Nodalities

THE MAGAZINE OF THE SEMANTIC WEB

talis®

www.talis.com/nodalities

May/ June 2008

INSIDE

Semantic Web as Blue Ocean opportunity

Ian Davis and Zach Beauvais lift the lid on Talis' conviction that the Semantic Web offers striking opportunities to create new markets.

Whole industries arise out of the 'ocean' of the global marketplace. The oceans, according to the article by Kim and Mauborgne, turn red as they are bloodied by the competition necessary for companies to make a profit. The competition is necessary, in the bounded ocean, because there is only so much market to go around, and everyone wants a piece. But, looking back over the past 100 years or so, we can see that a plethora of industries we now think of as commonplace simply didn't exist. Think of automotive industries, biotech, petrochemicals or any Dot Com. When a disruptive idea hits the ocean, within the Blue Oceans metaphor, the traditional competition becomes irrelevant, because no one is offering a similar service or product.

Many of the fundamental principles behind the Semantic Web make it a tantalising Blue Ocean opportunity. From its ability to be a global platform to the fact that it can appear behind the software of businesses of any size, coupled with its ability to harness a world-wide network effect; nothing in the current market seems able to offer relevant competition, and we could be well on our way to swimming for a time in a blue sea clear of the bloodied murk of one-upmanship and limited competition.

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;

Semantic Technology Conference
San Jose, California 18-22 May 2008

European Semantic Web Conference
Tenerife, Spain 1-5 June 2008

LinkedData Planet Conference
New York City, New York 17-18 June 2008

Dublin Core Conference
Berlin, Germany 22-26 September 2008

International Semantic Web Conference
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

EDITOR'S NOTES

Welcome to this second issue of Nodalities Magazine. Since we released our first issue a few days before last month's World Wide Web Conference in Beijing, it's been truly gratifying to see the response in person, on various blogs, and in terms of a growing subscriber list for the free print edition.

Speaking of Beijing, the Linked Data meme that was our cover story last month certainly made its presence felt there. Sir Tim Berners-Lee pointed to its importance in his conference keynote and elsewhere, remarking that 'Linked Open Data is the Web done as it should be.' The Semantic Web itself featured in many of the conference sessions, and a discussion panel on the last afternoon tackled 'Commercialising the Semantic Web' to show that a wide range of organisations are making a business built upon semantic technologies work. One of those businesses is UK-based Garlik. Garlik CTO Professor Nigel Shadbolt joined my panel in Beijing, and Garlik CEO Tom Ilube offers further insight on p.5.

Elsewhere in this issue, Talis CTO Ian Davis and Zach Beauvais share our belief that the Semantic Web offers what W. Chan Kim and Renée Mauborgne would doubtless recognise as a perfect Blue Ocean opportunity, and on p.9 Talis CEO Dave Errington paints a picture of our corporate journey over the past four decades.

Nadeem Shabir introduces the notion of open and closed worlds on p.10, illustrating ways in which building applications on the Semantic Web calls for a change of mindset amongst your development staff, and on p.7 David Peterson talks about the ways in which a small group in Australia is putting the Semantic Web to work in tackling real-world problems that affect us all.

This month's podcast transcripts sit well together, casting light on the uses to which two of the world's biggest news organisations put the Semantic Web.

Thank you for reading Nodalities Magazine, and if you haven't done so already please do visit our website and subscribe to receive future print issues by post. After many years of promise, the Semantic Web is finally delivering substantive benefits across a wide range of business areas. We intend to track and celebrate this growth in Nodalities Magazine, and I welcome your suggestions as to where we should be watching.

paul.miller@talis.com

Continued from front page.

Scale and Network Effect

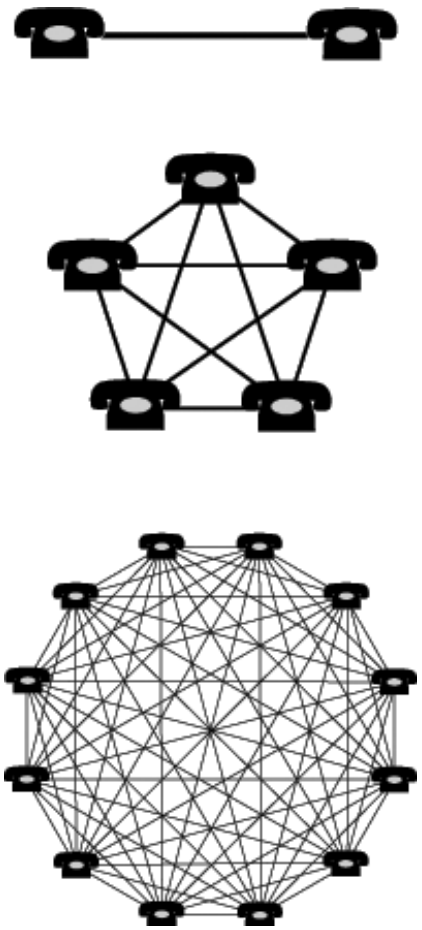
Publishing a document on the Web and assigning it an address has a specific value to the publisher and to their intended audience. While this value might appear obvious, it has a deceptively hidden value which is much higher than the simple addition of an extra resource to the network. The side effect of publishing in a linked network is that the addition actually increases the value of the entire network exponentially, because everything can link to it, but it can also potentially link to every other piece of the network. The more connections made, the more value to the publishers and the other members of the network.

