Nodalities

THE MAGAZINE OF THE SEMANTIC WEB

talis®

www.talis.com/nodalities

July / August 2008

INSIDE

Review of Semantic Technology in the Financial Industry

By Eric Hoffer

Try to quantify the number of ways you've seen or heard semantic technology or the Semantic Web being explained, and you'll probably fall short. This may also be the case for those explanations making it come to pass—so far. And you've likely heard as many opinions about that being the case and why. "Great," you're thinking' "this is going to be another attempt at explaining the same thing—what it is and what it's used for—or what should be done differently to make the explanations clearer." It isn't. Well, not really, but sort of.

Not long ago, Semantic Universe (the organizers of the Semantic Technology Conference) asked if I'd help arrange a track and panel focusing on the use of related technologies in the area of financial services, at last month's fourth annual conference.

continued on page 3

SUBSCRIBE NOW

Nodalities Magazine is made available, free of charge, in print and online. If you wish to subscribe, please visit www.talis.com/nodalities or email nodalities-magazine@talis.com.

MEET US AT

Talis staff will be presenting and attending several events over the next few months, including;

i-Semantics
Austria, 3-5 September

Dublin Core Conference
Berlin, Germany 22-26 September 2008

European Semantic Technology Conference (ESTC)
Vienna, 24-26 September

International Semantic Web Conference (ISWC)
Karlsruhe, Germany 26-30 October 2008

Defrag
Denver**,** Colorado 3-4 November 2008

For further information visit: www.talis.com/platform/events

Nodalities Magazine is a Talis Publication

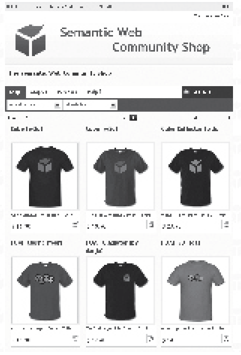
ISSN 1757-2592

Talis Group Limited
Knights Court, Solihull Parkway
Birmingham Business Park B37 7YB
United Kingdom
Telephone: +44 (0)870 400 5000

www.talis.com/platform

The views expressed in this magazine are those of the contributors for which Talis Group Limited accepts no responsibility. Readers should take appropriate professional advice before acting on any issue raised.

EDITOR'S NOTES



Semantic Community Shop

Welcome to the Summer 2008 edition of Nodalities Magazine. I'm Zach Beauvais, newly-dubbed editor and a researcher here at Talis. If you haven't seen past editions, or you're interested in receiving future issues, don't forget to sign up online at www.talis.com/nodalities.

Well, the Semantic Web meme this year seems to be: "Semantic Technology is here!" 2008 has certainly seen a growth in Semantic technologies; not just in the theoretical spheres of techie discussions, but in the actual adoption and investment in Semantic Web applications. This issue covers several real-world perspectives on the Semantic Web, from telling a Semantic Web story and the marketing of new technologies right the way through to academics discussing the benefits of Linked Data and applicable elements of the semantic stack for the benefit of education and research.

From the academic side, we have a discussion by Ben O'Steen from Oxford University's library services in which he describes the need for machine-readable data for the curation of digital resources. Oxford's libraries are beginning to "cherry pick" useful elements of Semantic technologies in order to organise and administer their vast repositories of digital resources. Also from an academic perspective, Juan Sequeda (A Ph.D student at the University of Texas and an invited expert on the "RDB2RDF W3C Incubator Group") explains the increasing necessity for relational data sources to be described in RDF.

From a more enterprise perspective, Greg Boutin gives us his view on telling a Semantic Web story to the private sector, while Eric Hoffer runs through the economic side of networks in a "Well, not really, but sort of" explanation of the Semantic Web. This edition also sees a contribution from Alex Iskold—member of the Semantic Web Gang and founder of AdaptiveBlue—in which he sheds light on the myth and realities of semantic search.

The Semantic Web community is certainly not short of interesting people. When they get together, impressive conferences and Semantic Web events are often the result. Eric Franzon, an organiser of SemTech in San Jose, gives a run-down of this past Semantic Technology conference and a quick look forward to 2009. I'd also like to mention the opening of the Semantic Web Community Shop which started as a "thank you" to SWEO project contributors. If you've ever wondered how clothing could be represented in RDF, this could be your chance to find out. They sell a range of Semantic Web community ware, including T-Shirts and emblazoned laptop bags with your favourite graphic representation of all things semantic!

Continued from front page.

This seems the first organized foray of turning the "semantics" discussion to a number of business verticals (in addition to Financials, parallel vertical focuses also include Life Sciences and Government). The question was: Who in the particular space is doing what with the technology and related approaches?

The process of arranging the panel was almost as revealing as the panel discussion itself; it offered an opportunity to step back and look at the whole of the semantic concept within a context that, itself, has many nooks and crannies. Through examination of the objectives within some of the nooks, and in some of the crannies, a useful perspective on adoption and its path emerges.

> *There seems to be a perfect storm brewing for a sea change down the pike in terms of financial information filing and consumption, which could be game-changing in the financial space*

The most obvious take-aways from the Financial Industry Sidebar were threefold:

- From the number of participants in the discussion, their level of inquisitiveness, and the intensity around parts of the conversation, people in this domain are clearly interested;
- The companies making the most visible use of semantic tools and approaches right now in the financial space are information providers; and
- There seems to be a perfect storm brewing for a sea change down the pike in terms of financial information filing and consumption, which could be game-changing in the financial space.

Backdrop

With the intention of the panel session being to provide a backdrop and to begin a discussion around how semantic technologies are being employed or considered for use within the area of finance and financial markets, it is necessary to frame a realm of finance, what is "semantic", and the kinds of concerns we might look to investigate. Open for discussion would be: related benefits and advantages, hurdles, needs, which components of "semantic" technologies are being used for what purposes, are they standards compliant or proprietary, are those efforts in-facing or out-facing…

Before going down this path, however, we'll skip right over the question of: What is "semantic" by making the assumption that the definition for this purpose could include anything from among: NLP and entity extraction, ontologies and taxonomies, use of RDF, OWL, SPARQL and/or related semantic methods. It is necessary to frame what we mean by "financial"—continuing a familiar issue of disambiguation.

As with the definition of "semantic", there are surely many views on what falls within the scope of "financial"—from corporate perspectives to personal ones, from mergers and acquisitions to payment systems, from research to risk management, from securities to physical real estate, or even the financial aspects of health care… In the spirit of Semantic Universe's desire to zero in on a digestible realm for a discussion—yet with a wide enough net to capture interesting and related topics to engage the group, the field was set to include related use in any of the following realms from within Finance: Investment Banking, Lending, Insurance, Trading, Asset Management, Securities Research (equity and credit), Brokerage.

In one regard, for example, Transaction Management drives the Trading realm. Insurance and Lending activities focus significantly on risk assessment and management. Investment Banking, as well as Equity and Credit Research and the Ratings worlds all involve intensive information harvesting, storage, classification and modelling to facilitate their analysis. Asset Management needs to efficiently digest massive volumes of unstructured information to identify important details and trends.

The Mechanics

A common element among these areas, besides their being focused on the value, possession and movement of money, is that they all have (and are driven by) specialized analytical needs, involving massive amounts of information, and requiring unique and flexible structures for that information so that it can be modeled and processed privately, securely, reliably and quickly—in support of decision processes. Along with the plumbing and wiring of these processes, the raw inputs to these processes (information about the companies, instruments, economic trends…) are essentially the lifeblood of the financial markets. In addition to examining what might be going on within companies in these spaces in their achieving these ends then, core to this discussion then is also examination of what information (and related service) providers are doing—either proactively or reactively—around their support of the financial space.

> *As with the definition of "semantic", there are surely many views on what falls within the scope of "financial" - from corporate perspectives to personal*

What follows is a run-down of some of what was exposed during the organizing process and at the session itself:

Dr. Christian Halaschek-Wiener, CTO of investment firm Clados Management LLC, facilitated the dialogue during the session at the conference, by acting as moderator. To this, he brought his interest in creating semantic representations of unstructured information, and his experience in developing content syndication frameworks using OWL to represent published content and perform finer-grained filtering than would otherwise have been possible using mainstream frameworks such as RSS. In his work, we see one of the primary issues with applying semantic technologies to the financial domain: its need for realtime performance in its information resources—addressed by him in his previous academic work in which he developed incremental reasoning algorithms, and demonstrated through his collaboration with a major financial news provider.

> *While there may have been trepidation at the highly traditional enterprise level, they were able to increase the reach and effectiveness of Credit Analysis knowledge management, information discovery and integration*

Transaction Enablement and Management

While Ioachim Drugus, Main Architect at SemanticSoft, was not able to join us for the conference, he shared some insights in advance regarding what they are doing in EDI (Electronic Data Interchange) and Transaction Management / Enablement, which drives the Trading realm. The former, in support of transactions in ongoing commercial situations, tend to be more stable and require less flexibility than the latter, for Financial Markets, which are more often one—off transactions. Due to divergent standards (UN/EDIFACT and X12), with a variety of industry specific flavors, and some also adopting XML counterparts, the plethora of data types and sources makes reconciliation and interoperability quite a task. Their SemanticServer is a system for distributed management of data in repositories on different hosts, with strong metadata management and SPARQL search. There are some prospects for semantic technologies which parallel EDI, but we'll come back to that later.

Insurance and Credit

At a higher level, Jonathan Mack, Senior Technical Architect with Guardian Life Insurance, shared his perspective on the place for semantic technology in enterprise architecture at big insurance. His presence indicated a recognition of the value and possibilities of semantic technology inside the insurance enterprise. His message was, while perhaps radical for the conservative insurance world, conservative for the evangelists of the semantic world (pragmatic overall). Using SOA experience as a model to demonstrate adoption issues for semantic technology, he emphasized that the approach for the latter needs to be more about the solution to problems than the means to the solutions, and the interface over the technology—with the collateral being the addressing the enterprise issues of complexity, data integration and interoperability. A model, perhaps, for all who seek adoption in their own enterprise.

Engaged with one of the major Credit Ratings Agencies—still at the conservative end of the financial spectrum—John Robert Gardner, VP and Director of Content Management Services with Digitas, described some of his work. In short, he guided development and implementation of taxonomies and ontologies (using RDF and OWL) for the Agency to broker interaction and bridge between legacy enterprise systems for Credit Ratings. While there may have been trepidation at the highly traditional enterprise level, they were able to increase the reach and effectiveness of Credit Analysis knowledge management, information discovery and integration.

