SANDIA REPORT
SAND2014-2498
Unlimited Release
Printed March 2014

Information Findability:
An Informal Study to Explore Options for Improving Information Findability for the Systems Analysis Group

Nora K. Stoecker

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2014-2498
Unlimited Release
Printed March 2014

Information Findability:
An Informal Study to Explore Options for Improving Information Findability for the Systems Analysis Group

Nora Kathleen Stoecker
Policy and Decision Analytics Department

Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-0421

Abstract

A Systems Analysis Group has existed at Sandia National Laboratories since at least the mid-1950s. Much of the group's work output (reports, briefing documents, and other materials) has been retained, along with large numbers of related documents. Over time the collection has grown to hundreds of thousands of unstructured documents in many formats contained in one or more of several different shared drives or SharePoint sites, with perhaps five percent of the collection still existing in print format.

This presents a challenge. How can the group effectively find, manage, and build on information contained somewhere within such a large set of unstructured documents? In response, a project was initiated to identify tools that would be able to meet this challenge. This report documents the results found and recommendations made as of August 2013.

ACKNOWLEDGMENTS

Many people shared information during this exploration of the issue of "information findability" and the effort to identify tools that could help staff effectively find, manage, and build on prior work that is so easily lost in mountains of unstructured content stored in multiple archives.

Thank you to Keith Almquist, George Backus, Travis Bauer, Lozanne Chavez, Rick Craft, Tim Eriksson, Mark Foehse, Heidi Herrera, Jerry Johnson, John Mareda, Judy Neff, Nancy Orlando-Gay, Thor Osborn, Jon Rogers, Steve Schafer, Jason Shepherd, Wendy Shaneyfelt, Judy Spomer, Brian Vanover, and Don Waye.

Contents

Tables

This page intentionally left blank

1. INTRODUCTION

Consider it an embarrassment of riches: for the contemporary enterprise, progress is not inhibited by a lack of information but by a lack of easy access to that information.
—Daniela Barbosa, Dow Jones, 2008

A Systems Analysis Group, sometimes called Systems Studies, has existed at Sandia National Laboratories (Sandia) since at least the mid-1950s. The group's major purpose has been and still is to provide in-depth, objective evaluations of technical and business issues to support the leadership of the Lab, its U.S. Department of Energy (DOE) customer (and the predecessor organizations), and other government agencies as requested. This study focused on the Systems Analysis Group based in Albuquerque, New Mexico.

Much of the group's work output (reports, briefing documents, and other materials) has been retained, along with large numbers of related documents. Over time the collection has grown to hundreds of thousands of unclassified documents *that we know about*,[1] most of them in unstructured form. The set of documents includes final reports, multiple drafts, and supporting reference documents and other data found as Word documents, text files, PowerPoint slides, PDF files, Excel spreadsheets, and other formats.

Most of the collection is stored in electronic form in thousands of folders contained in one or more of several different shared drives or SharePoint sites, with perhaps five percent of the collection still existing in print format. The collection of classified documents is similarly large, unstructured, and primarily stored in electronic form.

Many staff members in the Systems Analysis Group find it difficult to effectively search for and find relevant information from within this large and growing collection of documents. They often rely on the memories of long-term staff members to point them to relevant information, however even these staff members admit to difficulty in quickly and easily finding needed information from within the various archives. In addition, as long-term staff members leave the department or retire from Sandia, they are no longer a resource to new staff members.

This presents a challenge. How can the group effectively find, manage, and build on information contained somewhere within such a large set of unstructured documents? In response, a project was initiated to identify tools that would be able to meet this challenge. This report documents the results found and recommendations made as of August 2013.

[1] For purposes of this study, the document collections are those found on shared network drives, the 0240 Library SharePoint sites, and print documents in the 0240 library collection. An unknown number of other documents reside on personal hard drives, team SharePoint sites, email accounts, and other locations.

2. OBJECTIVE AND METHODOLOGY

2.1 Objective

The main objective of this study was to identify tools that would enhance "information findability", or the ability of the Systems Analysis Group to find, manage, and build on prior work as represented by information contained in the hundreds of thousands of unstructured documents currently stored on department-wide shared drives and in the Group Library SharePoint sites.

Related to the objective was a desire to find appropriate tools that were either in use at Sandia already, or that could build on related Sandia applications, ideally with enterprise support.

2.2 Methodology

A literature search on related subjects was conducted and an effort was made to identify Sandia information findability efforts, either enterprise-wide or Sandia pockets of effort, through informal discussion with other Sandians. Results of both were used to identify potentially relevant tools, to formulate a general assessment of each, and to identify other relevant issues.

2.3 Criteria

The "ideal" system would:

1. Search across multiple sources of content, for example
    a. Local drives and shared drives
    b. Outlook, SharePoint, and other sources
    c. All document types
2. Find specific documents and open selected documents in original format.
3. Find documents that are *about* specific subjects and open selected documents in original format.
4. Create sets or collections of documents.
5. Annotate or tag selected documents.
6. Identify themes within any given repository large or small.
7. Summarize any given document, regardless of format.
8. Identify duplicate documents.
9. Work on both the Sandia SRN (Sandia Restricted Network) and the SCN (Sandia Classified Network).
10. Work as a shared resource across the group.
11. Build on applications already in use at Sandia.
12. Integrate well with other Sandia applications.
13. Support additional features useful for conducting thorough analyses and studies.
14. Support knowledge transfer.
15. Be relatively easy for all staff to use.

No one tool was found that would meet all of the desired criteria.

3. SANDIA EFFORTS

An effort was made to identify enterprise-wide or Sandia pockets of effort related to information findability through informal discussion with other Sandians.

3.1 Enterprise-wide

It became apparent, in discussion with the manager and staff of Sandia's Knowledge Systems department and others, that there is no enterprise-wide effort directed at improving information findability in the sense defined for this study. There also does not appear to be a currently deployed enterprise-wide or enterprise-supported tool or toolset that could be used to enhance information findability.

Sandia's enterprise-wide search engine is IBM's OmniFind. In late 2012 there was an expectation that Sandia would upgrade to OmniFind 9.1, described as having a better query function and a "content analysis" tool. As of mid-2013 the upgrade had not occurred. Sandia may now be testing another IBM product, Content Analytics 3.0, for use as a future enterprise search engine.

3.2 Pockets of Effort

Staff members in the Knowledge Systems department were familiar with one related effort within Sandia, the development of the Citrus toolset.[2] They were not able to identify other specific Sandia pockets of effort directed at information findability.

Through a very informal networking process, twelve Sandians were identified who were either using an existing toolset to find information or were attempting to develop a process or toolset to find information. Several of these Sandians were aware of the Citrus toolset and its developer whether or not they had ever used Citrus. None were aware of other ongoing efforts or tools beyond their own.

3.3 Potentially Relevant Tools

The author learned that at least one other department had embarked on an effort to develop its own application, and another was establishing a wiki to collect and share relevant documents. In addition, nine potentially relevant tools or toolsets, listed below and summarized in Table 1 and Table 2, were identified through discussion with these individual Sandians, or through other Sandia reports. Although there may be other related tools or applications in development at Sandia, no such information surfaced during the admittedly informal series of interviews conducted as part of this study.

The nine tools identified are

- Brainstorm SharePoint site search tool
- Beyond Compare
- Citrus toolset (Clementine, Durian, and PaperMinder)

[2] Citrus consists of a set of text analysis components, including a search function, developed by a team in Sandia's Analytics and Cryptography department. There are users throughout Sandia, but the toolset is not, at this time, being supported as an enterprise-wide application.

- Copernic
- dtSearch
- IBM Content Analytics
- IBM OmniFind (Sandia's enterprise search tool)
- SAS
- SABIAS, a custom search solution proposal from Sandia's Knowledge Systems department.

