

American Laboratory

March 2007

Volume 39, Number 6

www.americanlaboratory.com



AFM/Raman Opens New Horizons for Research and Industrial Characterization

Multistep Thermal Characterization of Polymers Using GC-MS

Improving Refinery Laboratory Productivity 20% While Increasing Accuracy With LIMS

Semantic Searching Comes to R&D: The Importance of Metadata

by Joe Peden

Today's business and research communities are faced with the stark reality that the velocity with which data are being created will continue to accelerate at exponential rates over the next decade. It has become vital for organizations to identify logical content relationships among the vast amount of data. Organizations are realizing that in order for their teams to work collectively and not as groups of individuals, they must create an automated knowledge management environment that fosters sharing and collaboration. Semantic searching is seen as the Holy Grail, but if semantic searching is to become a reality, metadata tagging will have to be exploited in a more sophisticated, automated manner than that used by organizations today. It starts with addressing the need to dynamically add metadata to all files. One of the reasons metadata are receiving such attention is their role in facilitating improved information seeking, i.e., business searching.

The three levels of enterprise collaboration that metadata directly impact are shown in Figure 1. The first, and most basic, tier has very little sharing and collaboration capabilities. Keyword searching is mostly used for internal information, with only a small percentage of files indexed and

virtually none of the files with any consistent layers of metadata applied. Without these two key aspects of infrastructure—indexing and metadata—the level of “content intelligence” does not impart any real added value to existing information.

The second tier of enterprise collaboration accessibility begins to offer knowledge management value to the end user. The business goal of providing improved knowledge management, however, is unobtainable without extensive user involvement. This level supplies basic data relationships, but, without useful and relevant metadata, it is not possible to offer semantic and multifaceted (multicomponent) searching and filtering.

The goal is to provide activity-oriented semantic queries that are based on rich metadata that have been fully indexed with multiple layers of domain-specific metadata. This has not been feasible until recently, because organizations only focused on indexing and searching for structured data stored in one of their many company databases. The reality is that almost 75% of R&D data are unstructured. Thus, the majority of data are unable to help a company create an enterprise knowledge management platform.

What are metadata?

Metadata are structured units of information that describe, explain, locate, or otherwise make it easier to retrieve, use, or manage an information resource. Metadata are often called data about

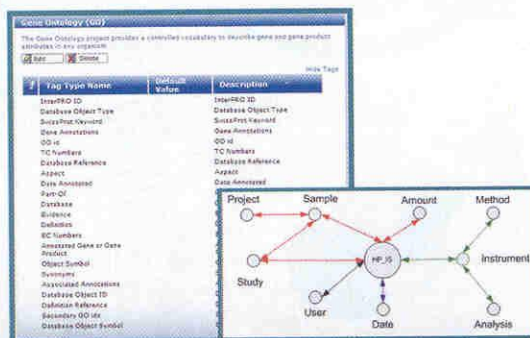


Figure 2 Standardized metadata ontologies create organizational structure for increased interrelationships of content.

data or information about information. The need for metadata is not, however, limited to the laboratory. For example, “12345” are data, and with no additional context have little meaning. When “12345” are given a meaningful name (metadata) or “ZIP code” value, one can understand (at least in the U.S.) that, by placing “ZIP code” within the context of a postal address, “12345” refer to the General Electric plant in Schenectady, NY. Traditional information retrieval (IR) technology has been based almost purely on the occurrence of words in documents. Some browser search engines can even augment this in the context of the World Wide Web with information about the hyperlink structure of the Web. Thus, we begin to see the power of metadata.

In fact, many of us already exploit metadata on a daily basis. Consider the songs on an iPod or MP3 player. The user does not search or organize the music by the words in the song or even by the file name, but rather by metadata (ID3-tags) such as title, artist, and genre. The gene ontology example (see Figure 2) shows the types of metadata available to improve knowledge management and the ability to search for content. Metadata offer multiple ways to relate the vari-

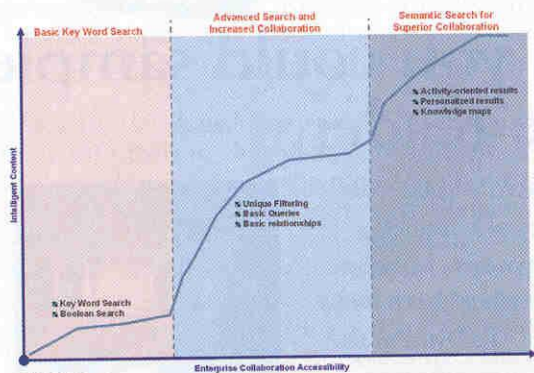


Figure 1 Three levels of enterprise collaboration accessibility.