John emphasises that: "the complexity of financial services data and relevant information context is so vast that the industry is rightfully wary of "taxonomy" as a way to address scaling and integration complexity for both the customer and researcher-facing portions of their information network. Ontologies, in fact, present a more addressable course of approach, appearing to stakeholders as an incremental and less rigid organizational structure yielding/supporting multiple taxonomies for different uses."

Another sign that semantic technologies are being employed deeply within this corner of the financial markets, Clark & Parsia LLC's recent release of Pellet, an OWL DL reasoner, was sponsored by Moody's KMV, which is using Pellet and OWL for internal R&D projects related to MDA and data integration efforts between different parts of their overall analysis service offerings.

Big Financial

Large banks and Securities Industry wire-houses were on the minds of many of the attendees of the panel session. Some speculated that their lack of physical presence for the session might be an indication that they're not making use—or if they are doing so, they prefer to keep it quiet, perceiving it as a competitive advantage. Contrary to both of these suspicions there is activity in this area, and while they were not able to make it to the conference, they did share some of what they're doing.

In one such case, David Palaitis of Citi explained during the planning stages for the panel how, for a fixed income group, they were using RDF and ontologies to improve a legacy Regulatory

Compliance system. The intention was to enhance the process of relationship discovery between clients, contacts and analysts. Until then, knowledge workers had to make menu and checkbox selections to discover relationships held within a massive database. By defining a simple ontology for analysts, and contacts, for example, and loading the database into a triple store, they enabled an intuitive navigation and discovery process. Among key benefits of using micro ontologies, David emphasized their enabling rapid application prototyping and proof of concept.

While Shahin Nassiri of JPMorgan was also in the end unable to attend, he shared, in materials submitted for his presentation and participation, that they are making use in for knowledge management purposes by "defining a business process taxonomy and mapping [their] application suite to that taxonomy using OWL. Previously, that type of information was captured in a combination of documents, which were not easy to maintain, modify or perform queries against, and especially not easily used in conjunction with an inference engine. The purpose is to have supporting material for strategy decisions by senior management."

> *Among key benefits of using micro ontologies, David emphasized their enabling rapid application prototyping and proof of concept*

Related to banking and finance, on the consumer side of the equation, are the efforts of Garlik. While Tom Ilube, CEO of the UK based identity monitoring and management company, had to leave the conference early and could not, therefore, participate in the session, he shared that they have embraced semantic underpinnings in their architecture from the ground up. From the outside, they give off a Web2.0 look and feel, and you wouldn't appreciate all the heavy lifting going on behind the scenes. This is an embodiment, on the consumer product side (though serving enterprise too, in that client companies provide as a service to their own customers), of what Jon Mack of Guardian had described (discussed above) as stealthy utilization of the necessary tools without a whole lot of jumping up and down about their use, and with the focus being on solving the problems.

Big Information

Particularly relevant to the processes and workings of these financial institutions, and their interaction with the markets, is the flow of raw materials into their systems. For this reason, a look at their information providers, is worthwhile.

At Thomson Reuters, they paint a strong picture of adding semantics and structure to their traditionally unstructured, textual news information. Thomas Tague, who runs their ClearForest unit's Calais initiative (spearheading strategy, product development and partner relations) emphasizes open use of metadata tagging to establish and leverage meaning and connection among and between information. They hold out their NLP tools and capability for anyone to use: Calais enables submission of documents, to be returned semantically marked up; and Tagaroo, a blogging tool that provides images and links which relate to the post an author is working on (paralleling some of the capabilities of Zemanta—whose Andraz Tori is also featured in this issue of Nodalities). By openly enabling use of entity extraction and identification within context of financial news, one potential vision is to leverage the accumulated semantics to expose subscribed content that relates to whatever else a user is working on.

Given Christine Connors' title at Dow Jones (Global Director of Semantic Technology Solutions for the Enterprise Media Group) it is pretty clear that the company is serious about their focus on leveraging semantic technology. Christine explained that a large part of her role involves overseeing development of taxonomies and the metadata used to add value to their news and financial information products. She also oversees product development of Synaptica (their software product for creating and maintaining enterprise taxonomies) which is aligned to a full spectrum of standards for controlled vocabularies: Z39.50 through to RDF, SKOS, and OWL. She noted their having helped to move some of their data—feed clients from XML to SKOS. In addition to helping their customers better manage data, Dow Jones is eating their own cooking—by migrating their management processes out of their legacy systems and into Synaptica to internally take advantage of its capabilities as well.

Targeted Information Services

A notable observation Christine Connors made about the discussion was that there's an interesting delineation between the "focus on semantics—for—machines (in support of algorithmic trading for example)", a large focus for Dow Jones' semantic attention to date, "and semantics—for—people (for purposes of business/competitive intelligence). Calais' attention in terms of outward effort seems more around the latter—but generation of metadata through this approach can also be used to drive the former. With Dow Jones' recent acquisition of Generate, whose g2 crawls the web for news and blog entries, extracting information about people, products, companies, it seems they too will be playing in both sandboxes.

This brings us to another, more subtle if not controversial area—the identification of patterns and trends, and the leveraging of sentiment that might be derived from news stories and press releases. Capabilities in these areas, combining NLP with pattern recognition, and then comparison to those patterns—or in essence, benchmarking gleaned from unstructured information—go directly to efficiencies in determining which material to give ones attention to. In a sense this extends (or blurs) the delineation Christine noted, making it “semantics—for—machines, for—people”, useful for decision support throughout the financial space, from Corporate Strategy and Risk Management, to Research and Asset Management domains, for example.

This too is the focus at FirstRain. COO, YY Lee joined us and shared their use of concept extraction, pattern detection and qualitative analytics, as applied to web content, in provision of their qualitative analytics information service platform for the Asset Management community. Rather than being a primary or raw information provider, as was the case with the two previous examples, FirstRain adds value by enabling customers to search and analyse volumes of unstructured information which they would not otherwise have the capacity to handle, and enabling the detection of trends and details they may otherwise have missed. Other players who seem to be using this approach, and targeting similar needs, appear to include SkyGrid, and perhaps Jodange.

Netbreeze co—CEO, Leo Keller, also joined us in the panel, and his description of their approach is very similar, using NLP and concept extraction, with automated ontology navigation and management to identify patterns in series of unstructured text. Their offering seems less a standard platform for a particular realm (as FirstRain seems for Asset Management). Instead, they target a range of issues within the financial space—such as Anti Money Laundering, Fraud Detection, Early Warnings, Risk Analytics—with numerous large European Financial Institutions as clients. In Private Banking, network extraction and analysis aids in the Anti—Money Laundering tools, and combined with business intelligence is used for identification of new potential clients. They apply monitoring of financial and non—financial events and evaluating in context for Investment Banking, Asset Management and Hedge Funds needs, in the way described above for using web based content to dynamically establish and compare to relative benchmarks.

Tributaries and the Inputs to Market Workings

Circling back to our earlier discussion around EDI, the value of its having enbled automation in commerce is unquestioned. So too, some say, can financial reporting standardization for the entirety of financial workings of companies, and financial markets. To be sure, the SEC is now mandating that XBRL be the business reporting format for public companies, phasing in over the next few years. Despite the many challenges faced as a result, panel moderator Chris Haleschek noted that “it is clear that the financial reporting standards groups (specifically XBRL and RIXML) see the traction in the Semantic Web community and are interested in its potential”.

Eric Cohen of PriceWaterhouse Coopers, one of the original founders of XBRL and founder and chief architect of the XBRL Global Ledger Framework, joined our discussion and outlined some of the potential utility of having standardized formats for representing financial information. The standards groups have in essence taxonomised the financials of industry verticals, for consistency within any industry segment. While XBRL is simply encoded in XML, the W3C and the XBRL standards community have begun to work together. If all goes well, we could see a melding of the “SME—tises” from the Semantic Web standards world and the XBRL community. Embracing of the best of both worlds, with neither having to reinvent the wheel, could change the fabric of financial information processing and consumption.

> *“it is clear that the financial reporting standards groups (specifically XBRL and RIXML) see the traction in the Semantic Web community and are interested in its potential”*

Speaking of transformation, this is just what Elmar Drewitz of DrewITz Consulting in Munich has been focusing on with his work on XBRL to OWL mechanization. Elmar was not able to attend, but to summarize what he’s described: Embedding XBRL taxonomies into OWL ontologies, and augmenting with Semantic Web rules (stronger than OWL DL) facilitates the creation of applications such as the transformation of financial reports from national GAAP to IFRS in the European context (or even to US GAAP, though the SEC recently decided to allow submission of financials to it using IFRS) [1]. With EU companies having to move to IFRS starting in 2009, and US companies to XBRL for GAAP, such mechanization will ease at least some of the transition process issues.

Big Picture

While many struggle to better understand ways that semantic technologies can add value—the value propositions if you will—we’ve seen not just data integration and interoperability, but examples enabling more, deeper and better quality research, bridging qualitative and quantitative analysis, greater efficiency, and improvement of user experience. Significantly, we’ve also seen these not just within financial institutions themselves, but in the feeders of the information that fuel the financial markets and its ongoing effort to leverage ever more minuscule value differentials in the fewer and fewer milliseconds before equilibrium momentarily returns.

[1] http://tinyurl.com/6fsqjc

Eric Hoffer is Co—Founder and Director of Second Integral.
LinkedIn http://www.linkedin.com/in/erichoffer
Blog: http://www.secondintegral.com/axonomics

SemTech 2008 Highlights

By Eric Franzon

"Semantic Technology is here." That was the resounding message delivered at the 2008 Semantic Technology Conference ("SemTech") that took place 18-22 May in San Jose, California. At previous SemTech conferences, there was a lot of discussion around semantic technologies 'coming soon'. This year, we saw enough solid case studies, experienced enterprise-grade tools & services, and heard major announcements by large organizations that it was clear that semantic technologies have arrived in the business world. As Alex Iskold (AdaptiveBlue) put it, "There was a consensus that many technologies are nearly ready or ready for prime time and that 2008 is the first year when Semantic Web is coming out of the stealth mode."

Following are some key highlights and takeaways from the more than 150 sessions available to the SemTech audience.