It's this quirk that causes the fabled Network Effect. The act of participation makes the entire network more useful for everyone. The classic example is the telephone network. At some point the network reaches a critical mass and the benefits that accrue from being part of the network outweigh the costs of participation. The system starts to snowball as more people get involved and the value accruing to the existing participants grows faster and faster.

Tying this exponential growth into another set of laws (c.f. Moore's Law) governing increasing productivity of processing per dollar, means that the thresholds of data processing necessary for entry to the scale advantage is constantly lowering.

Blue Oceans

So we have a situation where costs of providing, consuming and processing data are plummeting and the value derived from networking that data is increasing. This kind of value innovation is a strong indicator of the emergence of a blue ocean - one where competition is irrelevant because no one knows the rules, and no two companies are offering the same product or service.



Network Effect (Derrick Coetzee)
en.wikipedia.org/wiki/Image:Network_effect.png

Blue oceans empower companies to achieve breakout growth by capturing new demand rather than competing for existing market-share. Companies can create blue oceans by innovating the value they provide to their customers. By making a leap in value, companies can make the competition irrelevant as large markets are unlocked. The key to doing that is to find a market with new value and devise ways to reduce the costs of providing that value. Opportunities become viable that would have been uneconomic under previous conditions. These are exactly the conditions being set by the Semantic Web.

By making a leap in value, companies can make the competition irrelevant as large markets are unlocked

When companies seize a blue ocean opportunity the result is the creation of a new industry, an entirely new and uncontested market with huge potential; and the Semantic Web seems to fit that bill.

Limitless Potential

Every aspect of society and economy that has been touched by the Web has been changed radically. Even the way we interact socially is very different to 15 years ago. We have instantaneous communication with global reach that means we can stay in touch like never before. The Web has affected commerce, driving down costs again and opening new markets. It's hard to think of an area of life that hasn't been affected: learning and teaching; how we discover news or plan our travel - even our health systems are changing.

These changes are being brought about by the Web's ability to link documents together into a single information space. The Semantic Web extends this to link anything and as such the potential for radical change in our society is vast. Because the Semantic Web is an extension of the World Wide Web, it harnesses the existing infrastructure and the networked effects of the existing online sphere. Its potential is as uncapped and global as the Web's was 20 years ago.

Walled Gardens and Isolated Content

Walled Gardens exist wherever a set of information needs to exist within a system to be used. In essence the data inside the system can be used, linked, and stored but it cannot be accessed from outside the system. Early examples include AOL and Compuserve, which controlled all the information customers could access in order to charge for

premium content and services. Following an initial success, these systems gave way quickly and completely to an open system; the Web.

The primary disadvantage of a 'walled garden' or closed approach is that it limits or even precludes the possibility of a network effect. Because nothing outside can link to the walled content, the two-way benefit cycle is broken and the actual value of adding information can only grow linearly as a few more bits and pieces are added to the system. In contrast, anything published to the Web both benefits the publisher and the rest of the network as more connections are made possible. It is also complicated and costly for data isolated in a walled garden to be used by software from outside. Data written for one system (garden) is difficult to use and effectively needs translating before external software can make use of it.

The Semantic Web approaches data management from an open perspective, in such a way that data published anywhere, provided software has the right to access it, can be used. This effectively positions all linkable data so that it can be scaled by the Network Effect. By eliminating walled gardens and by linking islands of data together, the Semantic Web also makes it possible to drastically reduce the cost otherwise accrued whenever disparate data is introduced. Software written for the Semantic Web can 'read' any data on the network, so translating and importing is no longer necessary.

The open network outpaces the closed.

A Long Way to Go?

Although we have seen that technology can grow exponentially, society adapts to it linearly. Often the technology itself suffers a backlash. This can happen when technology races ahead of social norms. For example, here's a recent posting by a concerned user on tribe.net:

Can someone please tell me why my bio and all of my tribe friends are listed on a site I have never been to or heard of? I didn't think this was Tribes style. I feel cheated and betrayed. If I wanted my profile to be farmed out, I would join Facebook.

Blue Oceans empower companies to achieve breakout growth by capturing new demand rather than competing for existing market-share

In this case, tribe are publishing profile data which was freely available via tribe's profile pages. However, most people assumed it was hard to access. So, the expectation of the user was compromised by the service provided by the technology vender.

Entering your postcode into a site is one thing, having it plotted on a Google map of your neighbourhood together with your photo, date of birth and marital status is quite another.

