

Siderean Software (Offline 2008)

© 2013 by Stephen E. Arnold, www.arnoldit.com

Siderean offered a semantic search system. Enterprise search embraced semantic technology, triples, rich metadata, and standards as competitors pursued proprietary solutions.

Author's note: This is an unpublished, preliminary draft of a description recycled in various monographs and articles I have written.

Siderean opened for business in 2001. The company sought to use semantic methods to deliver a more effective enterprise search solution. Siderean was one of the early adopters of Lucene for keyword retrieval. The firm was an early entrant into the SGML/XML data management approach to information retrieval. The company's approach combined traditional content processing with semantic methods closely hooked to SGML (Structured Generalized Markup Language) and XML (Extensible Markup Language) to deliver metadata, lists of suggested or "See Also" content, facets, and rudimentary reports about the frequency of certain terms in the processed content. The "rich metatagging" gave Siderean an early lead in what has become the "metadata business."

The company's focus on semantics, data management, and facets proved to be a challenge for the firm's marketers. The company attempted to educate potential licensees about the meaning of structured documents and semantic concepts. Throughout its existence, the company sought funding to develop its sophisticated system. By late 2007, the company was faced with growing competition from a number of companies offering semantic functions that could be added to an existing search system. In 2008, Siderean went offline.


This information is a rough draft and is frozen.

Introduction

Siderean’s Seamark Navigator is a data management and faceted-search system that supports the World Wide Web Consortium’s standards for tagging content. On the surface, Siderean’s search solution is similar to Endeca’s implementation of “guided navigation.” However, Siderean’s engineers have woven XML and the principles of the Semantic Web into the innermost workings of the Seamark Navigator system. Few search vendors have embraced the concepts of the Semantic Web with such fervor.

As more content becomes available that uses the type of structure and tags recommended by the W3C or in structured formats like XML, Siderean could be positioned to deliver the full benefits of semantic methods for search. Siderean’s approach to search requires that content be structured. An unstructured document such as an email with a Word document attached to it must have metadata assigned to it in some way. Siderean uses the Resource Description Framework or RDF in its content processing subsystem.

Table 1: Siderean Seamark Navigator: A Bird’s Eye View

Product Thumbnail	
1 Search Brand	Seamark Navigator
2 OS Supported	Solaris, Linux, Microsoft
3 Est License Fee	Pricing begins at \$75,000
4 Functions	Provide faceted search for structured and unstructured information
5 Claimed Features	Uses features that exploit the World Wide Web Consortium promulgations for the Semantic Web and the Simple Knowledge Organization System for content that carries metatags comparable to field names in database systems
6 Downsides	Tag maintenance and resource demands may be a concern in some implementations
7 Similar To	Dieselpoint, Endeca, Fast Search, MarkLogic
Product Close Up 	The semantic system can perform deep indexing of unstructured documents. For content in Extensible Markup Language or a traditional database, Siderean’s system can deliver traditional search and retrieval, content parsing or “slicing and dicing” so reports or compilations can be generated, and clustered and faceted views of related information. The company emphasizes the importance of semantic methods, relying on jargon like “triples,” OWL (Web ontology language), and tagging (indexing).

Like Endeca, Siderean emphasized the advantages of “faceted search”; that is, showing a person who runs a query, hot links to related content. This “See Also” function exposes information that otherwise might not be located by

the user. Siderean demonstrates its approach to faceting at <http://www.siderean.com/fooddemo.jsp>.

Siderean's content processing assumes that content will be provided to the system in one of the formats supported. These include content in standard database tables, SGML, XML, or a tagged stream like information in an RSS feed.

Siderean's faceting system does not require a customized list of terms or a dictionary of entities. Seamark Navigator discovers categories or "tags" by analyzing the information in the mark up, database tables, and content. The system also extracts a file's name, the date attached to the file, the document's time stamp, and information from indexing words and phrases in the documents and tables. Siderean uses all of these items and groups related information in meaningful categories. Like Endeca's system, the Siderean content processing subsystem requires sufficient resources. Without robust processing power and disc space, the system's throughput can bog down.

Consider the figure below, which comes from Siderean's online demonstration of its search system:

The screenshot shows the Seamark search interface. At the top left is the logo "seamark™" with the tagline "The Enterprise Navigator For Digital Information & Components". To the right is a search box with a "Go" button. Below the search bar, it displays "13,204 items". The interface is divided into four columns of faceted navigation options, each with a subcategory list and a "more" link.

by Course	by Season	by Main Ingredients	by Cooking method	by Cuisine	by Source
Appetizers 1222	Christmas 345	Cheese 1712	Advance 2016	American 1497	Bon Appetit 6444
Bread 497	Easter 137	Citrus 1184	Bake 3768	French 566	Epicurious 98
Condiments 685	Fall 3048	Dairy 2602	Broil 332	Greek 138	Epicurious Television 18
Desserts 3093	New Year's Day 108	Eggs 1190	Fry 305	Italian 769	Gourmet 5681
Hors d'Oeuvres 724	Picnics 179	Fruits 2099	Grill 609	Jewish 168	House & Garden 276
Main dish 3958	Spring 3168	Herbs 2147	No-cook 493	Kid-friendly 486	Jewish Cooking in America 21
Salads 1145	Summer 3060	Nuts 1356	Quick 2791	Low-fat 590	La Parilla: The Mexican Grill 18
Side dish 1747	Superbowl 217	Poultry 1125	Roast 652	Meatless 1501	Letter from France 56
Soup 747	Thanksgiving 657	Tomatoes 1111	Saute 1008	Mediterranean 163	Red, White & Greens 15
Vegetables 746	Winter 3063	Vegetables 2705	Slow-cook 457	Mexican 330	Wisdom of the Chinese Kitchen 13
6 more	3 more	20 more	5 more	7 more	296 more

The difference between a bare Google-style interface and the Siderean interface is dramatic. There is a search box, but the system displays the categories into which the indexed content "fits". The idea is that a user can see at a glance that this recipe database contains 13,204 items. The recipes can be explored by Course, Season, or any of the four other top-level categories.

Under each category, subcategories appear with the number of documents in each category. For example, under Course, there are entries for Appetizers with 1,222 documents, Bread with 407 documents, and so on.

“Each of these pinpoint navigation applications allows you to search the way you think. The power of this approach becomes immediately obvious upon use.”—Bradley Allen, founder of Siderean Software

The user can review the categories which are in bold face, any of the subcategories, pick one, and click to get a list of results. At any point the user can launch a query to narrow the number of documents in any one result.

The question is, “Does the average user want to have so many choices before beginning a search for information to answer a specific question?” Consumers may prefer a restricted results list or a point-and-click application that provides search “training wheels” to reduce confusion.

Each of the items in the recipe search page is a facet is, therefore, a field that descriptively names and identifies an aspect of a recipe or other document. A faceted search is characterized by some type of overt display of categories or topics. The user does not have to figure out how to formulate a query and then key it into a search box. The user recognizes what he /she wants and clicks to dive directly into the needed information.

One aspect of the Semantic Web endorses the notion that any Web-accessible content should carry this type of “deep” indexing or metatagging. When a document has this type of indexing applied to it, systems such as those developed by Siderean and a handful of other companies, can provide opportunities for a licensee to develop interfaces that provide the user with helpful entry points to indexed content. If the source content is structured, a system like Siderean can “slice and dice” information so that snippets can be combined into a report or a roll-up document.