Pre-conference

The conference kicked off with a pair of Sunday evening sessions designed to serve as a jumping-off point for the week. First, Dave McComb, president of Semantic Arts and a SemTech co-chair, delivered an "Introduction to Semantics" that focused largely on semantics in enterprise computing. Dave presented the challenge we have created while "building systems for 50 years with just a casual approach to what the information meant." This, he said, leaves us 'drowning in data'. He then laid out a case for how semantics may help us out of this situation. Along the way, he defined many core concepts that would prove useful as the conference progressed. Much of this material is available as a recorded webcast at **www.semantic-conference.com/news.**

Dave was followed by Ivan Herman, the Semantic Web Activity Lead of the World Wide Web Consortium (W3C), who gave an encouraging "State of the Semantic Web". Encouraging both because of the wonderful worldwide community that has emerged in recent years; and due to the sheer volume and quality of standards, tools, and other resources that are available (triple stores, reasoning engines, middleware, converters, Semantic Web browsers, development tools, search engines, wikis and CMS systems). Of course, there is still a long way to go and Ivan spent some time devoted to the important topic of what has not yet been accomplished.

> As semantic technologies move more into the business world, the need to reach 'mainstream' audiences increases

Mix of Business and Technical Topics

Monday was Tutorial Day, and the program provided opportunities for people from both business and technical realms to explore core ideas, concepts and case studies. Some Tutorials covered foundational and technical topics such as Common Logic, Ontologies, Proof & Trust, Taxonomies, RDF, and SPARQL. Others addressed business applications: "Semantically-Enabled Service Oriented Architecture", the "Semantics in Social Networking", and "Semantic Mashups" and "Financial Services XBRL". The diverse educational and technical agenda continued throughout the conference. Following are some highlights.

Experience from the Cutting Edge of the Semantic Market

Nova Spivack (Radar Networks), delivered the opening keynote and talked about Twine. The focus of Twine is "interests". It's a different type of social network that allows users to store, track and share areas of interest. A key element here is that everything in Twine is generated from an ontology. All of Twine.com comes from an application ontology: user interface, graphic elements, etc. Additionally, while the data is also based on an ontology, the folks at Radar are starting to bring in other ontologies and will eventually allow users to create their own ontologies.

The Next Generation of Semantic Development Tools – The Challenge

Eric Miller (Zepheira) gave the second keynote to present the work Zepheira has been doing around the idea of "Repurpose, Reuse, Remix". He demonstrated Exhibit from MIT SIMILE. Exhibit is a software service for rendering data. By sending data to Exhibit, one can create a faceted navigation system without building a separate database. From an interface standpoint, Exhibit can be styled in different ways. Eric also discussed other best-of-breed tools that Zepheira uses to create interfaces, manage data, etc. For an example of a system built on Exhibit and some of these other tools, see **http://www.semantic-conference.com/scheduler.** Eric challenged ten companies to try out these tools, to come back to SemTech 2009 with reports, and to share what they have learned. This was an exciting demonstration of how semantic technologies can be used to put the power of data in the hands of end users.

Taking Semantic Technologies to the Masses

As semantic technologies move more into the business world, the need to reach 'mainstream' audiences increases. This has not been an easy task, and many at the conference spoke about the difficulty of 'selling semantics' to these groups. Thomas Tague, of Thomson Reuters, sat on the panel, "Taking Semantic Technologies to the Masses", and quoted one of his customers: "If you have to explain it, I don't want it.". Also on this panel were Carla Thompson (Guidewire Group), Josh Dilworth (Porter Novelli), Mark Johnson (Powerset), and the consensus was that companies need to get away from semantic buzzwords, and discussions of the underlying technologies. Business users don't want to hear about RDF, OWL, etc. The marketing panelists also agreed that it is often marketing hype itself that is getting in the way as products and services are unable to deliver on grandiose claims. Managing expectations about what semantic technology can and cannot achieve is critical. Another conclusion from this panel was that User Interface is key. Without a good UI, even the best products will fail. In a later panel, Tom Gruber, (Stealth-Company.com), went so far as to call intelligent design the 'killer app' of semantic technology.

Managing expectations about what semantic technology can and cannot achieve is critical

Venture Capitalist Panel: Investor Opportunities and Pitfalls

Some of the top VC's in the semantic space assembled in a panel for an open Q&A session with the SemTech audience.

Stephen Hall (Vulcan Capital) indicated that he thinks there are tremendous opportunities in the space. He pointed out the vast amount of unstructured information in the world and that amount is increasing rapidly, thereby creating more need for semantic technologies to automate processes and assist with the 'drowning in data' problem.

User Interface is key. Without a good UI, even the best products will fail

Amanda Reed (Palomar Ventures), pointed out that VCs do not think about semantic technologies as one market. Rather, the ones who truly comprehend the power of semantic technologies fund companies for which there is market need and opportunity. They do not fund solely because a company uses semantic technologies. VCs are looking to invest in companies that are solving real problems.

Building on that idea, Eghosa Omoigui (Intel) indicated that one of the strategies is to look at solutions based on a theme. There is a great diversity in Semantic Web technologies and problems, so by adopting semantic technologies as a theme, VCs can build a portfolio of companies that collectively offer solutions along the entire stack. In this scenario, one needs to look for the best company in each niche,. That is, one with a clear proposition and superior technology. Eghosa closed this with a friendly warning: if someone says, 'we only invest in one company in the space,' run, because that means they don't understand the space.

Rising Stars of the Semantic Web

This CEO panel included Barney Pell (Powerset), Alex Iskold (Adaptive Blue), Ian Davis (Talis), Nova Spivack (Radar Networks), and Tom Gruber (Stealth-Company.com), and was moderated by David Scott Lewis (Startech Global).

Each panellist gave a brief demo and then time was given for questions. In answering the question, "What is the best opportunity that would get you a quote in the New York Times?", Nova said that staying away from the 'unsexy' would help. Unsexy would be talking about triple stores. 'Sexy' on the other hand, is typically found in applications. With a great application to talk about, the press, and the general public, in turn, will pay attention. Alex gave an example of an application that might get attention: the intersection of iPhone and the Semantic Web. That is, an application that understands your contexts as you move around. Tom spoke about "the interface", saying that companies need to address basic human needs. Ian agreed, suggesting an application dealing with travel. Last but not least, Barney put forth search advertising and publishing as an area that will grab attention.

In discussing why some of the big analyst firms were not in attendance, Nova suggested that these firms don't think of "Semantic Web" as a separate space – a theme we heard from the VCs earlier in the week. The analysts are still learning about semantic technologies and for now just want to know how it's going to help the categories that they are already tracking.

Closing Panel: Bringing Semtech Back to the Business

Here are a few gems from the panellists as they discuss the business value of semantic technologies.

James Hendler (Rensselaer Polytechnic Institute): I think we have several different business models that have been evident at this meeting. What you've seen with companies like Radar, Metaweb, Powerset, Garlik, and many others, is that this stuff is really coming true now. It's not the researchers showing 'cute' technologies like it was at the first of these conferences.

Jeff Pollock (Oracle): The value that semantic technology brings to some area should be defined in terms of business value that people are already willing to pay for and should be adding value to areas that are already generating revenue.

'The key message is 'simplify'

Jonathan Mack (Guardian Life): If you improve how the web works, you're going to get buy-in from business…figure out how to improve the web experience and how to develop tools that make it easier for the next person to build stuff. The key message is 'simplify'.

Christine Connors (Dow Jones & Company, Inc.): There are only a handful of cases where you can make a true justification for your return on an effort for a standalone semantics project. It generally has to be part of a larger goal. Consider semantic technology as one of the options for how you're going to build it out. Start examining them; start bringing them into the business; start exposing people to them. And then, as you continue to do that, you'll work them into your organization.

Ivan Herman (World Wide Web Consortium): There's a large variety of applications, requirements, tools, etc. out there, and a wide variety of communities; each with their own needs. For example, Health Care/Life Sciences is a community with a very different set of needs than, say, the XBRL community or all of these various software startups operating in the Web 2.0, 3.0, 2.5 – whatever we are calling it. Each of them have their own business model; I don't think there is just one.

Stephen Hall (Vulcan Capital): As an early stage venture investor, one of the more interesting dynamics we see (and most of our view is consumer web-centric) is that semantic technology is positioned to disrupt and even destroy value before they create value. An example is LinkedIn and FOAF. If FOAF enters the mainstream, Linked In is dead. Once data becomes open, portable, and semantics can expose that data in a structured way, these existing systems and services that we have decrease in value.

John Gilman (Blue Shield of California): (speaking on security in semantic systems) Ontologies can offer a very powerful level of security. In my view, security belongs with the data – that's what you're trying to secure. Because of the ability of ontologies to infer things about the user that's requesting the data and infer the match between the two, security becomes rather simple.

Once data becomes open, portable, and semantics can expose that data in a structured way, these existing systems and services that we have decrease in value

About the Semantic Technology Conference

With over 1,000 attendees in 2008, SemTech is the world's largest annual gathering of users, technologists, publishers, developers, specialists, innovators, and entrepreneurs targeting comprehensive, real-world business applications using semantic technologies. The next Semantic Technology Conference is scheduled for June 14-18, 2009.

Eric Franzon, Vice President of Semantic Universe, organises the Semantic Technology Conference, and is currently preparing the programme for 2009. http://www.semantic-conference.com ericaxel@wilshireconferences.com.

New Web Cambrian Explosion

Andraz Tori, co-founder of Zemanta, discusses the importance of applications on the Semantic Web.

By Andraz Tori

Everyone is getting on the Semantic Web bandwagon: creating, curating and publishing both linked and annotated data. Nearly every day, mashups using Semantic Web and web services from big platform providers are born. By opening up the data the new platform has been quietly established.

Definitions of the Semantic Web are as elusive as definitions of Web 2.0—it means different things to different people. However, to me it seems that the most obvious and direct consequence has been mostly overshadowed by bigger visions. **Semantic Web technologies make services that were prohibitively expensive to create cheap and possible—be they big services or small.**

Data acquisition for services is becoming increasingly affordable because data is now (getting) well organized, exchangeable, accessible and in some cases even free. When looking at things on the scale of the web, there are two big data sources that are now at the disposal of every developer, entrepreneur, artist and investor.

First is Wikipedia and its semantic derivatives such as dbpedia and Freebase. The second is data from social networks opening up gradually, but surely. One describes topics of the world and our lives at large, while the second manages to describe our lives at a micro level. The third benefit occurs when the rest of the data sources map their information to one of those two. Zigtag and Faviki are initial examples of services trying to map a Wikipedia view of the world to the web at large (and I strongly suggest trying them out).