So, while we want to encourage the publication of machine-readable information and the lowering of barriers, we need to be mindful of the social norms around the use of the data. It's this mismatch between the pace of technological change and society's capacity to adapt that gives rise to the adoption curve.

It's also why we believe that there's still a long way to go. However, with its inherent ability to enable global-level software, and the potential to benefit any kind of industry; we feel it is an opportunity not to be missed.

Talis

In the coming years we expect that large numbers of companies will seek new ways to conduct business as the Semantic Web changes the nature of information exchange across the world. Talis aims to support these companies by providing technology and infrastructure that can dramatically lower the complexity and costs of storing, indexing, searching and augmenting large quantities of data. Our Platform is designed to unlock the potential of the Semantic Web and make it possible for any business to participate and succeed in the coming Blue Ocean.

Find out more about the Talis Platform:

Businesses:
www.talis.com/platform
blogs.talis.com/nodalities

Developers:
n2.talis.com/
blogs.talis.com/n2

W.Chan Kim and Renée Mouborgne, 2005. Blue Ocean Strategy: how to create uncontested market space and make the competition irrelevant.

Ian Davis is Chief Technology Officer at Talis, responsible for the ongoing development of the Talis Platform. Zach Beauvais is a researcher in the Platform Division at Talis.

ian.davis@talis.com
zach.beauvais@talis.com

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/

Social Networking Demands Social Verification

Tom Ilube, CEO of UK-based Semantic Web startup Garlik, discusses the role that Semantic Technologies play in combating the increasingly serious threat of identity abuse.

By Tom Ilube

The Dangers of Digital Impersonation

The other day I received one of those spam emails inviting me to assist the executor of a recently deceased gentleman's will, to transfer $24m abroad in return for a generous cut. We are so used to these emails nowadays that I was about to delete it without thinking when something prompted me to take a closer look at the sender's name.

It was an odd name, not the usual 'made up' looking name so I searched for it online to see what I could find. To my surprise it turned out to be the name of an Arizona-based lawyer with an impressive track record and a clear and credible profile on the Web.

This struck me as an interesting new development. The email fraudsters in this case had deliberately selected the identity of a perfectly respectable lawyer to use for the scam. He has no idea (yet!) that his name is appended to possibly millions of scam emails hitting in-boxes around the world as we speak.

Recently I was involved in a UK TV program on online dating. It appears that fraudsters are creating fake identities on online dating sites in order to build relationships, over a period of weeks or even months, with 'lonely hearts' and then having won their trust they hit them through one scam or another. One poor gentleman actually went to Heathrow Airport with chocolates and flowers to await the arrival of his 'Beyonce-lookalike' beauty, having sent over several thousand pounds for her travel expenses. Needless to say she didn't arrive.

Again the interesting development here is that many of the fraudsters are smart enough to use, either in whole or in part, the identity of a real person with a real digital presence in order to instill confidence and build trust.

This sort of digital impersonation is growing rapidly and as fraudsters realize how much more effective it is to establish trust relationships by exploiting real identities rather than inventing fake ones, Garlik expects this to become a serious challenge in the age of social networking.

Web Dynamics Predict the Problem

This development should not be a surprise to us. In expressing the foundations of the emerging new discipline of Web Science, Sir Tim Berners-Lee, the inventor of the Web, has articulated a compelling view of the dynamics of Web-based phenomena.

It appears that as a Web-based phenomenon moves from its relatively closed 'micro' stage of use (perhaps between a well defined, trusted user community) to an open, macro stage (with tens or hundreds of millions of Web users) one almost always sees an explosion of unwanted and potentially damaging behaviour.

For example, as email moved from the exchange of messages between small groups of trusted academic, government and even technical communities to today's open Web-scale email environment, so a wave of spam has hit us to the point that perhaps 80% of all email flying around the internet is unwanted spam. Similarly, as Google emerged out of the tech communities where it was initially adopted into the mainstream with billions of searches per day, so an entire industry of spoofers, spammers and link-farms has emerged.

The fastest growing area of Web activity today is the rise of the social networks and a range of social activity – from Facebook, mySpace and LinkedIn to online dating and picture sharing. Social networking sites are the latest phenomena and over the past couple of years they have moved from their closed micro stage to their Web scale collaborative macro stage.

This is a real challenge because the pain threshold in social networking when their equivalent of spam hits is likely to be much lower than other areas. If I use an online dating site, only to find that one in twenty, perhaps even one in ten of the people I try to interact with are not who they claim to be I will walk away from that site. My guess is that a 5-10% level of impersonation fraud will send even the biggest such site into a rapid downward spiral as real users vote with their feet. The penetration levels may vary by type of site (Garlik is currently conducting research to gauge consumer sensitivity to this) but in any case I believe it will be far below the tolerance levels that we have to other types of spam. Once our spam filters are in place we happily use email despite the 80% spam levels. Our confidence in email itself is not significantly undermined. But would you really put your personal details in to a social networking site if you knew in advance that 80% of the other 'people' in it were fakesters and fraudsters – even if there was some spam filtering equivalent? I doubt it.

