conception to consumption, a process has to be in place to endure that every thing is going as planned and according to the expectation. The figure 1 illustrates the process of converting Data into Knowledge.

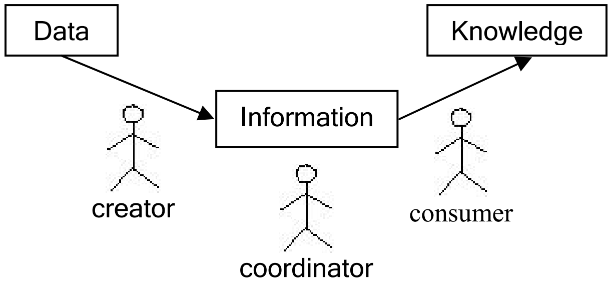

Figure 1: Converting Data into Knowledge

Technology: With the rapid technology growth, a powerful PC running a GUI is common in most of the offices. Imaging, database, networks and desktop applications are some examples of the technologies associated with the Document Management Space.

4. EDMS Technologies

The technologies that make up the EDMS can be categorized into six distinct functional groupings namely: repository, conversion, indexing and searching, creation, workflow and distribution [8].

4.1. Repository

This is the place where documents/objects are persisted and restored for use. DBMS and/or file systems are the most commonly used repositories. When you consider a repository, it consists of a database and/or a file system as the backend, a server engine in the middle and a client system as the front end.

With the current operating systems optimized to handle files, most repositories are built with both a file system and a database. While the file system is used to store the actual document or object, the database is used to store the Meta data and rules. The database engine or a complete separate application can act as the server application. By the introduction of the server in the middle layer, the client becomes thin. The client system can be a stand-alone application. Now most of the systems provide a web based thin client as their front end.

Library Services (Check in/Check out services, Verification of privileges, Access Control), Version Control (Version numbering, Linear or Branched versioning, Creating and controlling versions, Archiving) and Configuration Management (Virtual Documents) are the primary functions provided by the repositories. The secondary functions provided by the repositories are conversion, searching and indexing, and workflow.

Open document management API (ODMA) and Document management alliance (DMA) are two main repository standards. ODMA is working towards a set of interfaces between desktop applications and document management client software and DMA is working towards a set of interfaces between document management servers.

4.2. Conversion

Most of the documents are comprised of text, images and multimedia objects. Since, each of these formats are fundamentally different, they have to be converted using different tools and techniques.

Conversion is generally regarded as a non-value adding process. But, conversion that helps to improve the performance of searching and retrieving is considered as value adding process.

Some of the common standard formats are:

Text	– ASCII, SGML, HTML
Graphics	– CGM, IGES, TIFF, GIF, JPEG
Multimedia	– MPEG

Document conversion process is described by figure 2. Here, the source format is the format of the original documents and the target format is the format required/preferred by the end users. A filter is used in the middle to find the best way to do this conversion.

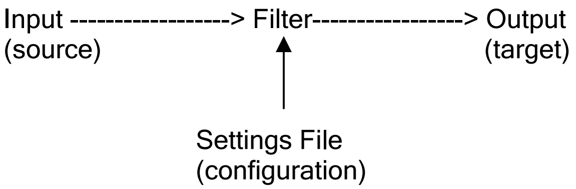

Figure 2: Document conversion

4.3. Indexing and Searching

4.3.1. Indexing

With the growth of the document repositories and online document sets, the speed of information retrieval becomes critical. Indexing allows a way of this by breaking up a document into more granular down to word level. Inversion of terms is one popular way of indexing. This method will have a sorted list of all the keywords in a file with pointers pointing to the actual location.

Normally the documents are fed to the indexing process initially. Some of the most popular indexing engines such as Open Text, Verity, Fulcrum [2] can accept most major word processing documents.

4.3.2. Searching

Any EDMS should provide the searching facility to the user. The user must be able to search for a particular document or a document containing particular information without delay.

Information retrieval effectiveness is measured by recall and precision. Recall is the proportion of the relative materials retrieved and precision is the proportion of retrieved material that are relevant.

The following types of search mechanisms are used.

Secondary search: Search within the results of the first search.
Semantics: Identify the exact definition of the word and then start searching.

Synonym: When searching for a particular word, search also for its synonyms (e.g. searching for Book or Publication).

Boolean: is the ability to merge together multiple words with operators like AND or OR.

Proximity: is the ability to search for multiple words grouped together.

Fuzzy word: is used when the word is not perfect due to some reason (OCR reading etc.). This uses character and string pattern searching to determine the correct word.

Concept: is the ability to pass information or news related to a concept.