Text information that is stored in a database is ideal for a Siderean-type search system. With little or no additional processing, the text from a database can be converted into a powerful text cross-tabulation system. Cross tabulations, one of the most popular features of spreadsheet users, is a way to organize or display the values or levels of one variable according to the values or levels of a second variable. In a more colloquial turn of phrase, metatagged content can be presented in a way that a standard key word result list cannot duplicate.

Siderean—like Convera, Entopia, Vivisimo, and other next-generation search systems—includes such features as “social bookmarking.” Users can post links to important documents, sites, or data that interests them. Others with access to the system can browse these posts and harvest the bookmarks for republication. The Siderean system allows users to tag any bookmark or item in the system so that these tags can be used to segment the documents by tags.

Siderean’s solution is a hybrid of an Endeca-style and MarkLogic-type of system.

“Our technology goes in typically in a matter of hours -- at worst several days - - and begins offering users the benefits of faceted navigation immediately thereafter. It's a significant step in the evolution of information management and a clear departure from the dinosaur-scale applications that we depend upon today.”—Bradley Allen, founder of Siderean Software

Company Background

Siderean's Web site (www.siderean.com) says:

Siderean Software provides turn-key, enterprise-class search and navigation software that permits you to “search the way you think.” Siderean's mission is to dramatically improve enterprise access to information.

In 2001, Siderean's pre-launch or stealth product code names were *bpAllen* and *Teapot*. The *bpallen* was a coinage from the name of the Siderean's founder, Bradley P. Allen. The *Teapot* code name evoked, perhaps unintentionally, the notion of reading tea leaves to get an insight into the meaning a user sought with a query. Coincident with the company's emerging from stealth mode in 2003, the flagship product became Seamark Navigator.

Siderean is an interesting neologism. According to the company, mariners in the central Pacific Ocean developed what is called *sidereal navigation*. Canny sailors just looked at the stars, sniffed the wind, and steered to their destination. At least, that was the theory. Less intuitive Spanish and Portuguese explorers embraced crude mechanical instruments.

Sidereal navigators used a couple of dozen star-rise and set points on the horizon as a “star compass.” *Seamarks* were currents, schools of fish, kelp beds, and islands. The idea was that certain winds would push the vessel along a predictable path. We know that adventurers using these methods did get from A to B. There is less information about those lost along the way.

What's this have to do with search?

Siderean chose this root because its system gives users a precise means of navigating the digital realm without having to know exactly what they are looking for before undertaking a search.

By 2005, Siderean asserted that its Seamark Navigator would support search and retrieval, personalized information delivery, and search-based applications for customer support.

Management

Bradley P. Allen is the founder and Chief Technical Officer of the company. Bradley P. Allen began his career as a member of the research staff at Carnegie-Mellon's Robotics Institute. He worked at Inference Corporation and was the creator of CBR Express, one of the first case-based customer problem resolution products. Before founding Siderean, he developed Web Compass, a metasearch engine available in the mid-1990s.

Other Siderean senior managers include Robert Petrosian. He was eventually replaced by Michael Schmitt. The management in 2006 included:

- Ivan Ivankovich, vice president of finance

- Jack Berkowitz, vice president of product development
- Robert MacGregor, chief scientist.

By mid-2006, Siderean raised \$6 million from early-stage venture capital firms. Investors included:

- Clearstone Ventures Partners in Santa Monica, California
- InnoCal, a private venture capital firm funded by institutional and private investors, focusing on investing in early stage information technology with the majority of its investments located in Southern California
- Red Rock Ventures (Palo Alto, California) specializing in seed and early-stage information technology investments.

Secret Sauce: Two Cups of Semantics

Siderean embraced RDF (Resource Description Framework), and metadata models, triples, and other semantic jargon.

At some point in the near future, most Web pages and standard office documents will have “Semantic Web” tags that identify the structure of each document. Seamark Navigator can work on any tagged content, but the system has been designed for organizations with a large volume of structured text in a database or a flat, well-formed XML file.

For most technologists, RDF is synonymous with the Semantic Web. Any information object tagged in conformance with Siderean’s system can support unstructured data, but the Seamark Navigator’s core strength is its ability to manipulate tagged information objects. Text, multimedia, and hybrid documents conforming to the RDF model can be handled equally well. Text documents are also indexed so that key word queries can be supported.

Siderean’s customers understand the value of structured content and building to the guidelines of the Semantic Web.

The company’s technology is specifically tuned to process tagged data. With tagged content, the Siderean system could perform what amount to cross tabulations of text. The idea was to provide users with a system that worked “the way people think,” an assertion that may be a bit of a stretch.

Siderean stresses that its technology automatically organizes almost any type of digital information—whether structured or unstructured, wherever it resides—into intuitive groupings. Properly tagged content can be presented in categories. For some users, browsing groups or clusters allow a user to grasp quickly the scope of what information is available. By scanning the suggested categories and clicking a hyperlink, Siderean displays content in that topic. Siderean emphasizes that a user of Seamark Navigator can retrieve information in a way that is “intuitive, almost instinctive.”

Probabilistic models have been said to be the models of language acquisition. If we look at human possession and acquisition of language, whether words, sentences or text, a human tends to have different behavior with respect to different sorts of structures.”-Anna Maria Di Sciullo, Delphes

According to Mr. Allen, founder of the company, “In late 2005, Siderean announced its UIMA-based product, Seamark MAPP.” IBM introduced the open standard UIMA so that content from different systems could be accessed without special and often expensive connectors. Siderean’s semantic system includes:

- A process-oriented framework for collection, extracting, and organizing metadata before feeding it into Seamark Navigator
- An architecture built on IBM’s UIMA standard to permit scaling
- Out-of-the-box adaptors for Microsoft SharePoint, RSS, Web, and file system access. Siderean has slated additional adaptors for release throughout 2006
- A built-in metadata extraction tool as well as plug-in support for third-party UIMA-compatible products
- A procedure for incorporating business rules for analytics.

There is little doubt about the engineering effort Siderean has invested in its Seamark Navigator. The company solved the difficult problem of allowing a user to move up and down any category by developing a special XML-based markup language used within the Seamark Navigator server.

Customers

Siderean’s Seamark technology is being used for both internal enterprise and public applications. Representative implementations include:

- Environmental Health News archives from Environmental Health Services (<http://www.environmentalhealthnews.org/archives.jsp>)
- Fortunoff (www.fortunoff.com)
- Resource Connection (<http://resource.smartdesktop.org/rescon/>) by the Indiana Humanities Council
- Vacation Search (<http://www.beachhouse.com/advsearch/search.asp>) from Beachhouse.com

The Indiana Humanities Council has built a portal, called SmartDESKTOP. The system was developed by Indianapolis-based development consultant MindGent LLC and is available initially to select teachers from central Indiana schools. By tapping into the portal, the teachers can access a growing database of information to support their classroom lessons. Access to SmartDESKTOP should be extended to teachers throughout Indiana by the start of the fall semester.

Siderean technology is part of Indiana’s ResCon, or the Resource Connection, one of three components of the SmartDESKTOP program, which also includes an electronic plan book and an assessment module. ResCon provides a searchable database of materials from major state and national muse-

ums, cultural organizations, government offices, and other sources. About twenty-five major state institutions and ten national organizations act as resource partners. The ResCon database includes more than 1,300 resources at present, with new items to be added each month. The public version of the Resource Connection can be accessed at <http://resource.smartdesktop.org>.

SchemaLogic Deal

In May 2005, Siderean and SchemaLogic formed a partnership. Siderean provides the Navigator system, and SchemaLogic provides its metatagging system. The combination allows a licensee to perform value-added indexing, metatagging, search, and “pinpoint navigation” from a hybrid system.