Is This True Semantic Web?

For the true crusaders of Semantic Web this is a "little semantics goes a long way" approach. But it can also be seen as bottom-up approach which will end up using more detailed ontologies for specific purposes while still being connected to the larger sphere of the web and leveraging its potential. Lately even Cyc maps some of its knowledge to Wikipedia.

Not all of Semantic Web data will be free, but it seems that easy exchange, augmenting and repurposing of data will lower the barrier for competition and thus drive prices into the ground. As creating and maintaining complex datasets becomes easier and almost implicit, more people and companies will be doing it. Along with the encyclopedia market is another interesting case. Microstock photo providers managed to undercut prices of large photo providers with the help of both web architecture and the social fabric which made it cheap to aggregate supply from low-cost sources.

Data acquisition for services is becoming increasingly affordable because data is now (getting) well organized

One thing that I hope is yet to happen is social networks competing themselves into opening the data for repurposing to third party applications (with appropriate mechanisms for privacy protection in place). While still being walled garden, Facebook is a very interesting case. You might not have noticed, but Facebook's platform is just one step away from being a vision of automated agents come true. The data is in there, the development platform is standardized, developers with millions of ideas are all there. The question only remains: How far Facebook will let applications act on their own?

Beyond Web 2.0

The second web Cambrian explosion (Web 2.0) came into existence largely because of the abundance of cheap infrastructure for both hardware and software. Suddenly, startups didn't need to buy large servers up-front. Instead, they could rent a few and scale up when needed. And as importantly, software infrastructure became cheap to acquire and build upon. Open source has created a huge, diverse and extremely low-cost software stack that many startups are leveraging. Not only that, but it is standardized enough that developers don't need long lead-in times to start crunching useful services out. The trend is going even more in the direction of "infrastructure as service", and we are seeing companies like Amazon and Google not only offering their CPU cycles and storage, but whole database engines, queuing services and similar. Web 2.0 was enabled by cheap hardware and software infrastructure, this time web

Open source has created a huge, diverse and extremely low-cost software stack that many startups are leveraging will be enabled by cheap data. Cheap not only in sense of not having to pay for it, but cheap when acquiring, using, storing and reasoning with it. In other words: **Data will become cheap to find and to use.**

Killer App?

While it seems that universal access to data makes many things possible and cost-effective, it also makes it more difficult for some business models—especially with the "free" mentality of the modern web. But I will not go into the business models of web startups, since so much is being written about that already.

Aside from artists, everyone is asking the tough question: "What is going to be the 'killer app' of the Semantic Web?" Unsurprisingly, answering: "Semantic Web is the killer app" does not satisfy developers and entrepreneurs, much less investors! It might satisfy visionaries, but visions rarely put bread on the table.

For investors it is sometimes hard to accept the fact that the killer application is the distributed platform itself and that there seems to be no gatekeeper emerging that would cash-in on the "Semantic Web" as a whole. Bets are placed all over the space, from search engines that 'understand' your questions, to social networks using semantics; but the most of success currently seems to be in vertical search engines.

> *Data will become cheap to find and to use*

However there is a problem a lot of these companies have stumbled into. **Semantic Web is sold on the promise of computers understanding the human world and affairs on one level or another. And when you tell that to common users, it is unbelievably easy to completely disappoint them.** Just look at the media coverage of some recently launched startups in that field.

Dreaming of Apps

So are there any applications that might come into existence and not disappoint? What if we manage to establish Semantic Web technologies as ambient to the fabric of the web? Which works best when users don't want to pay too close attention, but want improvements across the board?

Well, I can think of some and I am sure you can too. My question is why we are not seeing more work in that area?

Lets start with social networking: a social network such as LinkedIn knows where I am, who I work for now and in the past, and has a list of my business associates. It knows all about my education and career. Add a bit more information and it could easily know hobbies, attended conferences, and my connections by type. This could be pulled from other networks such as Facebook. **Why doesn't LinkedIn let me state the direction I want for my career?**

It could then use that information to automatically discover who I should get to know to achieve short and long term career goals, couldn't it? And, while we are at it, why wouldn't it look at mine and his/her calendar and schedule a time and place for a meeting? Oh, and since I am meeting someone I don't know, it can offer a list of topics to break the ice with - matching hobbies, friends and maybe interesting facts about places we both worked or people we both knew. I'd just love to answer a simple yes/no questionnaire every week and select between a few people and have a meeting setup automagically. Could it not consider just the people already in a position to help, but also people that are expected to go into those positions?

Long term benefits of this far outweigh a few mismatches that might happen. Is anyone doing research on predicting career moves based on social network dynamics?

The general idea here is that social networks have semantic data that can be leveraged by 'learning machines' for great recommendations of how to develop one's career and life. You could call it "automatic career steering" or "managed real-life social networking".

If you want to get a bit wilder, think about your mobile phone. It is following you around, knowing where you are and who you communicate with (discovery of proximity via Bluetooth), it could even record all the audio on and off the phone. Indexing this audio is difficult, but with good context—geographical, social and cultural—the tasks becomes a bit more tractable. Maybe it could work in some specific situations: it could listen to your phone conversation and put the meeting you just verbally scheduled into your calendar. Maybe it could provide an instant recall for the book you know somebody mentioned a week ago or maybe it could even create automatic meeting minutes for you. But all of this is hard to get right every time. That's why every morning, your computer would ask you a few questions about conclusions that came out from processing all your yesterday-data while you slept.

> *Social networks have semantic data that can be leveraged by 'learning machines' for great recommendations of how to develop one's career and life*

Getting real (with Zemanta)

The previous paragraph is quite a stretch of imagination, but some things are already doable today. Imagine you are an author writing an article, blog post or report. Right now your computer is a tool to let you input the text and put a bit of design in it. But why couldn't computer at least try to figure out what you are talking about? And just give you some addition material on the side; unobtrusive, but possibly useful. It can make mistakes, but it could bring real benefit enough to be perceived as useful. What could a computer suggest?

Well, it could establish relations between your writing and other semantic sources. We already mentioned two – Wikipedia as world at large and your social network as your microworld. There is high possibility that parts of your writing are going to map to those two sources. Maybe you want to know something more about the term you just mentioned or maybe you want to offer your readers a chance to read about it themselves and insert a hyperlink to the further reading. Natural language processing is capable of doing those things today.

> *Computers can automatically discover (free or commercial) images that might illustrate what you are saying in your text!*

Maybe you would like to know who else has written about a topic? Maybe someday I'll write something that agrees with John C. Dvorak for once and I'd like computer to warn me about that extraordinary fact!

Computers can automatically discover (free or commercial) images that might illustrate what you are saying in your text. And you can pick and include them with one click. And published tags can be automatically suggested to make discovery of their content easier. Gradually even more types of suggestion can be implemented as the "understanding" of text gets better.

All these are things we do today at Zemanta. We are trying to hide all the complexities of the process of "understanding" the text and matching it with the semantical sources. The user never has to know. He or she just wants suggestions that make writing the process more efficient and the end product better. And when a user selects specific suggestions, maybe we can even sneak in (with his permission) semantic annotation that come handy later on when semantically-capable search crawlers like SearchMonkey come along. Currently this technology makes most sense for bloggers, since they can use all those suggestions to make their blogging easier and in some cases more profitable. However, maybe other people want to do different stuff with it, and that's why our API was born. It is currently in testing and open to anyone who sends a mail.

Maybe others have better ideas than us. Why wouldn't this technology come handy in any CMS and in any application where you have to author text? Or in your online word processor, or maybe in an email program? When I type "Hi mum, I am on holidays in Tenerife" in an email I want the computer to suggest me a photo of Tenerife where I am at. Mum would love it!

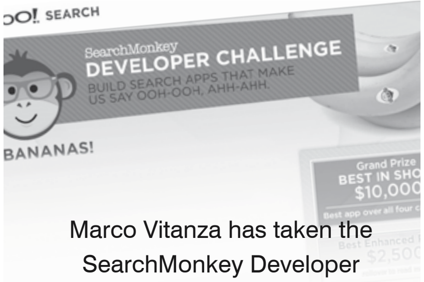
Andraz Tori is the CTO and Co-Founder of Zemanta, an applied-semantics startup aimed at blogging.

Notices

umbel

From the world of Linked Data comes the public release of UMBEL—a lightweight ontology for relating Web content and data to a standard set of 20,000 subject concepts.
More at http://www.umbel.org

Sir Tim Berners-Lee has made mainstream appearances in the UK explaining and promoting the Semantic Web: first on BBC Radio 4's "Today," <http://bit.ly/23sE3O> and also in an interview with the Guardian <http://bit.ly/2OaO6d> .
Talis coverage at: http://bit.ly/1BkYAK.



Marco Vitanza has taken the SearchMonkey Developer Challenge Grand prize - worth $10k - for his Blogspot Infobar.
Other winners included: David hinckley (Best Data Service), Greg Schechter (Best Enhanced Result), BooRah (Best Infobar) and StumbleUpon for Innovative Structured Data

Cherry-picking the Semantic Web

Ben O'Steen discusses Oxford University's efforts to improve access to their research output with the help of semantic technologies

By Ben O'Steen

Implementing ideas from the Semantic Web makes the discovery, reuse and more importantly, curation of digital resources easier.

Libraries and institutions have long realised that to make their resources as accessible and as relevant as possible, the resources need to be put on the web and in as free a form as possible. Researchers can go online with minimal technical know-how and they are able to find information they may not have found otherwise. However, enabling this is not as trivial as it might seem to the user - it requires massive amounts of computing power and engineering. Users take advantage of powerful supercomputers every day, often without realising it - massive computing clusters are put to task, reading and cataloguing all the web pages that can be found, thereby powering the search engines we all use. Large teams of highly skilled people are employed to write the code which controls how the information is catalogued and how this catalogue is then searched.

Yet, with all the time, money and sheer individual effort spent on search engine technology, how often does the first hit match exactly what we are looking for? How many useless pages do we have to look through before we find something relevant? Why are we in this position in the first place?

"Garbage in, Garbage out"

The majority of websites are written with only human readers in mind and this tends to cause a lot of the contextual information on the page to be unreadable - "garbage" - to a web crawler. Whilst natural language-processing and other related algorithms can extract some form of meaning from the text, these approaches are based around the idea of 'recovering' the information, rather than 'reading' it.