Tag Names		Tag Values	
Internal Standard Report			
Data File Name	A:\75394-12.D	Page Number	1
Operator	MSM	Vial Number	1
Instrument	5890-25	Injection Number	1
Sample Name	12	Sequence Line	1
Run Time Bar Code		Instrument Method	QXYPO.MTH
Acquired on	14 Jun 93 10:04 AM	Analysis Method	HP_CHEM.MTH
Report Created on	09 Jul 93 11:31 AM	Sample Amount	12.05
Last Recalib on	26 FEB 93 09:05 AM	ISTD Amount	0.074
Multiplier	10		

Ref Time	Area	Type	width	Ref#	Amount %	Name
2.270	205	PB	0.047	1	0.0239	VC1
2.720	2572	BV	0.043	1	0.300	C2H5CL
3.781	85	PV	0.063	1	0.0134	CLM2
4.291	183	VV	0.055	1	0.0307	1,2-DCET
4.543	524	PV	0.063	1	0.0862	1,1-DCEa
5.122	508	BV	0.053	1	0.0824	1,2-DCEC
5.326	832	PB	0.062	1	0.153	CLM3
6.476	2311	BV	0.078	1	0.489	CLM4
7.578	629	PV	0.072	1	0.128	CHLORAL
7.989	1896	BV	0.069	1	0.387	C2H4CLBR
9.066	24285	BV	0.075	1	4.025	1,1,2-TCEa
12.862	856	BV	0.067	1	0.160	1,1,2,2-TRCEa
17.228	51148	PV	0.075	1-IR	6.143	*** N C11H24 ***

Time Reference Peak	Expected RT	Actual RT	Difference
13	17.210	17.228	0.1%

Figure 3 Leveraging the metadata from an unstructured instrument data output begins the layering process.

ous file content to a multitude of facets, as displayed in the interrelationship chart in Figure 2.

Metadata in R&D

Consider an output file from an Agilent 5890 gas chromatograph (Agilent Technologies, Santa Clara, CA) (see Figure 3). The file layout is representative of many of the files produced by laboratory instrumentation. The results section contains and displays the analytical results for the sample analyzed. The user finds information such as retention time, type of peak recorded on the chromatogram, relative amount of each compound, and name of the compound producing the specific peak.

All of these data are meaningless unless they can be put into context. The researcher must be able to associate these experimental data with the actual sample that was injected into the chromatogram that produced the results. The metadata for the report are found in the header section. The metadata show the sample identity, identity of the user of the instrument, instrument used, and instrumental method followed to produce the results. All of this information provides a framework on which the recorded results are based, but since these data are not normally accessible to search tools (because the file is unstructured), users can only locate the file if

they know the file name. Unfortunately, the file's metadata header does not store any information about the study or

the project the study relates to. Thus, in addition to the metadata that could be obtained natively from this instrument file, there is also a need to apply additional related metadata. Attaching unique metadata is the only way to add context and relevancy to such unstructured data files. Figure 4 is an example of how to associate additional metadata values as defaults for each file when created from their native data source. The key is that, together, these tags now enable activity-oriented searching.

Today's solution

The ideal solution would be one that supports full text indexing and the capability to search by file names or any string or word contained within them, even if such words are not marked as keywords. It must also enable the organization to search through structured and unstructured files, particularly binary files, not just by file name, but also by all of the associated file metadata. The solution should also support easy parsing of meta-

EPA Method 1664A

Results you can take to the bank.

Horizon Technology offers a fully automated alternative to expensive, labor intensive LLE methods for extracting oil & grease from aqueous samples. Using a combination of advanced liquid handling and Solid Phase Extraction (SPE) disks, the Spe-Dex® 3000 Extraction System delivers the accuracy and consistency of data laboratories need to comply with Method 1664A QC specifications. And at costs below traditional Freon 113 methods. To discover how the Spe-Dex 3000 Automated Extraction System can help your lab meet its business objectives through lower operating costs, higher sample throughput and reduced solvent usage, visit www.horizontechnic.com or email sales@horizontechnic.com.



www.horizontechnic.com

Web access AL20.com?2448

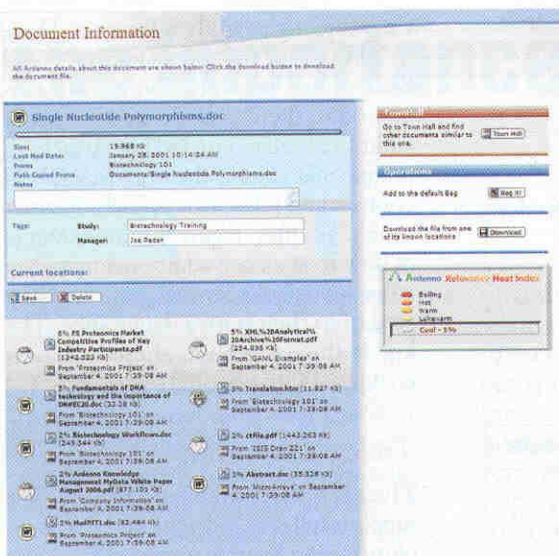


Figure 4 Access to the metadata makes researchers much more efficient.

