

TeraText

© 2014 by Stephen E. Arnold, www.arnoldit.com

The system provides a single consistent search model, independent to the physical data representation. Although the technology dates from 1980, TeraText is a robust, capable information access system.

When I last updated this report on TeraText, SAIC, a large government-centric services firm, was little-known outside of defense and governmental markets in Australia and the US. SAIC attracted some unfavorable publicity in 2007 when Vanity Fair Magazine published "Washington's \$8 Billion Shadow."¹ This story was a follow on to a 2003 Business 2.0 article "In the Company of Spies." TeraText, a search and content processing system, has been used as an information platform for decades. TeraText, a system with its origin in a 1980 system called Titan, may have been the first search-application platform.

The TeraText system reached version 5 in 2007, and it continues to be a product offered longer than any other enterprise-grade search systems. The product can be used as a content management system, a search system, an authoring and versioning system, and a platform upon which other applications can be built.

I learned that TeraText will be offered through a company called Leidos, a unit of SAIC. Leidos offers a number of information-centric products and services. What is interesting is that over the last thirty-five years, TeraText has remained firmly attached to the Australian technology community.

The work of TeraText's developers has influenced, directly or indirectly, other Australian search and retrieval systems. These include the ISYS Search Software system and the Funnelback system. TeraText's early support for SGML and then XML has remained a core technical innovation referenced in patents for search systems that use similar systems and methods.

Understanding the TeraText approach and how the company built specific applications to serve the needs of its different government customers was a road map for the later activities of Autonomy, Convera, and Fast Search & Transfer among others.

Author's note: This is a 2008 draft. It will not be updated.


Stephen E Arnold, February 17, 2014

¹. See <http://vnty.fr/1ak1DkL>

Introduction

The TeraText system uses a combination of library standards, SGML and XML, and proprietary technology to store and retrieve information. The system handles terabytes of structured, semi-structured, and unstructured data. TeraText is an information management system with search, analytics, and reporting functionality. The content management component permits versioning so a user can retrieve a document from a specific point in time. The system can index and search from multiple document repositories as well as from other content sources, including Web sites to which the TeraText crawler has access. More remarkable is the fact that the technology dates from 1980 and is available for licensing as I write this in 2007.

Table 1: TeraText: A Bird's Eye View

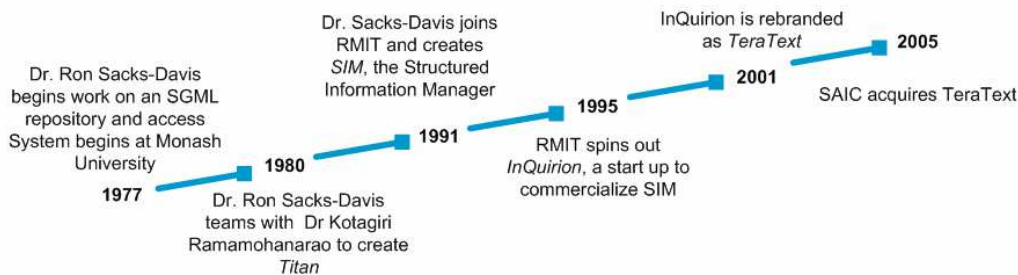
Product Thumbnail	
1 Search Brand	TeraText Database System (TBS) and TeraText Document Management System (TDMS). Modules include an email archive system, MPS (a metadata publishing system),
2 OS Supported	Linux, Solaris, Itanium, Microsoft
3 Est License Fee	Fees can begin as low as \$5,000 but typically TeraText is a six or seven figure system. A custom price quote is required because the license fee has to be calibrated to operating system, number of CPUs, TeraText modules required, engineering service, support, and maintenance.
4 Functions	Search and document management. XPath is used for mapping XML data to the conceptual model. Leading applications include intelligence gathering, technical documentation, legislation management, publishing and knowledge management. When Dr. Ron Sacks-Davis developed his idea, the approach was fresh and married access to content, search, and manipulations of metadata and content snippets in one large, complex system.
5 Claimed Features	With a search capacity of 47 million pages per second per CPU, processing up to two billion documents every four seconds, while simultaneously loading data, a single query can search diverse information collections in an organization, without the user understanding all of the information collections available. TeraText uses high performance compressed inverted file indexes.
6 Downsides	The system implements traditional document management, records management, and search functions with support for such standards as SGML, XML, Z39.50, etc. Index overhead is 1.1 times the size of the source content.
7 Similar To	The system is similar to iPhrase, MarkLogic, and other repository-centric systems with standards support and document management systems. For document assembly, TeraText competes with other robust content management systems.
Product Close Up	 <p>TeraText is a general purpose information processing, management, and search system. The system stores processed content in XML (Extensible Markup Language) and in the information's original format. A representation of the content is created and queries run against these indexes. The system supports ISO 23950 / Z39.50. The search features include words, phrases, adjacency, word distance, sentence, paragraph, fuzzy match, stemming, and truncation. The system also can return selective information, multiple delivery formats, stored filters, index exploration, and sorting. TeraText implements chained version control so that a user can track the changes to a document through time.</p>

A key goal of TeraText is to support flexible information architectures, not mandate a single architecture.

TeraText exhibits its “library” orientation; that is, the system focuses on organizing and retrieving information that has been processed to conform to a consistent structure. The TeraText products have been designed to manage extremely large quantities of heterogeneous content. The system updates its indexes in near real time as TeraText simultaneously supports thousands of concurrent users. TeraText search is one important facet of an enterprise information management system. TeraText has made it possible for licensees to build applications on the TeraText framework. Licensee applications include:

- Defense and intelligence systems
- Legislation publishing, versioning, and tracking
- Technical documentation
- Web portals
- Online publishing.

Some competitors offer similar functionality, but most of these vendors have entered the market decades after TeraText became commercially available. Therefore, TeraText is an important system to understand and consider when a procurement team needs a robust, scalable, flexible information management and retrieval system.



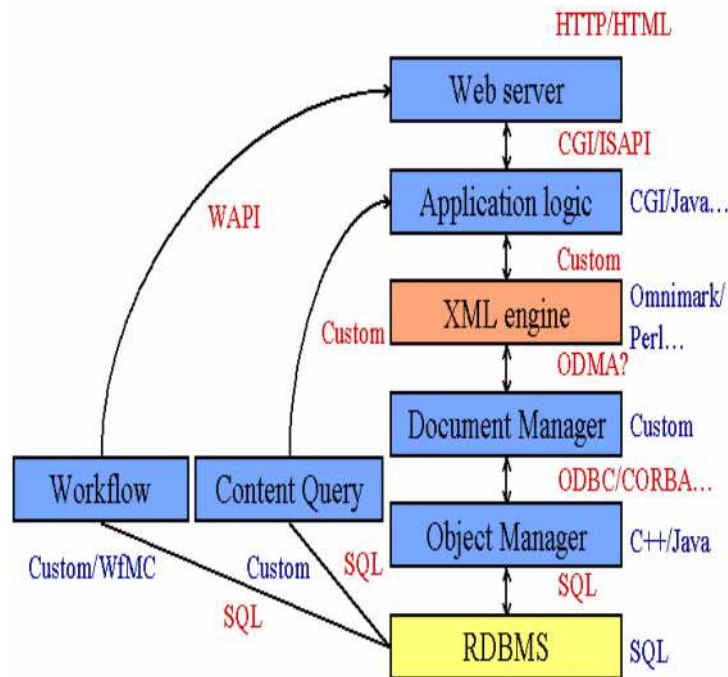
TeraText: An Innovator from the 1980s

In 1980 at Monash University, Ron Sacks-Davis teamed with Dr. Kotagiri Ramamohanarao to develop a text search system called Titan while working at Monash University in Melbourne. This means that chunks of today’s TeraText are a quarter century “young.” Some of the work on TeraText was funded by the Australian Research Council (ARC).²

². Australia has been a fertile ground for information retrieval. ISYS Search Software, Funnelback, and YourAmigo are other Australia search systems. However, some of the ideas for TeraText may stretch back to 1980.

