

SemioTagger and Skyline--EContent Decision-maker Review

By Robert J. Boeri - January/February 2005 Issue, Posted Feb 22, 2005

<http://www.econtentmag.com/?ArticleID=7623>

All Content Copyright © 1998-2005 EContentmag.com - All Rights Reserved

Entrieva SemioTagger and Skyline

Purpose: A categorization and indexing engine and viewer that organizes unstructured text to allow it to be viewed and leveraged as business intelligence.

Starting Price: \$75,000 per CPU plus 20% annual maintenance for which you receive SemioTagger, the Entrieva Software Development Toolkit, one Taxonomy Workbench seat, and 27 subject-area taxonomies. \$50,000 more buys Skyline, SemioMap, and SemioDiscovery.

Reviewer's View: Entrieva offers a comprehensive product line providing information categorization, discovery, and notification software. The company's products combine to create enterprise solutions that discover and categorize information intelligently. Real-time notification via cell or wired phones, PDAs, email, and similar systems is optional. SemioTagger, Entrieva's core product, is a categorization and indexing engine based on patented technology that extracts implicit and explicit knowledge from online sources and documents. Other Entrieva products, including SemioSkyline, provide views into SemioTagger's results. These products include browser-based tools for viewing and analysis and a knowledge workbench for integrating Entrieva's systems into an enterprise solution. Entrieva has a network of channel and technology partners, helping customers build enterprise solutions or incorporate Entrieva systems into their own product.

I examined SemioTagger and its results using SemioSkyline interactively. To do so, I approached this review with the mindset of a member of an IT-business project team. My fast-growing theoretical company has 500 employees. In two short years we've managed to build several discrete content repositories, each with a different folder or navigation scheme: an intranet, our email system, shared network drives, and an enterprise content management system. We know we need a solution that will allow employees to find what they need quickly, but we are afraid that a search solution alone might just produce overwhelming, irrelevant search results—which would not promote the sort of subtle knowledge discovery we believe is critical if we are to use our content effectively. We also know we need a supplement to search systems that will help us classify our content and reveal important

information relationships. We also need help categorizing content in our various repositories that will even provide a similar way to tag, view, and navigate them.

Our needs reflect those of many firms and departments or divisions today and many taxonomy solutions are becoming available to help solve those needs. Taxonomies, after all, are merely consistent data models or classification schemes, and many are in the public domain, available for the asking. The trick is customizing and applying them to dynamically growing and changing content.

I approached Entrieva with a simple test case: I gave them a quarterly CD of information I build and use as background for my *EContent* Information Insider columns. All told, I store several thousand pages of information each quarter, mostly as PDF files searched by Acrobat. I have a good classification scheme by vendor, product, and subject (and a hobby area of tech products just for fun). The subject section under XML has several dozen folders. Like everybody else, I do not always store information—in my case, press releases or stories—consistently. My only hope is to index and search the content, but my full-text searches often yield dozens of meaningless results. My challenge to Entrieva: show me something I don't already know about my content.

Show and Tell

Entrieva jumped at the chance to show me what SemioTagger and SemioSkyline could reveal in this modest test. With only a couple hours required to set up the demonstration, they applied a standard information technology taxonomy to my content, yet they allowed the system to tweak the results based on names I'd given my folders. Most importantly, SemioTagger applied its noun-phrase analysis to develop relationships I hadn't expected. Entrieva then let me examine the result via SemioSkyline, an interactive browser-based view of SemioTagger's analysis. I was very impressed.

Semio added Sony to my predefined vendor list, based on what Semio found in the "hobby" section of the CD, which I didn't even expect to be part of the test (I've been saving reviews on digital cameras). Moreover, when I browsed the IT categories and found Artificial Intelligence, I said to myself "I know there isn't anything there...that is so 'retro.'" Wrong. In fact, SkyLine showed several links to AI content, both explicit and subtle. When I reread the EOS Rebel camera review in my hobby section, I could see why Semio suggested it fit as AI content. The review referred to features like "Eye-controlled focusing," which are indeed a kind of neural network intelligence.

