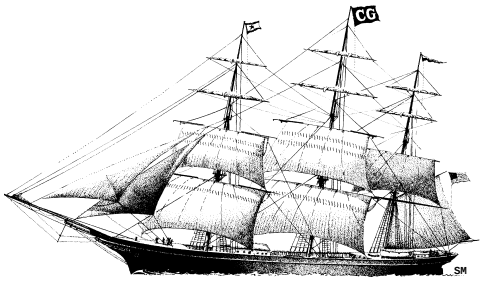
THE CLIPPER GROUP
Navigator™



Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2006065 July 26, 2006

Civilizing the Information Environment — Kazeon Starts with Indexing

Analyst: Anne MacFarland

Management Summary

It is easy to fall into nostalgia for the days when we seemed to know more. Of course, most of us had access to far fewer sources of information, and our knowledge was actually pretty limited – but we were less prone to embarrassing exposures of our ignorance. We how have much more about which we can be informed. The information to run a business of any size far exceeds anyone's ability to remember – which is what information systems are for. However, when it comes to information stored in files, what is called *unstructured data*, we are still substantially disadvantaged. File structures are created by applications and individuals for data's immediate operational use, and perhaps reuse, but not for reference. Machine-generated information organization is often incomprehensible to people unless they are very familiar with the application that generated the information. Even the naming conventions of file systems tend to be subjective. **Information systems have become an obstruction to the knowledge of information that underlies its proper use. With automated indexing, it can also be the solution.**

Efficient use of information, like any other operation, starts with an accurate and detailed knowledge of what you have to work with. If the assets are electronic unstructured data, knowledge of the capacities of storage used by each department, while interesting to the storage administrator, does not promote better information use. It is indexing that can tell you more. Indexing of file system information is a lightweight, non-intrusive process that, in some cases, gives enough metadata, or information about information. In many cases, you will expose more business value, and identify confidential information to be protected, by analyzing and indexing the content.

Once you have discovered, through indexing, the characteristics of what you have, you can sort it by search (to determine what all consists of), or divide it into categories via classification. Classification for data management will change as the organization and its addressable markets change. With indexing, you have a set of data attributes that can be used and re-used over time in different classification schemes. While the characterization in some areas will grow insufficient (think of the rapid evolution of vocabularies and areas of interest in a field like Life Sciences), the basic characteristics can still identify the information that deserves a more detailed indexing to address new concerns. If you begin by assigning categories, you are stuck with the assumptions (and limitations) of your first efforts.

Indexing underlies the functionalities of search, file recovery, information grooming, legal discovery, regulatory compliance, data governance, and intelligent information lifecycle management. Kazeon, a company based in Mountain View CA, has for some time leveraged its automated indexing expertise to offer an Information Server to promote the intelligent use of electronic data that underlies most businesses today. For more details, please read on.

IN THIS ISSUE

Why Indexing is the Place to Start

Databases have their own finding aids. Most enterprise files do not. These unstructured sources of information are the documentation of the enterprise and a significant source of insight about it. Due to litigation, they cannot be ignored. With the pace of innovation in competitive markets, eschewing the insights they can provide is downright stupid.

Indexing may be done to data in place, working non-disruptively from file system information, and then attributing the files (via XML) with characteristics that support intelligent IT and business information strategies. Easily accessed file system information exposes where the information came from, what supersets of information it is part of, and, perhaps, its relevance. Textual analysis, properly done, yields up not just key words, but their context – their affiliation, by analyzing their placement in a sentence with other concepts. Indexing of files can also be done as part of their ingestion into a repository. This may be the time to do more detailed indexing to expose additional metadata elements to ready information for reference use. If the content indexing is done by clustered appliances, the impact on server and storage systems is minimal.

Indexing is the Basis of Search

Search engines work as a two-stage process[1]. The first stage of search is to crawl through the target information with small programs, sometimes called spiders that index it. In the second, time-sensitive, query stage, they use these indexes to get the quick turnaround that users demand. Indexing enterprise information is not as straightforward as indexing Web information. Optimizing for relevance is not as important as fully documenting all the file data that exists. File system information, and, if needed, text parsing are used. If enterprise-naming conventions are disciplined, the amount of useful information that can be easily gleaned increases.