“I believe that TeraText is to documented knowledge, what Oracle is to transactional and tabular data.”—Dr. William P. Hall, TeraText

ARC initially funded a Centre for Knowledge-Based Systems at RMIT and the University of Melbourne.³ Titan morphed into a system called SIM, an acronym form of Structured Information Manager to identify the system with Structured Generalized Markup Language. Work on the system continued within the RMIT Multimedia Database Systems Research Centre. In 1993, InQuirion forged a strategic alliance with Ferntree Computer Corporation in Melbourne. Ferntree was acquired by GE Capital. Under the alliance, Ferntree was responsible for marketing the product in Australia, while RMIT continued the research and development of TeraText. To commercialize the technology, RMIT formed a company in 2001 owned by the university and the professionals working on the system. The company was named InQuirion Pty Ltd. In late 2001, InQuirion management inked a deal with the privately-held services firm Science Applications International Corporation (SAIC). The low-profile, publicity-shy company has ranked at the top of information technology services suppliers to the United States Government.



The TeraText stack which licensees and application vendors build content processing and information access solutions. Source: Ron Sacks-Davis and Alan Kent, "TeraText Database System," SAIC white paper, 2002, page 20.

³. From 1991 to 1994 Ron Sacks-Davis was Research Director of CITRI, an IT collaboration of the Royal Melbourne Institute of Technology (RMIT) and Melbourne Universities.

“We're all about solving hard problems. Terrorism is a hard problem.”—

Steven Rizzi, SAIC TeraText. Source: Paul Kaihla, In the Company of Spies,” Business 2.0 Magazine, May 1, 2003.

Under the terms of the 2001 deal, California-based SAIC has the exclusive rights to distribute the Titan-SIM-InQuirion system as “TeraText” in North America through a newly-formed unit called TeraText Solutions. InQuirion would retain the intellectual property in Australia and continue to develop and research the products.

In October 2005, SAIC bought InQuirion Pty Ltd. SAIC is a savvy outfit. The company recognized that TeraText provided software licensing revenues as well as on-going training, engineering, and support services revenue. By 2005, TeraText had captured significant customers in the Australian government and the US defense intelligence sector. TeraText, like Fulcrum Technologies’ and ISYS Search Software’s system, is among the longest-lived information retrieval systems still marketed in 2007. Many newcomers to information retrieval assert their systems are more modern, but TeraText and its longevity prove that large systems tailored for government-centric content processes can establish a foothold in the market and persist for decades. TeraText is now in Version 5. This suggests that the product has aged gracefully over the last 25 years.

Selected Executives

SAIC does not provide a list of the management team in place at TeraText. Individuals identified as having a role in TeraText include:

- Ron Sacks-Davis, the developer of the SIM (Structured Information Manager) technology that became the database upon which TeraText was built. Dr. Sacks-Davis served as the managing director of InQuirion
- Alan Kent, once the chief technology officer at InQuirion, the precursor of TeraText. he was the first Ph.D. from Royal Melbourne Institute of Technology (RMIT University) in Melbourne, Victoria. His thesis was titled “File Access Methods Based on Descriptors and Superimposed Coding” (1988)
- Dr. Timothy Arnold-Moore, Marketing Engineer for SAIC from 2005 to the present. Dr. Arnold-Moore received his Ph.D. from RMIT. His dissertation was “Information Systems for Legislation.” He is an attorney with a degree from the University of Melbourne. He developed the EnAct system for Tasmania and used TeraText (then InQuirion).⁴ From 1999 to February 2000, he joined TeraText. He helped augment TeraText database system with facilities to support the authoring process. In addition, he worked on the TeraText document management system which was designed to manage technical documentation for Royal Australian Navy frigates with TeraText partner Tenix Defence Systems. He

⁴. This application of InQuirion/TeraText made use of chained deltas; that is, tracking sequences of modifications to documents. This is a repository function useful to legislative authoring and editing.

“The vast torrent of data being carried by the Internet has created unprecedented demands on information management systems – and an Australian technology is meeting the challenge by processing up to two billion documents every four seconds.”—National Survey of Research Commercialization Years 2001 and 2001, page 55 at <http://bit.ly/1dtJtvv>

continued to support the company’s legislative drafting customers. Dr. Arnold-Moore was involved with the TeraText workflow enactment service based on the Workflow Management Coalition (WfMC) architecture and Document Version Management capability based on the Document Management Architecture (DMA 1.0). Dr. Arnold-Moore was also responsible for the selection of relevant standards, the initial architectural design of this project and performed a number of project management and implementation tasks.

- Steve Rizzi, based in Annapolis, Maryland, TeraText office, SAIC corporate vice president.
- William Wolf, assistant vice president of TeraText products.

TeraText’s offices have been listed as Melbourne, Australia, locations in Maryland and Virginia, and in various cities around the world.

Financial Performance

The companies fused to yield the TeraText content processing system do not reveal their revenues.

In 2004, RMIT revealed the following:

In 2003, it recorded revenues of \$2.8m, an increase of 86 percent on the 2002 figure. Expenses remained relatively low, giving a positive cash flow result.⁵

At the time of the SAIC buy out, ArnoldIT estimates that TeraText revenues were US\$10 to US\$12 million. Revenues of TeraText are estimated to be in the \$6.0 to \$9.0 million range for 2006.

The majority of the TeraText revenue is for classified government-related projects. Additional revenue is generated via maintenance, engineering services, and technical support.

However, the structure of SAIC is not suited to the sale of commercial software products. The firm’s focus is on the sale of engineering and other professional services. It is possible that the business units involved in TeraText sales and service delivery will be rationalized. Similarly, TeraText remains an Australian project; therefore, the center point for TeraText expertise is in Australia. SAIC is a US entity with governmental work delivered from the Washington, DC area. TeraText is, therefore, likely to be a support or ancillary product.

⁵ RMIT University, Annual Report 2003, page 57. See <http://mams.rmit.edu.au/knyw70ablybvz.pdf>.

Selected Clients

TeraText's clients include:

- Australian Department of Defence
- Australian Research Council
- Australian Tax Office
- British Columbia Archives
- Canadian Department of Justice
- Commonwealth Scientific and Industrial Research Organization (CSIRO)
- Music Australia
- National Library of Australia (Kinetica Search Service)⁶
- Picture Australia
- New Zealand Navy
- Royal Australian Navy
- Standards Australia
- State Government of New South Wales
- State Government of Victoria
- Tasmanian State Government
- Tenix Defence Systems
- US Department of Defense
- US Department of Homeland Security.