4.4. Creation

Initially, document creation focused on creating paper output only. But, with the introduction of online document management system, features like document content, format and electronic structure also started to have an impact.

The following has to be considered when an online document is created:

- What the reader sees
- Hyperlink information
- Visual conventions
- Document sizing
- Number of files needed for online manual
- Purpose and content of information
- How the information will be used

Many fail to realize that documents are not mere static collection of texts, but are capable of embedding organizational knowledge and transmitting across the organization through the use of tools.

We could find three types of knowledge in documents, namely: structural knowledge (how a document is constructed), domain-specific knowledge (what the document contains) and contextual knowledge (how document sections relate).

4.5. Workflow

Workflow is the movement of a document through a series of steps to achieve a desired business objective. Workflow seeks to eliminate wasted time such as the time documents spend sitting in an in-box, the time taken to gather information to take action and time spent in moving documents from one person to the next.

Though workflow is not part of document management, having this feature enrich the functionality of the document management process. A document management repository controls documents, while a workflow engine controls the review and approval process.

For a document management system to give a meaningful solution a workflow must integrate properly with users, security, versioning, attributes, documents and relationships.

A workflow comprise of 4 primary elements, namely: process, actions, people and document.

Process: the sequence of steps necessary to reach an end objective, the business process. The two fundamental types of process are structured and ad-hoc.

Actions: what's to be done at each step? The two classes of workflow used as part of an EDMS are Review &Approval and Other tasks.

People: who is to accomplish these items? Normally this is specified by roles rather than by individuals.

Document: the focus of the process.

The key to implementing the workflow is not technical, but rather the people and their resistance to change.

There are four types of workflows, namely: sequential, parallel, branching and time drive.

Sequential: Linear set of steps. Each step is dependent on the completion of it's previous step

Parallel: Document can be passed to multiple people for action at the same time. This introduces an issue of reconciling the results of each parallel path.

Branching: Is a conditional type of workflow where paths to be taken are chosen based on a criteria.

Time driven: A time period is defined for a step, along with a action to be performed if the time period is exceeded.

4.6. Distribution

Distribution is the act of delivering the needed information, usually in document form, to the end user who will use or consume it. The format of the document will vary based on the organizational style and user needs.

Although many think that electronic document is the mode of distribution when it comes to EDMS, it is actually false. Many readers still prefer paper documents than electronic documents. Hence, EDMS should continue to support output in paper form. Nevertheless, electronic document delivery has many advantages over the traditional paper delivery. They are:

- Lower cost of document distribution
- Easier maintenance and updates to the documents
- Faster access to documents
- Greater nonlinear access to information
- Customized views of the document set
- Better quality of presentation

Distribution of documents takes place in four different ways as online publishing, offline publishing, repository viewing and web access. In the typical online publishing, documents are viewable by the consumer. In the offline publishing method, the documents extracted by the repository is published for use by the consumer. In the repository-viewing method the consumer searches and retrieves documents from the repository. Finally in the web access method (figure 3) the consumer uses a web browser to search and retrieve documents.

Document viewing is done for electronic paper, online documents and native formats. Tools are available for viewing of such documents. Important features of these tools are searching, zooming, hyper linking, annotations, outlining/tables of contents, bookmarks printing and integration capabilities.

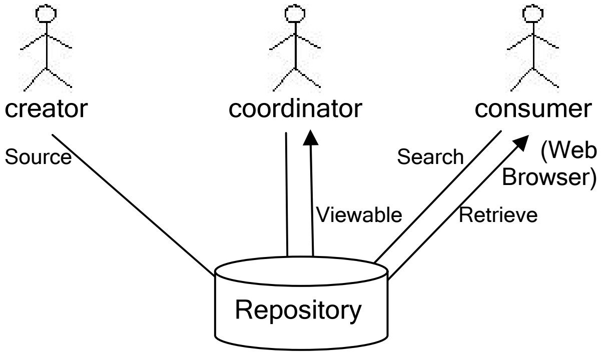

Figure 3: Web Access Method

Currently most of the document management tools have a 2D interface for the user to display the grouping of the documents. Good example is the Explorer of windows, which has a tree structure to display the directories and within it, it displays the document titles or thumbnails.

One interesting thought towards this is to explore the viability of providing a 3D interface for this purpose. The 3D interface can have a 3D plane, and the thumbnails of the documents can be arranged on this plane by the user. This design is expected to exploit the humans' natural capacity for spatial memory and cognition. The user has the advantage of placing the important documents close to him and the less important documents away from him. Finding a document is just like locating something in a house or room. Some amount of research has taken place in analyzing the performance and user preference of this approach. Research done by Andy Cockburn and Bruce McKenzie shows that 2D interfaces performed slightly better than 3D when it comes to storage and retrieval times. But, the users preferred 3D interfaces than 2D [3].