According to Steve Ardire, chief strategist for SchemaLogic:

“In semantic enterprise search, the vagary of language, with each word having many meanings, requires enterprise metadata management to narrow down or understand the specific meaning and topic. We believe our partnership with Siderean will enable information architects to effectively organize information in ‘human terms’, and thereby improve an organization’s agility to quickly respond to changing business conditions. With more relevant and accessible information, knowledge workers can collaborate more effectively, which will accelerate decision-making and improve governance over data silos.”

The licensee, therefore, will use two companies’ systems to deliver a more effective search solution. Licensees, of course, may elect to use either Siderean or SchemaLogic for a search solution, “to fully address the issue of findability of relevant information, effective enterprise search uses a clear representation of knowledge to retrieve, organize and display results that are driven by metadata and enterprise information management workflow.” The company is expanding its partner programs.

Seamark’s Architecture

Seamark Navigator is a server-based suite of software. The Seamark Navigator processes the documents and their metadata building indices to support facets.

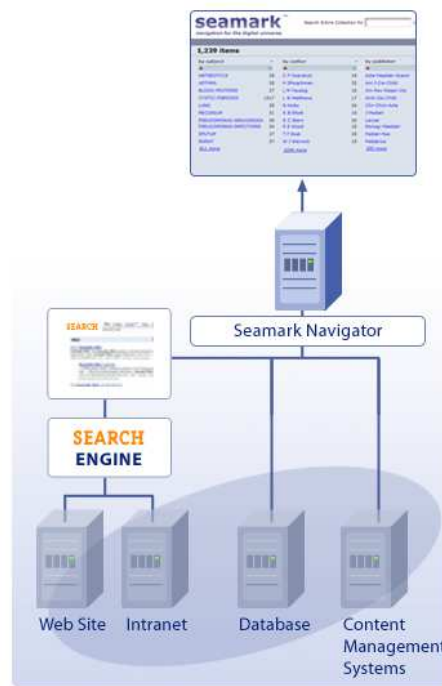
The key premise of Siderean’s approach is that no processed information is hidden from the user. A Google search box, by contrast, shows nothing of what’s been indexed. In the recipe example above, none of the recipe information is hidden. The user can see categories such as “By Season,” a list of subtopics, and the number of documents in a specific category at a glance. In contrast, a search box such as the standard Google interface does not provide an overview of the information available from the system. Siderean allows

the user to pick what's important. In the Google-style interface, Google uses click streams to determine what's important. For business-related queries, link analysis and other click stream techniques are generally inadequate.

Siderean argues that “lack of visibility for scope is fast emerging as the urgent information and content management problem.” Users don't know what's available in a system. Typing a query and scanning a list of results consumes time and inhibits business success.

Keyword and search box information retrieval systems don't allow an overview of information assets. As the volume of structured and unstructured information has grown, enterprises are discovering that keyword searching is a partial answer to many employees' information needs. When a user is able to identify all the content available, the benefit of a single view becomes apparent. Siderean's facets give the user what the company calls a “contextual framework for discovery and selection.”

Siderean then approaches search by solving the problem of “hidden scope.” The company exposes available content and the necessary navigational functions to create a different type of information access and management tool.



The Seamark Navigator provides all functions in a single server or a server cluster. A typical installation appears in the block diagram.

What Seamark Navigator Does

What Seamark does is systematically examine the various data sources to which it is introduced, discovers both the explicit and implicit structure or organization in the data, and produces a metadata description of its content and characteristics. The system automatically generates a browsable, prototype application based upon that description.¹

As more content is processed by the system, the choices presented users dynamically change. For example, a Seamark application that includes an RSS data feed as one of the data sources could not only alert the user that new information has become available, but that new categories of information are available, as well. Mr. Allen notes, “Not only can enterprises envision and define entirely new, strategic business applications based upon previously unavailable, aggregated sources of data, but the generated applications—and the information they present—are completely dynamic.”

An organization that has a word list that include “See Also” and “Use For” references can integrate these connections in Seamark Navigator. In an organization, a user could look at a geographic category and the system would know that an office was in Chicago in the state of Illinois. The system administrator can set up the system to allow a user to annotate and tag data, thus adding a social or folksomic dimension to the Siderean system.

The Seamark Navigator can be installed as a turn-key system. Siderean can configure the system and deliver it ready to process content on the licensee’s premises to streamline deployment. The basic Seamark Navigator consists of one or more servers that perform the functions needed to make content accessible.

According to Bradley Allen, founder of Siderean, “more typically, licensees simply download the code, install, and configure the system without anything more than telephone and email support.”

The diagram shows a simplified view of the Seamark Navigator.

There is minimal setup required by the licensee. Siderean says that within a matter of hours, the Siderean system can begin to process content, including information in databases and from live data feeds. The variable in getting Siderean up and running is the amount of content. Document processing can require significant time. Unstructured content must, of course, be converted to structured content. Automating this process is possible; however, some manual editing may be needed to handle certain types of content such as an Adobe PDF converted to HTML and then converted to well-formed XML. The content is organized into a searchable, cohesive whole. Once documents have been processed, users can navigate that information using the category and subcategory displays. Licensees can use the default category display or modify it by changing its associated style sheet. A keyword search function can be placed on any Siderean category display.

The key point about Seamark Navigator is that it is a lightweight solution for faceted navigation.

The turnkey system integrates various and disparate data sources (both structured and unstructured from both inside and outside the enterprise). When

¹. Certain types of content require additional analysis. To accomplish this, Siderean uses tools and technology from SchemaLogic.

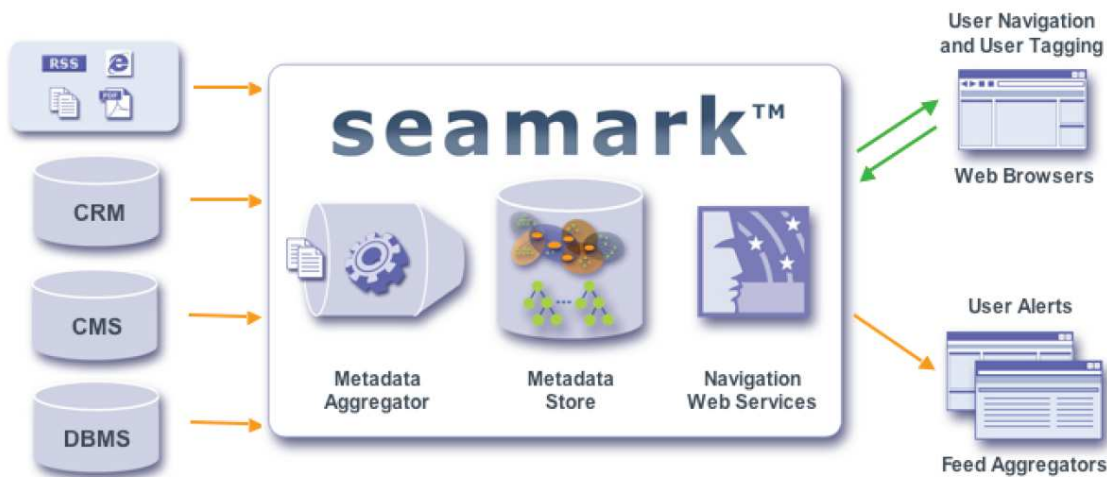
new content becomes available to the system, the indexes are updated and the category displays automatically changed to reflect the new categories, subcategories, and content counts. The Web-ready, Seamark-generated application can be used “as is”; refined as necessary for look, feel or function; incorporated into a Web page; or linked to other applications as a Web service.