The distinction between recovering and reading parallels the distinction between code-breaking and reading the code with a key; A code-breaker may eventually break the code, but it is far easier for the message to be readable in the first place.

Unlike mathematically-based codes, languages keep evolving and changing and are contextual; a phrase's meaning can depend on who is 'saying' it, to whom and when. A successful method that 'breaks the code' for one piece of text, may not work well on a second text. If the recovered information's accuracy cannot be trusted, this affects every service that might re-use it.

Linked Data or lack of it

There is an issue with the way that information on the internet is interconnected, or rather, the way it really isn't connected in a way that we can make sense of. The basic, crude and vague link mechanism is used everywhere, but it can be very hard indeed to qualify what this link means—a link on a webpage can be used to cite a work, to reference data, to criticise, to agree or disagree, to refute or support, to convey a emotion, or simply to point to a funny picture of a cat.

In my opinion at least, the easiest way to improve both the accuracy and relevance of search engines is for content providers to represent knowledge in a global way. A description of a resource, and a context of how that resource interrelates with other resources in a global network need to be represented. Even a cursory glance at what the Semantic Web is all about should indicate that it and the standards around it are a tight match to what we seem to need.

The idea of making the pages of a site machine-readable - that is, making the information presented to a human reader similarly understandable by a machine - is not a new one, but is rarely implemented as a priority.

The Oxford University Research archive has a remit to store, preserve and—if copyright and/or IPR are not issues—to disseminate the research publications and other research output produced by members of the University of Oxford. Content includes copies of journal articles, conference papers, theses and other types of research publications.

As the lead developer on this project, I feel a strong responsibility to make sure that the work is as discoverable and as re-useable as possible. The quote that I often think of is that "the coolest thing that will be done with your data, will be thought of by someone else." Data on the scale of the web is really only going to be manipulated by computer agents, and so it follows that to provide discoverable and re-useable content, the archive's interface will have to be machine-readable.

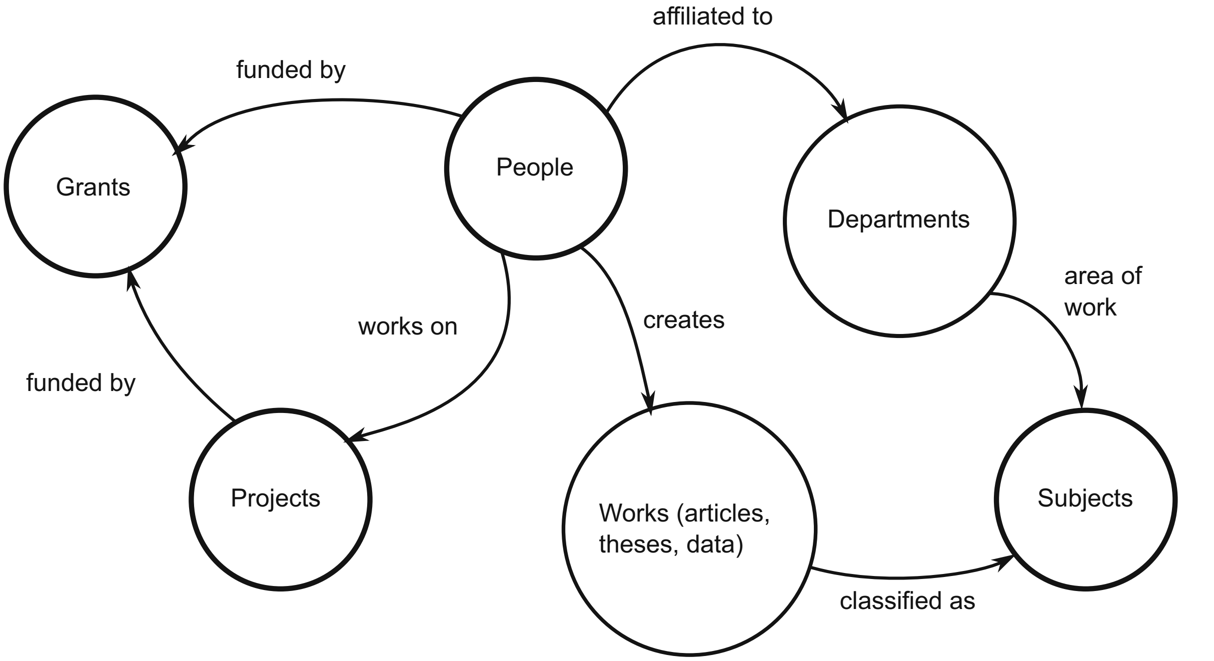
Oxford University's Research Archive - making the archive machine-readable

The archive's architecture has been consciously designed so that the open access research can be re-used and easily found, and a good deal of how this is done is through Semantic Web concepts and technologies.

The developmental approach we have taken is that of a gradual adoption of useful semantic concepts rather than trying to do everything at once. We believe that these ideas will make the content easier to find and curate. It is not necessary to adopt every Semantic Web standard at once - it is possible to cherry-pick the most useful parts and techniques and implement those first.

The most important thing to realise was that the research material and the information about that material can be modelled by linking one ‘thing’ to another. ‘Things’ include authors, editors, departments, grants, the research itself, and so on - the ‘metadata’ of a work, as well as the work itself. The way in which this modelling can be captured is through the Resource Description Framework (RDF) standard.

RDF is machine-readable, and uses “vocabularies” of nouns and verbs that are intended to be globally accessible. (These vocabularies are better known as ontologies.) With it, the archive is able to express ideas such as ‘This person is an author, and they have authored this work and co-edited works x,y and z. Work z is a derivative of work y.”



Basic modelling of research information

By making these simple ideas both human- and machine-readable, we have found that it has some very beneficial knock-on effects:

• Treating people, departments and so on as things in there own right, it becomes much easier to adopt the idea of distributed authorities of information; this puts the control back in the hands of the people, subject specialists and departments and will lead to a more efficient way to maintain these lists.

• Taxonomies described using an RDF ontology called SKOS are very readily reuseable and more precise than simple text phrases - the cataloguing scheme is bound to the subject, rather than having to be qualified elsewhere.

• The internal structure of each digital object in the archive is exposed as an RDF document makes backups and other preservation activities very much easier - this description has been proved sufficient to migrate content from one repository to another, without loss of contextual information and without needing access to the internal systems of either.

• Linking the key players together in this way makes it easy to explore the connections between them - the mapping between subject taxonomies can be exploited to find works which may use different wordings (or languages even) in describing their topics, but are about the same thing.

• Information can be repurposed - portals and web interfaces become just views on the set of data - the emphasis on people means that a ‘social’ networking tool is just as possible as a departmental listing of research using the same data.The use of semantic ideas and technology, while they break away from traditional methods, they enable so much benefit in terms of enhancing discovery, re-usability and also curation, that to not look at what the Semantic Web can do for you may be a very bad move indeed.

Ben O’Steen is developing the digital object architecture behind the Oxford University Research archive and other related projects at Oxford.

2009 Semantic
Technology Conference

June 14-18, 2009 — San Jose, California

SAVE THE DATES

2009

June 14-18, 2009
San Jose Fairmont, San Jose, California

www.semantic-conference.com

Marketing the Semantic Web

Greg Boutin walks through his thoughts on telling a Semantic Web story to the business world.

By Greg Boutin

Time to Deliver

The so-called Semantic Web is a set of technologies intended to facilitate data analysis and transformation by microprocessors, thereby extending the functional reach of machines. As such, it's capable of enhancing a wide range of existing applications, and enabling new ones we have yet to picture. Given the breadth of possibilities and interests the Semantic Web covers, it will come as no surprise that there are many ways it can be marketed.

> *The industry has used this as a stepping stone towards enhanced awareness among technologists and early adopters*

Building general awareness of the possibilities offered by those technologies is critical to getting to the so-called tipping point. In that respect, the Semantic Web has so far done a pretty good job. With the endorsement of the W3C, and Tim Berners-Lee as its top rep, it had the perfect seal of approval to start with. The industry has used this as a stepping stone towards enhanced awareness among technologists and early adopters. Just in the past months, there were articles about the Semantic Web in the tech sections of The Economist, the International Herald Tribune, the New York Times, the Financial Times, and in Scientific American.

Driving user adoption has proved to be a bigger challenge. Invented by the web equivalent of rocket scientists, the initial offering was too complex. For the most part, it remains quite daunting even today: simply including semantic capabilities on one's website remains overkill for most webmasters, for example. Semantic Web technologies have managed to buy time by delivering a few narrow benefits, first in vertical B2B markets (see W3C Semantic Web case studies, or the offering of Zepheira), and more recently, in the consumer market, through services like Powerset, a semantic search engine for Wikipedia; UpTake (previously Kango), a semantic travel search engine; and of course Twine, the most imaginative semantic service so far, but still a semantic mashup between social networking, RSS feeds, and del.icio.us, whose key benefit appears to reside in its bookmark recommendation engine. Among larger corporations, Reuters has been the first one to buy into the concept with its OpenCalais service.

> *Building general awareness of the possibilities offered by those technologies is critical to getting to the so-called tipping point*

In a positive development, Semantic Web technologies are increasingly infiltrating existing value propositions: Yahoo is testing the waters with its Search Monkey platform and, having acquired Powerset, Microsoft is expected to rapidly incorporate semantic capabilities in its search engine. I was told that Google already makes use of some sort of semantic indexing capabilities in its search engine, although I've seen little evidence of it so far. I'd make a wild guess here based on limited evidence I collected, and say that the search gorilla does not have a search monkey in the making, and instead is taking a statistical approach to tackling semantics, which might lead to some interesting development.

All in all, the Semantic Web has had difficulty delivering on all those promises it made. And so everyone is now eagerly awaiting a killer value proposition to catalyse market adoption and bring those technologies into the mainstream. Corporate search, travel search, wikipedia search, recommendation engine: those are all nice-to-have apps, but not killer apps. And so, at this stage, marketing the Semantic Web is not so much about PR and nice brochures as it is about building marketing right into the products, to get that killer app out and silence the cynics.

But how do you go about building a killer app for such complex lab technologies?

My Personal Recipe For Killer Value Propositions

For all its long history and numerous gurus, marketing surprisingly remains much more of an art than a science, and it is especially wasteful and ineffective at producing killer value propositions. In my previous roles, I have been grappling with a number of theories and concepts supposed to achieve just that. Most recently, I used the Pragmatic Marketing framework, which I implemented through a modified Agile Programming process adapted to fit our marketing needs.

