Managing Unstructured and Semi-Structured Information in Organisations

Dr Ashley M. Aitken
School of Information Systems, Curtin University of Technology
A.Aitken@Curtin.Edu.Au

Abstract

This paper considers software systems for the management of unstructured and semi-structured information in organisations. It catalogues and explains information challenges facing organisations, namely the need to lower barriers to information, to provide better version control and collaboration, to provide more consistent productivity tool functionality and more powerful search capabilities, and to reduce information maintenance costs. It catalogues and discusses current (partial) solutions to these challenges (including file servers, web servers, content and document management systems, wiki webs, portals and databases) and points out their limitations. It describes the characteristics of, and sketches, a more complete solution to these challenges (namely a centralised information repository accessible from a rich client application and from Web browsers, on-line and off-line, that allows easier creation and publishing of information, easier access and collaboration from inside and outside the organisation, finer grain access control, automatic versioning, treats different information similarly and with consistent functionality, full content search, and fine-grained reuse of information).

1. Introduction

Whilst many people and organisations have, in the past, focussed on the well-defined problem of data management, and many now focus on the seemingly more intangible problem of knowledge management, there has been little real progress over the years in the management of unstructured and semi-structured information within organisations. Although there are numerous challenges surrounding the management of such information (and not all of them are technical), it is seen by many of the corporate research agencies as a critical area for improvement, productivity gains, and significant cost savings in organisations going forward.

The goal (and contribution) of this paper is to catalogue major information management challenges facing organisations, to catalogue current partial solutions to these information management challenges, and finally to sketch what a more complete solution to information management may be like – in particular, what characteristics it would likely have. This paper is not based upon any extensive survey of information management within organisations but rather evidence obtained from working with a number of small, medium and large organisations, discussions with others working in similar organisations, and personal experience of the author with information management in his workplace. The critique of current solutions is also based on the generally understood limitations of these solutions and the author's experience with, and investigation of, these solutions.

Before the information management challenges are discussed in more detail it is important to set the context for the discussion. In particular, and firstly, that the paper is focussing on information management not data management, secondly that the paper is focussing on enterprise information management not personal information management, thirdly that the paper is focussing on the productivity aspects of information management not professional presentation aspects, and finally that the paper is focussing on something between the formally structured information in a database and completely unstructured information, which is generally referred to as semi-structured information. Although information management is definitely not just a technological problem, technological solutions, particularly solutions involving software systems, will be the focus of this paper.

In section two, the paper will catalogue and explain some of the main information management challenges currently facing organisations. In section three, it will catalogue and explain a number of contemporary (and partial) solutions to some of these challenges and, in particular, evaluate and discuss the various advantages and disadvantages of these solutions. Finally, in section four it will consider the characteristics of a more complete solution to these problems. The paper doesn't offer a complete solution,

and perhaps there never will be one (for technological or other reasons) to the information challenges organisations face. However, it does aim to assist organisations when looking for (or evaluating) a better solution to their information management needs now and in the future.

2. Information Challenges

This section catalogues and briefly explains some of the significant information challenges facing organisations.

2.1. Lowering Information Barriers

There are currently many barriers to creating, sharing, and accessing information within organisations. Most information within organisations is stored in documents in files on file servers, in pages on a Web site, in specific corporate applications, or on (any of the many) PCs (at the organisation and possibly the homes of employees) and laptops. To find information stored within a document in a file it is generally necessary to find out on whose computer or on what server that file is stored and then where in the file system it is stored. It is easy to store information in files using traditional productivity tools but then these files are difficult to find and manage on personal computer or servers. On the other hand, the Web generally provides easy access to information (although links need to be organised and maintained) but creating and installing Web pages has generally been an extended and complex process.