The Seamark Navigator may consist of multiple servers or clusters. The configuration depends upon the volume of content and the frequency of changes to that content.

The Three Components of Seamark Navigator

Seamark Navigator, like most search systems, consists of separate software modules. The three core modules may be enhanced by other third-party software when additional functionality is required to make the information available to the system. Two examples include identification and generation of additional metatags via the SchemaLogic system or the inclusion of a translation subsystem from Systran or BASIS Technologies.

In its default form, the Seamark Navigation component consists of three major subsystems. These are illustrated in the diagram below:



The first module is the metadata aggregator. It acquires or receives content from the various sources. The metadata aggregator performs document processing, maps metadata to documents and indexes the words and phrases in a document. The system uses available tags whether XML or RDF. As part of the document processing function, the system automatically seeks and finds relationships among documents so that a user can run a query and see related

information. These are matrix implementations of search. (Additional information about the transformation functions performed in the metadata aggregators appear in the following section of this document.)

The second module is the metadata store is the equivalent of the index of the documents. The metadata components exist as tables, and the key word index is an inverted index. This component of the Seamark Navigator contains the data and code necessary to display the categories, subcategories, and document counts for the content processed by the system.

The third module provides Navigation Web Services. In addition to handling interaction with browsers, this module receives the query or the user's click stream, converts the query to a form that is understandable to the metadata store, obtains data from the metadata store, formats it, and sends the data to the user. Certain higher-level tasks such as filtering and alerts are handled by the query processing and results display and supports user-generated tags and such functions as filtering and distributing alerts about new documents.

Platforms Supported

Siderean can process data from such sources as:

- Third-party information feeds in News XML
- Databased information via JDBC drives
- content with XML/RDF tags
- RSS
- Flat files generated from any system.

Unstructured data must be processed and tags inserted prior to processing by the Siderean Navigator.

Siderean supports Windows, Linux, and Solaris operating systems. The user accesses the Siderean system via a standard Web browser. Other methods of access Navigation Server supports includes:

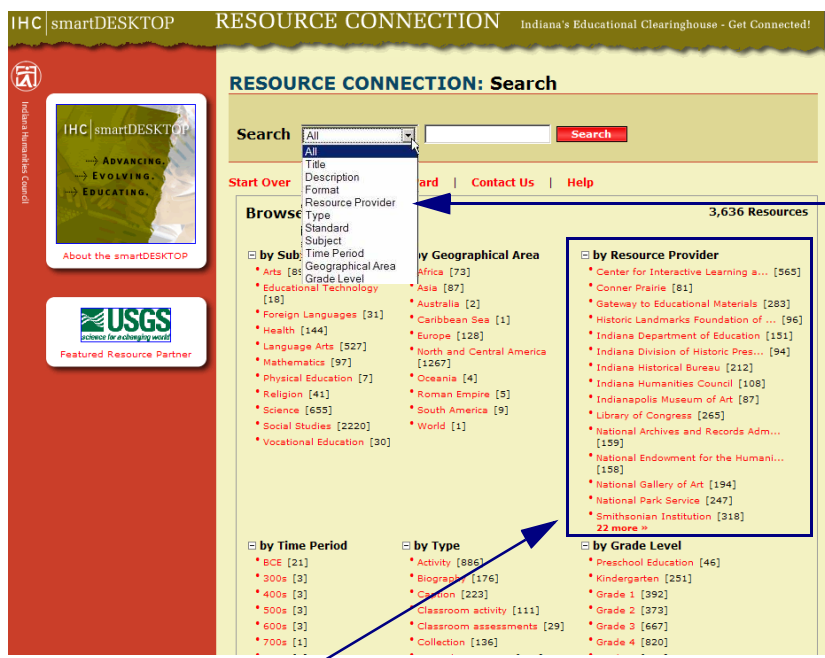
- Web Browsers
- RSS Aggregators
- SOAP
- ASP/JSP.

The system supports the following databases:

- MySQL
- Oracle
- Microsoft's SQL Server
- Hypersonic

A Bet on Semantics

The interface for Indiana's application illustrates what is possible with the Siderean system. The figure below shows the main query page for the Resource Connection service in the portal. Several features jump out at the user. First, the user can run a query across all of the resources indexed by the system. The interface also permits a drop down menu so that the user can limit the query to a particular slice of the collection. The user can also scan the categories of educational resources and access materials directly by clicking on a subcategory.



A drop down menu makes allows a user to restrict the query to a specific category or content type.

The user can click on a specific publisher's or provider's content. The results come from that source.

Siderean allows the user to search the content without having to back track or enter special commands to make the search function target a specific category. The system generates a list of results the user has selected. At any time, the user can select another category or resubmit the query against another category or the entire collection. Up and down navigation of topics requires pointing and clicking.

Second, in the screen shot above, the user can get a bird's-eye view of the content indexed by the Siderean system. No topics or content is hidden. Under each main heading such as "By Subject" or "By Grade Level." Under each heading are descriptive labels for subgroupings of information. For

example, “By Grade Level” allows an educator to click on Grade 2 and gain access to the resources appropriate for that level.

Third, as the user “drills down” into a subcategory, headings appropriate for the topic appear. At any point in the content exploration process, the user can return to the overview display or launch a new query.

The benefit of this approach is that a user can follow a topic that is analogous to how a library patron can browse for shelved books in a traditional library. A user can browse a topic, backtrack, follow a suggested topic, and locate the needed information without figuring out how to use the card catalog. Siderean assumes that the user will feel comfortable with a system that leverages the library research method.

The Siderean system allows the licensee to control the number of facets and other aspects of the interface enabled by the Siderean system. Most of these adjustments are handled through style sheets or by editing the templates provided with the system. A licensee may discover that displaying too many facets and make a quick change to simplify the results display. A specialist looking for information is likely to find the richly faceted interface a useful adjunct to keyword searching.

However, some users may be overwhelmed by large number of categories and subcategories. Users familiar with technical, medical, and scientific information find that categories and subcategories are time-savers.

Documents processed by Seamark Navigator are transformed into tables. The secret sauce of the sideways, upwards, and downwards navigation comes from looking up pointers. Seamark does not provide much information about the size of these tables, however. Some search system vendors, like Blossom Software, create indexes that are about 20 to 25 percent of the size of the source documents. Siderean’s tables are likely to be as large as the source documents.

Organizations seeking to license the Siderean or Endeca system will want to test these faceted systems to ensure that the resources necessary to perform the transformations and the storage to hold the tables are robust.

XML itself is a verbose way to express a document. Fast Search and other vendors create a proprietary index format to manage both the size of the index and help ensure quick response to user queries.

When Siderean processes an RSS feed, that feed is transformed into the RDF format. Metatags are added from the Dublin Core and SKOS vocabularies. These documents are then made navigable in the Seamark system using the `dc:subject (tag)`, `dc:creator`, `dc:publisher (site)`, `dc:moderator (feed)` and `dc:date` as the facets.

For a short period of time, Siderean indexed the links and comments on the Delicious.com site.

At the fac.etio.us home page you'll see all five facets exposed: tag, creator of the tag, site tagged, the feed it was found in (del.icio.us, del.icio.us/popular, and Brad's feed), and the date the tag was created. You can click on any, but let's say we click on one of the entries in the list of tags: music. We are taken to a page that lists all the bookmarks tagged "music," but are also shown a list of all the other tags given to all pages tagged with "music," all the people who have tagged a page "music," all the feeds that contain bookmarks tagged "music," and every day in which someone has used the "music" tag. Each of these tags is itself clickable.