Selected TeraText Partners

SAIC does not provide a comprehensive list of its partners. A few SAIC partners have been mentioned, and these are:

- eG Innovations- a provider of real-time performance monitoring and proactive triage solutions for IT infrastructure, licenses TeraText for some of its enterprise solutions.
- Hewlett Packard- This tie up is due in part to SAIC's offering an Itanium version of TeraText. If the Itanium technology becomes a dead end, Hewlett Packard may seek other search-and-retrieval partners.

⁶. In 2007, this organization explored using Lucene, the open source search system, as the retrieval engine for the certain digital content. See libraries Australia Advisory Committee Meeting, 18 April 2007, a report of a meeting on April 18, 2007.

“Google isn’t doing a bad job indexing huge volumes of information, but RMIT University in Melbourne, Australia has developed an even better application, now being marketed under the name TeraText. It is able to concurrently index and retrieve terabyte volumes of text. —

William P. Hall,
Tenix Defense at
<http://bit.ly/1eRPdQj>

- Stellent—the TeraText search system can make content in a Stellent system searchable.
- Tenix—an Australian company that specializes in defense and related projects.

TeraText in Action

The case examples illustrating the use of TeraText technology are sparse. One of the reasons is that TeraText is a component in much larger systems. Another reason is that the majority SAIC’s work is for government entities involved in defense and intelligence activities. TeraText has a number of unclassified projects; for example, in conjunction with the Canadian system integrator Irosoft Inc. InQuirion is undertaking a project for the Canadian Department of Justice in Ottawa.

Other projects include:

New South Wales Parliamentary Counsel Office

NSW Parliamentary Counsel Office engaged InQuirion to build a Web site that makes current NSW legislation available online. This is an interim Web site that will contain a subset of legislation only, and will be upgraded when a full set of the NSW legislation becomes available. The interim Web site went live in June 2002. As part of Australia’s Legislation Information Management System (LIMS) project Department of Justice require an integrated XML-based Content Management and Delivery System (CMDS). The main features of the CMDS will be: version control, workflow management, low-level XML component management, extensive search-and-retrieval facilities including “point-in-time” capabilities, Internet and intranet delivery, English and French interfaces and searching capabilities, and remote accessibility of all user-level features from a standard Internet browser. TeraText Database System (DBS) and TeraText Document Management System (DMS) are used as development and delivery platforms and InQuirion also provides technical and legislation consulting expertise. Irosoft is undertaking development of CMDS using TeraText DBS and TeraText DMS. Irosoft also provides system integration and project management services.

Canadian Department of Justice

InQuirion itself is undertaking a project for the Canadian Department of Justice in Ottawa. As part of their Legislation Information Management System (LIMS) project the Department of Justice requires an integrated XML-based Content Management and Delivery System (CMDS). The main features of the CMDS will be search in English and French by time interval, version control, workflow management, low-level XML component management, extensive search-and-retrieval facilities including “point-in-time” capabili-

ties, Internet and intranet delivery, and remote accessibility of all user-level features from a standard Internet browser.

The Canadian system will use the TeraText Database System (DBS) and TeraText Document Management System (DMS). InQuirion also provides technical and legislation consulting expertise. Irosoft, a Canadian electronic document consultancy, has responsibility for the development of content management and development system using TeraText DBS and TeraText DMS. Irosoft also provides system integration and project management services. Irosoft is located in Saint-Laurent

NEW COMPONENT DETAILS	
Component type	TeraText Content Server
Host IP/Name	192.168.10.2
Nick name	content server
Port number	80
Agentless	<input type="radio"/> Yes <input checked="" type="radio"/> No
Internal agent assignment	<input checked="" type="radio"/> Auto <input type="radio"/> Manual
External agents	LPGS

Add

The most recent version of TeraText, Version 5, incorporates graphical interfaces. Configuration files and original code are required to tailor TeraText to specific licensee requirements.

ANZAC Joint Ship Project

Tenix Defence is one of Australia's largest defense contractors. The company integrates TeraText technology into projects that require document management, search, and content processing. The particular task involved managing and accessing the content associated with building 15 warships, vehicles, and other items.

By 2003, William P. Hall was describing an exemplary use of TeraText technology for a large Australian and New Zealand naval project.⁷ Mr. Hall describes this use of TeraText as a component in "the ANZAC ship project." In 2004, a detailed diagram of this project became available.⁸

Checkpoints for the system include:

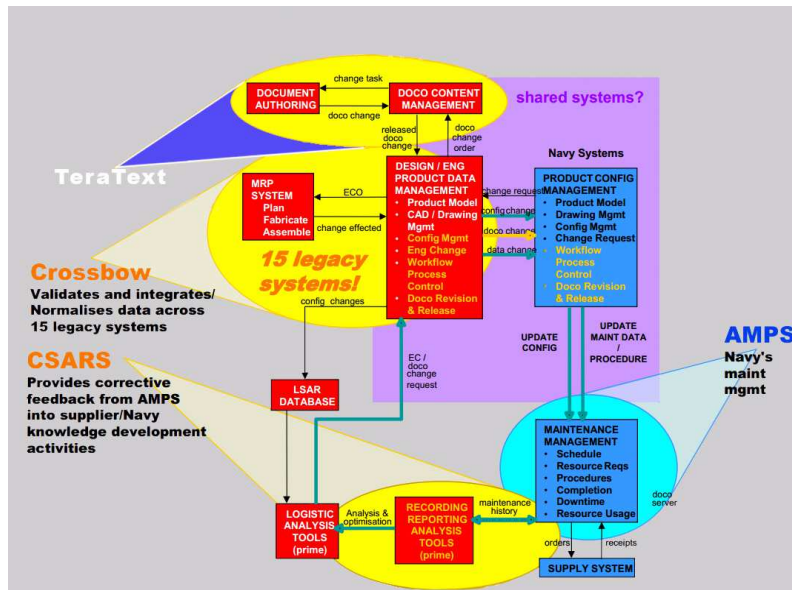
- Documents, drawings, and data indexed and stored in a repository

⁷. See William P. Hall, "Managing Maintenance Knowledge in the Context of Large Engineering Projects: theory and Case Study," *Journal of Information & Knowledge Management*, Volume 2, Number 2, 2003, pages 117-133.

⁸. See William P. Hall and Paul Brouwers, "The CMIS Solution for Tenix's M113 Program, Matrix One Innovation Summit 2004. See slide 10.

- Document management, including versioning and access to related documents such as engineering change orders, drawings, and specifications
- Support for user annotations

This diagram prepared by Dr. Hall warrants three observations: [1] TeraText, like other search and content processing systems, is one component in a far larger system. For TeraText to integrate into complex defense systems, TeraText can be tailored to specialized environments involving other large, sophisticated systems. [2] TeraText functions as the accessible memory for a network of distributed systems. Search is important as an access point, but search is essentially a utility. [3] TeraText was one of the first vendors to provide



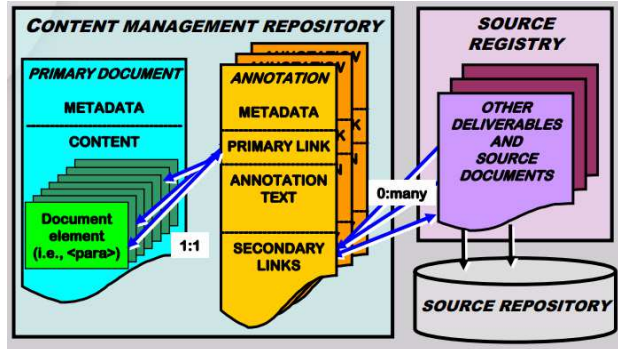
an enterprise-class, secure federated content and text management system with analytics, connectors, and near real-time search. Many of today's "innovators" are, in actuality, TeraText followers. Note: I have highlighted the TeraText component.