Another problem facing today's organisations is that their staff generally want access to all of their information from their work PC using standard productivity tools and access to the same information on the intranet (when they are away from their desk or to share with others) and on the Internet using a Web browser (when they are working outside of the organisation). It is also becoming more common to share particular corporate information with customers and suppliers, as a way of extending the organisation and providing a better service to customers, and getting better service from suppliers. It is thus necessary to consider how organisations can make information available for viewing and editing from both traditional desktop applications and from more contemporary Web browsers, inside and outside of the organisation, and to staff, customers, and partners. The challenge is not just about lowering barriers; it is also about increasing access to information.

2.2. Providing Better Access Control

Another challenge with information management in organisations is controlling access to information, i.e. who can read specific pieces of information, who can edit specific pieces of information, who can create or delete specific pieces of information. Currently, access control within organisations is generally file-based and, as a result, generally too coarse-grained, i.e. giving access to larger pieces of information than would optimally be the case. Often it would be more beneficial to be able to give certain users the ability to read certain parts of a document (specifically, certain pieces of information) and other users the ability to edit certain parts of a document (again, certain pieces of information). These are not functionality that should need to be implemented in custom applications; it should be a standard part of the information management and productivity infrastructure.

2.3. Better Version Control

Version control enables information users to see how information has changed over time and allows them to refer to (or revert to) previous versions if needed. Version control should be an important part of any organisation's information management strategy. One of the benefits of such versioning is that it can remove the reluctance of information producers to make major changes to information. Unfortunately, however, most version control for file-based information within organisations is manual and brittle. Some productivity tools provide facilities for version control within a document (and developers often use source code version control systems) but such facilities seem not to be widely understood, valued, or used by users. Most users make copies of files as a basic form of version control. This is ad-hoc, extra work, inefficient and error prone. Organisations need to provide more automated, perhaps enforced, and definitely integrated version control for all information within the organisation. With the growing auditing obligations on organisations, and requirements for information to be saved and its changes tracked and traceable, the completely uncontrolled information environment present in most organisations these days represents a significant challenge, risk, and potential liability.

2.4. Better Support For Collaboration

Individuals seldom work alone in their information processing tasks. In particular, they often work on documents with their colleagues inside an organisation and, more and more frequently these days, with others

outside the organisation. Collaboration can be an important facilitator of increased productivity and quality. However, there is little available within current information production tools and infrastructure that provides for effective collaboration amongst two or more people. Unfortunately, when traditional files start flying back and forwards between collaborators (especially when there are more than two) tracking and merging concurrent changes becomes very difficult. It is thus necessary to consider how organisations can provide better support for collaboration within groups of any size and including people from within and outside of an organisation (where access to a shared file server, for example, may not be as easy).

2.5. Consistent Productivity Tool Functionality

Information is information whether it is presented in a word processing document, in a spreadsheet, on a slide, in an outline, or any other way. Most current productivity suites tend to separate these different presentations of information by providing a different productivity tool for each type of presentation (e.g. we typically have spreadsheet applications and word processing applications). This does allow for specialisation in each of the tools but it has the disadvantage of making it harder and more cumbersome to share and reuse information across the tools and also for consistent use of the functionality available from each tool. It shouldn't be necessary to use a different tool to put a table into a word processing document with access to all the spreadsheet like functions that should naturally be associated with tables. Further, the application controlling this information should be consistent across all the different information formats within. One should be able to use formulas, not just in spreadsheets but also between paragraphs of text in a word processing document, or between a number on a slide and information in spreadsheet. All information should be treated consistently from within one common user interface. Having different applications makes it harder for users – requiring them to learn how to use multiple applications and also not allowing them to integrate the information as easily and consistently as possible.