Kazeon's IP (and patents) lie in parsing file information, and also in cracking and parsing content. Earlier this year, Google chose Kazeon's *Information Server* technology to provide the advanced search of information stored on enterprise systems. Network Appliance has, for some time, used Kazeon to power its *SnapSearch* capabilities.

Indexing Promotes Intelligent Information Lifecycle Management

Indexing also adds considerable intelligence to information lifecycle management. Tiers of storage of various vintages are inevitable in a data center, and using older, less performant storage to house information that is infrequently used is a great idea. However, identifying infrequently-used information by where it is located is unnecessarily complex, when you can have indexed information and use search and classification to organize it. Then files can be replicated, migrated, retained, or deleted according to their attributes, not by their physical location. Replication and migration processes can be used more sparingly, and their targets more precisely defined. Encryption can be used with the information that deserves it – but not with adjacent files that don't. Equally importantly, the multiple copies of files that exist in today's systems – a source of confusion and, frequently, security risk – can be brought under control.

Indexing Underlies Faster, Better, File Recovery.

Recovery of a particular file is a pain because it is based either on location (which may not be obvious) or on parsing an entire environment, which is tedious. If the file has been automatically indexed, a search will uncover all the versions of a file – including the versions to which pointers have been deleted by user *delete* operations. It is then easy for the end user to identify the needed file, recover it, and save it properly.

Kazeon's *Information Server* has always worked with EMC's *Centera*. Now Network Appliance has chosen to integrate Kazeon's technology with its *SnapVault* software. SnapVault replication keeps files in their native formats, as opposed to "backup" formats. This allows easy file recovery from backup files. It also makes indexing – and the re-indexing that may be needed in the legal discover process – easier.

[1] For a more detailed discussion of search, see **The Clipper Group Explorer** dated July 28, 2005, entitled *Enterprise Search Adds a User Dimension to Business Information Organization*, available at http://www.clipper.com/reserach/TCG2005048.pdf.

Indexing Facilitates Legal Discovery, Regulatory Compliance and Data Governance

These three issues (or, for many, pain points) relate to the obligations to larger societal forces that business operations entail.

1. Government regulations require that key kinds of data for publicly-traded companies and also in some industries be retained for some years.
2. Legal discovery demands that information requested should be found promptly.
3. Data governance is a set of policies aimed at reducing institutional risk of exposing customer information, product information, operational information or any kind of confidential information inappropriately.

In order to achieve all of these mandates more easily, the knowledge of what you hold, given by indexing and classification, is indispensable.

What you build on top of indexing can be considerable. Indexing makes the development of retention policies to fulfill government regulations more nimble. It is a far less expensive approach than migrating data in an emulation of the old days of file cabinets. Indexing may be used – repeatedly - in complying with legal discovery subpoenas. First, it gives a way to identify all files that are potentially relevant to an inquiry, winnowing by date, organizational domain, and parties of interest. Then, the ability to index content can reduce that bulk to a more manageable yet complete set of files that can then be put on retention hold.[2] Indexing also underlies file system data governance by allowing you to sort for particular kinds of confidentiality.

More important over the long term is the higher business value. Indexing, and the search and classification that indexing enables, can expose the extents and commonalities of enterprise information, revealing non-obvious organizational structures that otherwise would remain hidden.

Conclusion

The effectiveness of the use of information, or any other asset, depends on how much you know about it. In this time of data glut, indexing is a prerequisite to getting the right stuff – and no extraneous stuff – with which to make a new product offering or a business decision.

With indexing as a basis, search is possible, classification is a rational and repeatable process, IT administrative routines become better targeted and less needlessly complex. If you are looking for a way to use your information better, look at what Kazeon has to offer.

[2] Kazeon's partnership with Decru allows encryption and security to be imposed on all files identified in a legal discovery case.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Data Strategies and Information Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial "128" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** *Captain's Log*, **The Clipper Group Voyager**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, "*Navigating Information Technology Horizons*", and "*teraproductivity*" are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.