



Unstructured data: A management overview

By
Tony Pfitzner & Tyson Lloyd Thwaites
Allianza Services Pty Ltd

Abstract

Today we live in the 'Information Age.' Due to the explosion of the Internet, and the reducing costs of disk storage, overwhelming amounts of information are being stored by today's businesses. While there is no lack of data available, what are sorely missing are the tools and methods to manage this unstructured sea of facts and turn it into usable knowledge.

Unstructured data includes such things as documents, spreadsheets, presentations, multimedia files and emails and may represent 45% of all network data. In some organizations it may be as high as 85% (according to Merrill Lynch). This data includes a significant part of the knowledge base of an organization and needs to be properly managed if it is to be of long-term use.

This white paper looks at the issues surrounding Information Lifecycle Management (ILM) and outlines a practical and powerful approach to the problem of managing unstructured data.

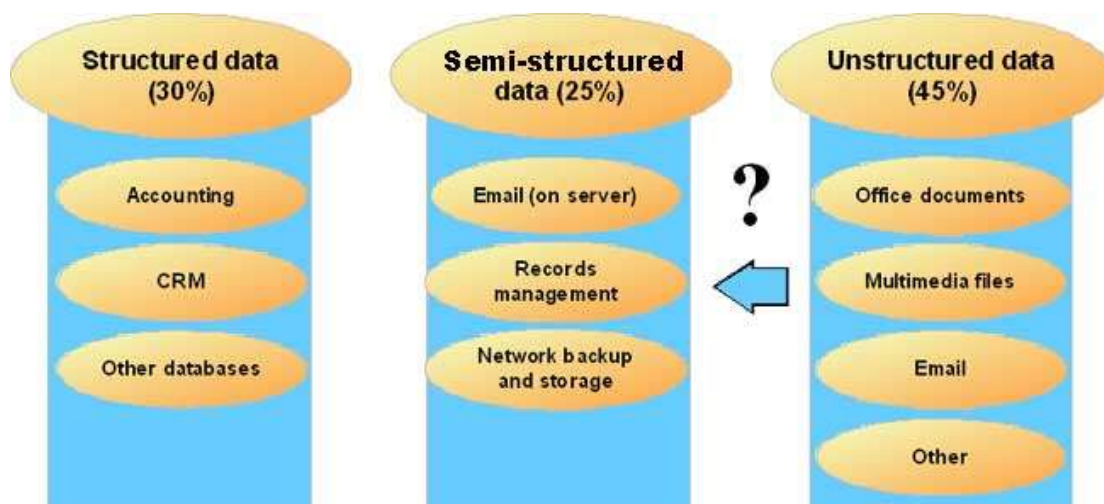
The problem of unstructured data

The information stored by an organisation can be categorised into two main types: structured, and unstructured. 'Structured data' refers to information that is formally stored in some kind of database or records management system. Each piece of data is assigned a specific format and has known significance.

'Unstructured data' refers generally to all the information that is not stored or indexed in any formal information system. The most common forms of unstructured data are text-based files, such as Microsoft Office documents, HTML, and other text documents.

'Semi-structured data' is a transition phase between the two. The information is indexed, for example by a records management system, but the content itself has no formal structure applied.

Unstructured data typically accounts for a staggering 85% of an organization's knowledge base. However, unlike structured data, there is no simple and reliable way to access, analyse, or search this information. The result, according to Mani Shabrang of Dow Chemical Co, is that "We are drowning in information but are starving for knowledge."



Levels of data structure. Few tools exist to move unstructured data into a more structured environment.

Cost, risk and lost opportunity

According to Gartner projections, approximately 30-40% of white-collar workers time is spent managing documents. Apart from the time wasted by these knowledge workers in retrieving information from the network, there is also the cost of having to recreate documents and the loss of intellectual property that cannot be located. This latter loss of information is sometimes referred to as a 'loss of corporate memory'.

In addition to this loss of time, there are at least three other areas that are impacted by ineffective management of unstructured data, and result in either increased costs or risk exposure. These include:

- Compliance and legal risk management.
- Cost of storage.
- Losses arising from a failure to efficiently utilize intellectual property

Network storage costs are increased when data that should have reached the end of its lifecycle, is kept on high-value drives. Examples include documents that have not been accessed for years, duplicate files, runaway log files, temp files, unauthorized multimedia files etc. The cost of storage is not just the cost of the storage infrastructure of servers, drives, and tape and optical backup, but the considerable administrative time that is devoted to managing this problem.

The solution

The problem of unstructured data has been recognised as one of the largest problems facing information managers today. One of the first challenges is raising awareness of the problem within an organisation. Many people are not aware that some of their most valuable data is spread across the unstructured network, or are not aware that anything can be done to streamline and organise this data.

Despite the growing profile of this issue, there is still a lack of tools available to effectively manage the problem. Many of the tools available are developed by storage vendors, and thus have a focus on how to get the unstructured data into a managed storage area. While this addresses the immediate problem of preventing data loss and optimising storage space, it does very little to help analyse and optimise the actual content.