5. Techniques

The eventual goal of document management is a single electronic system combining paper and digital information [10]. The document management domain is continuously undergoing changes exploring new techniques to cater to the demand of the industry needs. As for the immediate future of document management is concerned, it can be said that the focus will be mainly on areas such as web growth, file compatibility, structured documents, efficient document storage and graphical interface. Structured documents, XML databases, file manager interfaces and intelligent documents are new technologies used for the management of documents.

5.1. Resource Management Design Pattern

The main intention of the Resource Management design pattern is to manage multiple resources of the same type. To maintain the status of managed resources, a Resource Manager is implemented having lists. Resource Managers are commonly used to control access to any class of "sensitive" resource objects. Resource Managers maintain two lists of resources, i.e. free list and busy list. Free list contains all the resources that are free. Busy list contains all the resources that are busy. A resource allocation request is serviced by the resource manager by allocating a resource from the free list and putting it in the busy list. Similarly, a resource release request is serviced by the resource manager by inserting the freed resource to the free list.

A DMS can be considered as an instance of the Resource Management design pattern. In this particular instance, the documents are the managed resources and the Document Manager can be considered as the resource manager. Hence, the Document Manager allocates, tracks, controls, and de-allocates document objects. Clients cannot access documents directly. Instead, they must access these objects indirectly, through the Document Manager, using identification numbers provided by the Document Manager. In this way, illegal client requests can be detected.

Document Manager maintains free list and the busy list. When a document is requested by a client, the document manager first checks if that document exists in the free list and if it exists there, it retrieves the document and passes it to the requesting client. It removes the entry of this document from the free list and places it in the busy list. While maintaining the free/busy status of the documents, the Document Manager must be able to keep track of some additional state attributes of the document also. These can be facilitated by storing these attributes as a collection in the free/busy list along with the document information.

5.2. Structured Documents

While HTML has become the universal electronic delivery encoding of a document, it is neither presentationally rich enough to support high quality paper delivery, nor expressive enough to allow easy automatic transformations to multiple formats or to recombine the content. Generic structured markup like that provided by SGML or XML, opens many new possibilities for Document Management Systems (DMS's). A DMS that supports such standards natively is referred as a *Structured DMS*. The characteristics of a Structured DMS are as follows [1]:

`On-the-fly' creation of renditions: SGML and XML support delivery from single source in multiple formats. This is achieved by having a style sheet or filter for a class of documents. Instead of storing all these renditions, these can be created 'on-the-fly' as per the need.

Automatic transformations: This is the process of reordering document components or incorporating into new documents. For instance, entries from a dictionary identified as being legal terms might be extracted from a general dictionary to create a new legal dictionary.

Access control at element level: This transformation capability can be used to strip out certain components from documents before they are delivered to a user.

Access to elements (component versioning): Most of the DMS's manage complete documents. Component versioning is the approach of versioning the components of a document.

Intensional versioning: *A*llows a number of rules to be applied to a version set to derive a particular *variant*.

Human-readable description of changes: Because the logical structure of documents is explicit in structured documents, differences between versions of a given document (commonly termed *deltas*) can be represented in terms of operations on its elements--*syntactic differencing.*

Extended search capabilities: The markup in structured documents can be used for more sophisticated searching capabilities. Documents and elements can be retrieved based on structural relationships and a mixture of content and structure.

Document-based workflow: SGML and XML also provide a convenient syntax for describing data associated with business processes. Workflow *process definitions* (defining the sequence of activities that make up a business process) and *process instances* (recording the status of a particular business process in progress) can be represented as one or more structured documents. This facilitates mobility of the business process between systems and allows the process to be advanced between connections to a central server. Because these processes are just another SGML or XML document, they can be versioned like any other document in the DMS.

5.3. Storage: XML Database

The shift from SGML to XML has created new demands for managing structured documents. Many XML documents will be transient representations for the purpose of data exchange between different types of applications, but there will also be a need for effective means to manage persistent XML data as a database. The paper, *Requirements for XML Document Database Systems* [9], explores requirements for an XML database management system. The paper does not suggest a single type of system covering all necessary features. Instead it aims to initiate discussion of the requirements arising from document collections, to offer a context in which to evaluate current and future solutions, and to encourage the development of proper models and systems for XML database management.

5.4. File Manager Interface

Most part of the time is spent looking for information than actually using it. This problem is the result of the shortcomings of the modern desktop. File manager software is no longer an effective tool for managing documents. Tools for creation and information exploration are disintegrated. Key contextual information is hidden from the user (we call this *the hidden we*b). Search tools are impersonal.