There are dozens of other potentially useful, commercially available, tools or applications for enhanced search, text mining, and content analysis, as identified through a literature search. However, there was no indication through this preliminary effort that any of these tools was currently being used inside of Sandia or would readily build on applications already in use at Sandia.

Table 1. Summary of Tools Identified through Discussion with Sandians

Product	Description	Assessment
Color-code: Green = high perceived value to Systems Analysis Group information findability; Yellow = some value; White = little perceived value		
Brainstorm search tool	Internal Systems Analysis Group advanced "file search" application purchased as an add-on to SharePoint; launches from within the Brainstorm database. Automatically searches across shared drives and the Systems analysis Group library database, indexes files; saves the index in SharePoint and searches the index.	Searches within or across the Systems Analysis Group library database and shared drives. Does not point to local drives (uncertain whether that would be a capability). Allows for more refined searching than Internet Explorer but cannot organize results or do additional analysis.
Beyond Compare	Tool for comparing content in files and folders. Addressed in a Sandia Networks Grand Challenge LDRD Seminar Series talk, 2009.[3]	Did not test. Specialized focus.
Citrus toolset	Sandia-developed text analysis library and associated applications. Three components were reviewed in this study – Clementine, Durian, and PaperMinder.	Overall, strongest "analytical" capability of the tools reviewed (excluding SAS); meets the most "findability" objectives. Strong Boolean search capability; very strong ability to assess results in various ways but poor ability to manipulate straight search results. Refinement is possible. Used on various projects across Sandia. Related user applications show promise for the Systems Analysis Group's information findability objectives. Could benefit by Sandia IT adopting Citrus as an enterprise-wide toolset; a funding source is needed.
- **Clementine**	Citrus end-user application for basic indexing and search.	Downloaded to and tested on SRN local and shared drives. Downloaded to SCN. Not yet tested on SharePoint or Outlook files.

[3] Beyond Compare was one of two software tools demonstrated during a February 12, 2009 talk delivered as part of the Sandia Networks Grand Challenge LDRD Seminar Series, as announced in an internal Sandia email message dated February 9, 2009. The other tool was dtSearch.

Product	Description	Assessment
		Robust search capability. Achieves search and find objectives on local and shared drives; application-generated term cloud and summary can be selected for each document, but not readily saved. No ability to sort, tag, or easily save results elsewhere; does not provide additional analysis options. A very fast and useful search tool, but related application Durian provides both search and analysis options. However, the Citrus package provides both tools.
- **Durian**	Citrus end-user application for content analysis and "document investigation."	Downloaded to and tested on SRN local and shared drives. Downloaded to SCN. Not yet tested on SharePoint or Outlook files. Robust search capability, although takes more steps than with Clementine. Achieves search and find objectives on local and shared drives. Application-generated term cloud and summary can be selected for each document (but not readily saved). No ability to sort, tag, or to easily save all documents found in a category, though it is possible to save/load the bins, including the full content of the documents. Provides options to cluster results, a step toward the objective to identify themes (but refinement needed). Documents can be tagged at a high level (i.e., "interesting"; "uninteresting"; "review"; other); the application then identifies related documents within the larger set. Essentially no ability to sort results or to easily save them into new locations. Slow response with large sets of files; end-user interface is in need of refinement. Resources (funding and expertise) needed to address problems when running on very large sets of documents. Requires more effort of users than does Clementine, but provides much greater potential for analysis Resources needed to build out some additional capabilities useful for conducting analysis and studies.
- **PaperMinder**	Citrus end-user application to manage, search, and tag documents. Still in beta.	Did not test. Sounds promising, but it is unclear how this application links to the Clementine or Durian applications.

Product	Description	Assessment
		Resources (funding and expertise) needed to build out some additional capabilities useful for conducting analysis and studies.
Copernic Desktop Corporate	Commercially available file indexer and searcher. Has been used in the past by a member of the Systems Analysis Group. Current price: $59.95 per computer. Free trial available.	Claims to index, search, and retrieve unstructured text files – Word, WordPerfect, .txt, PDF, PPT, HTML, email files, other types of files, and the web, from personal drives and networked drives. Sounds similar to dtSearch, which is used by various Sandians. See below.
dtSearch	Used by several analysts across Sandia; addressed in a Sandia Networks Grand Challenge LDRD Seminar Series talk, 2009; included in a review of text mining tools in SAND2005-1219.[4] Also included in a review of COTS (commercial off-the-shelf) tools in SAND2010-7392.[5] Current price: $199 per computer for 1-4 computers; price drops as number of seats increases. Free trial available. There are currently two licenses within the Systems Analysis Group.	Strongest "search and find" capability of the tools reviewed; problems indexing very large filesets. Indexes, searches, and retrieves unstructured text files – Word, WordPerfect, .txt, PDF, PPT, HTML, email files, other types of files, and the web, from personal drives and networked drives. Allows for refined Boolean searching; can search within results for additional refinement; initial results will be sorted by date, hit count, or relevance; can then be sorted by title or any other listed field. "One-click" ability to save selected documents to a new folder, for manual analysis; can also create a report containing search results, for additional, manual, analysis. Software does not do additional analysis. Indexing works well on smaller filesets but times out on very large filesets (i.e., V-drive). Workarounds would involve creating several smaller indexes that could then be searched jointly. A strong search tool for finding documents; very good ability to sort results on various fields and to easily save selected documents into new locations, but lacks analysis capability.
IBM Content Analytics	Originally a content analytics add-on to OmniFind, Sandia's search engine, it is now marketed as Content Analytics 3.0, which includes OmniFind search capability.	Based on training documents – search function is more sophisticated and robust than the current OmniFind; allows for results to be exported. It is less intuitive to use and requires administrator permissions for many features. If similar to current OmniFind will not easily be pointed to local archives and drives if at all. Content analysis features seem intended for business intelligence, customer sentiment analysis, trend analysis, and other types of

[4] Chapman, Leon, Rossitza Homan, Jim Treadwell, et al. *Knowledge Discovery and Data Mining (KDDM) Survey Report*, SAND2005-1219, February 2005.
[5] Trahan, Michael and Mark Foehse. *A Toolkit for Detecting Technical Surprise*, SAND2010-7392, October 2010.

Product	Description	Assessment
		analysis similar to SAS.
IBM OmniFind	Sandia's corporate search engine; searches Techweb index, SRN/SON web content, and corporate policies, processes and procedures, and can search the Technical Library Catalog and FileNet.	"Search" only; does not support sophisticated search strategies; results cannot be easily organized or analyzed. Cannot easily be customized; uncertain if it can be pointed to specific local (departmental) shared drives and other information repositories.
SAS	Statistical analysis software with a text analytics module (TextMiner), typically used for identification of patterns and trends from within extremely large sets of files, i.e., "millions of pages." Pricing varies depending on various factors, but ballpark estimate is $10,000+ per license, plus possibly, annual renewal or maintenance.	Pockets of users throughout Sandia; the Systems Analysis Group is developing the capability. SAS is a powerful tool primarily for data analysis; identification of patterns and trends. Expensive difficult to learn; not suitable for individual desktop deployment across a department because of cost and learning curve. The Systems Analysis Group is building its text analytics capabilities; expanding staff "SAS-proficiency" is desirable although this tool is not recommended for the information findability objectives.
SABIAS (Systems Application for Browsing, Information Analysis, and Search). Custom Search Solution Proposal	Sandia's Knowledge Systems department, offering "enterprise level technical solutions for information management, information analytics, and search", proposed the development, maintenance, and technical support of a custom search and document analysis toolset. Proposal is for ~$60K for initial browsing/ collections/ search; ~$140K for other advanced features and tools.	As envisioned it would be built with the Lucene core package,[6] using other APIs or plug-ins, and will be written using Java. It would index documents from various locations, provide robust search capabilities, and include several desired information analysis tools. Additional development would be intended for advanced tools for analysts. This is a proposal to develop a completely new toolset. Because of similarities to the existing Citrus toolset, it would seem more effective to refine Citrus and add additional capabilities to it. The Knowledge Systems department staff may be able to do that.

