

Market Development

## Basis Technology embraces Lucene and new languages in latest text-analysis release

**Analyst:** [Nick Patience](#)

**Date:** 24 Mar 2010

**451 Report Folder:** [File report >>>](#) [View my folder >>>](#)

**Basis Technology** has released the next major version of its Rosette multilingual text-analysis platform, adding name-resolution modules into the platform that it had previously sold separately, enhancing its entity-extraction technology, adding more languages and embracing open source by enabling integration with **Lucene**. The company continues to win business in its core markets of government and OEM.

### The 451 take

If multilingual search or text analysis is required, Basis has been one of the go-to software providers for more than a decade now. Its customer wins tend to come slowly, since many of them are OEM deals, but they continue to come. It won't ever set the world alight in terms of its growth rate, but as e-discovery moves beyond the US to become more of a global opportunity, we anticipate further growth opportunities in that market, among others.

Rosette 7 is the latest cut of Boston-based Basis Technology's multilingual text-analysis platform. The software can identify 55 languages; enable search applications in 21; find names and entities in 15 languages; and resolve, match and translate names in nine languages. It has use cases in search engines (**Google** was an early customer); government intelligence (i.e., espionage, counter-terrorism, etc.) applications; and as a general text-analysis engine to power e-discovery, information governance and other applications that need to understand and manipulate text.

The new Lucene support isn't all that new – parts of Rosette have been integrated with the Lucene search libraries for three years, but Basis has rewritten the connector and has seen two of its major SaaS customers switch to using Lucene and Solr, including one that closed in the fourth quarter of 2009, replacing a commercial search engine with Lucene. Basis gets paid on a site-license basis when it works with SaaS providers, just as it did in its early Web deals, such as the one with Google.

The company has made good on its previous roadmap promises by integrating its name indexing and translation technology into the platform, whereas it was previously sold separately. Put together, Name Indexer and Name Translator can take a given non-English name and match it against a database of thousands of names in different spellings and languages to identify it correctly, even if the name is misspelled or spelled phonetically. For instance, if someone, upon hearing the name of the president of China for the first time, thought his name was spelled 'Hoo Jintow,' it promises to be able to match that to the correct name in normalized English form (Hu Jintao) or the actual Chinese characters in simplified Chinese. Both products have been around for a while, but are just now part of the platform.

The entity-extraction enhancements result in more accurate and faster extraction in Arabic, Chinese, English, Japanese, Persian and Russian. Additional extractors can also be written more easily now, the company claims, and have higher F scores (the way of measuring entity extractions, measuring the balance between precision and recall) of about 91-92% out of the box. Human beings are thought to be able to get an F score of about 97% identifying entities (people, locations, dates, etc.). With this new version, Basis claims to be able to build, in about 10 days, new extractors that can get F scores in the low 80s and then be trained more easily to get to the required threshold.

The new language support in Rosette sees new Afghan and Pakistani language support with Pushto and Dari, as well as improved language detection between Cyrillic languages and more accurate name indexing for Arabic, Chinese, Dari, Pushto, and Urdu. Customer count is now at about 75, with recent wins including **Ayna**, the most-visited Arabic search engine on the Web, and **NCB Capital** of Saudi Arabia as part of an anti-money-laundering application.

E-discovery has been another area of recent activity, where the use case is enhancing software and service providers' review and analysis capabilities – the most critical and costly part of the e-discovery process. Deals have been signed in the past year with **Clearwell Systems**, **Daticon EED** and **IPRO Tech**. The initial appeal there was to identify languages, but more

recently, Basis has gone back to sell its more advanced analytics; however, it reports most e-discovery providers aren't quite ready for that yet.

Altogether, Basis has about 50 government and 200 nongovernment clients. Of that total, about 75 are paying for support. The company finds that customers often hit a point where they're happy with what they have, and stop paying for support contracts. The average deal size is about \$250,000-300,000, and while Basis signed half a dozen or so very large six- and even seven-figure deals in 2009, it is also seeing more deals coming in at five figures.

## **Competition**

Basis has a few competitors, but not many. **Inxight** was its main one, prior to its acquisition in May 2007 by Business Objects, which was subsequently bought by **SAP** that September. While the company is certainly committed to the space, text analysis has been a slow burner within SAP since the acquisition, and we're not clear on how many OEM customers – if any – it has won since then.

**Teragram** is another company with extensive multilingual text-analysis capabilities. It was picked up by **SAS Institute** in March 2008, in part due to the purchase of Inxight, from which SAS licensed text-analysis technology. SAS has made headway in integrating the Teragram technology into its own tools, as well as continuing to sell on an OEM basis. However, in the wake of these acquisitions, Basis is looking increasingly more like the go-to OEM vendor for multilingual text analysis among vendors that may compete with SAP or SAS in some way.

**Content Analyst** is another company racking up e-discovery OEMs with its own text analysis, and has been pretty successful in winning customers over the past two years. However, its technology is statistically based rather than linguistically, and it learns through training documents rather than natively identifying languages like Basis or SAP. The statistical and natural-language-processing (NLP) approaches have different strengths and weaknesses.

**Attensity Group** is another text-analysis vendor, but it previously focused its considerable text-analysis assets on voice-of-the-customer applications – with only English parsing capabilities – and not so much on the OEM opportunity. It was at the center of an April 2009 rollup into the Attensity Group holding company, along with two German text-analysis and content management vendors. New European blood could mean more languages for Attensity, and it has already announced applications for risk and compliance, research and discovery in corporate and legal processes, and intelligence analysis for government agencies. **Linguamatics** offers NLP-based text-analysis software, but focuses on markets such as life sciences, pharmaceuticals and healthcare.