- Content and data security
- Federated results when data were not resident in the TeraText repository
- An operational knowledge model that provides a framework in which the individual system components and workflows are collected and organized.

The ANZAC implementation is an important milestone in search and content processing. The system delivers information access via tight integration with a content repository, multiple distributed data sources, and a number of different enterprise systems.

Unlike vendors who "glue" a search system to existing systems, TeraText implemented middleware that allowed new and existing systems to interact

in a transparent, largely seamless way because a larger framework was used to rationalize information across a distributed organization.



The integration of TeraText content functions with search makes it easy for a user to retrieve a base document, related documents, and annotations to any individual document. See page 16 of "Document-Based Knowledge Management in Global Engineering and Manufacturing Projects," no date.

TeraText Products

TeraText is a database system developed from first principles to process and manage hierarchically structured and linked information. The system implements a repository with the content stored in a native XML database. The XML works with many different structural formats (SGML, HTML, RTF, MARC, or binary formats at the file level.

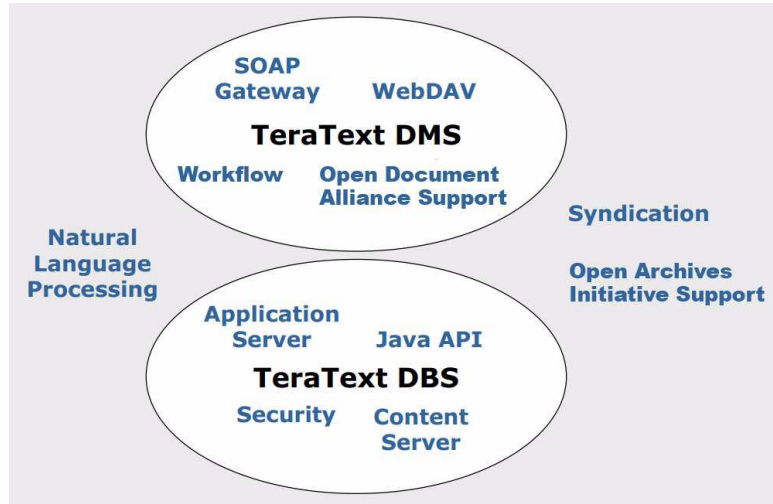
TeraText supports parallel searching of multiple databases; Google-like searching with phrase, synonym and spell-checking features; results sorting options; an alerting service and options to add content to collections, and online commerce.

For indexing hierarchically structured textual knowledge in SGML or XML, TeraText's indexing capability is able to respond to queries more quickly than object oriented or relational databases.

TeraText can index terabytes of text daily concurrently processing queries against the same multiple terabyte databases.

“What the company's [SAIC] really all about is the inspiration of individual entrepreneurs.... The company is extremely opportunistic.”—Steve Rizzi, SAIC VP, “Uncle Same Keeps SAIC on Call for Top Tasks, Baltimore Sun, October 26, 2003. <http://bit.ly/1hGKvVK>

A customer licenses the core system that includes the indexing repository server, a Web server, administrative tools, security and logging servers, and the proprietary application scripting language ACE.



TeraText's products are mixed and matched components of the two core systems: The database and the document management system.

A licensee can snap in the TeraText document management system. Workflow capability is included with the document management component. Regardless of product configuration, TeraText can:

- Capture/manage source documents
- Author, review, publish workflow
- Versioning and change management
- Validate elements of content
- Maintain
- Index, search and retrieve
- Support reuse

Key Products

SAIC and TeraText offer a number of products. The products incorporate a text database system with native XML support. The distributed architecture is engineered to deliver speed and scalability. TeraText asserts that it provides large scale capability that can process large flows of information and data in real time. (SAIC does not define “real time,” nor does the company provide response time data.) The idea is that any of the builds provide a flexible foundation for “a new class of solutions to complex information prob-

lems, accessible from a wide range of application development environments.”⁹ TeraText was one of the first, large-scale integrated information systems to support standards; for example, Z39.50 for searching and retrieving text from different sources. Some of the TeraText products not discussed in this report are:

- TeraText Content Server application (one or more instances)
- TeraText Advanced Search Interface Server application (single instance)
- TeraText Command Line Interface Server application (single instance)
- TeraText APIs
- TeraText Application Server application
- TeraText Database Design Interface Server application
- TeraText Security and Logging Server application
- TeraText Boot Server application
- TeraText Directory Server application

Brief descriptions of five major TeraText solutions appear in the paragraphs below.

TeraText Database System (DBS)

is a high-performance platform for storing, searching, and managing information. It was built from the ground up to be optimized for text. Its speed, scalability, and distributed architecture are unique. It was designed for semi-structured data in standards such as XML and SGML, but will store and index files in one of about 370 different standard office formats. Unlike relational database management systems (RDBMS), the TeraText DBS stores documents natively, handles large data collections, and provides complex text operators. The TeraText DBS differs from other XML databases as it can store and index multiple document types (not just XML), is highly scalable, and performs distributed searches. The TeraText DBS does more than text search engines as it takes advantage of structured fields, and stores and manages content, not just indexing it.¹⁰

TeraText Document Management System (DMS)

Utilizing the power of the DBS, the DMS delivers a platform for managing your corporate document assets. It provides flexible versioning models (including support for fragmentation of XML or SGML documents and sep-

⁹. See Alan Kent's 2002 presentation "TeraText Technical Overview." The document is no longer available on the InQuirion or SAIC Web site as of October 2007.

¹⁰. InQuirion offered an eCommerce capability. This shopping cart function allowed item selection, ordering, etc.

arate versioning of the fragments), easy definition of additional fields (meta-data) stored with the document or in the document, the ability to browse and search documents and components at a specified time, the ability to track which version of a component belongs with which versions of documents, and the ability to manage complex document life cycle or business process rules in a workflow management environment.

Searchable Archive for Files and Email (SAFE)

Safe is an enterprise-class search platform that enables government agencies and corporations to archive, store and search emails, files, and attachments in real time. TeraText Safe archives every email, both incoming and outgoing, along with its attachments immediately and automatically in a safe designed to be tamper-proof. The massively scalable archive is fully accessible from the time the system stores an email. TeraText Safe uses precision search capabilities originally developed for intelligence agencies Initiative Protocol for Metadata harvesting (OAI-PMH) standards to package information about changes on a Web site.