As a result, social networks are at the early stages of seeing a wave of unwanted and damaging behaviour. Online fraudsters are beginning to infiltrate these sites en masse. This gives rise to the real threat challenging the long term viability of these sites. The challenge of light weight, large scale verification.

Introducing Social Verification

There are many approaches to verification, both on and off line. From identity cards with biometric, to credit checks, digital certificates, virtual cards, open identities and so on the world is not short of approaches to verification. So why explore a new one?

I think it is worth looking at new verification approaches for two reasons.

Firstly, the social networking world in its current form of a limited number of huge walled gardens and its future form of a sea of ubiquitous and portable 'socially networked' identities is new. The whole area of identity is playing catch-up with the emergence of the web at all, let alone this latest manifestation of it.

Secondly, as an unashamed advocate of the Semantic Web, I think Semantic Web approaches bring something interesting to the party here that could enable us to tackle this emerging challenge before it really gains momentum.

It is instructive to stand back and look at how we verify identities in the off-line world. Obviously it depends on the context. If I am going to a secure area I need to produce some 'hard' identity credentials to verify my identity. However if I am introduced to you socially a 'softer' approach often suffices.

What we actually do in these social situations is verify each other socially. You and I discover whilst chatting that we know someone in common. A link is formed. I 'know' you a bit better, I'm a bit more comfortable with you. We discover a couple of other mutual acquaintances. Great. Your brother works somewhere that I used to work. I've never met him but it's another link. This is social verification in action. We establish our identities in a soft, loose way gradually over time, by identifying links and that emerging network leads to growing confidence and trust.

Coincidentally social networks, of whatever form, hold within them fragments of social graphs that enable this type of link discovery to happen more or less automatically. At least those based on semantic technologies are able to see and analyse the social graph that they have gathered through user interaction. Thus, the social networking community has in its hands the ability to offer a new approach to verification that is much more analogous to what people do in the real world. Garlik calls this 'social verification' and we believe that the key to realizing the benefits of social verification lies in adopting semantic technologies so that the social graph can be made visible and republished to aid verification.

The First Steps to Social Verification

All too often people start at the end of a huge vision and get stuck on how to get started. But the beauty of using a Semantic Web approach to tackling this problem and providing a solution is that it can start small and grow organically, from the bottom up, just like the Web itself.

So a relatively trivial example that hints at the direction this journey could take us as a Semantic Web community can be seen in an automated, semantic approach to handling the verification of comments on a busy blog. The blog owner has a simple and elegant Semantic Web based solution to staying as open as possible to comments from the ever changing community around him, whilst staying away from manual moderation and avoiding an unmanageable free-for-all. Only friends of colleagues and "friends of friends" of colleagues (i.e. up to 2-degrees of separation) can post comments as defined in the FOAF (friend of a friend) files of his colleagues. It is a kind of social verification that asserts that someone who is a verified friend of a friend, or colleague of a colleague (to 2-degrees), is likely to have something interesting to say in that particular context. The parameters are easy to flex to widen the verified community if he so wishes. Essentially he is using social verification to maintain adaptable choice rather than the usual 'all or nothing' solution.

FOAF files are an interesting starting point in this social verification context as they are expressions of fragments of the social graph that the individual has published. There also appears to be a manageable number of them, perhaps in the tens of millions at this stage. At Garlik we are scratching our heads about how and whether the implicit FOAF social graph could be utilized for a (limited) form of social verification.

For example, on receiving an email could I 'ask' the FOAF network "how confident should I be that this email is from a real person who I have a (strong/weak) connection with?" This approach to verification is about establishing semantically relevant confidence levels appropriate to the context of the interaction rather than black and white absolutes. Garlik aims to release a service that brings this type of social verification to life and to invite the Semantic Web community to explore and extend it.

In conclusion, we believe that one of the big challenges for the next stage of the Web will be fighting an emerging wave of 'identity spam' which, if left unchecked, could grow as quickly and aggressively as past phenomenon such as email spam, but with more devastating consequences for individual social networking sites that get caught up in its grip. However, we believe that an approach to context-sensitive verification that leverages the social graph, what we call 'social verification' offers one way of tackling this challenge. The Semantic Web and associated technologies provide a necessary and compelling platform to bring this to life and at Garlik we are doing our bit to make it real by taking the first few steps on what promises to be an interesting journey.

Tom Ilube is CEO of Garlik

www.garlik.com

Semantic Web and the Environment

David Peterson reports on work in Australia, leveraging semantic technologies in the fight against climate change

By David Peterson

BoaB is a small group with big ideas. For the past couple of years we have been playing "spin the facet" with semantic technologies; looking for the angles and surfaces that best intersect with our own talents and interests. There's so much hype and buzzword overflow that it's a full time job just sorting out interesting bits from re-packaged fluff. And when you're trying to find a commercial angle on the bleeding edge, the challenges just get that much more interesting.