Conventional file managers organize the documents based on the directory hierarchy. The computer directory tree is one of the oldest artifacts of the pre-web era and is virtually unchanged since its creation. After nearly 30 years, the only significant advancement in file management software is the overlay of a graphical interface on what is still a text-based directory.

The directory structure is a poor way for a human to organize documents, since we organize *contextually* as well as hierarchically. This problem is particularly apparent when documents contain numerous references, both to other user documents and to documents on the World Wide Web. Directories were simply never intended to highlight and manage the relationships between information *within* documents.

The paper, *The Personal Web* [11], describes the nature of the modern personal information space *(the personal we*b) and a tool that improves on conventional file management for organizing and exploring that space. It is based on the concept that a user's web experience should be as personal as possible, flowing easily between user and web documents, following various types of document relationship "links", and involving searches that take into account *who* is doing the searching.

Studies are underway to explore the possibility of providing 3D interfaces for the file manager. The main idea behind this is to give the user the feeling that he is placing a document in a physical location. Humans have a tendency to remember a position, which is similar to a position he deals with in his daily life rather than a hierarchical arrangement of files.

5.5. Intelligent Documents

An *intelligent document* contains knowledge about itself and its environment. It supports assembly of documents based on inputs given by the user [5]. An *active* intelligent document is able to construct and transform itself dynamically.

One of the basic problems in document management is to provide on-demand generation of individualized documents through dynamic *document assembly*. Document assembly composes new documents from an existing collection of documents. Naturally, document markup and structure contribute to the retrieval of the document fragments.

Automated assembly consists of three phases [5], namely: The user expresses his demands, Appropriate documents or document fragments are found and returned and The returned fragments are merged into a single uniform assembled document. Hence, the final document will be composed with information from various different documents.

5.6. Industry Standards for EDMS

Although ODMA and DMA have been the industry standards for Document Management for the past several years, most vendors are now looking into XML to become a standard for document management. XML promises to succeed where SGML failed, by being easier to implement. It is expected that, apart from virtually universal support, XML will also offer for the first time the opportunity to embed metadata intelligence *within* the documents themselves. To gain the maximum benefits of XML, all the documents should be converted to XML. But, this will be an expensive operation when we consider the millions of legacy documents stored in different vendor specific formats. But, XML can be effectively used to store meta data of the documents. In DM, the documents are generally stored in the file systems as flat files, but the meta data is usually stored in the database for frequent quick access. Indexing and searching capabilities are provided through this meta data. Currently, most of the vendors have their own vendor specific meta data format in XML without using a industry standard meta data format. This is mainly due to the fact that a meta data standard in XML doesn’t exist currently.

If a general detailed meta data format is defined in XML and if the industry accepts to use that as the standard to store meta data, then all the individual DMS can be considered as one large virtual DMS. Hence, search and index operations can be generalized. Any client will be able to look up the information in a DMS provided they have the proper security credentials to do the operation.

Since the current trend is to web enable all the applications, DMS is also not an exception. But the industry standards that exist for DMS, such as ODMA, are not optimized for web and multiple platforms. A general, industry standard, robust framework is needed for DMS so that any web based client will be able to connect to the DMS server to obtain services. Web services is one technology that can be seriously considered for this.

When a client request for a service from the DMS server, it has to pass the necessary information for this service and DMS should be able to return the results after the service is performed. For this, the client and the DMS server have to use a common language for communication. Due to XML’s flexibility and robustness, it will be one of the ideal candidates for this.

6. Conclusion

To conclude the paper we first summarise the benefits of EDMS and then address the current status of EDMS and its drawbacks.

6.1. Benefits of EDMS

The benefits of EDMS can be described as follows:

- Lower cost of document creation and distribution
  No material (paper) cost involved during the creation phase, and the documents can be distributed as a softcopy to whoever concerned without the need to make hard copies
- Improved, customized access to documents
  Users will be able to view all the available documents and open any of them by just clicking the mouse. Read, write permissions could be set per user by the owner or administrator
- Faster document creation and update processes
- Increased reuse and leverage of existing information
  Information in a document could be reused and leveraged by giving links in other documents or by the concept of virtual documents
- Better employee collaboration
  Inputs from all concerned could be accumulated in a single document in real time and the necessary changes could be done
- Reduced cycle times in document centred processes
- More complete regulatory compliance
- Refined managerial control and reporting
- Enhanced document control and security

The administrator could control documents by setting privileges and access restrictions.