2.6. Providing Powerful Search Capabilities

Probably one area where technology is (only recently) coming to assist with regards to information management in organisations is search. A key factor in reducing the replication and increasing the reuse of information is to make information easy to find. Traditionally, content-based search within organisations has been limited to slow file-by-file search on servers or personal computers. However, with the enormous success of Internet search engines (like Google) that can "search" the whole Internet in seconds for some specified keywords, people started to wonder why the experience of searching on a (much smaller) personal computer was so poor. This has led to a number of operating system-integrated and third-party content-focussed search tools (e.g. Google's Desktop Search and Apple's Spotlight) for personal computers and servers on par with the capabilities of Internet search engines. Although they provide a significant component of an information management solution, on their own they are not a complete solution, even for search. For example, what if individuals wish to search for information within an organisation that currently could be located or spread over many personal computers and servers? Also what is the use of searching for a term if you get back many different versions of the same file containing the term? As a result, providing powerful organisation-wide search capabilities is still a challenge for organisations.

2.7. Reducing Maintenance Costs

There is also, currently within organisations, a very high maintenance cost for information, primarily as a result of duplication of that information within many files and physical forms within organisations. As most organisations cannot afford the time or money involved with this maintenance there is often a lot of inconsistent and out-of-date information "floating around" within organisations. For example, some organisations keep important information (e.g. company policies, staff contact and phone lists) in files on central file servers. As well, these companies often make this information available on a Web page, email it to all staff, and even print out copies to keep in various places. Although, in this case, it is relatively easy to distribute the information in different forms, it is now more expensive to update and to keep consistent across the organisation. For example, if any of this information changes, it is necessary to update or replace all the various instances of the information that are distributed across the organisation. Duplication of files, and the "copy and paste" of information within an organisation is usually done to make access to the information easier (e.g. it is no longer hidden away somewhere on a file server). However, the flip side of this is that maintenance now becomes much more expensive.

3. Current (Partial) Solutions

This section will discuss a number of the current (partial) solutions to information management within

organisations. The solutions will be briefly described and explained, and then some of the important advantages and disadvantages noted. Table 1 summarises the evaluation of these solutions according to the challenges discussed in the previous section.

3.1. File Servers

File servers (or more generally file systems) provide a hierarchical store for information that is stored in files. They provide a way to organise and control access to information. They also provide relatively easy access to information (as long as you know the particular file server that contains the file that contains the particular information you are looking for), make it relatively simple to share information with others (as long as the others have access to the file server) within the organisations and can also locate the file, and now allow fast full-content indexing and search systems. However, these file servers are usually not accessible from outside the organisation (for security reasons), usually don't have any built-in version control (although version control file systems do exist), and because the information is stored in files they don't enable fine-grained access control or reuse or easy collaboration. These disadvantages often lead to a high maintenance cost for information stored in files on file servers (and file systems). They are, however, currently the most common and default solution for information management within organisations but they are very far from a complete or optimal solution.

3.2. Web Servers

The rise in popularity of the World Wide Web has seen web servers become a key information publishing system within organisations from the Intranet or more widely from anywhere on the Internet. A web server basically responds to requests to serve particular documents stored on the Web server. The advantages of a web server for information management is that it provides a centralised repository for information (like file servers), enable easy access to information (e.g. via links), and can offer full content search. The disadvantages are that it is generally speaking more difficult to publish information on a web server (although web content management systems and similar currently assist in this regard and will be discussed next), and the information container is generally still the file (although more and more content is being kept in databases). Further, the basic web server is not focussed on production or collaboration (although applications may be run within the web environment to provide these facilities).

3.3. Content Management Systems

Content Management Systems (CMSs) are generally web-based systems for managing the publishing of web-based content. They add to standard web server facilities to make it easier for regular users to publish information, and give administrators better control of when and to whom information is made available. The unit of information within a content management system is often a chunk of information like a piece of text, an image, or other media. Such chunks of information are often presented within the context of a page template. Content management systems are often used by organisations in the publishing business (e.g. news organisations) but they may also be used within an organisation as a means for publishing and managing information [1]. Content management systems thus have the advantage of making it easier to create information for publication on the web server, allow easier access to that information, and provide slightly finer-grain access control. However, they do not generally provide powerful collaboration or integrated productivity tools, or allow for the fine-grained reuse of information.