To tackle those challenges, we're cultivating connections across sustainability sectors and the green economy. To that extent, we work almost exclusively with scientific research and natural resource management organisations. These are going to be big growth areas in the next few years.

What we have been finding is that semantics really start to make sense once you've got a reasonable knowledge of the domain and the sense of scale you're working in. For example, our work with marine and tropical research centres is based on over a decade of background work in the space. We know the language and we know the problems. So our solutions can find a natural sweet spot. That's a problem for many start-ups; knowing when to stop coding and when to start deploying.

I think we are settling down to three broad approaches in the projects we take on.

First, **we put the user in the centre of the information space**. Establishing the personal layer first is critical if users are going to be able to navigate the myriad of personal preferences, role-based contexts and organisational rules which will result in meaningful collaboration. To maximise the user experience we've been experimenting with social networking platforms, like Drupal, and tuning them for professional applications.

We have been asked to build a facility that allows climate change professionals to meet and share ideas so that, as a community of practice, they support better research and share resources more efficiently.

> *Semantic tools are the only practical way to find useful connections between assets*

Within this domain, there are huge differences in the way people collect, store and exchange their data as well as their knowledge. The barrier to entry must be lowered; they don't have time for triple stores, OWL, FOAF and any other buzzword. Users can participate by simply sending an email to the system the network then suggests linkages by parsing out natural language. A flexible metadata framework is also essential to allow datasets, photos, and models to be registered and seamlessly accessed. Naturally, there are multiple vocabularies, differences of scale values, etc. Semantic tools are the only practical way to find useful connections between assets.

These assets and the people that are associated with them form a Web of relationships that can be visualised in a number of interesting ways. Amongst many visualisations, users can see a timeline of assets filtered by organisation, person, and role relations. This will show a pretty good overview of who's who and who's doing what. Geo-referencing people, events, and data allow for further powerful visualisations; think Processing and Prefuse.

A second strategy is **providing deep tools and talent for "unique collaborations"** that are aimed at sparking a shift in an identified problem domain. BoaB is working with a large Australian science organisation to set up a sort of hothouse where the brightest cross-disciplinary minds can gather for a once-off jam session. We're providing a line up of tools and remote talent to help these researchers move fluidly across datasets, ontologies and access protocols with a minimum of turbulence.

This sort of thinking has brought about Hack Climate Change (HC2). This ground breaking event has secured support from the Great Barrier Reef research community. They'll be coming to Boston with some of the toughest knowledge synthesis issues on the planet. It's an open call to the Semantic Web community put your skills on the line for a good cause and the recognition. HC2 is scheduled for July this year. Visit the website for further details. www.boabinteractive.com.au

The third key area we have identified is **shifting legacy knowledge assets into the semantic** space. There are a lot of very interesting and very lonely RDBMS projects out there. The information encoded in these collections is staggering far too valuable to be left gathering dust. Semi-automatic tools

like D2RQ and Triplify offer one way. Another that shows an amazing amount of promise is Semantic MediaWiki. Alone or together, this cutting edge tech lends a new lease on life for those isolated bits of data.

We hope to see a steady stream of these science resources being funded for semantification – bridging the gap between the old, closed world into the rich tapestry of Sir Tim Berners-Lee's Giant Global Graph.

I am a firm believer that semantics can truly deliver that next step in the evolution of the Web. It began with networked bulletin board systems giving the idea that information can exist in an open environment. Social networking drives the point home even more my information and my identity are utterly important; much too important to be constrained by any one network. Removing those constraints is what linked data accomplishes bringing about the first real chapter of the Semantic Web.

There's a way to go before we're kicking the revenue goals we need for long-term profitability, but we're confident that we're on the right track. To get this far, we've been fortunate to work with some very generous thought leaders. In particular, I'd like to thank John Wilbanks, Alan Ruttenberg, Dr Renato Iannella, Prof Jane Hunter, Prof Iain Gordon, and of course Tom Heath.

David Peterson is a Systems Architect at BoaB Interactive in Townsville, Queensland.

david@boabinteractive.com

www.hackclimatechange.com

www.boabinteractive.com.au

Predictable Mavericks

Talis CEO Dave Errington looks back at the company's evolution over four decades, and discusses the new opportunities offered by the Semantic Web.

By Dave Errington

We are often asked how did you get here from there? As an organisation Talis will be 40 years old next year, entering venerable old age as a software business, but therein lies a story that underpins the ethos and values of this software business.