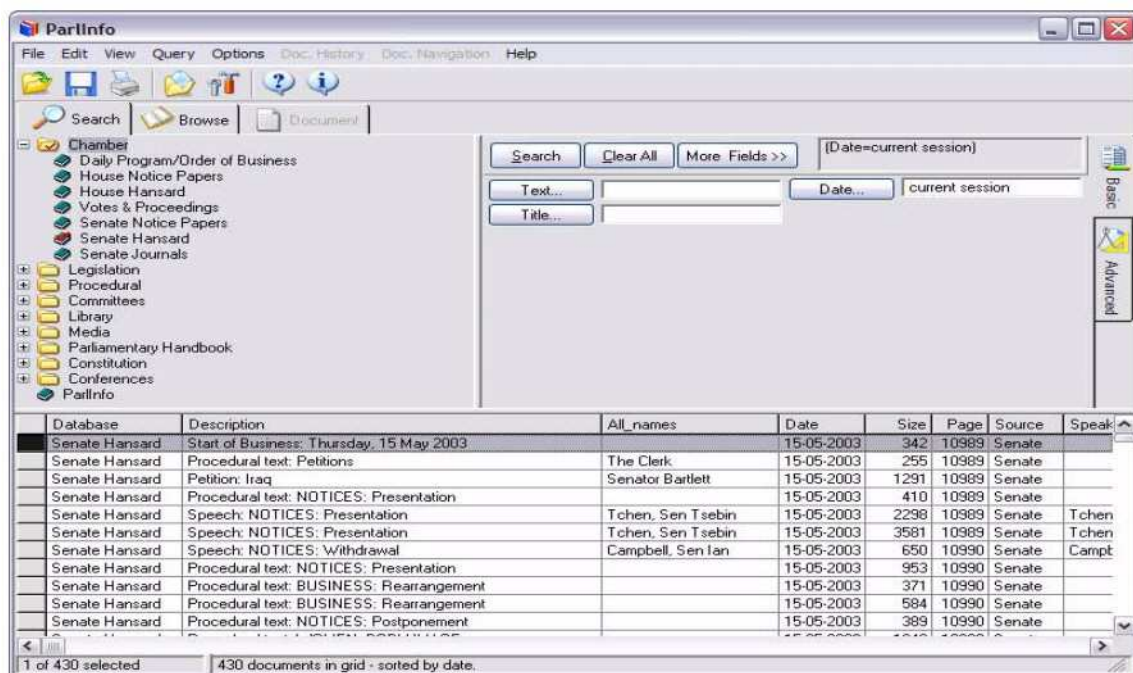
TeraText Safe helps a licensing organization meet compliance requirements by providing a tool set to collect and maintain a searchable archive. This search capability can also increase productivity by delivering information to users, wherever they can access a web browser or PDA. Fast search combined with powerful security features such as access control and audit logging support information security, forensics, and e-Discovery activities to reduce the IT staff's workload of manually searching email back up tapes.

TeraText Safe is a new product built on the proven TeraText Database System. Originally developed for the intelligence community, TeraText DBS delivers a proven platform for storing and searching large volumes of information. From this core capability, SAIC developed an enterprise-class tool to store and search billions of emails, files, and attachments. Years of email. Seconds to find.

- Enterprise scalability
- Real-time search
- Protect privacy of users and security of information
- Mobile — web or PDA access
- Compatible with existing email servers — Exchange, SendMail, PostFix
- Federated architecture — store your data locally but search globally
- Rapid deployment for efficient implementation

TeraText for Legislation

This product adds a set of tools to the document management system (DMS) to help manage the process of drafting and publishing legislation and assists governments (including Canada, Australia, and Papua-New Guinea) to manage and even automate many of the drafting and publishing steps for these very important documents. This tool set is the basis of the very successful EnAct/TeraText system deployed in Tasmania and Papua New Guinea. The TeraText DBS satisfies numerous guidelines that assist government agencies to satisfy their Section 508 compliance requirements. A Voluntary Product Accessibility Template for the TeraText DBS can be emailed on request. Additional information is available on Section 508 Solutions.



An interface for a legislative implementation of TeraText. Licensees can configure the search functions and the layout of the controls to meet the users' requirements. Source: Shirley White, et al, "All Aboard ParInfo Search," Government of Australia, 2006.

“Among the fruits of that vision are two of SAIC’s most technically advanced products: TeraText and Latent Semantic Indexing (LSI). They’re data-mining programs—some of the most powerful in existence. Both are central to enabling intelligence agencies to sift the immense volumes of data they now collect.” —Paul Kaihia, “In the Company of Spies,” *Business 2.0 Magazine*, May 1, 2003.

Information Analyst Support System (IASS)

This is the TeraText real time information system to support decision makers on a 24x7 operational status.¹¹

The system handles data, documents, and fielded data. Billions of records with millions of new records added each day are within the capacity of the system. In addition to the native TeraText data management system, IASS can support different database management systems. The idea is that TeraText uses the optimal RDBMS for each particular data set.

IASS incorporates a messaging architecture. The content is “tied together.” XML is used for flexibility and interoperability.

IASS incorporates a comprehensive set of analytic tools. TeraText delivers what Steve Rizzi, vice president of TeraText calls “tailored data fusion.” An exemplary IASS can support 4,000 users, five billion XML documents, and deliver outputs with two second latency. The total number of sources “fused” is 186 with 60 different data types, 84 separate RDBMS tables, and 66 text databases. The user does not have to specify a data set or collection.

TeraText is able to address the inadequacies of traditional RDBMS systems.

The IASS is able to:

- Add new file types and sources as needed
- Scale to handle growing amounts of information
- Load text documents that can arrive in bursts
- Update the searchable index within seconds of receiving a document
- Support multiple languages
- Permit complex searches; that is, standing queries, Boolean, natural language, and fielded inputs.

TeraText can transform input content into XML. Some XML documents exhibit variances in encoding; other data are processed via TeraText filters.

The search function is tightly coupled to XML. TeraText is a type of middleware. There are databases and content repositories. There is a search component. The pivot or hub of IASS connects the content and the content processing function to the user requiring search results or outputs.

The architecture is distributed and uses commodity hardware. TeraText calls its hardware approach “Data Power.” The servers sort result sets and handle the presentation of query results.

¹¹. This information appears in a presentation from 2002 or 2003 by Steve Rizzi, the vice president in TeraText’s Annapolis, Maryland office. See <http://bit.ly/1f9EWNs>. The material is part of SAIC’s Lighthouse information program.

TeraText includes a “hardware assist” called DataPower XA-35. This accelerates the compression and decompression of XML content. This engineering innovation reduces query latency often associated with processing billions of documents. The payoff of this hardware/software combination is:

- DataPower XA-35 provides a performance increase of XSLT transformations of 10 to 50 times
- The solution integrates with industry standard load balancing software and hardware
- DataPower XA-35 supports World Wide Web Consortium standards related to XML processing; for example, the built in XML parser, support for XPath, and XSLT
- The system includes a graphical interface
- The system operates without spinning discs
- The system supports three modes: Co-processor, proxy, and in-line.

The system incorporates record-level security with role-based access control.

Search and Retrieval

TeraText offers a range of text search capabilities. Users or scripts can query the indexes by words, phrases, word adjacency, word distance, sentence, paragraph, fuzzy match, and lemmatization. Results can be displayed to return selective information, different delivery formats, different sorting options, and other operations.

TeraText generates bibliographic metadata for content processed by the system; for example, author, subject, and title. The system maps the bibliographic data to concepts in the original data forms; for example, an “author” may be the “From” of an email and “Title” may be the subject of an email or the title of a document.

TeraText supports Boolean, natural language, fielded, and stored queries.

The basic search functions include:

- Proximity operations; that is, NEAR, WITHIN, SAME, ORDER
- Range operators; for example, STRING and NUMERIC
- Fuzzy match
- Lemmatization (stemming)
- Limit operations

- Custom case folding (ignore case or acknowledge case), punctuation stripping, transformations, expansions (an acronym can be expanded), modular lexing¹², and relevancy algorithms

The screenshot displays a search interface with a blue header and navigation tabs. The search results are summarized as follows:

- Search Report:** Total of 528 Results Matched For Search: **shipping** Using Mode: SPECIFIC
- Within:** Fonds Descriptions, Textual Records, Visual Records, Moving Image Records, Cartographic Records, Indexes: Library, Web Pages, Combined Genealogy Indexes
- Fonds Descriptions Index:** Summarized results 1 - 10 (of 14 total matches within this index).