[6] Apache Lucene is open-source search software; the previously listed Sandia Citrus Toolkit uses Lucene and other open-source software as well.

Table 2. Tools Comparison to Criteria

Product		1 Search Across multiple sources	2 Find Specific Documents	3 Find docs *about* specific subjects	4 Sets of documents	5 Annotate/tag documents	6 Identify themes within repositories	7 Summarize documents	8 Identify duplicates	9 Work on SRN & SCN	10 Work as shared group resource	11 Build on current SNL applications	12 Integrate with current applications	13 Support for rich analysis	14 Support knowledge transfer	15 Easy for all staff to use
	Brainstorm advanced search tool	X	X	?						X	X	X	X			X
	Beyond Compare								X	X						
Citrus Toolset	**Citrus - Clementine**	X	X	X				X		X	?	X	?			X
	Citrus - Durian	X	X	X	X	X	X	X	?	X	?	X	?	X		
	Citrus - PaperMinder					X				X	?	X	?	X		
	Copernic Desktop Corporate	X	X	X	?	?			?	X						X
	dtSearch	X	X	X	X	?			X	X		X				X
	IBM Content Analytics	**Not available to test**														
	IBM OmniFind	?	X									X	X			X
	SAS	X		X			X					X		X		
	SABIAS	**Proposal only; not available to test**														

4. PRELIMINARY ASSESSMENT

Of the nine tools identified through discussion with Sandians, two may offer the most potential value for improving information findability within the Systems Analysis Group. Those two are the Citrus toolset for a combination of search and analysis capability, and dtSearch for a robust search capability but without additional analysis options.

4.1 Descriptions of Nine Tools

Brainstorm file search application – This is a SharePoint add-on, probably Microsoft FAST. It is an advanced search feature that launches from within the Systems Analysis Group's library SharePoint site (Brainstorm). Although it cannot be used to organize results or do additional analysis, it allows for more refined searching than the basic Windows search function. It also can search across the library catalog and across the group's shared drives simultaneously. Although the tool is supported by Systems Analysis Group Information Technology (IT) staff it has not been widely used, possibly because many staff members did not know about it. Efforts are being made to raise awareness about the tool and to make it easy to access. [Assessment based on actual use.]

Beyond Compare – This is a specialized tool that is used to compare content in files and folders. It was addressed in a Sandia Networks Grand Challenge LDRD Seminar Series talk, 2009.[7] At the time Sandia had a site license, but it no longer does. Although the ability to compare content would be useful for identifying duplicates, and for editing multiple versions of a document, it does not seem useful for the broader information findability objective. [Assessment based on a review of the product description.]

Citrus toolset – This is one of the two toolsets recommended for further consideration and development by the Systems Analysis Group. It is a Sandia-developed text analysis library with associated applications. The related user applications are called Clementine, Durian, and PaperMinder. (See descriptions below.)

This toolset has been in development for approximately five years. Although the developers use it in part for ongoing research in the content analytics field, it is also used on various projects across Sandia. The related user applications show promise for the Systems Analysis Group's information findability objectives, though additional refinement would be desirable, as would enterprise support help to ensure stability for core functions and to enhance the end-user interface. There is currently no financial cost to use Citrus, though additional development for Systems Analysis Group-specific requirements would require funding. [Assessment based on actual use.]

> ***Clementine*** – This is a Citrus end-user application for basic indexing and search. It offers a robust search capability but no ability to sort, tag, or easily save results elsewhere; does not provide additional analysis options. It is a very fast and useful search tool, but related

[7] Beyond Compare was one of two software tools demonstrated during a February 12, 2009 talk delivered as part of the Sandia Networks Grand Challenge LDRD Seminar Series, as announced in an internal Sandia email message dated February 9, 2009. The other tool was dtSearch.

application Durian provides both search and analysis options and commercial tool dtSearch makes it easy to process and save results elsewhere. Clementine is more straightforward to use for searching than the Durian search function, and could be enhanced to add additional capabilities. [Assessment based on actual use.]

Durian – This is a Citrus end-user application for content analysis and "document investigation." It offers a robust search capability, although it takes a step or two more than with Clementine. Term clouds and summaries can be automatically generated for each document, facilitating quicker understanding of the content of each, although these summaries cannot be saved for use elsewhere. Documents can be tagged into general categories and can then be used by Durian to identify related documents from within the larger set. There are also options to identify duplicates, and to cluster results, a step toward the interest in identifying themes within large collections of documents, although these features did not work consistently well on large sets of files. [Assessment based on actual use.]

PaperMinder – This Citrus end-user application is still in "beta", intended to manage, search, and tag documents. It is unclear how or if PaperMinder will link to Clementine or Durian. [Assessment based on a review of the product description.]

As stated, the Citrus toolset shows much promise for the Systems Analysis Group's information findability needs. Some problems were experienced during testing, for instance, indexing large sets of files took a long time and it is uncertain how long it will take to update the indexes; clustering efforts frequently failed during the initial testing efforts (with the results being reviewed by the developers and a potential fix recently delivered); additional desired features will need to be developed, which will require resources; and the end-user interface is in need of refinement. Nevertheless, the Citrus toolset offers features not readily available in the other tools and has users across Sandia, an indication of its potential.

Copernic Desktop Corporate – This is a commercially available file indexing and search tool which was recommended by one of the Sandians interviewed and sounds very similar to dtSearch, which is used by several other Sandians. It is designed to index, search, find, and retrieve unstructured text files – Word, WordPerfect, .txt, PDF, PPT, HTML, email files, other types of files, and the web, from personal drives and networked drives. During a trial download, however, the software immediately began to index the drives, without being directed to do so. This seemed undesirable on a computer connected to a number of shared drives so the trial was stopped. A second trial on a personal home computer (with the same "automatic indexing" result) showed that Copernic offers robust Boolean search capabilities, which would improve the group's ability to locate specific documents or documents *about* specific subjects. It did not appear to be quite as strong a tool as dtSearch, however. [Assessment based on actual, but limited, use.]

dtSearch – This is the second of the two toolsets recommended for further consideration and development by the Systems Analysis Group. dtSearch is a commercially available file indexing and search tool which is used by some analysts across Sandia; it was mentioned in a presentation

about "Networks Grand Challenge LDRD", 2009, was included in a review of text mining tools in SAND2005-1219,[8] and was included in a review of COTS tools in SAND2010-7392.[9]

Like Copernic, dtSearch is designed to index, search, find, and retrieve unstructured text files, for instance, Word, WordPerfect, .txt, PDF, PPT, HTML, email files, other types of files, and the web, from personal drives and networked drives. In addition to Boolean searching, it allows the user to search within results for additional refinement; initial results will be sorted by date, hit count, or relevance and can then be sorted by title or any other listed field. There is a "one click" ability to save selected documents to a new folder and users can also create a report containing search results, for additional, albeit manual, analysis.

dtSearch offers the strongest "search and find" capability of the tools reviewed and is fairly easy to use. With dtSearch it is easy to identify specific document locations and to sort and save results; however, it lacks additional sophisticated analysis capability and has problems indexing very large filesets, although there are workarounds. As a search tool, its out-of-the-box capabilities would improve the group's ability to locate specific documents or documents *about* specific subjects, and to quickly save them to new locations for additional analysis, perhaps with Citrus tools. Currently there is no way to build additional capabilities into the tool. [Assessment based on actual use.]

IBM Content Analytics – This is an enterprise search tool. Originally a content analytics add-on to OmniFind, Sandia's current search engine, it is now marketed as Content Analytics 3.0, which includes OmniFind search capability. It is currently being reviewed as a possible upgrade to Sandia's enterprise search function. Based on a review of training documents, its search function is more sophisticated and robust than the current OmniFind, and it allows for results to be exported. But is less intuitive to use, and requires administrator permissions for many features. If it is similar to the current OmniFind, it will not easily be pointed to local archives and drives, if at all. The content analysis features seem intended for business intelligence, customer sentiment analysis, and trend analysis, similar to SAS statistical analysis software, but possibly easier to use.