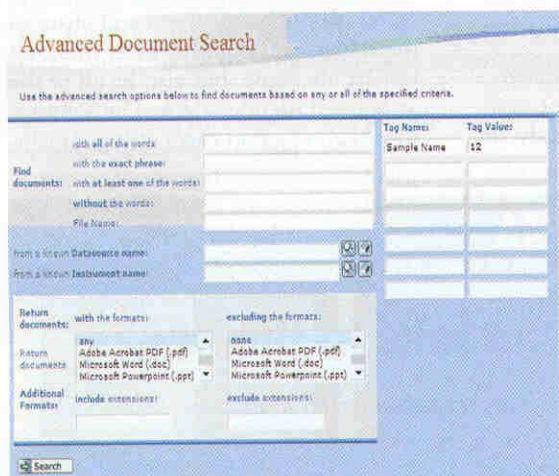


Figure 5 Extensive searching using the layers of metadata offers new power to the researcher.

data from files and or style sheets, so that metadata tags produced by other systems can be automatically associated and leveraged. The solution must permit the attachment of file management tags to files, such as folder names and paths, and allow for the attachment of additional metadata tags not stored in the file or from the file management system to be attached to files. Finally, all of this must be done automatically, in the background, without the need for human intervention. Making this a task for an individual researcher will doom the efforts to failure.

The following section explains how this is possible. Consider the Agilent 5890

the user to browse on the metadata tag names and a value for that tag, it would provide a much more structured search, without the parsing overhead and ongoing maintenance issues and costs.

Semantic searching

Metadata give activity-oriented semantic searching the data context and content relevancy needed to display pertinent and related files in a multidimensional manner. Figure 6 shows that, when a file is selected, it can be presented at the center of a graphic that displays relevant files related to the selected file. The graphic is unique because the closer the files are displayed

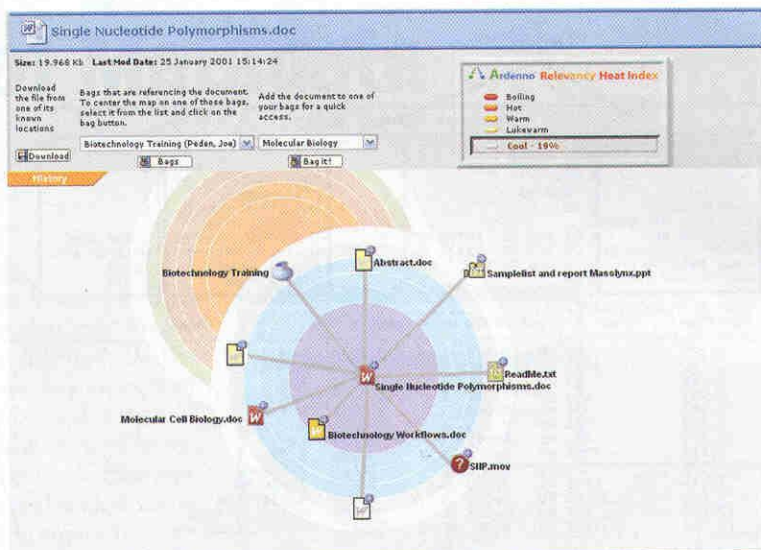


Figure 6 Viewing search results in a multidimensional manner offers more usefulness than can be viewed in a traditional "flat" data list.

gas chromatograph (file output shown in Figure 3). By applying metadata about its sample content, we are now able to search through the unstructured file to locate the string of "sample name equals 12" (see Figure 5) or the project, or study name, that was associated to the file's metadata. The exploitation of the file's metadata opens up an entirely new world to the understanding of data relevance and enables activity-oriented semantic searching. One might reconsider the continuing value of parsing files at all. If the system allows

to the central file, the more relevant are their content and metadata tags. This would not be possible without the application of metadata.

The icons of the various files are also displayed in different colors. This is not due to the format of those files, but rather to their relevancy heat index. The more frequently a file is downloaded and opened, the more intense is its relevancy heat index color. This color-coding is applied to each file and provides a quick visual cue of a file's importance. In a world of exploding data volumes, anything that helps users focus their efforts saves time.

Conclusion

The future of improved collaboration and knowledge management will be based on an organization's ability to maximize its use and creativity around a file's metadata. Only those organizations that devise a way to automate the process will succeed. Those that automate the metadata process will be able to persuade the user community to accept and leverage the available metadata. Semantic searching is the Holy Grail for business searching, and it is all possible today.

Mr. Peden is President, **Ardenno Solutions, Inc.**, 1101 Beverly Dr., Garnet Valley, PA 19001, U.S.A.; tel.: 610-203-8475; U.K. tel./fax: +44 161 973 0857; e-mail: joe.peden@ardenno.com.