The results are presented in a table with four entries:

Item	Call Number	Creator-Author	Title
1	MS-2701	R.P. Rithet and Company	R.P. Rithet and Company fonds
2	MS-1969; MS-1696; P/F/Si3	Silver Spring Brewery ; Coast Breweries	Silver Spring Brewery fonds
3	MS-1247	Cascade Transportation Company	Cascade Transportation Company fonds
4	MS-1083	Douglas Lake Cattle Company	Douglas Lake Cattle Company fonds

This implementation by British Columbia Archives provides different search features on tabs: basic and advanced search across text and images. The user can set preferences for the TeraText search system and access help with the command syntax. Note that the TeraText system does not generate summaries of the documents in this result screen example.

- Support for wildcards in queries; that is, #, #n,?, and ?n
- Hit highlighting
- Saved searches and result set caching
- Pass a query to return changes since a certain date
- De-duplication.

¹² TeraText jargon for translation components that a licensee can “snap in” to handle a particular language operation.

”Most information in organizations resides in semi-structured, primarily textual documents, not in structured, organizational repositories. The volume of reports, submissions, emails, contracts, policy documents and similar documents in most large organizations is beyond the capacity of most systems. The TeraText suite of products provides the foundation for a range of solutions for large-volume, high-complexity collections of documents.”—About TeraText, SAIC

The syntax for TeraText search looks like this:

To **search** for the terms libraries or librarian (but not library or librarians) in the controlled terms index, and then all words beginning with automat that appear in the abstract field:

```
FIND ct=librar### AND ab=automat?
```

To find the intersection between the results of two previous searches:

```
FIND S3 AND S4
```

To **search** that combines a result set with a **search** for the term university:

```
FIND University and S4
```

To return records 1, 3, and 5 from a result set.

```
FIND S3(1,3,5)
```

To return the first 100 records from a previous result set.

```
FIND S3(1-100)
```

To return all but the first 100 records from a result set.

```
FIND S3(101-)
```

A single conceptual model is layered over the content that TeraText processes. Processed content is stored in a repository and indexes are generated for key words, concepts, and metadata, including versioning information.

A single query can be applied to multiple collections of content.¹³ Because the processed information is maintained in its original form, the semantics of the source are preserved. Within a collection of content, multiple models can be supported. A licensee can develop an internal standard for a content type.

Licensees can modify the system to meet specific indexing and query processing functions using C++, Microsoft Dot Net, Java and the TeraText application programming interfaces. The system supports Unicode and multi-lingual content. Pre-defined word parsers are part of the system’s core functionality.

TeraText offers an index term exploration function. A user can examine terms in indexes. The user can form an opinion about the words used in the indexed content. Among the data available are:

- Term frequencies
- Misspellings or spelling variations for terms
- The TeraText scripting language allows the licensee to define complex extraction rules from the support data types, not just XML

A user can determine how and what words contributed to a result of a complex query.

¹³. The system supports ISO 23950, Z39,50, and GILS (Global Information Locator Service).

Indexing

The TeraText system uses compressed, inverted file indexes.¹⁴ Dr. Ron Sacks-Davis refined an insight hit upon in the 1970s. The precursor to TeraText (InQuirion) is the result of years of university-supported research. The approach reduces disk traffic and are CPU efficient. TeraText incorporates advanced algorithms to improve query performance while sidestepping unnecessary decompression of data. TeraText's indexes with full term position information using 1 percent more disk space than the original, uncompressed source information. Some competitive approaches require ten times the index storage space. TeraText is storage efficient. In 2002, according to Alan Kent, InQuirion's chief technology officer:¹⁵

The combination of high performance indexing and distribution has been used in a production environment to load 400 1 kilobyte XML documents per second, fully indexed, support a one billion record collection, and deliver search times of two to three seconds. [These benchmarks are from] a content Server on a single CPU that can easily handle 10 to 100 gigabyte collections with sub-second queries.

Metadata can be extracted at query time so that metadata and content are synchronized. The system also supports embedding the metadata in a content object, but this approach increases storage overhead.

The TeraText Content Server supports XML natively as a single field type of a physical record. The field types supported by TeraText include:

- Scalar types such as integer, float, ASCII text, Unicode text, binary, datetime
- Complex types such as XML, SGML, MARC, RTF, and PDF

An important architectural ability of the TeraText system is that as new data formats become accepted, support can be added without the need to change the TeraText semantic search mode.

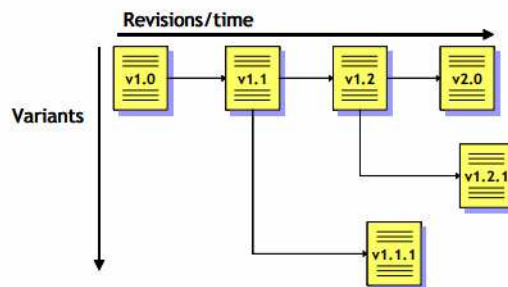
¹⁴. The TeraText databases can be queried like a database. This allows a licensee to obtain analytics about the index, entities, and documents.

¹⁵. Alan Kent, "TeraText Technical Overview," InQuirion (TeraText), 2002. No longer in print.

“InQuirion employed 25 full-time developers and continued to grow its business until its eventual outright purchase by SAIC in October 2005. Since then TeraText has been put to considerable use by US intelligence agencies in pursuit of terrorists. The system's applications for national security are incredibly significant, particularly in the global war on terror, and in some instances could be the difference between life and death.” —
Newsletter from America: TeraText Leads the Way in the USA, February 2007. Source: <http://www.australiandefence.com.au>

Point in Time Searching

The user of a TeraText system can run a query for a specific point in time. The idea is that a document that may undergo numerous revisions exists in different versions during the drafting process.



Systems such as Fulcrum Technologies or MarkLogic do not offer native support for point-in-time search. TeraText offers this function to licensees.

The user can retrieve a document variant from the TeraText repository showing the contents as they existed at the date or point in time the user specifies. Within the TeraText architecture, pointers and metadata for documents and their changes at specified times are retained. The point-in-time document version is reconstituted from the XML in the TeraText repository.

Metadata

TeraText performs metadata harvesting. The system supports OAI (the Open Archives Initiative) protocol. The OAI method allows data to be aggregated from multiple smaller collections into larger collections. These aggregations can then be searched. According to TeraText, the reason for harvesting metadata is that while distributing queries is good, performance can suffer if too many servers are involved. TeraText supports:

- Scalable central collections
- Distributed collections
- Harvesting aggregated collections.

Natural Language Processing

The TeraText NLP system adds new concepts to text that were not explicitly mentioned in the source document. The system performs Latent Semantic Indexing. SAIC has invested in technology, which performs advanced concept identification via deal for software developed, in part, by Bell Communications Research.¹⁶ Latent Semantic Indexing (LSI) is a statistical information retrieval method that searches text collections for specific concepts. LSI technology makes it possible to reduce the dimensionality of cer-

tain metadata. SAIC has not made publicly available details of the systems and methods used to enrich metadata via the firm's LSI procedures.