Delicious repurposed in the Siderean index from May to August 2005.² Clicking on a Delicious category and then exploring a subcategory shows the tags extracted and assigned to a Delicious entry. Inspection suggests that the Siderean-generated tags are likely to require storage roughly equivalent to 50 to 70 percent of the source document's size. With the decreasing cost of storage, large indexes are not a problem. Licensees of a facet-based search system will want to plan for adequate temporary storage space to handle temporary files during index and table updates, adequate random access memory for the systems to eliminate the delays of accessing discs during certain steps in the update process, and sufficient processor capacity to ensure that system latency remains low.



The Fac.etio.us service demonstrates the value of a Seamark Navigator approach. Siderean's half-year test makes the narrowness of the Delicious.com service evident. The metrics attached to tags provides clear evi-

². The site is located at <http://www.siderean.com/facetious/facetious.jsp>. Verified on January 15, 2006.

dence that social tagging schemes are dependent on a small number of terms, many of which are too general to make a refined search possible. Siderean's metatags provide an extremely useful way to access the Delicious.com content. One hopes that Siderean's follow on service, now in development, becomes available in 2006.

Metatags and Indexes

Siderean makes use of two complementary, yet functionally different ways, to index a document. Siderean also makes use of traditional key words and any metadata associated with a document. This is similar to double entry bookkeeping in that multiple views of the data can be used without performing extra work.

Siderean uses the words and phrases in a document to create a searchable index based on Lucene, the open source search library. The index allows a user to search for a particular word or phrase; for example, search for documents that contain the phrase *airport terminal*.

Siderean also uses any metadata attached to a document or about the document like its date and time. If a document has a metatag *Author*, then Siderean's system can display the category *Author* and list the names of the authors of documents.

Siderean's system allows text to be cross tabulated in a way roughly analogous to how Microsoft Excel's pivot table function allows numbers to be viewed from different vantage points. According to Robert Petrosian, Chief Executive Officer of Siderean:

Seamark Navigator reveals the full scope of digital assets and cataloged items available to enterprise users and customers, presenting a discovery and precision search framework for navigating to the most pertinent choices. Ideally suited for a wide variety of markets and organizations, Seamark Navigator improves overall business performance by providing a clear, contextual map of available digital materials, quickly directing users to all relevant assets or goods, leading to better individual and group decisions, content usage and purchasing.

Siderean's product description emphasizes that the company is among the first to provide an A result set can be displayed by one of these tags and other documents where the same tag may be used can be displayed as a related concept.

Seamark Navigator Functions

Siderean uses metatags to create the categories into which information is placed. Seamark Navigator will use available metatags. If a document does not have a metatag such as “Location,” Seamark Navigator cannot discover and assign a location from unstructured text.

Human-created taxonomies such as the original Yahoo! categories required an editor to determine if a new category was needed. Once the category was created, Yahoo! could populate that category with human and software processes. Siderean’s system can accommodate the outputs of a text mining system that performs entity extraction and metatag assignment - for example, SchemaLogic’s system, among others. To integrate a system from another vendor such as Inlight, for example, some scripting may be required. Seamark Navigator’s support for Web services makes such integration a comparatively easy job.

Once the metatagged content is available to the metadata aggregators, Siderean’s approach is very reliable. A lack of consistency in the metatags used can lead to some inconsistencies when the category view displays documents. If assigned metatags are not consistent—for example, a document has a metatag for “Geographic_Location” and others have the tags “Département”, “State”, and “Province,” Siderean’s system will use these tags as found in the document. An editor can create a list of synonyms and instruct Siderean to map any of these terms to the tag “Location”. The challenges are cost and time. Major inconsistencies will require manual editing of the metadata repository’s tables or re-indexing the document with these normalized metatags.

A semiautomatic or manual process can add significant cost to a search system. A procurement team will want to factor the time and cost of document normalization; that is, generating consistently metatagged content. Creating synonym lists, “teaching” the indexing system, and checking that documents are categorized as intended can derail some search implementation schedules.

For the purposes of this discussion, let’s assume that the Siderean system has access to content that contains consistent metatags. These can be either XML documents or exported “reports” from a database.

Siderean uses the tags in the documents to create one set of index pointers. The text of the document or content object is also parsed and indexed by key word and phrase. The two streams of index terms are mapped to each document. If the majority of the content is well-formed XML, Siderean can map at the structural unit of a document; for example, by section or paragraph.

These tags are the guts of what Siderean refers to as “Semantic Web principles.” The “Semantic Web [sic],” according to Siderean, “a collection of

metadata about the regular Web.” Metadata are used to describe existing documents, Web pages, concepts, databases and file types on the Internet so that software applications gain an understanding of what the content means.” Siderean applies this notion to documents in an organization by capturing the file name, the document date and time stamp data, and other information about each individual document.

Siderean, therefore, creates a richer index, because it has the document index plus any other metadata the system has been able to obtain. Many search systems, such as Blossom Software’s, for example, use only the words and phrases occurring in a source document, ignoring tags and any other metadata included in a document.

- Siderean, i411, and Endeca, among others, take a different approach. With consistent tags from many sources, a faceted-search system allows some useful functions. It is the use of additional tags and more metadata that encourage some search engine developers to distinguish their systems by describing them as next-generation or, in the case of Siderean, based on the Semantic Web. Other distinguishing features include:
- Arranging documents by time. Google searches for the most part are not displayed in chronological order. If a time tag is available, Siderean can display a result set in chronological order.
- Locating a specific set of documents which share a common attribute such as a particular chemical compound or characteristic; for example, the recipes in which cilantro appears as an ingredient.
- Displaying information by source; for example, specific publication or type of publication.

To some, the use of the phrase Semantic Web may not make clear that Siderean’s functionality depends on the quality of the metadata available to its system.

Siderean’s use of the phrase Semantic Web is intended to make clear that each type of metadata becomes a facet. The relationships among resources (concepts, objects) are described using the Resource Description Framework (RDF) or some tools, such as the Dublin Core or any other consistent tagging system. RDF is an XML-based language for describing Web resources. It has been designed to express a range of semantic meaning by stipulating that an “author” identify:

- A resource; that is, the item such as a particular document
- A predicate; that is, an attribute or facet of the resource such as the abstract section of a technical paper
- An object; that is, an attribute value such as type of chair.

Faceted classification allows the user to progressively filter a large data set through the user's classification choices until he or she has a manageable set

of items to browse. The user gains significant insight into the nature and scope of the collection based on the facets or characteristics presented to guide his / her exploration. Instead of sifting through a pre-determined hierarchy, the items are organized on-the-fly, based on which of their inherent qualities are of interest to the user. Guidance is provided by showing the user both a summary of items in the results (as a regular search engine would) as well as the number of matches available at each category choice.

A keyword search is always an option for the user.

Result of the query shows first 12 chairs, each hot linked to a product description.