Based on a very quick review of some introductory training materials, it would not appear useful for the Systems Analysis Group's specific information findability objective, but certainly shows promise for achieving more focused results for enterprise search. [Assessment based on a quick review of introductory training documents.]

IBM OmniFind – This is Sandia's current enterprise search tool. It searches Sandia's Techweb Index, SRN/SON web content, and corporate policies, processes and procedures, and can search the Technical Library Catalog and FileNet. This is a "search" only tool; it does not support sophisticated search strategies and results cannot be easily organized or analyzed. In addition, it cannot be customized and it is unclear if it can be pointed to specific local (departmental) shared

[8] Chapman, Leon, Rossitza Homan, Jim Treadwell, et al. *Knowledge Discovery and Data Mining (KDDM) Survey Report*, SAND2005-1219, February 2005.
[9] Trahan, Michael and Mark Foehse. *A Toolkit for Detecting Technical Surprise*, SAND2010-7392, October 2010.

drives and other information repositories for local use. [Assessment based on actual use and discussion with Sandia staff.]

SAS – SAS is a powerful statistical analysis software package with a text analytics module (TextMiner), typically used for identification of patterns and trends from within extremely large sets of files, that is "millions of pages." Training examples indicate it is intended for business intelligence, customer sentiment analysis, and trend analysis as are many commercial statistical analysis applications, including IBM's much newer Content Analytics tool. However, it is expensive, difficult to learn, and not suitable for individual desktop deployment across a department because of cost and learning curve.

There are pockets of users throughout Sandia, and the Systems Analysis Group is building its text and data analytics capabilities with SAS and other tools. It might be possible to use the SAS text mining module capabilities for analysis of sets of textual data retrieved through other tools, but SAS software itself is not recommended for the group's information findability needs. [Assessment based on efforts to learn and use the software and on discussions with Sandia users.]

SABIAS – This is a proposed "Systems Application for Browsing, Information Analysis, and Search" solution, developed by Sandia Knowledge Systems department staff in response to discussions about text mining capabilities and advanced features for information analysis. As envisioned SABIAS would be built with the Lucene core package,[10] and would be written using Java. It would index documents from various locations, provide robust search capabilities, and include several desired information analysis tools. Additional development would be intended for advanced tools for analysts. The proposal is intriguing, but because of similarities to the existing Citrus toolset, it would seem more effective to build on Citrus. [Assessment based on a presentation and discussion about the SABIAS proposal.]

4.2 Potential Issues

Four potential issues emerged during this search for information findability tools: access, cost, learning curve, and security.

Access – Most of the tools reviewed, including both of the recommended tools, are designed to be downloaded onto individual computers. The software can be used to index shared drives, but the index results are saved to each individual computer. This indicates that each staff member would need to download a copy of the software onto his or her own computer or computers, and would need to run the indexing function and the search and analysis functions individually as well. The initial review and testing did not explore if or how a team-wide or group-wide database of relevant results could be established. Further investigation is needed.

Even if it was possible to have a team- or group-wide resource, Sandia policies, procedures, and practices limit the access that each individual staff member has to both unclassified and classified information.

[10] Apache Lucene is open-source search software; the previously listed Sandia Citrus Toolkit uses Lucene and other open-source software.

No one member of the Systems Analysis Group staff, for example, would be able to search across all the documents in all the files found on the group's shared drives and certainly would not have access to information stored in others' local drives or email files. So no one member of staff would be able to know with certainty that he or she had found all the information relevant to any given problem.

Although it might be beneficial to be able to offer a team- or group-wide location for information findability toolsets, this would not remove the need to adhere to access control limitations, an issue that is unrelated to any specific hardware or software.

Cost – Commercial search tools such as dtSearch and Copernic come with a price of $50 to $200 per license/per computer. The tools cannot be modified to add needed capabilities not already built in, but they do allow for very precise searching and finding of relevant information from within a very large set of documents, email messages, web pages, or other sources. For the cost of per-computer licenses, an investment of IT staff time and resources to explore solutions for indexing and access issues, and an investment of staff time to learn how to use the tools, analysts could add a powerful search capability to each of their computers.

Extremely powerful commercially available analytical tools such as SAS are typically expensive, and/or must be paired with other vendor products. They are generally not suitable for widespread deployment across a department because of the cost and the steep learning curve.

The Sandia-developed Citrus suite of search and content analysis tools is available at no financial cost to all Sandians who choose to download the software, but would require an investment of time to learn how to use it effectively. Funding and other resources would also be required to address additional desired needs. The result could be a very powerful tool for the Systems Analysis Group and other similar Sandia organizations.

Learning Curve – None of the identified tools are as easy to use as existing search tools such as the Microsoft Windows default search option, Sandia's enterprise search, or Google basic search. Those tools, on the other hand, do not enable sophisticated search, findability, and analysis of results.

There is a learning curve involved with effective use of an indexing and search tool like Clementine or dtSearch, or of a content analysis tool like Durian, or of a statistical analysis tool like SAS, in return for the increased capability. The Citrus developers advise that the learning curve for Citrus tools is comparable to Google when the features are comparable. For example, they say, when Citrus is deployed as a web app, in which users search through a web browser and do not get to choose individual indexes, the learning curve is similar to that for Google.

Because no one staff member can search for, find, and analyze all the group's content, every staff member would need to invest time and effort in learning how to use the selected tools, which has proven to be a barrier in the past.

Security – One of the benefits of adding powerful search and content analysis tools is the ability to pull together and analyze content from a disparate, or at least previously unconnected, set of documents, to gain new insights, spot trends, and otherwise build on prior work. One of the related risks, at Sandia, is the possibility of inadvertently creating a classified document by associating otherwise unclassified content. The same risk exists now, whenever content from two or more sources is assembled, although an improved ability to find, compile, and "massage" content could, conceivably, increase the opportunities to do so. Analysts will need to exercise awareness and caution, just as they do now.

5. SUMMARY AND RECOMMENDATIONS

The Systems Analysis Group in Albuquerque has retained its work output over the decades, with a current unclassified collection of hundreds of thousands of documents stored primarily in shared drives, a similarly large classified collection, and uncounted other documents and additional information retained in local hard drives, email files, miscellaneous SharePoint sites, and other locations.

Many Systems Analysis staff members find it difficult to search for and find relevant information from within this large and growing collection of documents. They often rely on the memories of long-term staff members to point them to relevant information, however even these staff members admit to difficulty in quickly and easily finding needed information from within the various archives. In addition, as long-term staff members leave the department or retire from Sandia, they are no longer a viable resource to new staff members. This set of circumstances is not unique to the Systems Analysis Group.

The main objective of this study was to identify tools that would help the Systems Analysis Group effectively find, manage, and build on its prior work, specifically by being able to search for, find, and analyze information contained in this vast number of unstructured documents and other sources. Such tools could also prove useful to other groups at Sandia with similar needs.

Related to the objective was a desire to find appropriate tools that were either in use at Sandia already, or that could build on related Sandia applications, ideally with enterprise support. This would avoid the proliferation of yet another stand-alone application in isolated use within Sandia.

This report identified nine tools or toolsets for further review, although dozens more commercially-available tools and applications exist, and there may well be other efforts at Sandia. No one tool was found that would meet all of the desired criteria. Though there are issues related to access, cost, learning curve, and security, these issues are not unique to the tools under consideration, and do not adversely affect consideration of these tools.

Sandia Knowledge Systems department is exploring IBM's Content Analytics 3.0 as an enterprise search engine, recognition that Sandians need a more robust search capability and the ability to better analyze content. However the enterprise solution might not be customizable to local department archives and local needs, even if it is adopted.