Security

TeraText offers one of the most robust content processing security fabrics available. Oracle Secure Enterprise Search asserts that it has robust security; however, TeraText is engineered to ensure that the system can conform to any client's security infrastructure. TeraText provides tools to manage permissions, authentications, and security for the system and authorized users. TeraText integrates with existing organizational authentication databases.

A standard TeraText system supports record and element level security. The system can hide records or fields of records. Teratext can strip out unauthorized parts of XML and SGML documents. There is content server authentication and the mechanism is configurable.

The Content Server also supports comprehensive logging of system activity so that the licensee has information about:

- Information about operations (searches etc.) and who performed them
- Information about records returned and to whom, allowing copyright etc. obligations to be observed

A number of specialized functions are available; for example, "P.AUTHORIZED_USERS and T.MASQUERADE. For clients requiring TeraText's advanced security functionality, a document called *SAIC TeraText DBS Security Target* is available to authorized prospects and clients.

Application Server

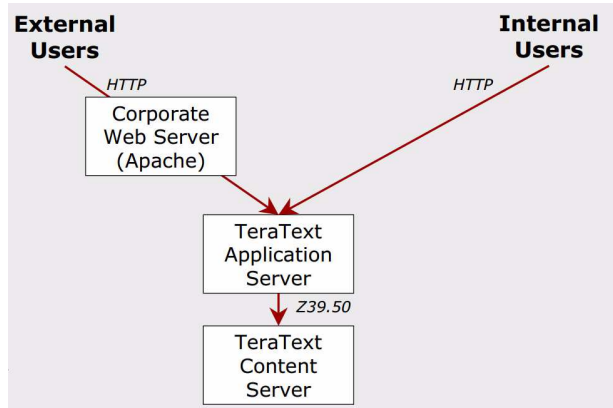
To simplify application development, the TeraText product range also includes an application server with access to the numerous TeraText libraries; for example, log file analysis, interface, analytics, relevance, content assembly and display, etc. The TeraText Application Server is a high performance, multi-threaded server. In production with hundreds, up to thousands, of concurrent users.

The Content Server allows applications to be developed in any programming language with a Z39.50 client library available. The TeraText content server is shipped with C++ and Java APIs.

The server supports HTTP allowing applications to be developed with a Web interface. The focus is applications with a Web interface, not a Web site. The

¹⁶. SAIC acquired the LSI technology from Bellcore (Telcordia) in 1997. The LSI technology was spun out of SAIC under the "Content Analyst" brand. See <http://bit.ly/Lh7wnb>

server can be accessed directly by Web browsers or linked into an existing site using Apache's reverse proxy support.



The Content Server is a text database system with support for XML, not an XML database system with support for text. TeraText implements a format independent conceptual model used for searching against, rather than XML. This gives the ability to search other data formats. TeraText uses the SML Path Language to map XML data to the TeraText-generated conceptual model.

Applications can be developed in ACE, a TeraText Scripting Language. ACE is an object oriented language with integrated support for the TeraText Content Server. Wolf, another TeraText component, is also provided to reduce development and deployment time. Wolf is an object oriented programming methodology. With WOLF, a licensee can build large-scale applications. WOLF allows multiple developers to build modules concurrently.¹⁷

There are two forms of integration in the TeraText application server: ODBC access of a Z39.50 server via SIMBA or Z39.50 access of an ODBC server, but integration is required. A third party software like ZBig can be used. TeraText recommends harvesting data from an ODBC database into a TeraText content server. The TeraText Application Server allows applications to

¹⁷. See Alan Kent, TeraText Technical Overview, InQuirion, 2002, page 60.

access both Z39.50 and ODBC repositories, but does not perform protocol translations.

The screenshot shows a web application interface with a navigation menu on the left and search results in the main area. The navigation menu includes links for Home, Support, Contact Center, Developer's Corner, Related Sites, Sample Sites, Demo, Press Releases, Events, and Site Map. The main area displays the title 'Biographical Directory of the U.S. Congress' and search results for 'content=(engineer)'. The results are presented in a table with columns for Member Name, Position(s), Born, and Died. The table lists 10 members, including James George Abourezk, Walter Hugh Albaugh, Truman Heminway Aldrich, John Miller Baer, John Courts Bagby, Michael Bilirakis, Walter Preston Brownlow, Felix Campbell, William Leighton Carss, and Bertram Tracy Clayton. The interface also shows pagination controls indicating 83 results and a 'next' link.

Member Name	Position(s)	Born	Died
ABOUREZK, James George	Representative, SD Senator, SD	1931	
ALBAUGH, Walter Hugh	Representative, OH	1890	1942
ALDRICH, Truman Heminway	Representative, AL	1848	1932
BAER, John Miller	Representative, ND	1886	1970
BAGBY, John Courts	Representative, IL	1819	1896
BILIRAKIS, Michael	Representative, FL	1930	
BROWNLOW, Walter Preston	Representative, TN	1851	1910
CAMPBELL, Felix	Representative, NY	1829	1902
CARSS, William Leighton	Representative, MN	1865	1931
CLAYTON, Bertram Tracy	Representative, NY	1862	1918

TeraText makes it possible to provide a user-friendly, browser-based interface for end users. The screen shot shows the results generated from a directory corpus.

Performance

TeraText uses Z39.50 standards. The company asserts that this approach, plus TeraText distributed architecture, allows low-latency performance.

For licensees using SOAP, TeraText places SOAP gateways near applications to improve throughput.

According to a report from the Australian National Library:¹⁸

The new database has much improved performance characteristics and on its current hardware platform can support up to 350 users

¹⁸. Tony Boston, Bemal Rajapatirana, and Roxanne Missingham, "Libraries Australia: Simplifying the Search Experience," National Library of Australia Staff Papers, 2005, page 3 at <http://bit.ly/1apsFRY>

doing simple searches with below two second search and present times and up to about 1,000 concurrent users with below five second search and present times. This represents a database throughput of about 100 searches per second. The increase in response times by number of concurrent users is approximately linear.

Standards Support

The TeraText Database supports the following standards:¹⁹

- ANSI Z39.50
- Common Command Language/Z39.58
- Extensible Markup Language
- Open Database Connectivity
- Rich Text Format
- Standard Generalized Markup Language
- Unicode.

Programming Languages

TeraText supports

- ACE (Algebraic and Calculus Expressions)
- C++
- Java.
- Visual Basic
- Wolf (a proprietary language).

¹⁹. See <https://www.escidoc.org/JSPWiki/Diff.jsp?page=TeraText-DBS&r1=2&r2=1>

ArnoldIT Opinion

The TeraText system is a platform purpose-built to manage large repositories of text-based information, build information access applications, and support sophisticated operations on the data processed by the system. At the heart is a high performance text database system capable of scaling on multiple axes (single large databases, or via distributed architectures). TeraText provided a comprehensive information framework decades before Autonomy, Convera, Fast Search & Transfer, and Oracle offered similar functionality.

TeraText is designed to support a wide range of information infrastructures, allowing searching of both heterogeneous and homogenous collections of data. Unlike the Web-centric and somewhat limited Vivisimo, TeraText operates on federated data and unstructured information. Its basic design principles are firmly grounded in the expertise of the library community, but packaged for commercial application. TeraText is one of the high-water marks in the information access sector. Instead of seeking short cuts, the system applies a conceptual model to data and unstructured information. The system then operates on those content objects and their constituent elements. Comparing the Google Search Appliance to TeraText and its SAIC's allied Latent Semantic Indexing technology is like matching a high school wrestler with a world champion gold medalist in Greco-Roman wrestling.