The basic requirements for managing unstructured data are:

- The ability to analyse, categorise and report on unstructured data
- The ability to identify intellectual property and reclaim wasted storage space
- The ability archive files, or promote them to records management and other systems, according to user-defined rules
- The ability to monitor file usage and enforce company network policies

While many solutions are aimed towards a very specific subset of these requirements, such as optimisation of storage, the successful solution must address all of these issues at some level.

It is also important that, wherever possible, the solution should integrate with the organisation's existing ILM systems and policies, rather than trying to replace them with another solution.

Allianza Discovery

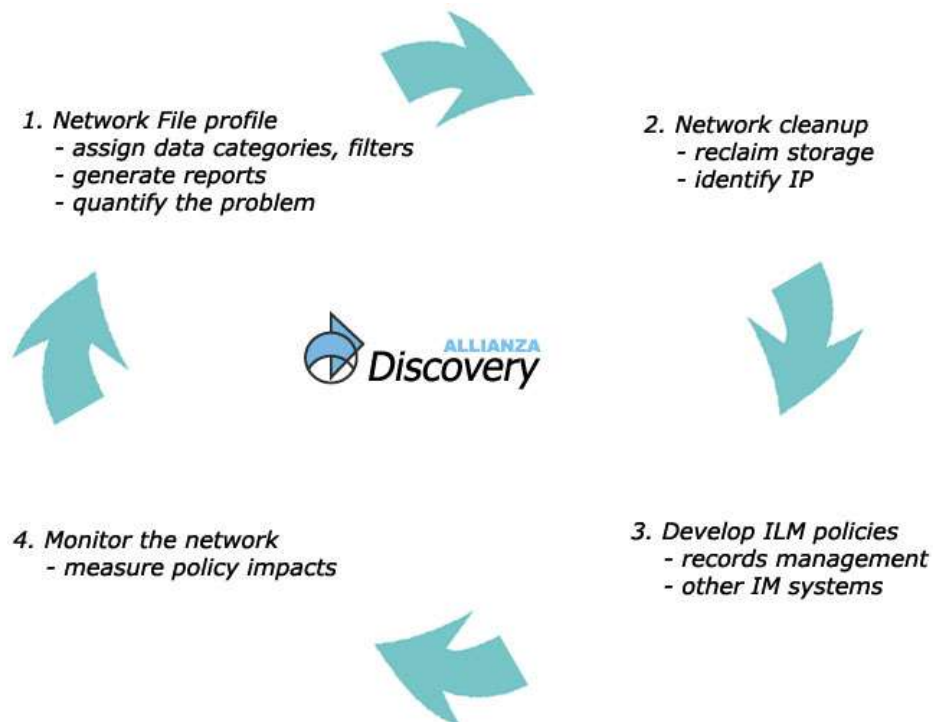
Allianza Discovery is the latest offering from Allianza Services Pty Ltd, an Adelaide-based technology company specialising in Information Lifecycle Management. It is a flexible platform that provides a powerful framework for managing unstructured data.

Allianza Discovery uses a distributed architecture to gather exhaustive information about the network, with continuous monitoring to ensure up-to-date information. Detailed reports provide System Administrators with accurate and insightful statistics, helping them to analyse current file usage trends and facilitate storage planning.

The solution also includes easy-to-use tools to analyse the unstructured data according to user-defined rules, and provides the ability to perform actions on files. This could mean moving important files into a records management system, and nominating other files for movement to lower cost storage, archiving, or deletion.

Allianza Discovery is designed to evolve to suit the needs of each specific organisation. Users can define policies on file naming and location, and can tune the wasted-space detection algorithms to line up with company IT policies. By using an extensible architecture based on best-of-breed technology, custom functionality can be easily incorporated to address the exact needs of the corporate environment.

Allianza Discovery is not simply another document library or storage optimisation tool. By providing a platform to manage unstructured data, the organisation is able to grow and manage its network with confidence, both now and in the future.



A management plan. Allianza Discovery integrates with the existing ILM systems and policies.

Future

The 'Holy Grail' of Information Lifecycle Management is the complete automation of the entire unstructured data management process.

The ideal system will monitor the network, automatically enforcing policies on file naming and storage availability based on how valuable the content is. Intelligent analysis tools will suggest which files should be imported into structured data systems, and which should be downgraded to low-cost storage or deleted.

Enterprise-wide full text searching will also be a key to harnessing the full value of the knowledge contained in an organisation's unstructured data. The search process should be customisable to allow it to intelligently search each unique corporate environment.

Contact us

For more information, please visit our website at <http://www.allianza.com.au>.

To organise a demonstration of our software, or for all general enquiries or questions, please do not hesitate to contact us:

Allianza Services Pty Ltd
Level 1, 122 Pirie Street
Adelaide SA 5000
Ph: +61 8 82325525
Fax: +61 8 82328252
Email: enquiries@allianza.com.au