SAS is a powerful statistical analysis package, with a text mining module, in use at Sandia. However its intended purpose, cost, and complexity are barriers to its use for the information findability needs identified in this study. Most other applications in use at Sandia do not offer strong search or content analysis capabilities.

The SharePoint advanced search tool (Brainstorm) does not support content analysis, but it does offer a stronger search capability than the basic Windows search function. As deployed within the Systems Analysis Group it enables staff to search across multiple shared drives and SharePoint sites simultaneously, which is a benefit.

There are two tools that have been used by at least some staff at Sandia for a number of years and that appear to be particularly strong:

- dtSearch, a commercially-available tool, emerged as offering the strongest "search and find" capability of the nine tools reviewed.
- Citrus, a Sandia-developed toolset, emerged as offering the strongest analytical capability and a strong "search" capability.

The following recommendations were made to the group.

At a minimum, staff should take advantage of the Brainstorm advanced search features for better results than they can get by using the standard Windows search function.

If the Systems Analysis Group is interested primarily in improved "search and find" capability, this report recommends that the group partner with other users in the Lab for education and training on how to use dtSearch to the best advantage, and to make licenses available to every staff member in the group, with appropriate IT support for issues identified in section 4.2 of this report.

If the Systems Analysis Group is interested in expanding the ability of its staff to find and then *analyze* relevant information in various ways, this report recommends that the group direct some funding to add selected additional capabilities to the Citrus search function and also to explore options for identifying and adding user-specific analytical capabilities and end-user interface improvements, thus keeping the "search, find, and analyze" capabilities together in one tool. This option would also require training for staff and IT support for issues identified in section 4.2 of this report.

Either or both of the preceding tools, if used by staff, would support the Systems Analysis Group's objective to improve the ability of its staff to find, manage, and build on prior work as represented by information contained in its huge collection of unstructured content. In addition, investment in either or both can, potentially, serve as a catalyst for broader Sandia-wide adoption and support.

The preceding recommendations were made in August 2013, as part of a project to identify "information findability" tools for the Systems Analysis Group. Staff members have been solicited to test the various tools and there is an ongoing effort to provide use cases and training.

A related effort to evaluate and recommend an appropriate standard process for archiving Group work product and making that content accessible is being planned in support of the overarching goal to effectively find, manage, and build on information contained somewhere within very large sets of unstructured documents.

6. GLOSSARY

Annotate – Add descriptive notes to a content record.

Boolean – Method of searching a database or text in which logical expressions (like AND, OR, and NOT) are used to limit and specify the search criteria.

Brainstorm – Name for the Sandia Systems Analysis Group's library SharePoint sites and related search tools.

Cluster – Automatic organization of search results into groups of related content.

Content – Information or data contained in a document.

Content Analysis – Automated methods of deriving useful information from text/textual content. See text analysis/text mining.

Content Record – An individual document within a collection of documents.

COTS – Commercial off-the-shelf products.

Document – Traditionally meant word processed files but now commonly means any file produced by an application (text, charts, graphics, images, or any item that can be individually selected and manipulated).

End-User Interface – System of commands, graphics, or other tools that enable an individual to communicate with the system.

Enterprise – Large organization or the entire organization.

Enterprise Search – Simultaneous, automated search of multiple web pages or other content repositories across an organization or enterprise.

FileNet – Sandia's internal file sharing document management system.

Index/Indexing – Structured compilation of terms (i.e., keywords or taxonomy terms) associated with a document or other body of content; helps users find content corresponding to a chosen term. An index can be created by a human indexer or it can be automatically generated.

Information Findability – The degree to which information contained in huge numbers of unstructured documents stored across a department or across the enterprise can be found, managed, and reused.

Query – Search; a word or string of words entered into the search box of a search engine.

Repository – A virtual or other location where documents or other data are stored and maintained and from which they can be retrieved.

SCN – Sandia classified network (internal).

SON –Sandia open network (external).

SRN – Sandia restricted network (internal).

Structured Content/Structured Files – Documents or information sources that have a predefined or inherent structure or assigned fields (i.e., title, author, abstract, keywords, publication date, etc.), or other hierarchical tags or elements.

Tag – Keyword or index term; "to tag" is to assign a keyword or set of keywords.

Techweb – Sandia's internal website.

Term – Word or phrase found in a document or in an index.

Term Cloud – Stylized way of visually representing occurrences of search terms or words, making it easy to discern the subjects covered at a glance. Also called tag cloud, word cloud, and text cloud.

Text Analysis/Text Mining – Automated methods of deriving useful information from text/textual content.

Unstructured Files – Contain textual data with no predefined or inherent structure or assigned fields (i.e., title, author, abstract, keywords, publication date).

7. ANNOTATED BIBLIOGRAPHY

To support the objective of identifying existing tools that would enhance "information findability", or the ability of the Systems Analysis Group to find, manage, and build on prior work, the study examined industry information sources directed at end-users/practitioners, to include trade journals, blog posts, and white papers, and also two Sandia technical reports that contained assessments of commercially available tools.

Because one of the toolsets described in the body of this report is a Sandia National Laboratories-developed toolset, the study also examined some text-mining focused technical reports from U.S. Department of Energy national laboratories. Summaries of those reports are provided.

Industry Sources

Arnold, Stephen A. "Open Source Search: Clarity or Confusion?" *Online*, Jan/Feb, 2012.

The author addresses the challenges of "open source search" and provides a table of open source search systems and software.

Arnold, Stephen A. *The New Landscape of Enterprise Search: A Critical Review of the Market and Search Systems. (*Pandia: Oslo, Norway), 2011.

Report (eBook) available for purchase. As reviewed by Deborah Lynn Wiley for *Online* magazine, Nov/Dec 2011: This report reviews six enterprise vendors: Autonomy, Endeca, Exalead, Google Search Appliance, Microsoft FAST Search Server, and Vivisimo. Arnold tells "what works, what doesn't, and what to watch out for." Other aspects of enterprise search are not covered.

Arnold, Stephen A. "Redefining Search: The Quiet Enterprise Revolution." *Information Today*, June 2011.

Primarily an overview of enterprise search *services* provided by Search Technologies Corp. and CEO Kamran Khan. Khan states "lack of attention to detail in preparing the data set for search is often the root cause."

In more general terms Arnold states "services are an important part of enterprise search deployment" and "the challenge of enterprise search is less about technology and more about understanding what problem to solve and then using the appropriate off-the-shelf system to end the pain." Arnold also briefly mentions other services companies: Flax (U.K.), Lucid Imagination, New Idea Engineering Inc., Raritan Technologies, Inc., and Comperio.

Balakrishnan, Karthnik and Janine Johnson. "Text Mining in the P&C Back Office." *Claims Magazine, March 2012.*

Article directed to property & casualty (P&C) insurers; describes using text mining to "uncover critical insights from unstructured data sources" through review of claims operations, customer feedback, financial records, etc. Acknowledges that setting up a text mining system can be challenging.

Barbosa, Daniela. "Finding the Right Recipe for Organizing Enterprise Metadata." *DM Review*, December 2008.

"Consider it an embarrassment of riches: for the contemporary enterprise, progress is not inhibited by a lack of information but by a lack of easy access to that information."

The article addresses taxonomies (structured hierarchies of metadata) and "folksonomies" or tagging (collaborative categorization using freely chosen keywords). A hybrid approach can provide great benefit to the enterprise.

Boeri, Robert. "Enterprise Search or Content Management?" *EContent*, May 2010.

This piece addresses the issue of organizations looking to "enterprise search" (ES as a shortcut to solving information findability problems rather than pursuing enterprise content management (ECM) systems or waiting for content management interoperability services. In the opinion of three findability experts, ES provides great value but needs to work with ECM to yield the best results.