I also noticed that Skyline reported the metadata title rather than the less-meaningful file name. (For every document I'm careful to supply "title" as an Acrobat metadata field.) I searched for "XML and PDF," and found dozens of suggested concepts to browse that I'd never considered, including an article I'd saved and an email from my editor about the article. My search for Sarbanes (as in "Sarbanes-Oxley") resulted in a suggested category of IT/Law and Regulation. SemioSkyline reported finding no titles or abstracts that matched, but did find full text search results in three documents.

All these findings strongly suggest that the Semio products do indeed go beyond mere keyword matching and can extract significant value from document content. They met the challenge of "showing me what I didn't already know" about my content. I only wish they offered a consumer version that I could use to manage and analyze my own content.

Picturing Success Stores

Entrieva suggested a visit to HighWire Press (www.highwire.org) to see Semio categorization in use, integrated with Verity's search system. Go there and you will see Semio's categorization of over 15 million articles from 1,500 PubMed journals, logically grouped into more than 54,000 categories. You can navigate to the articles starting with four general science categories further subdivided into trees of categories. If you prefer a visual approach, click on the TopicMap Java applet and you'll see a pop-up window with a dynamic, spidery map of categories, giving you a sense of context for each category. Categories themselves expand into your area of interest as you drag them towards the center of this map. Hover over a category that interests you, and you'll see how many documents are in the category. Select the category and return to the main page and you'll see the documents themselves. If instead you want to search by content, you'll be presented with the Verity search integration. Select from one or more related categories, then refine the search by author or keywords. This site demonstrates the power of dynamic navigation as a complement to full-text search, a capability clearly needed when browsing such a large repository of articles.

Entrieva also described a completely different approach to using its tagging tools, applied recently by the advertising agency Saatchi and Saatchi. Procter & Gamble retained this agency for a recent launch of a new skin-care product. The agency used Entrieva to analyze marketing slogans, tag lines, and product positioning found on 25,000 Web pages describing competitive products. This helped Procter & Gamble discover the areas of strength and weakness in several competitive products, such as Johnson & Johnson's Neutrogena products, in essence leveraging the marketing expertise of the competition. This allowed them to brand their new product persuasively and sharpen the focus on its target markets.

Caveats and Competition

This is a dynamic, changing field with vendor consolidations likely and there are a few issues to consider when deciding if Entrieva's are the right categorization solutions for your organization. First off, even Entrieva is careful to say that its products work with whatever search or content vendors you have, supplementing them and not displacing them. Still, some major search vendors are working on their own categorization/ classification systems. If you prefer to buy integrated solutions instead of best-of-breed, be sure you know your own search vendor's plans. Remember too that no system completely automates the building and application of classification schemes. You still must allocate human resources to fine-tune classification schemes. Entrieva claims that after a three-day course, librarians or business professionals can develop an effective taxonomy for immediate use, perhaps by

building on the more than two dozen ready-made taxonomies that Entrieva supplies with its product.

Cross-licensing is the highest form of flattery and as a testament to Entrieva's products, major search vendors and CM giant Documentum (EMC), have licensed or integrated Entrieva technology. As content search and management vendors extend their reach, they have to decide whether to develop technologies such as categorization and taxonomy management or to buy those that have already demonstrated competitive advantage. But it seems significant that Documentum, for example, integrated SemioTagger into its Content Intelligence Services, recognizing the unique value provided by that system.

Sidebar: Key Features at a Glance

SemioTagger

- Is deployed as a SOAP Web service.
- Operates with any portal, viewer, database engine, or business application via standard SOAP protocols.
- Adds value to existing search and content management systems, helping you to "stop searching and start finding" what you need.

SemioSkyline

- Using SemioTagger's analysis, provides a quick, detailed view into online document content or database and search engines.
 - Gives a hierarchical view into this content by content, category, or suggested search terms.
 - Lets you explore relationships interactively.
-

Sidebar: Business Profile

Entrieva is a three-year-old venture-backed company formed by the merger of Webversa and Semio Corporation, itself founded in 1996. Entrieva's indexing and categorization technology is widely used by end users and partners, including OEM accounts of major search and content management vendors. Since its general availability in June 1999, Entrieva's flagship product, SemioTagger, is used by more than 100 companies in the Fortune 1000. Entrieva is headquartered in Reston, Virginia and headed by president and CEO Tom Lewis. Without discussing specifics, Entrieva said that sales doubled from 2002

to 2003, and that 2004 was much better than 2003.

www.entrieva.com