The screenshot shows the Fortunoff website interface. At the top, there is a navigation bar with links for Checkout, Order Tracking, Fortunoff Credit Card, Help, and Contact Us. Below this is a secondary navigation bar with categories: Fine Jewelry, Tabletop & Giftware, Home Store, Bedroom & Bath, Baby Fortunoff, and Gift & Bridal Registry. A search bar is located on the left, and a 'GO' button is next to it. Below the search bar, the text 'You are here: Home Page > Product Search Page' is displayed. The search results section shows 'Keyword' and 'Search' buttons, with 'Search Within Results' and 'Reset Search' options. It indicates '261 items matching result' and a checked box for 'Keyword contains chair'. A 'Narrow Your Search' section is present, with filters for Category, Brand, and Price Range. The Category filter shows 'Fortunoff' with sub-categories: Baby Fortunoff (11), Home Store (234), Bedroom & Bath (4), and Tabletop & Giftware (67). The Brand filter lists: Agio (2), Cast Classics (1), Hillsdale (4), Hillsdale House (2), Hooker (1), Interactive Health (1), Legacy (1), Powell (2), Samsonite (4), and The Source (10). The Price Range filter shows: \$100 to \$200 (34), \$1000 to \$2000 (19), \$200 to \$300 (32), \$2000 to \$3000 (8), \$50 to \$100 (14), \$500 to \$600 (10), \$600 to \$700 (7), and \$700 to \$800 (12). The main product listing area shows 'Showing 1 to 12 of 261 items | Next' and 'Matching Result' with 'SORT Name' and 'Price' options. The first row of products includes 'Royal Rocker Rocking Chairs', 'Beth Dining Chair Cover' (Price: \$4.99), and 'Martini Swing'. The second row includes 'Dallas 5 Piece Dining Set', 'Yorkshire 7-Piece Hand-Painted Dining Set', and 'Outdoor Furniture Covers'. The third row shows partial views of other products.

The user can jump to a specific price range knowing the number of chairs in the category before clicking.

Key Differences

Siderean Seamark Navigator is a good case to examine to understand some of the features of a semantic search.

Siderean straggles several “worlds.” Publishers implementing SGML or XML publishing systems have a way to locate specific segments of content easily. For the professional library researcher, Siderean’s approach provides a wealth of features that permit complex queries and browsing of related

content. For the average searcher looking for information on an eCommerce site, the ability to flip between product categories is a shopper's delight.

In 2003, Siderean was one of a handful of search systems embracing the Dublin Core methods. The conference proceedings were made available with Siderean as the search-and-retrieval system.

Siderean also was, since its inception, a member of the World Wide Web Consortium (W3C). Seamark's use of Web Services, via the open standard HTTP/XML communication protocol of SOAP, contributes to some of Siderean's integration claims. The idea is that Siderean's open standards approach reduces up front development costs and significantly lowers the total cost of ownership for customers.

Configuring a Semantic Search System

For those unfamiliar with the configuration of a semantic system, graphical interfaces are not available for some operations. I want to provide a glimpse of what scripts look like and make clear how the customization and maintenance costs can become a significant part of an information technology budget.

Putting Content on the Navigation Server

There are three steps to preparing metadata to enable site search using Seamark Navigator:

- Defining the facets used to describe site resources
- Creating the descriptions of the site resources
- Loading the descriptions into the Seamark Server.

Most readers of this report will not have first-hand experience with RDF tags. The system is not in use in most companies at this time.

Defining Facets

RDF Schema is used to define the facets and facet value taxonomies that are used to describe a given resource. For purposes of illustration, we will use a Web document residing on an Intranet.

Facets can be either flat, taking literals (strings or numbers) as their values, or hierarchical, taking values that represent concepts in a given taxonomy. Concepts are defined using the `rdfs:Class` tag. For example, to define the concept of a *cuisine*, the Siderean engineer would use the following RDF:

```
<rdfs:Class rdf:about=" http://www.siderean.com/recipe-  
demo#Cuisine" >
```

“It's time for computers to allow you to find information the way you think.—
Bradley Allen,
Siderean founder

```
<rdfs:subClassOf rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
<rdfs:label>Cuisine</rdfs:label>
</rdfs:Class>
```

The Siderean engineer uses the `rdfs:subClassOf` tag to specify taxonomical relationships between two concepts. The example below defines *Italian cuisine* as a subclass of *cuisine*:

```
<rdfs:Class rdf:about="http://www.siderean.com/recipe-demo#Italian">
  <rdfs:subClassOf rdf:resource=" http://www.siderean.com/recipe-demo#Cuisine"/>
  <rdfs:label>Italian</rdfs:label>
</rdfs:Class>
```

The engineer then defines the facets using the `rdfs:Property` tag. For example, to create a facet that will inform a user that a Web resource describing a recipe is from a particular type of cuisine, the engineer would code:

```
<rdf:Property rdf:about="http://www.siderean.com/recipe-demo#fromCuisine">
  <rdfs:domain rdf:resource=" http://www.siderean.com/recipe-demo#Resource"/>
  <rdfs:range rdf:resource=" http://www.siderean.com/recipe-demo# Cuisine"/>
  <rdfs:label>from cuisine</rdfs:label>
</rdf:Property>
```

Creating Resource Descriptions

Once the engineer has defined the facets for the search application, the next step is to create a description for each resource on the site that we want to make available to search. The code below uses the conventions specified in the standard Dublin Core. These tags provide generic document resource metadata in addition to the facets previously defined explicitly for the site:

```
<rdf:Description rdf:about="http://www.epicurious.com/run/recipe/view?id=105501">
  <dc:title>MOZZARELLA, GREENS, AND GARLIC BRUSCHETTA</dc:title>
  <dc:creator>Gourmet</dc:creator>
  <dc:date>April 1994</dc:date>
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
```

```

<recipe:eatenFor rdf:resource="http://
www.siderean.com/recipe-demo#Appetizers"/>
<recipe:eatenDuring rdf:resource="http://
www.siderean.com/recipe-demo#Summer"/>
<recipe:hasIngredient rdf:resource="http://
www.siderean.com/recipe- demo#Cheese"/>
<recipe:usesMethod rdf:resource="http://
www.siderean.com/recipe-demo#Broil"/>
<recipe:fromCuisine rdf:resource="http://
www.siderean.com/recipe-demo#Italian"/>
</rdf:Description>

```

Since RDF is used as the metadata format, any method that produces RDF can be used to provide descriptions for Seamark Navigator. Metatags are embedded in Web page HTML of many sites, and in order to reduce rework, the values stored in these tags can be extracted and re-expressed in RDF. The Siderean system includes crawlers that can extract key phrases from the text contained in site resources. These key phrases can be stored as values in `dc:subject` tags. Siderean's crawlers can be used to generate hierarchical facet values in resource descriptions. The trick is to classify pages by determining where these pages appear in the click paths in the site. For Web applications in which pages are generated from a relational database, support is provided in the Siderean Navigator administrative interface to allow the administrator to import the database's schema data into RDF format.

Loading the Metadata

Once the Resource Descriptions describing the site have been created, the SOAP protocol can be used to load the RDF into the Seamark Server. First, the operation `addModel` is invoked to create an RDF model for the site. Next, the Seamark Server provides two operations, `addRDFStream` and `addRDF` that allow RDF metadata about the site to be loaded into the server's database and query engine.

Siderean allows the system administrator to decide whether to push RDF to the Seamark Navigator with `addRDFStream` or providing a url from which the server can pull RDF with `addRDF`. These operations can be one-time or on a regularly scheduled basis.

Administration pages alternatively, provide an easy-to-use Web interface accessible to authorized users for the upload and scheduled import of RDF metadata files.

XRBR

Seamark Navigator uses an XML document format called XRBR to represent a query and its corresponding set of results. *XRBR* stands for *XML-based Retrieval By Reformulation*. An exchange of search queries and responses as a user drills down on resources of interest is thus modeled as a sequence of XRBR documents. XRBR is similar to the functionality of multidimensional OLAP applications for looking at data from different angles.