Chapman, Leon, Rossitza Homan, Jim Treadwell, et al. *Knowledge Discovery and Data Mining (KDDM) Survey Report*. SAND2009-1219, Sandia National Laboratories, February 2005. [Unclassified]

The authors surveyed the commercial KDDM market and summarized findings about data mining, text mining, clustering, visualization, and XML conversion tools. In addition they also ranked the tools using advanced collaborative environment (ACE)-specific criteria, and identified four research and development areas critical to the implementation of decision support systems on ACE.

The text mining tools reviewed were ClearForest, Docyoument, dtSearch, Enkata, InFact, LexiQuest, PowerDrill, Predictive Text Analytics, Readware, SemioMap, SmartDiscovery, VisualText, and XML Miner.

Cisco, Susan. "As Ye Index, So Shall Ye Retrieve: Findability's Critical Success Factors." *AIIM E-Doc Magazine*, March-April 2008.

Some of the success factors, as identified variously by a dozen "seasoned professionals" in the field, are standardized information infrastructure (consistent taxonomy; standardized metadata; consistently structured; semantically rich metadata); end-user perspective (understanding end-users' requirements and their search context; end-user confidence); and full content and cross-repository searching (integrating indexing terms, metadata, full content searching, and relevance

ranking). For the enterprise, legal factors such as discoverability are an issue, and the human factor cannot be overlooked.

Earley, Seth. "Content Curation: Contributing to Improved 'Findability'." *Information Outlook*, December 2011.

Earley argues for content curation of digital assets by corporate librarians, and addresses three aspects of content curation: taxonomy and content tagging; governance; and content life cycle management. No specific tools are mentioned, although there are tools for semantic analysis, taxonomy management, auto-categorization, and indexing. In addition, Earley calls for life cycle management to review for R.O.T (redundant, outdated, and trivial) content that has no business value and is not needed for compliance purposes. He touches on content audits.

Gantz, John, Angela Boyd, and Seana Dowling. *Cutting the Clutter: Tackling Information Overload at the Source*. IDC white paper, March 2009.

What's the volume? IDC estimates that there are more than 100,000 pages of paper per information worker in the workplace.

What's the cost (in time)? IDC estimates as much as 20 hours a week per worker:

- Reformatting from multiple formats to a one document format
- Searching for but not finding information
- Recreating content
- Publishing the same content to different audiences using different applications
- Moving documents from one format to another
- Acquiring archived records with little or no automation
- Dealing with version control issues

"Are we doomed to run through our work days on an accelerating treadmill of documents, emails, messages, spam, and lost information? Are we facing insurmountable odds in the war on information overload?" This study says no, and points to technology solutions and the need for processes – with a nod to Xerox.

Heck, Mike. "Data Mining for Your Desktop," *Network World*, July 26, 2010.

Heck states that many people want features that go beyond Microsoft Windows, such features as: multiple query methods, auto categorization, and clustering. He then reviews six desktop search products: Copernic Desktop Search, dtSearch Desktop, X1 Professional Client, ISYS Personal Edition Search Software, Exalead Desktop Search, and Google Desktop. However, most of the article addresses the search capabilities of each product. Clustering is barely addressed.

Hunt, Julie B. "Findability Inside The Firewall – Still Trying To Find The Information We Need." *Highly Competitive* blog, January 31, 2011.

Hunt's premise is that "inside the firewall, locating content and people is better served by 'findability' … it doesn't matter how great the information or content is if the people who need it

can't figure out how to get to it." Employees need to decrease time spent searching for documents or subject experts and duplicating previous work, and increase time gaining insight, building on prior work, and producing new knowledge and products. However, the "search" function in most content management solutions is inadequate, and other enterprise applications, including search, are often siloed.

"Enterprise search" works across silos, but is expensive and content and information must be optimized to support better search results. However "Enterprise 2.0" is evolving toward the "findability paradigm."

Hunt defines enterprise findability as "ease in locating information and resources for specific worker needs while reflecting relevant aspects of content attributes, context, worker roles, people/social resources, and business processes." She also refers to Mixon's definition: Findability is "the art and science of making information available to users," and "incorporates governance, architecture, design, and navigation."

Hunt briefly covers Baynote, adaptive case management software (Isis Papyrus), Attivio's "active intelligence engine, and others. She refers to a number of other people, with links to their work.

Huwe, Terence. "Meaning-based Computing: Text Analysis Takes a Great Leap Forward", *Online Magazine,* September/October 2011.

Huwe defines "meaning-based computing" as something that unites the power of modern search protocols with recent advances in text pattern recognition, language-context analysis, and "sentiment analysis." He describes some use cases, and identifies three related areas for digital content managers to watch: taxonomies, digital repositories, and collaborative workspaces and functions.

Langenkamp-Muenkel, Julie. "Snapshot on Information Applications: Making Information Available." *Information Management*, July/August 2010.

Fifty-eight percent of respondents in a Ventana Research survey believe it's 'very important to improve information availability." The leading drivers of information applications projects are decision-making, (60%), performance improvement (58%), competitiveness (50%), and revenue generation (38%). This document is a one page summary of the report.

MacFadden, Gary. "Enterprise Search and Information Access: Revisited." *Enterprise Management* blog, May 11, 2012.

In this blog post, MacFadden asserts a "growing user demand for more nimble ways to manage and access a variety of intra-organizational data along with merging externally created data that support a single department … ." These "unified access platforms" are expected to contain:

- The ability to index and search all major enterprise repositories
- The ability to include web content, social media and other external sources
- An intuitive user interface supporting faceted search, SQL queries, and data visualization

- Support for content analytics with business analytics (BI) tools integration
- Easy-to-use Google-like experience for business users, with no programming required
- Multiple access points from desktops, collaboration suites to web browsers and
- Ability to deploy departmentally and easily scale and migrate across the enterprise
- Portability of data and content between various repositories (CMIS, XML, and others)

MacFadden provides short reviews of eleven vendors. Enterprise search vendors are Attivio, Coveo, Endeca (Oracle family), and IBM (intending to acquire Vivisimo. Also mentions IBM (Omnifind), Lucid Imagination, and Microsoft (FAST internet search, paired with BING). Other vendors cited are Autonomy (acquired by HP), EMC, Google, SAP, and Symantec.

MacFadden, Gary. "In Search of a Strategic Imperative for Managing Enterprise Content: Lessons from AIIM, ECM Vendors and the User Community." *Enterprise Management* blog, April 4, 2012.

"Evidently, the connected world's desire to easily access content and information from virtually anywhere at any time reigns supreme, and a legion of solution providers is eager to build the backend infrastructure and content management systems to allow that to happen."

MacFadden references AIIM studies, including State of the ECM Industry 2011: How well is IT meeting business needs? Essentially, the study concludes: "Users are keen to attack content chaos and drive cost reductions, but are also looking for more effective staff collaboration and engagement."

He also addresses enterprise content management (ECM), which includes collaboration solutions (i.e., SharePoint), Document and Intelligent Capture, Content and Records Management, Message Archiving, Content Classification and Analytics, Search and eDiscovery, Case Management and Workflow; as well as Backup, Data Loss Prevention, Policy Management and Security elements. MacFadden says "Lack of interoperability between ECM solutions and limited access across the enterprise to business-critical siloed content repositories increases complexity and risk, reduces ROI and saps productivity – not to mention making for a stultifying user experience."

"For the most part, content management solutions are still sold and bought departmentally – because vendors are more than happy to sell users what they insist they need and some technology innovations are just too good to pass up regardless of the downside on the back end. But user organizations are beginning to feel the pain (cost, complexity, risk) that siloed solutions and piecemeal technology adoptions engender, hence the new-found emphasis on enterprise-wide information governance and strategy, along with complementary technology."

MacFadden provides a short overview of ECM vendors:

> ASG, Crawford Technologies, EMC (and its Kazeeon and Captiva solutions); Fujitsu (scanners); Hyland Software (OnBase); IBM ("integrated content and data management space" – FileNet; Datacap; PSS Systems); Kofax (document capture); Novodynamics; Nuix

(Australian-based); Recommind; Symantec (including Data Insight, for improving data governance of unstructured data including documents, spreadsheets, and emails). (Note - provides a link to IBM's Watson.)