- Improved productivity/Reduced headcount
- Better customer/Client satisfaction
- Quick and easy access

6.2. Current State

The EDMS industry is at crossroads in its own lifecycle. The industry is made up of a highly fragmented group of products with no single integrated vendor or framework for automating the entire cradle to grave document life cycle. Enterprises are currently trying to overcome this issue on an ad-hoc basis with no clear vision or path to the future for solving the complete document lifecycle problem.

6.3. Drawbacks of EDMS

Although document management has many advantages and benefits, it also has its risks and drawbacks.

Problems with the current state of the document management are as follows:

- Technologies are too difficult and take too much time to implement. In some cases the solution takes longer to implement than the life span of the technologies.
- Lack of standardization. Solutions involving documents are usually not compatible with one another
- Pseudo-standards have emerged that are still vendor specific.
- Difficulty in managing documents independent of the application.
- Typical document solutions are implemented in phases. The technology to create, manage, and archive these documents must be as modular as the implementations.
- The idea of plug-and-play has never been implemented past most marketing departments.
- Integrators have limited resources to learn new tools.

When there exist ways to publish documents online, many organizations rush headlong and put the documents online without proper document creation and management processes in place. Online distribution without proper document control tends to create problems, since it could lead to wrong information or outdated information being sent to a large group by mistake. A well-designed EDMS should address this problem effectively, e.g. date stamps each document.

Another drawback with EDMS is that it is driven by documents and technology, but not the end users. Hence, the focus on the end user is lost and this tends to give a negative impact on the overall system. Document create process can address this problem by making the end user contribute towards it.

Most of the organizations, when provided with an EDMS, tend to dive in and publish all their documents, irrespective of their state, quality and the need. This leads to information overload. Hence, a process should be in place for document control and this should actively involve the authors and users. Finally, the end users, especially the new users, might react negatively to the introduction of the system because the learning curve is large and they are used to the less challenging print based media.

References

1. Arnold-Moore Timothy, Fuller Michael, Sacks-Davis Ron, “System Architectures for Structured Document Data”, Markup Languages Vol. 2 No. 1, 2000, pg. 11-39, accessed on 101103, http://www.mds.rmit.edu.au/~msf/papers/MT99.html

2. Bielawski Larry and Boyle Jim, “Electronic Document Management Systems: A user centered approach for creating, distributing and managing online publications”, Prentice Hall Computer Books, November 1996.

3. Cockburn A. and McKenzie B. “3D or Not 3D? Evaluating the Effect of the Third Dimension in a Document Management System”, Addison-Wesley, Proceedings of ACM CHI'2001 Conference on Human Factors in Computing Systems, Seattle, Washington, March 31-April 6 2001, pages 434-441, accessed on 1011003, http://www.cosc.canterbury.ac.nz/~andy/papers/chi01DM.pdf

4. Condon Thomas A., Roberts Doug, Nash Dawn, “Understanding EDMS: A guide to efficiently storing, managing, and processing your organisations documentation”, white paper, 2002, accessed on 101103, http://www.sdichicago.com/insidetheitstudio2002Q2/edmsfeature.pdf

5. Document Management Research Group, Department of Computer Science, University of Helsinki: Structured and Intelligent Documents, accessed on 101103, http://www.cs.helsinki.fi/research/rati/sid.html

6. Functional Assessment of Hummingbird Enterprise, July 2002, accessed on 101003, http://devx.newmediary.com/abstract.aspx?&scid=231&docid=37933

7. Johns Hopkins Center for Information Services: Document Management Systems Recommendations, June 2002, accessed on 101103, http://it.jhu.edu/divisions/nts/status/systemsrec.html

8. Meyers Scott and Jones Jason, "Document design for effective electronic publication", Proceedings of the 5th Conference on Human Factors & the Web, June 1999, accessed on 101103, http://zing.ncsl.nist.gov/hfweb/proceedings/meyers-jones/

9. Salminen A. and Tompa F.W., "Requirements for XML document database systems". In E.V. Munson (Ed.), Proceedings of the ACM Symposium on Document Engineering (DocEng '01) (pp. 85-94). New York: ACM Press, 2001, accessed on 101103, http://www.cs.jyu.fi/~airi/presentations/XMLdatabases.ppt

10. Smith Brady, "The Future of Document Management", February 2002, accessed on 101103, http://www.arches.uga.edu/~cpeter/Future.htm

11. Wolber David, Kepe Michael, Ranitovic Igor, "Exposing Document Context in the Personal Web", Proceedings of the International Conference on Intelligent User Interfaces (IUI 2002), San Francisco, CA, 2002, accessed on 101103, http://www.usfca.edu/~wolberd/papers/iui2002Final.pdf