Siderean says that it developed its XRBR schema because existing RDF query approaches assume that the desired result format is RDF documents. RDF generates results designed for computers, not for humans. Implementing drill down via RDF requires additional programming. RDF handles single documents well, not large result sets.

XRBR solves these problems. It returns descriptive, human-readable information about the resources matching the user's query. Facets with literal values such as *Location* can be matched using a variety of operators including those supporting free text queries against the text content of the facet or resource. XRBR can match a query to hierarchical facets explicitly.

XRBR search requests are made in the form of an `xrbr:query` statement, and the results are passed back as an `xrbr:results` construct.

SOAP requests containing `xrbr:query` searches can be made from such SOAP-enabled client applications as Java servlets, Dot NET applications, and Perl). The Siderean server's default search application applies XSLT style sheets to transform XRBR search results into virtually any type of page layout.

No matter what the client will be, the first step in setting up a search application for the Siderean system is constructing an initial XRBR query. Note that XRBR is iterative. The iteration handles the drill down cycle through function. XRBR also reformulates results so that the slices of the data can be displayed using the style sheets included in Siderean. A user can drill down, roll up, or jump to other categories to explore the information in the system related to the user's need. The basic Siderean function requires that the facets be defined ahead of time. No facets are discovered by the system. However, Siderean can use any metatags that exist within the data set. Obviously, for the system to function, all information objects ideally will have the same tagging structures. Incorrectly tagged information may not appear in certain slices.

The `xrbr:query` contains three parts:

- 1 An item-type attribute, which specifies the RDF type of the items to search.
- 2 Dimension elements, which describe the facets predicates to be included in the search results. The use of the term *dimension* to describe Siderean

facets is Siderean's way of expressing the cross tabulation of text by available tags.

3 A where element specifies the search criterion.

Here's the syntax for an XRBR query fragment that states the item type attribute. The code specifies a recipe as the item-type. Then the code requests the first recipes from the result set, sorted alphabetically by title.

```
<?xml version="1.0" encoding="UTF-8"?>
<xrbr:query xmlns:xrbr="http://www.siderean.com/2001/10/
xrbr/"
  start-index="0"
  sort-dimension=" title" sort-direction="asc"
  item-type="http://www.siderean.com/
  recipedemo#recipe" result-set-size="10">
```

The query then continues with dimension statements. Each describes how to use metadata facets on the search pages. In this case, each recipe's magazine article title is one of the facets returned with search results.

```
<xrbr:dimension name="title" Predicate="http://
www.siderean.com/recipedemo#title">
  <xrbr:hint label="title" search="yes" sortable="yes"
  />
  xrbr:return />
</xrbr:dimension>
```

The code to specify that the facet *Cuisine* should appear in search results with up to 10 ten suggestions of cuisines sub-types that are in the collection is:

```
<xrbr:dimension name="cuisine" Predicate="http://
www.siderean.com/recipedemo#cuisine">
  <xrbr:suggestions count="10" />
  <xrbr:return />
</xrbr:dimension>
```

The code generates the facet shown on the right. The facet name is in bold face. The ten subcategories of Cuisine appear as hyperlinks with the number of documents in each subcategory shown in parentheses. The user can scan the list of subcategories and with one click access the recipes in that category; for example, *Mexican*. The licensee does not have to write code for this function. The drill down is generated on the XRBR generated by Siderean's default search application.



Cuisine: (10 of 17)
VIEW MORE LESS ALL

- [American](#) (1497)
- [French](#) (566)
- [Greek](#) (138)
- [Italian](#) (769)
- [Jewish](#) (168)
- [Kid-friendly](#) (486)
- [Low-fat](#) (590)
- [Meatless](#) (1501)
- [Mediterranean](#) (163)
- [Mexican](#) (330)

In addition to searching by drilling down into facets or dimensions, XRBR supports free text search in one or more particular dimensions. XRBR's *where* statement tells the Siderean what text to look for. To search for all recipes that contain the word balsamic in the recipes' article titles would look like this:

```
<xrbr:where>
  <xrbr:queryterm dimension=""title""
  value=""balsamic"" isliteral=""yes""/>
</xrbr:where>
```

Search Results

XRBR search results guide further search and navigation. The XRBR search results are structured in three parts:

- 1 `xrbr:query` represents the query sent to the Seamark Navigator query engine. Having a copy of the original query allows the client to maintain state in non- synchronous environments.
- 2 `xrbr:preview` represents the selection of facets and facet values that can be presented to the user to provide search look ahead and guide them in the refinement of their query. As we have seen, the presentation of facets or characteristics that define a collection is instrumental in helping users navigate unfamiliar sites and guides them as they drill down searching for items of interest.
- 3 `xrbr:resultset` represents the result set, sorted according to instructions in the `xrbr:query`. By grouping results into one or more `xrbr:tab` tags, a large set of results can be represented and navigated without forcing a huge amount of information to be explicitly containing in a single `xrbr:request` document.

The default Siderean server's search application takes the three parts of an XRBR search result and applies one of three different XSLT style sheets to it, depending on the context:

- Start Page style sheet is applied if there are no query terms in the *where* clause of the query
- Search Page style sheet is applied if there are query terms and the user is not in endgame mode. Endgame is Siderean's jargon for an action that has cleared out facet data.
- Endgame Page style sheet is applied if the user is in endgame mode.

A licensee can use XML utilities to process the document directly for presentation in a GUI. Another option is to use the search results from the document to automate various agent-based search activities such as updating a standing query.

Implementation Considerations

Like Fast Search's ESP system, optimal installations require support from the vendor's own engineers. Siderean is no exception. Before deploying the Seamark Navigator, the procurement team will have addressed most of the pre-acquisition issues. These include infrastructure, interface, security, and content to be indexed, among others.

However, due to the dependence of Seamark Navigator on tags, a number of implementation considerations must be scrutinized prior to flipping the switch on a new Seamark Navigator system.

The Taxonomy Process

The licensee will want to consider that a faceted classification differs from a more traditional one in that it does not assign fixed slots to subjects in sequence. Faceted classification uses the metadata's tags as defined, mutually exclusive, and metatag "fields" to generate facets.³ In general, highly complex classification schemes can present challenges to users, require up front planning, and a mechanism for spot checking tags and modifying those that aren't in line with user needs.

If a licensee does not have a categorization system, some thought must be given to the subjects that will be covered by the licensee's content. Without digressing into a discussion of taxonomies and classification systems, the licensee will want to define the subject to be covered by examining existing classifications or thesauri, or titles or objects in the Seamark Navigator database or index.

The derived topics can then be decomposed into facets each with a distinct label. Ideally the subfacet items will then be organized so that they are in homogeneous, mutually exclusive groups that differ from their category by one characteristic easily recognized by the users.

The result is that within each facet, subfacets or more specific topics are listed. The breakdown can continue to a depth required by the content domain, so subfacets within subfacets can be created. The items in each subfacet are ordered from more general to more specific, complex or concrete.⁴

³. See, for example, <http://www.kmconnection.com/DOC100100.htm> and Wynar, Bohdan S. *Introduction to Cataloging and Classification*. 8th edition. page 320

⁴. For more information about this process, see <http://encyclozine.com/Reference/Library/Classification/>

Integration with Other Enterprise Applications

The enterprise information architect will find Seamark Navigator relatively easy to integrate into the licensee's corporate data infrastructure. Seamark Navigator arrives with open-standards-based interfaces to data sources via JDBC (Java Database Connectivity), RDF/XML (Resource Description Format in Extensible Markup Language), RSS (Rich Site Summary, among others).