Talis – then known as BLCMP – was formed as a non-profit business cooperative in the UK in 1969 to share the process of cataloguing bibliographic works among university and public libraries. This process was enabled by sharing a common (and large) database of rich metadata running on a time shared mainframe computer. The business model was in the form of an annual subscription to access the service, whereby records were created, shared and consumed. That application – after various technology transitions – still exists today, and is called Talis Base. Even in those early days, the notions of shared data, shared innovation, community collaboration and an annuity revenue model were to form the foundations of the organisation for the next 40 years.

Over the next three decades, following the phases and evolution of the software industry itself, new applications were created and made available, first on proprietary hardware, then on Unix based minicomputers and so called 'client server' systems and then via the World Wide Web. These applications formed a suite of modules that automate the process of acquiring and cataloguing books for subsequent circulation by large communities of 'borrowers' – both students in universities and citizens in the local public library.

In the mid 1990s the first Web native applications were created to enable those users with early Web browsers to search an institution's bibliographic catalogue and then to reserve and renew items. This of course was before Amazon and Google brought new user expectations for search and e-commerce.

Throughout this time, Talis continued to build close and long term relationships with its customers, taking them through various technology transitions, but remaining consistent with its ethos of sharing data, sharing best practice and a highly predictable annuity based revenue model.

The latest chapter of the Talis story began in 2004 when I joined and formed a new executive team whose first challenge was to restructure the company from a non-profit, library centric cooperative into a commercially focused software business capable of retaining, motivating and attracting the talent needed to carry the organisation forward for at least another 40 years.

Alongside this a more ambitious agenda was set. Yet again, we needed to transition our existing applications to take advantage of evolving software technologies but this time we were to take a wider view than the relatively small library domain and build a modern software 'platform' that would be capable of powering applications in many other domains. Thus the Talis Platform was conceived as a Software as a Service application platform that takes advantage of Semantic Web technologies to power existing and new applications.

Once again, the common threads of shared data and shared innovation, with deep and real community collaboration – all coupled to a subscription based business model – prevail. This time however the Web itself provides many of the community mechanisms to make the process of fostering a powerful peer community more achievable. A belief in the principle that 'markets are conversations' has led us to be active bloggers and podcasters in the markets and communities that we serve. We are also acutely aware that thought leadership without execution leadership will not sustain our position as we strive to be an enduring and impactful software business.

Our highly predictable business model has led to a recurring element of over 75% of total annual revenues. This, compounded by 98% subscription renewal rates and an average customer lifetime of over 15 years, leads to a highly visible forward revenue stream – assuming of course that the intrinsic success of our customers in their use of our applications is paramount in our approach to customer services.

All of these cultural attributes and financial metrics combine to position Talis well for future decades. However it is our people and their daily actions and recurrent behaviours that are the actual manifestations of our culture. This I would argue is what leads to a successful business, often measured merely in financial terms. Creating an environment for smart people to be highly effective usually results in an effective organisation. Choosing the right technologies in solutions that add enough value, aimed at market segments at the right time – with a dollop of good business luck – usually leads to a financially successful business. That is where we see the opportunity for Talis and the Semantic Web. It is certainly very early days and the next decade will see whether we are smart enough to capture the opportunity.

Dave Errington is Chief Executive Officer at Talis.

Open World Thinking

Nadeem Shabir explores the idea that Semantic Web applications require a change of attitude if developers are to realise their full potential.

By Nadeem Shabir

"Openness tends to be an inexorable movement through time"
Sir Tim Berners-Lee

For me, one of the major goals of the emerging Semantic Web is to bring knowledge representation capabilities to the Web. However, the ability to put knowledge on the Web, share it, reuse it and even remix by using standard Web based mechanisms is very challenging. My own background is in Artificial Intelligence so the area of knowledge representation is one that I am familiar with, yet I really struggled for some time with the notion of what the Semantic Web represented, and how it might emerge, until I came to a rather profound realisation: it's about how we think about problems rather than mechanics of how we build solutions. At the moment our thinking is constrained by the technologies we are familiar with and their associated limitations.

The Web in its current form as a Web of Documents is very different to what we envisage as the Web of Data, or the Semantic Web. Getting from one to the other is not about a technology change, which is where many of us get hung up. What it's really about is a Paradigm Shift. It's a completely different way of thinking about the problems we are trying to solve and the applications we are trying to build. Fundamentally, It's about Open World rather than Closed World thinking.

Classically, in Knowledge Representation, the Open World Assumption and its opposite the Closed World Assumption are used to express the extent to which knowledge within a given system is viewed to be complete. Completeness is very important in closed world systems since it allows inferences to be drawn from what is not recorded in the system, in other words if you fail to find an answer in the system then you assume the answer is "no" or "false", whereas in an Open World system the answer would be "unknown". For example:

Statement: "Richard" "is a citizen of" "England"

Question: Is Richard a citizen of France?

"Closed world answer" (for example RDBMS) answer: No.

"Open world" answer: unknown (Richard could have dual citizenship).