3.4. Document Management Systems

Document Management Systems (DMSs) are generally server-side software for storing, versioning, indexing and searching traditional documents [2]. A user uploads a document into a DMS where it can be annotated with further information (either manually or automatically or both) and made available for others to access. DMSs overcome some of the disadvantage of a file server but not all. For example, document management systems usually have built-in version control, can handle many types of documents produced

Table 1 - Evaluation of Current Solutions

Evaluation of Current Solutions	SOLUTIONS	File Server	Web Server	Content Management System	Document Management System	Wiki Webs	Portals	Databases	"Complete Solution"
CRITERIA									
Easy Info. Creation		YES	NO	YES	YES	YES	PART	NO	YES
Easy Info. Access		PART	YES	YES	YES	YES	YES	NO	YES
Fine-Grained Access Control		NO	NO	YES	NO	NO	PART	PART	YES
Automatic Integrated Version Control		NO	NO	YES	YES	YES	YES	NO	YES
Easy Collaboration		NO	NO	NO	NO	PART	PART	NO	YES
Fine-Grained Reuse		NO	NO	NO	NO	NO	NO	PART	YES
Integrated Productivity Tool Functionality		PART	NO	NO	NO	NO	PART	NO	YES
Full Content Search		PART	YES	YES	YES	YES	PART	PART	YES
Low Maintenance Costs		NO	PART	PART	PART	PART	PART	PART	YES

by many types of productivity applications, provide a centralised repository for information, and easy access to the files with full content searching. However, their disadvantages lie primarily in that they still use files as the container for information, and thus do not provide fine-grained access control, significantly better collaboration, or the ability to perform fine-grained information reuse. Document management systems are stuck in the traditional file paradigm.

3.5. Wiki Webs

A recent addition to the World Wide Web has been the use of wiki-wiki webs or more simply wiki webs [4], initially developed by Cunningham [3]. These are, in simple terms, web-based software systems that make web pages editable (generally in a relatively easy but also a relatively crude fashion). A user, with appropriate access rights, can create new web pages and edit current pages within the wiki web all from a Web browser. The most famous wiki web site is probably Wikipedia [5] – the collaborative, open, and free web-based encyclopedia. Wikis have found great popularity and use inside and outside organisations. The advantages they have are in the fact that it is relatively easy to create and access information via (almost) any web browser, most wikis have built in version control, allow collaboration, and full content search. The disadvantages of wikis are probably their web foundation. The web is not the most productive environment (for productivity work). As well, wikis don't really enable fine-grained reuse (although they encourage and support linking) or fine-grained access control. The unit of content of a wiki is most commonly a page (as in web page not physical printed page), and this is usually the scale at which access and reuse is managed.

3.6. Portals

Portals are a relatively new concept in organisations. They are not so much information productivity tools as starting points for information access and productivity within organisations. Portals are generally browser-based systems that provide a customisable "dashboard" for accessing information and applications within an organisation. Portals are, if you like, a focus point bringing together information in various formats and containers (e.g. emails, files, calendar entries) and targeting individual users or groups of users within an organisation. For example, it is possible for users to have their own customised portals into their organisation's information, or for a number of users to share a portal, as an enabler for collaboration, for example, on a project. Portals do have a lot of advantages in that they can give access to a shared repository of documents, calendar items, instant messaging, emails, contact details, and other corporate applications. However, they don't directly confront the challenges of fine-grained information control and reuse, and versioning. Portals are a good starting point for information work within an organisation but much depends on what is behind (or within) the portal as to whether or not they have any success at meeting the information management challenges facing organisations.