Mancini, John, ed**.** *8 Reasons You Need a Strategy for Managing Information – Before It's Too Late.* AIIM e-Book, n.d.

This e-Book consists of a collection of blog posts on related subjects, all with a theme of "eight" – eight reasons, eight ways, eight benefits, eight things, and so on.

See especially:
George Tziahanas. "8 Things to Consider When Aligning with Today's Growing Corporate, Legal, and Regulatory Standards."

- Need to address *both* explosive growth in the volume and complexity of information, with new forms of content and communication *and* more rigorous legal and regulatory environments.
- Few organizations are focused on qualitative risk, especially those that arise in unstructured content.
- Must be prepared to address the multi-channel nature of information.
- "Forensically sound" indexing and search capability.
- Manage content in place (not in a central repository).

George Parapadakis. "8 Things You Need to Know about Information Risk."

- Non-capture – the risk of critical information not being captured into the system.
- Loss – the risk of captured information being accidentally removed from the system.
- Malice – the risk of captured information being deliberately removed, corrupted or damaged.
- Attribution – the risk of losing the context and metadata describing the information.
- Unauthorized access – the risk of information being accessed by unauthorized persons.
- Unavailability – the risk of disaster or technical failures, preventing access to the information.
- Findability – the risk of information being lost due to lack of sufficient classification.
- Inaccessibility – the risk of information become inaccessible due to its medium or format.

Also many posts about various aspects of implementing enterprise content management systems and tools.

Manning, Vivian. "Desktop Search Tools: Like Google for Your Computer." *Attorney at work.com blog*, February 6, 2012.

Manning provides a short assessment of five desktop search tools: Copernic Desktop Search, dtSearch, Everything, Windows 7 Search, and X1 Professional. Her focus is on quickly finding a needed document.

Miles, Doug. *Big Data – Extracting value from your digital landfill.* AIM Industry Watch Report, 2012.

This report summarizes the results of a survey sent to 65,000 AIIM members, with 403 respondents. Graphs summarizing the responses excluded responses from suppliers of ECM products and services, and from organizations with less than ten employees.

From the report conclusions and recommendations section:

Although surrounded by hype, the ability to analyze and correlate big data and big content repositories is deemed to be very useful by our respondents. In particular, the linking of unstructured text or rich media data with structured transactional data would be very attractive to many.

The range of available analytic techniques across unstructured and semi-structured data is considerable, and this, combined with the need to present a unified data connection across diverse datasets, presents a challenge for most organizations. Expertise is scarce, and generic analytic product platforms are in their infancy, with most organizations resorting to a mix of open source products, in-house development, and best-of-breed applications. For many, there is also a security issue of both a competitive and a compliance nature.

Recommendations

- Ask blue-sky questions of your business such as "if only we knew…" or "if we could predict…" or "if we could measure…" Consider how useful that might be to the business *before* thinking about how it can be done or at what cost.
- Play those questions off against the data you already have, data you could collect, or data that you could source elsewhere.
- Include in your thinking structured transactional data, semi-structured logs and files, and text-based or rich media content.
- Incoming communications from your customers, outbound communications to your customers, and what customers (or employees) are saying about you on social sites can all be useful for monitoring sentiment, heading off issues and analyzing trends.
- Consider high volume streams such as telemetry, geo-location, voice, video, news feeds, till transactions, web clicks, or any combination of these.
- If your content is currently digital landfill spread across disparate file shares and content systems, consider how this could be rationalized prior to any big data projects.
- Content access for both search and analytics, and if necessary, content migration to dedicated big data storage, can be facilitated by unified data access products.
- Don't be tempted to rush into in-house developments or specific point-solutions without considering wider, more universal analytics platforms.
- Consider "software as a system" (SaaS) and cloud deployment as a faster way to acquire experience and focus activity.
- Big data is much more about the breadth of the analysis and the insights that can be achieved than it is about the size of the data and the underlying database technology.

Moore, Andy. "What Are You Looking For?" *KM World*, May 2012 (editorial).

From a MindMere survey of over 2000 business directors and managers worldwide, 52% said they "cannot find the information they are seeking using their own organization's enterprise search facility" within a reasonable period of time. Says Moore "It's not information overload that's the vexing problem, it's 'information under-managed.' There can NEVER be too much information."

However, enterprise search is NOT a single unified piece of software. And it must not matter whether data is structured or unstructured – it's irrelevant to the person needing the information. And. "web search and enterprise search are nothing alike!"

What's needed? Effective application of:

- taxonomies and ontologies,
- automatic classification,
- better metadata,
- better navigation, and
- a user experience that is intuitive and frictionless.

Owens, Leslie. "Can't Find What You Need? Join the Club." *KMWorld*, April 2010.

Owens summarized 2008-2009 data from Forrester Research that showed conflicting interests between information and knowledge management (I&KM) software decision makers and information workers (I-workers). Only 28% of I&KM software decision makers at the enterprise level ranked implementing an enterprise search strategy as important or very important (compared to 51% for consolidating e-mail systems and 36% for implementing an enterprise collaboration strategy). Yet I-workers "would rather have improved findability than other workplace enhancements" but are willing to spend time searching for solid information. Her conclusion "if I-workers can start a grass-roots movement for prioritizing findability, I&KM pros can build a strong argument for senior management to … make it happen."

Pelz-Sharpe, Alan. "Enterprise Search Is Not Easy." *Information Management*, September/October 2011.

This short article touches on the challenges of enterprise search, names the "big three": Google, Microsoft, and Apache-Lucene; and then goes on to name other search vendors, such as dtSearch, Endeca, Exalead, and Vivisimo. None is evaluated.

Sabo, Tom. "Text Analytics: Leveraging Unstructured Data in Analysis Efforts", SAS Institute Inc., PowerPoint slide set, 2012, with examples from SBIR/STTR data.

This SAS presentation summarizes "text analysis" with illustrative examples from SBIR/STTR data.

Schillerwein, Stephan. *The digital Workplace: Redefining Productivity in the Information Age*. Infocentric Research whitepaper, September 2011.

According to this whitepaper, studies show that employees lose one to two hours a day to unproductive efforts to locate needed information because company work practices reflect the industrial age rather than the information age. The paper goes on to outline a "digital workplace" framework that would need to bring together currently isolated systems while also taking in to account existing work processes and individual work practices. No such system exists today. This white paper presents food for thought for enterprise-wide information system planners.

Skjekkeland, Atle and Tony Byrne. "Enterprise Search: How to Get It to Really Work." *AIIM E-Doc Magazine*, September-October, 2007.

"In fact there is a growing recognition in the industry that what matters is not how searchable you make your information but how *findable*."

Although this article is a lead-in to promoting a then new AIIM Information Organization and Access certificate program, it does provide a nice, short description of "information access" as including a collection of technologies: enterprise search, content classification, categorization and clustering, fact and entity extraction, taxonomy creation and management, and information presentation (visualization).

Solomon, Marc. "SharePoint in Practice." *KM World*, June 2010.

Solomon asserts that KM practitioners have the opportunity to use the common frameworks and toolsets of SharePoint to build best practices in search, taxonomy, provisioning, and site navigation practices. He mentions two third-party add-on tools:

- Coveo indexing to upgrade SharePoint search (indexing capabilities)
- BA-Insight Longitude tool (search refiner; automated tagging; document previews)

See also "Laying the Foundation for Your Next SharePoint Deployment," *KM World,* April 2010 and "On the hook: The SharePoint Ownership Imperative," *KM World,* May 2010.

Trahan, Michael and Mark Foehse. *A Toolkit for Detecting Technical Surprise*. SAND2012-7392, Sandia National Laboratories, October 2010. [Unclassified]

"No single tool does everything."