The TeraText system is designed to provide a database solution for text. Many companies say they support XML. TeraText stands as one of the first search vendors to build on the precepts of SGML and XML. Although non-text "objects" such as Portable Document Format (PDF) files and multimedia are supported, the system requires that a record with metadata about the object be included in the TeraText index and database.

XML systems can provide a wide range of functionality, but they come with some storage, system overhead, and response issues. TeraText has been improving its performance with each release. However, in order to provide the type of response time that an average employee equates with Yahoo!, for example, a robust infrastructure and appropriate resources are required. Technical resources need to maintain, troubleshoot, and adjust the system to meet the specific requirements of the users served by the system, particularly in mission-critical deployments such as military intelligence.

Storage costs can be an issue, particularly when large volumes of data are pushed through the content processing system. TeraText is designed to provide a comprehensive document creation, management, search, and retrieval solution. TeraText uses a repository model and indexing overhead adds to the storage calculation.

The ideal application for TeraText is for a search engine that will operate with the complete TeraText enterprise information authoring, versioning, storage, and access environment.

Table 2: TeraText Search Checklist

Attribute	TeraText Asserts	ArnoldIT Comment
1 Platform	Linux 32 and Linux 64, Unix, HP-UX (Itanium), Microsoft Windows	
2 Keyword search	Yes	Boolean supported
3 Text mining	Analytic functions are available	TeraText offers a version called Information Analyst Support System
4 Automated indexing	Yes	System performs key word and conceptual tagging, metadata extraction, and advanced processes like latent semantic indexing
5 Personalization	No	Licensees can create role-based interfaces via standing queries or auto-generated reports
6 Workflow	Workflow components are provided with the document management system	Alerts are supported so new content can be pushed to authorized users
7 Interface	Support for browser-based access is available	When TeraText is integrated into large-scale systems, TeraText functions can be accessed via the enterprise application interface
8 Hosted service	No	
9 Administrative interface and tools	Configuration files and some graphical interfaces are provided	Licensees will require SAIC training and knowledge of standard programming languages
10 Application programming interface	Yes	APIs are available for document management, query processing, and middleware components
11 Professional services	Yes	Licensing and services are available from SAIC and partners
12 Security	Role-based, content access, and extensibility to almost any type of security environment	Many of TeraText licensees are involved in defense and intelligence. Robust support is provided by the platform
13 Connectors	Databases, Rich Text Format, Lotus Domino. Support for 200 files types provided with the system	Licensees can create filters to handle almost any file type
14 Support for structured data	Yes	Database support and support for structured content
15 Relevance ranking	Yes	System includes controls to tune the ranking
16 Video	Metadata only	Content objects can be stored on the distributed network
17 Federated search	Yes	TeraText calls this function "tailored data fusion"
18 Fielded search	Yes	
19 Content crawler	Yes	
20 Price	Begins at \$5,000 but large-scale installations can hit seven figures	Custom price quote required

A typical installation to handle Canadian legislative information will cost \$4 to \$6 million to set up an additional investment each year in support and maintenance. TeraText is used for even larger-scale installations for various military and defense solutions. However, libraries can license TeraText and use the system to provide citizen access to collections at a fraction of the cost of a large-scale war fighting implementation. TeraText is an enterprise solution that works best when the entire TeraText environment is used as middleware as illustrated in the ANZAC schematic presented elsewhere in this document.

Anticipated Benefits

TeraText allows a licensee to create a warehouse of content. Once in the warehouse database, the TeraText system can be used to deliver fine-grained, secure access to specific documents or content objects in the warehouse. The warehouse can reconstruct content, apply algorithmic methods to the content, components, and metadata, and support standard Boolean and Google-style queries.

The benefits of the TeraText approach include:

- Making it possible to provide thousands of users to access, share and retrieve documents in near real time
- Scaling to meet the needs of a licensee's business, allowing storage of a few gigabytes of text to petabytes text
- Providing a robust solution that is ready-to-deploy, not a partial system that the vendor will code on the fly.

In sum, the product is best suited for text corpuses that can be normalized into XML, in situations that require substantial concurrent usage, an approach that “puts all the content in one place”, and warrants specific security and access controls.

Possible Drawbacks

The database designer must define semantic mapping from physical representations to conceptual search model. XPath can be used for XML. TeraText includes a built in flexible scripting language to define complex extraction rules from the supported data types.

Other possible drawbacks of the TeraText approach include:

- TeraText is a solution that imposes a conceptual model on the content processed by the system. Organizations preferring a free-style approach to content access may be a poor match for the TeraText approach

- The easiest, fastest deployment path requires licensing the complete suite of products and hiring TeraText to be involved with the implementation as engineers and advisers
- Content not produced within the authoring module must be normalized into a structured format and imported into the TeraText database system, potentially creating a need for additional storage
- TeraText is not a Google-style appliance or a collection of open source components. TeraText, like other high-end enterprise systems, requires management commitment, funding, and resources to perform at optimal levels.

Net Net

TeraText is truly industrial strength and has been in the commercial channel for decades. The system supports a wide range of information infrastructures allowing searching of both heterogeneous and homogenous collections of data. The TeraText DBS is flexible enough to be used in a wide range of situations, including integrating third-party components for visualization and text mining.

The system exhibits acceptable performance under heavy loads. Licensees are not tied to a proprietary data formats. The lingua franca of TeraText is XML.

But TeraText is beginning to show its age, however. Like the Autonomy technology acquired from Verity in 2005, TeraText lacks some of the features available from companies that are focused on processing social media or telemetry data.

A large-scale installation—for example, the Australian Navy or the British Columbia archives—can hit seven figures with licensing fees, infrastructure, customization, and maintenance. Such systems are designed to scale to the multi-terabyte range.

Stephen E Arnold

Minor edits to a rough draft on February 17, 2014

Annex: SAIC Subsidiaries 2007

TeraText technology may be available from these SAIC units as listed at <http://www.secinfo.com/d14D5a.u2dTy.d.htm>

AMSEC Corporation
AMSEC LLC
AMSEC Subsidiary Holding Corp.
B D Systems, Inc.
Calanais Pension Trustee Co. Ltd.
Campus Point Realty Corporation
Eagan, McAllister Associates, Inc.
EAI Corporation
Hicks & Associates, Inc.
InQuirion Pty Limited
JMD Development Corporation
MEDPROTECT LLC
Opta Ltd.
Planning Consultants, Incorporated
SAIC (Bermuda) Ltd.
SAIC Engineering, Inc.
SAIC Engineering of North Carolina, Inc.
SAIC Engineering of Ohio, Inc.
SAIC Europe Limited
SAIC-Frederick, Inc.
SAIC Global Technology Corporation
SAIC Limited
SAIC Pty Ltd.
SAIC Services, Inc.
SAIC Venture Capital Corporation
Science Applications International Corporation
Science Applications International (Barbados) Corporation
Science Applications International Corporation (SAIC Canada)
Science Applications International Corporation de Venezuela, S.A.
Science Applications International, Europe S.A.R.L.
Science Applications International Germany GmbH
Varec Holdings, Inc.
Varec, Inc.
VCC Holdings Corp.