3.7. Databases

Although the name doesn't suggest it, databases are most commonly used to store information. Databases meet some of the challenges of information management within organisations in that they allow (in an appropriately written database application) fine-grained access control to information and (pre-specified) fine-grained reuse of information. However, much of this functionality relies on a custom database application. The database is mostly the engine for the persistent storage of its information. Generally though, as an information storage engine, the low-level structure of databases is too rigid, access control is quite difficult to setup and generally needs to be built into a custom application, they require experts to construct and modify the database itself, and have poor integration (if any) with productivity tools. In essence, a database is best thought of as just the low-level persistence mechanism; to meet the information challenges facing organisations a database is not enough. A database needs, an appropriate application that can, in a sense, bring the information to life. A database application (at least an organisational wide database application) is not the sort of thing that the average information worker would be able to develop, and yet it should be possible for an average information worker to be able build and structure information within an organisation.

4. Towards A More Complete Solution

A complete solution to the information management challenges that face organisations does not currently seem to exist (and will, as mentioned earlier, undoubtedly require more than just a technological solution). However, in considering the challenges to information management in organisations and the partial solutions currently available it is possible to describe desirable characteristics of a more complete solution, and even to sketch what a more complete solution may look like.

A more complete solution to the information management challenges in organisations would most likely have at least the following characteristics:

- It should enable the easy creation and publishing of information.
- It should enable easy access to information (from inside or outside the organisation).
- It should enable fine-grained access control.
- It should include automatic version control.
- It should enable easy collaboration between users (inside and outside the organisation).
- It should treat different types of information in a similar fashion.
- It should provide consistent productivity tool functionality.
- It should provide full content search (with knowledge of versioning).
- It should enable and encourage efficient fine-grained reuse of information.

A more complete solution to information management within organisations would most likely be a software system something like the following:

- It would most likely include a (logically) centralised repository for the organisation's information.
- It would most likely be accessible from rich desktop applications and, as well, from Web browsers.
- It would most likely enable on-line and off-line work and synchronisation (of off-line work) with the centralised repository.

It should be pointed out that although it is suggested that a more complete solution would have a centralised repository of information, this is in the logical sense. The physical implementation of this information repository may involve information distributed within the organisation or geographically at a number of different locations – as long as it appears to be a single logical information repository to users.

The implementation of such a more complete solution will probably require a paradigm change in thinking about information and in how staff generally work with information (particularly a move away from information contained in separate files). However, the productivity increases and costs savings from such a change could be immense and are probably well overdue.

5. Summary

More than ever before organisations have to work with information, it is like blood circulating around the organisation, and without it most organisations would certainly die. This paper has discussed a number of the challenges facing organisations with regards to information creation, access, and management (amongst other things). In particular, the paper considered the need to lower barriers to information (inside and outside) an organisation, the need to provide better access and version control, to provide better support for collaboration, consistent productivity tool functionality, powerful search capabilities, and to reduce the maintenance costs of information with organisations.

Of course, these information management challenges are not new, and organisations have been looking for solutions to the information challenges for quite a while. This paper has discussed a number of the most common solutions available today; including the somewhat default solution of file servers, the more modern Web servers, content management systems, document management systems, and the recently popularised wiki webs. All of these solutions address some of the challenges of information management within an organisation, but are often lacking in other areas, i.e. they provide only a partial solution.

Finally, this paper discussed what a more complete solution may be like, and particularly what would be the characteristics of such a solution. Although this more complete solution doesn't seem to exist yet, it does not seem out of reach of current software technology. It will, however, require a change to the paradigm in which people work with information inside organisations.

6. References

[1] Browning, P. & Lowndes, M. 2001, "Content Management Systems: Who needs them?" in *Adriane*. Joint Information Systems Committee (JISC).

[2] Doupnik, A. 2002, "An overview of electronic document management system product offerings", *Topics in Health Information Management*, vol. 23, no. 1, pp. 62-73.

[3] Leuf, B. & Cunningham, W. 2001, The Wiki Way: Collaboration and Sharing on the Internet, Addison-Wesley Professional.

[4] Schwall, J. 2003, The wiki phenomenon. From http://www.schwall.de/.

[5] Wikipedia, http://www.wikipedia.org/