The report summarizes a Laboratory Directed Research and Development (LDRD) project to extend analytical tool ThreatView and to develop a toolkit for detecting indicators of technical surprise in textual data sets.

In addition to reviewing analysis tools developed at Sandia National Laboratories and Oak Ridge National Laboratory, the authors reviewed several commercial tools: Analyst's Notebook (for building graphs), dtSearch (for desktop file search), Google Trends (for tracking news), Google Insights (for tracking and visualizing news), and (TextChart (for creating CHARTS). The authors also reviewed commercial support tools: Beyond Compare (for comparing folder

contents), Camtasia Studio (screen recorder or editor), MindManager (visualization tool), Mind View (mind mapping tool), and SnagIt (for screen captures). The report also states that the team was working with SAS Enterprise miner and SAS Text Miner.

The report concludes with the following statement – "We failed to find a Swiss Army Knife for our toolkit: no single tool does everything. By using multiple tools, each of which analyzes and visualizes the data sets in a different way, we are able to provide timely, relevant technology assessments with a high degree of confidence in our results."

Van Noorden, Richard. "Text-Mining Spat Heats Up," *Nature*, March 21, 2013.

The article summarizes efforts in the U.K., Europe, and the U.S. to resolve differences between publishers and scientists on data- and text-mining issues.

Related Technical Reports from U.S. DOE National Laboratories

Buttler, D.J., D. Andrzejewski, K.D. Stevens, et al. *Rapid Exploitation and Analysis of Documents*, LLNL-TR-517731, Lawrence Livermore National Laboratory, December 2011.

The purpose of the Rapid Exploitation and Analysis of Documents (READ) project was to develop technologies to make it easier to catalog, classify, and locate relevant information for analysts, who are "…overwhelmed with information. They have large archives of historical data, both structured and unstructured, and continuous streams of relevant messages and documents that they need to match to current tasks, digest, and incorporate into their analysis."

The report describes the READ project's research efforts in infrastructure development and deployment, enhanced keyword search, personalized and collaborative clustering of search results, word sense disambiguation, mixed-context entity co-occurrence modeling, and topic coherence over many models and many topics.

Dunlavy, Daniel, Timothy Shead, Patricia Crossno, and Eric Stanton. *ParaText – Scalable Solutions for Processing and Searching Very Large Document Collections: Final LDRD Report*, SAND2010-6269, Sandia National Laboratories, September 2010.

This report summarizes the results of a two-year LDRD project called "Scalable Solutions for Processing and Searching Very Large Document Collections" LDRD, which ran from FY08 through FY10. The goal was to investigate scalable text analysis; specifically, methods for information retrieval and visualization that could scale to extremely large document collections.

The team developed ParaText, available through the open source Titan toolkit. Techniques and components developed as part of ParaText were also used by the networks grand challenge LDRD. (Kegelmeyer, 2010)

Dunlavy, Daniel, Dianne O'Leary, John Conroy, and Judith Schlesinger. *QCS: A System for Querying, Clustering and Summarizing Documents,* SAND2006-5000, Sandia National Laboratories, October 2006.

This report describes the results of research to develop a hybrid information retrieval system – the Query, Cluster, Summarize (QCS) system. The tool was shown to be able to cluster results, provide a summary of each cluster, provide a summary of each individual document, and point to each document, but further research was needed to address several issues.

Kegelmeyer, Philip, Bill Cook, David Rogers, et al. *Network Discovery, Characterization, and Prediction: A Grand Challenge LDRD Final Report,* SAND2010-8715, Sandia National Laboratories, November 2010.

This report summarizes the results of a three-year LDRD Grand Challenge to research, develop, and evaluate relevant analysis capabilities that address adversarial networks. The grand challenge focused heavily on development of the Titan software informatics framework; Titan also played a major role in the "very large document collections LDRD." (Dunlavy, 2010)

During the course of the research, the team created three prototype end-user tools that were intended only to be used by the team in support of its research, but that also garnered attention in and of themselves. One in particular, P2 (the second prototype) has specific relevance for this study on information findability and its interest in features useful for conducting thorough analyses and studies. See the description of an analyst's workflow and ways in which P2 could aid the process on page 33 of the Kegelmeyer report.

P2 was designed to analyze unstructured text in support of the information ingestion needs of counter-proliferation agents. Its ability to identify documents of interest from within a corpus was tested against commercial tool dtSearch, with no statistically significant differences. However P2's purpose was to find documents, automatically organize them by topic, extract entities, and permit the analyst to interactively build and adjust a "hot list" of entities of interest. Problems that would need to be addressed were discovered, for example, P2's interface appeared to constrain how analysts could interact with their data.

Oehmen, Christopher. *Annotated Bibliography for the DEWPOINT Project,* PNNL-18370, Pacific Northwest National Laboratory, April 2009.

This bibliography covers documents, primarily from technical literature, deemed relevant to a PNNL project on Detection and Early Warning of Proliferation from Online Indicators of Threat (DEWPOINT). The focus is on topics, such as methods for indexing, searching, and organizing information data, information extraction, and information retrieval.

Patton, Robert, Christopher Symons, Bryan Gorman, and Jim Treadwell. *Knowledge Discovery, Knowledge Management and Enterprise-Wide Information Technology Tools Final Report,* ORNL/TM-2012/148, Oak Ridge National Laboratory, April 2012. [Unclassified]

The authors report on the results of a request by the Office of Naval Research Global (ONR Global) for a knowledge discovery and management utility for its science and technology (S&T) documentation, because existing search mechanisms for SharePoint and commercial enterprise search engines do not provide the "push" function ONR Global wanted to keep its knowledge workers up to date and informed of developments within the enterprise knowledge base.

Although a successful prototype system which integrated text-analysis algorithms of a tool called Piranha with Microsoft SharePoint was implemented at ORNL, the pilot at ONR experienced problems. Processing times were unacceptably long, and a large number of non-relevant documents were returned. The authors point out the need for "noise filters" or supervised learning algorithms to tag documents into recommended classes, however both would require extensive development. COTS solutions alone will not provide the full answer.

Steed, Chad, Christopher Symons, James Senter and Frank DeNap. *Final Report: Guided Text Search Using Adaptive Visual Analytics*, SERRI Report 89990-01, September 2012.

This is a report in the Southeast Region Research Initiative (SERRI) project on Smart Search Analytics. The SERRI project was commissioned by the U.S. Department of Homeland Security in 2006, and is managed by Oak Ridge National Laboratory. An objective of the visual analytics study was to help DHS fusion center analysts cope with information overload.

To that end, ORNL researchers developed a visual analytics system called Gryffin that provides "tight integration of semi-supervised machine learning with inferential visual representations." This report describes the purpose, architecture, workflow, and output. Of potential value to information analysts, the client interface provides three main views: a temporal view, a list-based textual view with a "sandbox" for capturing relevant records, and term-frequency view.

Stickland, Michael, Shelley Eaton, and Gregory Conrad. *Natural Language Processing-Based COTS Software and Related Technologies Survey*, SAND2003-2916, Sandia National Laboratories, September 2003.

This 2003 report examined then-current commercially available natural language processing (NLP)-based knowledge management software with potential applications for information retrieval, information extraction, summarization, categorization, terminology management, link analysis, and visualization for possible implementation at Sandia National Laboratories. Software from fifty-one vendor companies was reviewed. By design, the software surveyed was not intended for desktop installation, but instead for integration into a corporate computing environment. A number of the vendors are still in business, and have continued developing products in this arena.

Distribution:

2 MS0421 N.K. Stoecker, 0249
1 MS0421 R.D. Skocypec, 0240
1 MS0421 C.A. Ulibarri, 0249

1 MS0421 Group 0240 Library, 0241 (electronic copy)
1 MS0899 Technical Library, 9536 (electronic copy)


Sandia National Laboratories