Seamark uses a Web-services model that permits flexible integration with other enterprise applications. The Web services model is somewhat easier to implement than a pure services oriented architecture. Furthermore Web services are now understood to be highly scalable. Client interfaces include RSS, SOAP, and ASP/JSP. It is offered as a standalone platform under Linux, Windows and Solaris.

The licensee will want to ensure that staff assigned to the Siderean project have a good understanding of these technologies.

UIMA Support

Siderean supports IBM's Unstructured Information Management Architecture (UIMA). Each Seamark Navigator includes a Metadata Assembly Process Platform (MAPP). This set of scripts facilitates the use of metadata from other systems in the Seamark Navigator.

Seamark Navigator includes a spider to acquire content. Siderean calls this module a "harvesting back-end" for accessing and generating metadata. As noted above, the MAPP module will then normalize the metadata from content regardless of the system and format. Seamark MAPP is a flexible framework for generating and analyzing metadata from diverse sources including file systems, content management solutions, databases, Web pages, RSS feeds, and blogs.

Licensees with a large volume of these kinds of unstructured assets can use MAPP to organize information. Once processed, the Seamark Navigator can be used to display specific content with faceted navigation to different user groups. Seamark Navigator can slice the data so that specific navigational applications can be deployed to customer support, marketing, competitive intelligence, and business manage user.

ArnoldIT Opinion

Semantic technology provides a licensee with more options for displaying related content. The interface exposed to the user, however, often suffers from too many bells and whistles. Siderean's sample interfaces are more confusing than a simple search box.

Table 2: Seamark Navigator Checklist

Attribute	Siderean Asserts	ArnoldIT Comment
1 Platform	Solaris, Linux, and Microsoft	
2 Keyword search	Supported via Boolean, free text, and facets	The system uses Lucene, the open source search engine, for key word retrieval
3 Text mining	The Seamark Navigator index provides countable objects	A third-party component may be required for certain applications
4 Automated indexing	Yes	The system uses controlled vocabularies if available. Manual maintenance of the system-generated terms is recommended.
5 Personalization	No	
6 Workflow	No	
7 Interface	Licensee configurable	
8 Hosted service	No	
9 Administrative interface and tools	Graphical interface provided	Some operations require scripts and more formal coding
10 Application programming interface	Documented API available with sample code for integration of search and categorization.	Licensees can interact with the system by writing scripts that are used in the configuration of the system
11 Professional services	Technical support is available	
12 Security	Uses the operating system's security functions	Additional security functions require custom coding
13 Connectors	SGML or XML content and traditional databases with content in tables; HTML and RSS feeds	
14 Support for structured data	Yes	
15 Relevance ranking	Yes	The licensee can interact with the relevance ranking subsystem
16 Video	No	
17 Federated search	No	Processed content is stored within the Seamark Navigator system

Attribute	Siderean Asserts	ArnoldIT Comment
18 Fielded search	Yes	
19 Content crawler	Yes	
20 Price	Pricing begins at \$75,000	The company provides custom price quotations upon request

Computationally-intensive semantic systems can demonstrate a raging appetite for computing resources, bandwidth, storage, and subject matter experts. The larger the volume of content the Siderean system must generate, the more time the index update consumes. If third-party systems like Schemalogic's are integrated into Siderean, there are additional computational and storage demands. In theory, semantic technology delivers major benefits. The challenge is to demonstrate that the time, cost, and complexity of the system delivers a payoff to the licensing organization that can be verified.

Anticipated Benefits

Seamark can project digital content navigation into existing applications or Web sites, or be used through its own navigation user interface. Siderean founder and chief technology officer, Bradley Allen, said, "It's time for computers to allow you to find information the way you think." Siderean's Navigator technology dynamically organizes the available data to exploit a human's ability to recognize the information that's needed.

Other benefits of the Siderean approach include:

- Comprehensive support for Semantic Web and related World Wide Web Consortium standards
- Technology to allow a user to navigate upwards, across, and downwards through the facets of content processed by the Seamark system
- A turn-key approach that offers more flexibility than some of the search appliances available from Google and Thunderstone with the text cross-tabulation functions of the Seamark Navigator.

Possible Drawbacks

Faceted search systems evoke strong reactions in the information retrieval community. For an enterprise search system, faceted search can speed certain types of search-and-retrieval tasks. However, faceted search requires that the organization commit appropriate resources to generating and maintaining the tags.

One of the issues associated with faceted systems is the cost associated with tagging. Siderean is a system that can automate most, if not all, of the indexing and cross-linking if the source data contain tags generated by exporting

the field names from a database system along with the content or by manipulating well-formed XML documents. It goes without saying that if the content is not in a database and properly mapped to consistent field names or if the content is not in XML, additional time and money will be needed to prepare the content for the Siderean system. The licensee will want to understand the initial time and effort to tag each of the objects in the collection with the attributes. Many organizations are not familiar with SGML and XML. As a result, the variation between SGML and XML documents is often a surprise. Normalizing SGML and XML can be an expensive and time-consuming task. For some organizations, the effort may not be worth the benefits of the rich interface. For this reason, SGML and XML centric systems may be more suited for smaller corpuses. Large flows of content and frequent changes to already indexed content can cause the semantic system to become sluggish and possibly unusable.

Another issue to consider is the need to create an interface that does not overwhelm the user. Some search professionals will find that a large number of categories and subcategories may be easier for domain experts to use. Without careful interface design, the initial display of categories may confuse some users.

Most faceted classification search systems require an ongoing commitment to maintaining the classification system. Siderean licensees with structured and metatagged information will have only minor maintenance work to perform. Licensees without properly structured and tagged data will find that Seamark Navigator can work little of its magic on unstructured text. Automation can be used to extract metadata from databases.

Conclusion

Siderean's approach to search takes advantage of the human mind's ability to recognize the needed answer when presented with information that is organized into familiar groupings or categories. The company says that a human can discard the 90 percent of the information that is unimportant, drill down into the remaining 10 percent, discard 90 percent of that, and within two or three iterations arrive at what the person needs. Siderean's technology allows a searcher to see information organized into familiar contexts, not laundry lists of results.

Faceted classification is effective because it splits subjects into their component parts and allows retrieval on whichever attributes of the subject are important to the person who is searching. Special features include the combination of hierarchical browsing and searching, and the ability to switch between these two approaches as needed. In sites that show the number of hits for each option, it is clear to the user whether they should refine their search further or go straight to the item itself.

Nevertheless, there is some controversy associated with using faceted search for an enterprise search application. For this reason, Endeca (as noted elsewhere in this report) links its system to work processes and work flow. Other vendors such as i411 (not profiled in this edition of the Enterprise Search Report) pursue directory and other applications where structured data are the norm.

Siderean joins Endeca, Autonomy, Fast Search, and Entopia in a move from basic search to the wonderland of “knowledge management” via directed navigation or faceted search. The buzzword facet, as mentioned above, refers to the dimensions or multiple classes in the presentation of available data (such as color, size, and price range or sex, age range, location, and viewing preference). A user recognizes a meaningful category and can click to reach the needed information.

Siderean’s challenge is to make its approach to directed navigation simple and economical for the content provider to be able to arrange things in a way that not only makes it easy for users to find the precise information they are seeking. Siderean also allows its licensees to display a search box and follow links to other relevant data wherever it resides inside the system. In a Siderean system, a user can go from a parts listing, to the maintenance history for a certain part, to the contact information for the maintenance shop.

Stephen E Arnold

Minor edits to a rough draft on November 5, 2013