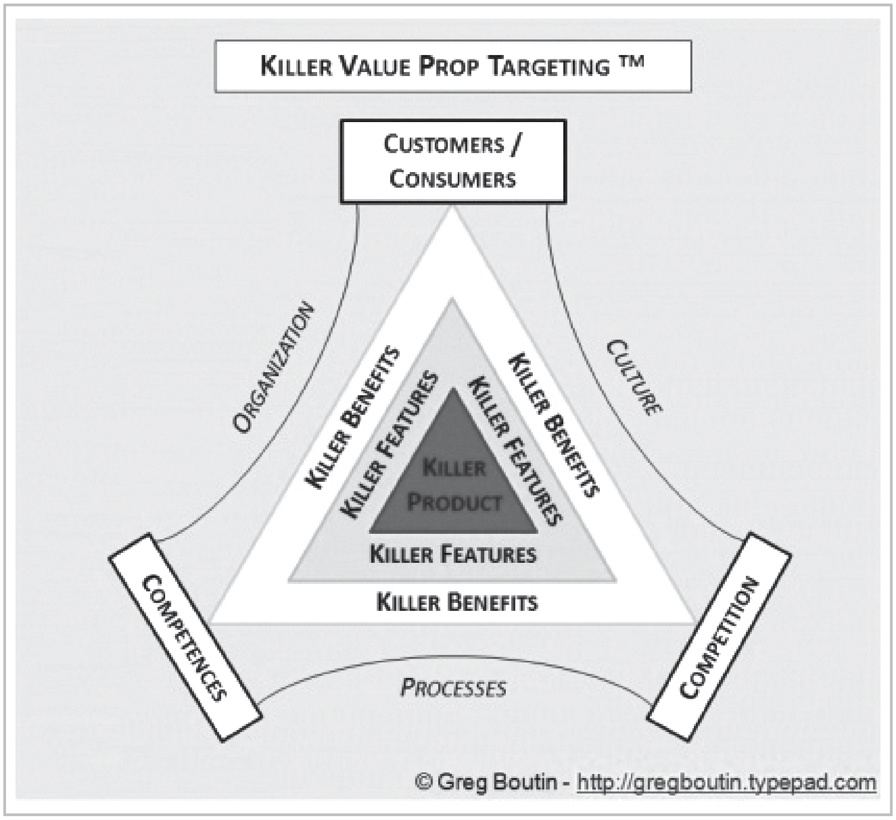
Quickly, the Pragmatic Marketing framework became just a nice-to-have checklist to assess whether we were covering all the basics, and to oversee the interactions between the different marketing pieces I was putting together. I found that Pragmatic Marketing was not designed to come up with killer value propositions. In fact, it itself claims to be dedicated to "managing and marketing technology products", implying the product already exists. Ultimately, what I learnt through all this is that common sense and logic, supported by some analytical firepower, a strong sense of the big picture, and the intellectual readiness to let the market show you what works and what doesn't, beat fancy theories and jargon every time.

With this in mind, the first piece of advice I have for anyone seeking to find the Semantic Web grail is to keep things simple. As I wrote last month on my blog, that generally involves simplifying and cutting things to the chase. Clarifying and focusing on simple benefits. This is notoriously difficult for many technical founders, and yet they must learn to let go, or risk falling into market oblivion. For more on this, I invite you to read the chapter on the Knowledge Curse in the excellent book Made-to-Stick.

A second piece of advice is to not define yourself too narrowly as a Semantic Web company. For one, you want to avoid marketing myopia, as the killer app may be one powered by "semantics + some other technology", and for two, it is generally more sustainable to define yourself by the benefits you provide rather than by the technology that supports you. If superior competing technologies come along, the brand that you have built by serving your customers will generally still grant you the best sit at the table.

To complete that work, finding an objective marketing lead and empowering her or him to deliver is essential. That person should be eager to do four things:

- learning quickly about your potential customer segments and finding out what they do;



© Greg Boutin - http://gregboutin.typepad.com

- learning quickly about your own competencies and resources and finding out what gives you an edge. You need to have a perfect understanding of your competencies, i.e. what will make you better than others at addressing those problems;

- learning quickly about your competitors (and potential partners) and finding out what could give them an edge

And the fourth, having your marketing lead continuously synthesize this into a benefit analysis that details out the main actions you intend to help users with. The key to success is to conduct all four marketing activities in a quick iterative manner so that, when you learn something about your customers for example, it is immediately reflected into your view of the competitive landscape, in the benefits you are working on delivering, and in the competencies you need to that end. Let me emphasize that further: the quest for a killer value proposition is about quickly iterating between knowledge of your capacities, those capacities of your competition, the needs of your consumer, and harvesting new insights from it every time into a coherent benefit analysis feeding your product development process.

Once the core benefits you plan to deliver are defined and prioritized, you need to turn them into features, bundle those features into products, and plan for the product development by having the Product Manager use those inputs to develop a first draft of the product requirement document.

> Sequence your launch(es) as much as possible, so that you test your proposed value

The key idea there is to sequence your launch(es) as much as possible, so that you test your proposed value propositions in the most cost effective way, rapidly seeing what sticks and what doesn't. Sequencing also helps by ensuring that your development always has an immediate deliverable to work towards. Lastly, and I'll keep the detailing of that for another time, build the product so people talk about it to each other (I found the book Word Of Mouth Marketing very useful in that regard). That means having emotional benefits too, something web start-ups often disregard at their own peril.

I have summarized the above in the chart above (pg 17). C'mon, it's not that fancy-pants, is it?

You'll see references to the concepts of Organization, Culture and Processes in the chart. Indeed, building a killer value proposition is not a process in isolation, it needs to be managed from the broader context of motivating people, focusing their efforts on the right task, organizing handovers and interfaces, and learning from it all, so the organization gets better and better at it over time. This is often a difficult one for start-ups with limited resources, and yet a key success driver. Taking an even broader look on the shared drivers of success start-ups, the book Blueprint to a Billion furthers the debate to board organization, alliances, management teams and other things important to crafting and launching killer value propositions.

Digitizing and Monetizing Existing Needs

One of the long-lived myths I'd like to dispel, and which will take us back to the Semantic Web, is one that I often hear from venture capitalists: "don't try to build a new market from scratch, go where the money is". Venture capitalists are not wrong, it's just that the way they express this idea is rife with potential misinterpretations. It leads too many start-ups to try and beat Google at the search game, for instance.

Creating new needs is close to impossible. Going after existing needs addressed by online players, too, is extremely tough. Common wisdom says that your technology needs to be 10 times superior to the current incumbent's technology if it stands a chance to overcome it. Incumbents have had time to look at the problem from all angles, and they have normally prepared themselves to all potential direct attacks. Instead, take existing needs that are neither digitized nor monetized, and monetize them by improving on the ways they are currently met. I didn't invent this; it is a key lesson from experience, also highlighted in a book called the Innovator's Solution (one of a few non-fluffy recent marketing books!). You'd be surprised by the number of tasks today that are still a real pain to accomplish and would greatly benefit from online assistance!

> *To get to a killer value proposition, take a technology, simplify it, focus it (on clear benefits), structure it, iterate it, sequence it, crashtest it, sample it out, evangelize it, and watch it grow!*

One such need I uncovered in my past explorations, to illustrate the idea, is that of synthesizing information. There are a lot of applications out there to search information, much less so to put it together in a coherent manner once I have found the data I was looking for. How do I do that today? I paste things from web pages into MS Word, or save them onto my hard drive. Then I review them, one by one, taking what I need and leaving the rest. I use my brain for the most part, and a few basic tools that are only a marginal improvement over paper. There are no tools yet that compete with my brain for those tasks. And what if I wanted to find out at a conceptual level where we agree and where we disagree on the results of the US policy in Iraq? Again, the optimal way today is to discuss it directly with you, or a long winded conversation online. I'm sure the Semantic Web could enable more productive ways of doing this. What if I wanted to connect on LinkedIn to all the bloggers having blogged somewhere about the semantic technologies conference? Or to get a timeline summary of all the dates and event present in a particular set of websites, together with a number of new visualization possibilities for existing web information. Pretty much impossible today. Wow.

Those are all existing needs that are not addressed optimally, and that semantic technologies can solve. Marketers can review existing processes to see where people spend time, and improve parts of those processes. Much easier than vying to take dominant companies out of their thrones. The bigger and more ubiquitous the need, the more "lethal" the value proposition. Just look beyond search.

In sum, to get to a killer value proposition, take a technology, simplify it, focus it (on clear benefits), structure it, iterate it, sequence it, crashtest it, sample it out, evangelize it, and watch it grow! Hopefully those thoughts will help a little to build the Semantic Web champion which I strongly believe will be instrumental in taking our technologies to the masses. The "Semantic Web" concept is complex and reflects a wide brochette of technologies and solutions: what we need now is a leader who will dispel that fog and finally leave the industry with a clear web forward!

Greg Boutin is a member of the Semantic Web Gang and an independent go-to-market consultant helping innovative ventures to turn their technologies into products that people love and pay for. He specializes in the Semantic Web, knowledge management and tagging spaces. You can find more about him and his services on the Ideas to Markets blog, http://gregboutin.typepad.com/ and connect to him on LinkedIn, http://www.linkedin.com/in/gregboutin.

Databases and the Semantic Web

Juan Sequeda discusses the data integration problem, and explains how RDF could help.

By Juan F. Sequeda

The Semantic Web as a database

The Semantic Web is about creating a web of data by integrating data that comes from diverse sources: HTML pages, XML documents, spreadsheets, relational databases, etc. In order for software applications to be able to make use of these diverse data sources, a primary objective is to make the Internet appear as one unified, virtual database. This is possible through RDF, which is a standardized format for representing data in a subject-predicate-object format. This facilitates the interchange and combination of data that is served all over the web, which we now consider Linked Data.

The term Linked Data has been coined to describe the method of sharing data on the Semantic Web: data that is only expressed in RDF. In April's edition of Nodalities, Tom Heath stated that "Linked Data is a style of publishing data on the Web that emphasizes data reuse and connections between related data sources". This same model is one that many businesses and researchers have been trying to accomplish over several years, with the goal of successful data integration. Has this problem been solved?

Solutions exist, but *the* solution is yet to take shape.

Solving the Data Integration Problem

One of the most difficult areas businesses face is in data integration. One of the best examples I've come across which illustrates the Sematic Web's solution to the data integration problem can be found in Ivan Herman's presentation on "Introduction to the Semantic Web (through an example…)". (If you have not yet seen this presentation, it is a must!)