Closed-world perspectives on data and meaning preclude the ability for further questioning. The answer to the question: "What don't we know yet?" is not possible to answer. It's capped at what you already know. This is a concept at conflict with a 'Web' metaphor or perspective. Asking: "What else is there? Where can I find it? What can I do with it?" are fundamental to the Semantic Web, and to the way in which we go about creating it.

In many ways the transition from Closed World systems to Open World systems, or from classical RDBMS' to the Semantic Web, is also about the transition from "structure up front" to "structure on the fly". No matter how hard we try we could never develop a relational database schema, up front, that modelled the entire world - it's just too complicated. The Semantic Web, as a Web of Linked Data, allows us to combine data from different sources to create structures on the fly.

The current efforts of the Linked Data Community attempt to ensure that Linked Data is published on the Web in a style that emphasises data re-use and the connections between related data sources. So if your current source of information doesn't contain a particular fact you can combine that data with another source which includes additional information about the same concepts. Sir Tim Berners-Lee was right when he said: Linked Data is essential to actually connect the Semantic Web. So for me the Semantic Web, as a Web of Linked Data, is not only an Open World system, but can only be achieved by thinking in an Open World way.

> *It's a completely different way of thinking about the problems we are trying to solve and the applications we are trying to build*

Let's consider a familiar example. One of the dangers or stumbling blocks we often face as software engineers is that, traditionally, the development of applications has always been approached in a Closed World mentality: We have domain experts from whom we elicit requirements, we then design databases to store the information the application will depend on, and then develop our applications as a tightly coupled view onto that data. Somewhere

down the line we decide we want to integrate with other systems which have been built in the same way so we have to use cumbersome integration solutions or find ways of exposing our data using Web Services. This works to a certain extent, but each integration brings its own difficulties and overheads.

Conversely we can approach the problem in a different way by thinking about the data as an entity that we want to publish and enable others to re-use. This means that we must think about the granularity at which we want to design our data, and then publish the data using existing ontologies (FOAF, SIOC, DC, etc.) if they are appropriate. If these ontologies are not appropriate then we need to design and publish new ones to describe our data, and enable others to re-use it. We can facilitate this further by linking to and interconnecting with other published data sets.

The Semantic Web, as a Web of Linked Data, allows us to combine data from different sources to create structures on the fly

We also need to approach this data design activity without specifically considering or focussing too heavily on the application we want to build so that we can get into the mindset of treating the data as an entity in it's own right - this helps to ensure that the data representation we design is not constrained to a single view. If we succeed in doing this our application becomes a single, contextualised, view onto that data. If we have succeeded in interconnecting our data with other published data sets then we are not only enabling re-use we are also encouraging the kind of unexpected re-use of information, that Sir Tim Berners-Lee referred to in his design discussion on Linked Data, he also recently stated:

"An application can run on a desktop or in my browser, it's my agent. It can access all the data, which I can use and everything's much more seamless and much more powerful because you get this integration. The same application has access to data from all over the place."

Perhaps if we move more toward a metaphor of an application as a view or a perspective, the Web of Linked Data will begin to make more sense. Instead of designing applications to do tasks with your data, engineers can begin creating views on the data. Essentially, instead of designing, using and selling an application to do something, they'd be creating, utilising, and selling their contextualised perspective onto this amorphous Web of Linked Data.

Nadeem Shabir is a Technical Lead at Talis, working on the development of several new applications that rely upon the Talis Platform.

nadeem.shabir@talis.com

This Week's Semantic Web

by Danny Ayers

Catch up every week at blogs.talis.com/nodalities/category/this-weeks-semantic-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/

WEB 3.0? IT'S WHAT HAPPENS WHEN YOU HAVE LOTS OF WEB, LOTS OF SEMANTICS, & LOTS OF SOCIAL INTERACTION WITH IT!

The semantic wave is here. Over the next decade semantic technologies will drive trillion dollar global economic expansions, transforming industries as well as our experience of the internet.

A new landscape is developing rapidly. Semantic technologies are being applied broadly across all industry sectors from consumer internet to mission-critical enterprise applications, and for new infrastructure from multi-core chips, to mobility, to web-scale cloud computing.

Along with investors, technology developers, and enterprises in the public and private sector, you want to know where the opportunities are, which to attack, what to build, what to buy, and why.

Project10X's new study provides answers. **Semantic Wave 2008 Report: Industry Roadmap to Web 3.0 and Multibillion Dollar Market Opportunities** charts the next stage of internet evolution and the role of semantic technologies in next generation applications and business models. This is business intelligence you can't afford to be without. To thrive in the coming business ecosystem— learn more at **project10x.com**!

PROJECT10X

Christine Connors talks about Semantic Technologies at Dow Jones

Christine Connors, Global Director of Semantic Technology Solutions at Dow Jones, talks to Paul Miller in this transcript of a recent podcast. See tinyurl.com/67fwg4 to hear the podcast and for links to resources mentioned during the conversation.