Ivan's presentation inspired me to adopt it for the Cyberinfrastructure for International Collaborative Biodiversity and Ecological Informatics Workshop where I went as a Semantic Web evangelist. It was no surprise to me that one of the main concerns was data integration. Consider, for example, businesses that need to integrate data after company mergers or biologist who need to share their data. For biologists, some standards created by the bioinformatics community to facilitate data sharing (TDWG, GBIF) exist, but it is not enough. To my surprise, these standards are slowly shifting towards Semantic Web technologies and RDF. At the workshop, I was able to "preach" the need of focusing to move towards RDF, instead of creating new biological data standards and protocols. This way, one can now imagine applying the Linked Data model to the bioinformatics field.

Data has to be stored in a relational database that is made with a schema

But the question remains: how can that Linked Data model be formed when all the existing data are already stored in relational databases? Furthermore, how can RDF data be created from relational database content?

Here is where Relational Data and RDF comes together.

Global Overview of Relational Data and RDF

The data that need/want to be shared are located in relational databases. Furthermore, internet-accessible databases contain up to 500 times more data compared to the static web; and three-quarters of these are managed by relational databases. Therefore if the Linked Data Cloud set wants to have more clouds, it is imperative to have a way to convert all the relational data into RDF.

To understand the relationship between Relational Databases and the Semantic Web, consider the following analogy. Relational data is to RDF as a relational schema is to an ontology (RDFS or OWL). Data has to be stored in a relational database that is made with a schema. (Wikipedia explains it: The structure of a database system, described in a formal language supported by the database management system (DBMS). In a relational database, the schema defines the tables, the fields in each table, and the relationships between fields and tables.) Likewise, RDF data is an instance of specific triple schema that is part of an ontology. Therefore to create RDF content from a relational database, it is necessary to use an ontology.

Now the problem is: how do we use these ontologies?

Ontology and Database mapping

One way is to reuse existing domain ontologies—such as FOAF or SIOC, since Ontologies are designed for reuse. If an ontology already exists, then a mapping between the old relational schema and the ontology has to be established. Imagine, for example, a database of contacts. The FOAF (Friend of a Friend) ontology represents the relationships between people. Because of this, the relation schema and relational data can be easily mapped to the FOAF ontology and RDF. This allows for the creation of a contact database bringing all the benefits of RDF relationships and the added bonus of always having up-to-date contact details.

Several tools have been created that have been able to successfully tackle this problem including D2RQ, R2O and others. Unfortunately, these current solutions are a bit complex. One has to learn a mapping language and an existing domain ontology must be used to map the database. Hence, it makes it harder for a database administrator to convert the relational database content into RDF and this hinders it from becoming a priority.

Direct Mapping

Another way to approach this problem is by direct mapping methods, which do not consider existing domain ontologies. A system like this could also facilitate translating SPARQL to SQL queries, so up-to-date relational data can be retrieved in RDF, instead of having an RDF dump of the relational data, which is the output of Ontology and Database mapping systems. Instead, these methods (semi)-automatically generate the ontology based on the domain semantics encoded in the relational schema. Once the ontology can be generated, the relational content can be mapped to RDF. Much work is still required in this area, however, to create more expressive automatic mappings.

Looking Ahead

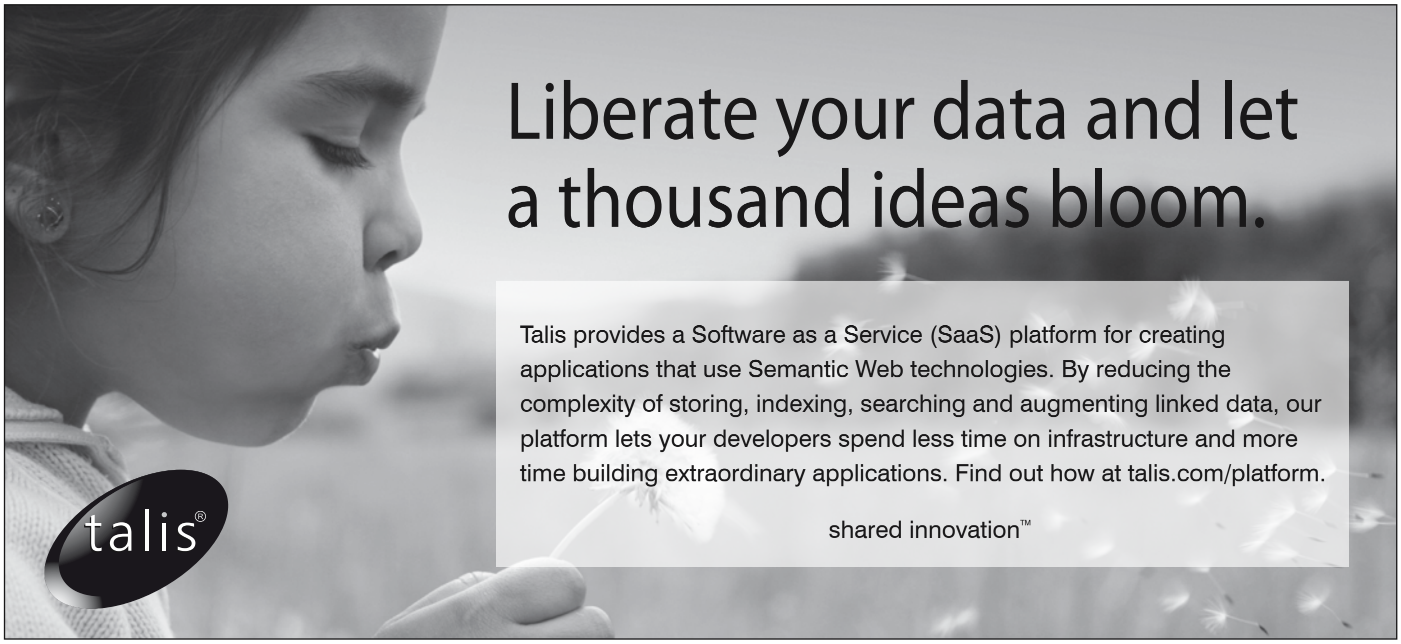
Due to the fact that the relational database is currently the main way to store data, the success of the Semantic Web hinges on creating effective mappings between relational databases and Linked Data content. This will enlarge the Semantic Web and allow for the creation of a Linked Data Cloud. It is rare that integrating a database with the Semantic Web accrues obvious or immediate benefit to the owners of a database. Thus, organizationally, integrating a database with the Semantic Web is not seen as a widespread, current priority. However, as discussed by Tom Heath in April's Nodalities, publishing Linked Data is among the largest of the themes established in the Semantic Web community. Hence, it is imperative for the community to make it as easy as possible to bridge relational database content and the Semantic Web.

Therefore, easy, automatic and effective mappings of Relational Data to RDF, such as Direct Mapping methods could facilitate the integration and publication of data. This could bring a much-needed change of priorities with immediate benefit for certain sectors such as business or biology. Although there are criticisms of Direct Mapping methods (such as a synthesized ontology would still require integration with some global domain ontology), in the larger context, the successful development of the Semantic Web will see the development of successful systems that implement ontology-to-ontology mappings. In this context, direct-mapping is the first step in a two step process, where both steps are the subject of intense research in automated methods.

Currently, a RDB2RDF W3C Incubator Group has been formed to tackle all these problems and offer initial recommendations of standardizations. The question remains whether if we will establish an easy and effective way of integrating relational database content to the Semantic Web.

Juan F. Sequeda is a Ph.D student at the University of Texas at Austin. He is an invited expert on the RDB2RDF W3C Incubator Group. His research interest involves integrating existing data in Relational Databases with the Semantic Web through Direct Mapping methods.

jsequeda@cs.utexas.edu and www.cs.utexas.edu/~jsequeda



Semantic Search: The Myth and Reality

Alex Iskold reflects on the state of Search and looks into the myths and realities of Semantic Search.

By Alex Iskold

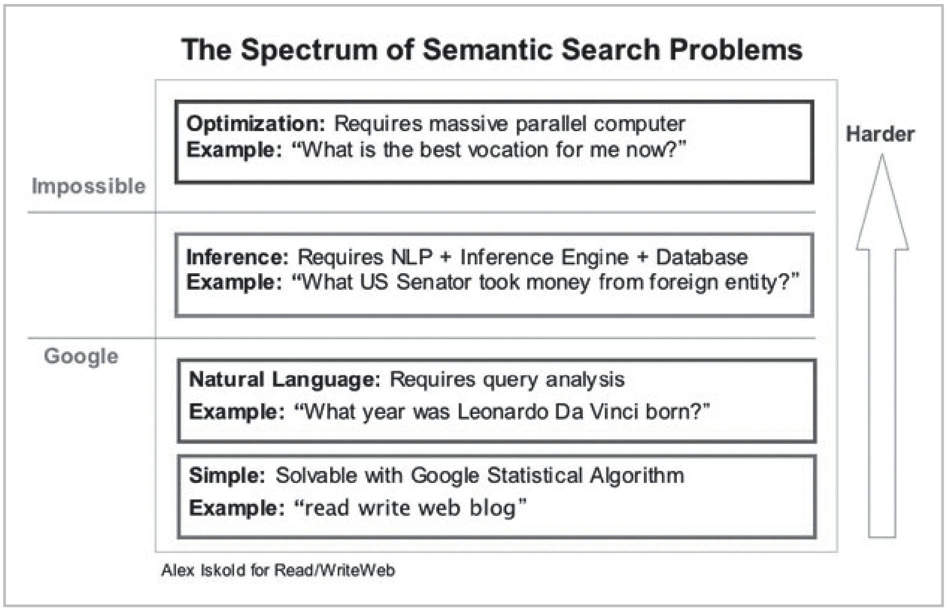
For a few years now people have been talking about semantic search. Any technology that stands a chance to dethrone Google is of great interest to all of us, particularly one that takes advantage of long-awaited and much-hyped semantic technologies. But no matter how much progress has been made, most of us are still underwhelmed by the results. In head-to-head comparisons with Google, the results have not come out much different. What are we doing wrong?

For example, when asked, “What is the capital of France?” both approaches come back with the correct answer - Paris. Also, a lot of queries that we are used to typing into Google in abbreviated form, come back with similar results if we type them using natural language. Clearly something is off. We all know that semantic technologies are powerful, but how and why? In this post we will show that the problem is that we are asking wrong questions.

> *The mistake is that semantic search engines present us with Google-like search box and allow us to enter free form queries*

The mistake is that semantic search engines present us with Google-like search box and allow us to enter free form queries. So we type the things that we are used to asking - primitive queries. It never occurs to us to type in What actor starred in both Pulp Fiction and Saturday Night Fever? or What two US Senators received donations from a foreign entity? We type simple questions, but this is not where the power of semantic search lies. Lets look at the spectrum of semantic technologies from Google, to SearchMonkey, to Powerset, and Freebase to understand what is going on.

As shown in the diagram above basic queries are easily handled by Google. Sadly, natural language processing gives little advantage when it comes to this category of problems. Google correctly answers the question about Leonardo Da Vinci’s birthday leaving no opportunities to improve the search by understanding the nouns and the verbs that user typed in.



The Spectrum of Semantic Search Problems

Impossible

Optimization: Requires massive parallel computer
Example: “What is the best vocation for me now?”

Inference: Requires NLP + Inference Engine + Database
Example: “What US Senator took money from foreign entity?”

Google

Natural Language: Requires query analysis
Example: “What year was Leonardo Da Vinci born?”

Simple: Solvable with Google Statistical Algorithm
Example: “read write web blog”

Harder

Alex Iskold for Read/WriteWeb

What Problem Are We Trying to Solve?

The first confusion in the space comes from the fact that semantic search is being positioned as the answer to all possible problems - from modern search, currently dominated by Google, to problems that are computationally impossible. The situation is made more difficult by the fact that right now there is only a thin range of problems where semantic search can clearly do better. This range is complex queries involving inferencing and reasoning over a complex data set.

Before looking at the problems that are perfect for semantic search, lets look at the hardest problems. These are computationally challenging problems that really have nothing to do with understanding semantics. The misconception has been perpetuated since early days of the Semantic Web that somehow, because we will annotate the web, we will be able to solve these super complex problems. This is simply not true. There are fundamental limits to what we can compute, and a class of problems that have an exponential number of possible solutions is not going to be magically solved because we represent data as RDF.

The good news is that there is a set of problems that are great for semantic search. These are the problems we have been solving so wonderfully with relational database. Way too often we forget that semantic technologies are here to help us represent relational data spread over the entire web - so it should be no surprise to us that it is relational queries that semantic search engines would excel at.

The Spectrum of Semantic Search Players

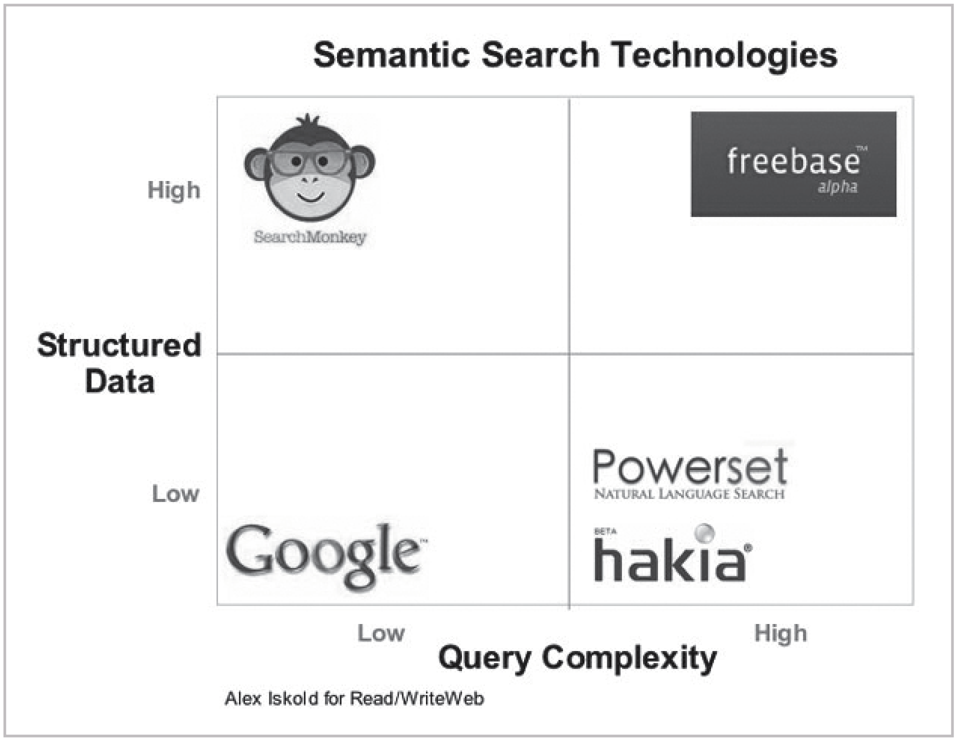
But semantic search is not just about the questions that we are asking. Because the web is just a bunch of unstructured HTML pages, semantic search is also about the underlying data. At its most structured extreme we find [| Freebase] - the semantic database of everything. Freebase is accessible via free text search, but more importantly via MQL (Metaweb Query Language). MQL is essentially JSON with wildcards. Using it you can construct any query against Freebase and the result will be the same query with answers filled in.

[| Powerset], in a way, is just a relational database. It operates against certain, structured information. On the other end of the spectrum is Google, which is all about statistical frequencies and very little semantics. The recently launched [| SearchMonkey] from Yahoo! is an interesting twist. It does not add anything to the result set, but instead uses semantic annotations to present a richer, more interactive and useful user interface.

Companies like [| Hakia] and Powerset are probably working the hardest. These companies are trying to simultaneously build Freebase-like structures on the fly and then do natural language queries on top of them. The difference is that Hakia is using (likely similar) technology to query over the entire web, while Powerset has (probably shrewdly) chosen to restrict the search to Wikipedia

Are Hakia, Powerset and Freebase All That Different?

This analysis brings up a question - which of these technologies are different and which are essentially the same? Lets get the easy one down first. Yahoo!'s SearchMonkey is no different from Google or any other search, as far as the core search technology is concerned. The difference is simply in the presentation layer. SearchMonkey is smart about creating a better user experience by letting publishers present the search results to the users in the best possible way.



Alex Iskold for Read/WriteWeb

But when it comes to Hakia, Powerset and Freebase the situation is much more complicated. On the surface all these products are different - Hakia lets you search the whole web, Powerset is restricted to Wikipedia (and Freebase!) and Freebase itself has two search interfaces - the search box and query language. Here is the problem - the natural language interface has nothing to do with the underlying data representation.

The fact is that all of these semantic search technologies allow people to type in arbitrarily complex questions and then interpret these queries and execute them against their databases. Fundamentally, Hakia, Powerset, and Freebase are databases. Fundamentally, all of them have some kind of Natural Language Processing that translates the question into a canonical query over the database.

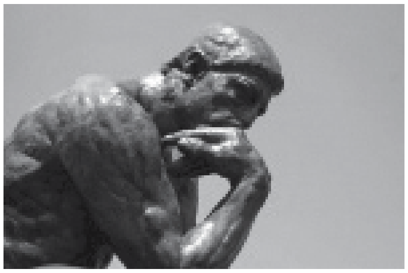
To gain insight into all of this, think about Freebase and its query language MQL. Unlike natural language, which allows all sorts of constructs, MQL is non-ambiguous. This JSON-like language allows users to construct precise statements against Freebase. The fact that Powerset allows natural language queries does not mean that inside Powerset there is no database. For sure, though, there is a similar kind of database as there is beneath the Freebase search box. What is really different about Freebase and Powerset is the data gathering approach and user experience.

Back to the Future: It's All About UI

Probably the most striking revelation about the semantic search space is User Interface. First, to go on the tangent, Powerset got it right by realizing that semantics needs to be surfaced in the UI. After a user searches Powerset, a contextual gadget, aware of the semantics of the results, helps the user complete the search experience.

> The search box that everyone is familiar with via traditional web search engines needs to go

Yet the biggest mistake that I think Powerset is making is also in the UI. The search box that everyone is familiar with via traditional web search engines needs to go. Having a simplistic search interface hurts Powerset and Hakia, and to a lesser extent Freebase, which is not positioning itself as generic search.



Think about the recent launch of Powerset. The company released a vastly better way to interact with one of the most important sources of information on the web - Wikipedia. But what did the critics say? Lets see if this is a Google killer. And the answer to that is "no."

But what if Powerset restricted what can be searched? What if instead of a search box there was another interface or what if they told users not to look up things that they can find easily on Google? Why is it that new companies are expected to improve on the algorithm that has ruled the web for over a decade? Instead, the expectation should really be to solve the problems that can not be solved by Google today.

Conclusion

Semantic search is an upcoming technology that has set the expectations way too high. We have all been misled into thinking that these technologies are here to dethrone Google by delivering better search results. Neither of those things are true. What is true, however is that semantic search is going to be big and it is going to help us answer questions that we simply cannot answer today - complex, inferencing queries asked over the entire web as if it was a database.

In order for these semantic search technologies to make a dent in the market, they need to clean up their messaging and most importantly, their user interface. Presenting a search box is both misleading and detrimental, as people associate it with the simplistic questions that Google solves without any problems. To really showcase semantic search, these companies need to come up with innovative UIs that will help users to understand the power that is being put at their fingers

Alex Iskold is founder and CEO of AdaptiveBlue and a member of the Semantic Web Gang. He is also a regular contributor to Read/Write Web.

"Closeup of The Thinker" by Brian - Progressive Spin: flickr www.flickr.com/photos/seatbelt67/502255276/ Diagrams by Alex Iskold for Read/Write Web.

Originally Published at Read/Write Web (http://www.readwriteweb.com/archives/semantic_search_the_myth_and_reality.php) reprinted by kind permission from Read/Write Web.

The Semantic Web Gang

The Semantic Web Gang is a monthly round-table podcast hosted by Paul Miller and featuring a regular panel of commentators on the Semantic Web. semanticgang.talis.com

ESTC2008

2ND ANNUAL EUROPEAN SEMANTIC TECHNOLOGY CONFERENCE
24-26 September 2008 Palais Niederösterreich Vienna, Austria

SAVE THE DATES

REGISTER NOW

www.estc2008.com/register

Europe's Business Platform of Semantic Technologies
Actively participate and maximize your benefit from ESTC2008

24-26 September 2008
Palais Niederösterreich, Vienna, Austria

www.estc2008.com