Paul Miller: Hello, and welcome to this podcast with your host Paul Miller. Today I talk to Christine Connors, global director of Semantic Technology Solutions at Dow Jones. Christine is one of the speakers at this year's Semantic Technology Conference, which begins in San Jose on the 18th of May.

We discussed the topic of her paper and explored some of Dow Jones' other work with Semantic Technologies. For more information on this podcast and others in the series, please see the accompanying show notes on our website. Thank you.

Christine, thank you very much for joining me for today's podcast. Before we dig into the detail around Dow Jones and what you're doing there, can you tell listeners a little bit about yourself and where you came from, please?

Christine Connors: Sure, I'd be happy to. As you said, my name is Christine Connors. I have, in fact, a library degree from Simmons. I was there in the late '90s. I was always sort of fascinated with information. In fact, I think one of my favourite programs ever was James Burke's "Connections" series. How all this data interweaves is a personal interest as well. I left library school and ended up at Raytheon, where I actually started in the libraries as a Special Librarian.

When you're in a small, special library you do everything. You do the cataloguing and the reference work and the research. But I ended up moving into information technology at Raytheon and working on knowledge representation, enterprise search; worked with our content management folks.

That's where I really got interested in Semantic Technologies. And I got to work with some really bright people there including Kevin Lynch of the missile systems unit. Then after that I went Intuit for a little over a year, doing online content management and the like in the small business division.

That was in Tucson, and while Tucson is a beautiful place it wasn't home. I'm an East Coast girl, so when I had the opportunity to move back to the East Coast, I did take it. I joined Dow Jones last November as global director of Semantic Technology solutions. So that's how I got here.

Paul: Great, thank you. Now, global director for Semantic Technology solutions; there's a trend here in this podcast, making fun of people with really odd job titles; [laughter] says the technology evangelist. So what does a global director for Semantic Technology solutions do?

Christine: I am in the enterprise media group at Dow Jones, which consists of the financial information services, the news wires, the Factiva suite of products... which is different from the consumer media group... which is where the Wall Street Journal, and Barron's, and Market Watch and such live. FEER is what I was thinking of, Far Eastern Economic Review.

The enterprise media group needed someone to wear a few hats, and that's what I'm doing. One of which is as the business champion for Synaptica, which is a product that we acquired with the Synapse Corporation about two and a half years ago. It is a vocabulary and metadata management tool.

That has got over a decade's worth of development into it. So it manages metadata but does so in adherence to standards. So everything from Z39.50 all the way through to RDF and SKOS; so that is a big part of my job. Another part is the Semantic Technology solutions for products within EMT.

I work on the R&D for the business case. I help the product champions and the PMs understand how they can embed Semantic Technologies into their products and use them for their customer's benefit.

We also have a consulting unit here at Dow Jones. And I do some work with them as well. So, for example, the National Library Board of Singapore is doing some fantastic work in this space. It's really great to be able to help customers solve their problems in the Semantic space as well.

So, that's what I do as global director. I love my job. It's been fantastic so far. It's been a good learning experience as well as a good sharing experience. We have a fantastic team all over the world doing really interesting work.

Paul: OK. So is this focus on Semantic Technology within Dow Jones a new thing with your post or is it something that Dow Jones has recognized as an important area has for a long time?

Christine: No, Dow Jones has been doing this for quite a long time. It's only recently though that they've come up with this title. My predecessor, Dave Clark, I think was director of global taxonomy. They recognized too that it was time to acknowledge the job as being more than that. So, Dow Jones has been doing taxonomies and metadata work for decades.

It is really the only way that Dow Jones could have managed such large quantities of information that we have coming through here on a daily basis. Because we do have the journalistic staff and the editorial staff creating content, but we are also a large aggregator of content.

It's the only efficient way to move data from one place to another.

Paul: Are you imposing your own standards upon this content as it comes in? Or are you taking existing taxonomies and things that exist out there in the wild and applying them to your own content?

Christine: We do a little bit of everything. It's really a hybrid approach on many fronts; on the editorial as well as the technological fronts. So, there is something called Dow Jones, formerly Factiva Intelligent Indexing... which is our own set of concept vocabularies. But there's also... and we do use things like NewsML and ITF... and also, of course, we're looking at XBRL. But we also have a large number of proprietary systems in house doing that work.

There are things that are pretty basic; geographical vocabularies, which are easy to predict. But there are also complex conceptual entities in the system as well.

Paul: Dow Jones is quite a big organization, part of News Corp now. How aware do you think people on the ground within the organization are the Semantic Technologies or the Semantic Web, or any of that, comes into helping them do their job?

Christine: I think that everybody here believes that the Semantic Web is our future. I believe that the folks on the ground are fully on-board with the importance of taxonomies, with meta data schema.

I think that we are migrating towards a fuller awareness of the power of the Semantic Web. The things like triple stores and ontological search, those are ideas that are well understood in pockets of the organization and are moving their way down the food chain to the editorial desks, as it were.

I think that things like RDFa and micro formats and some of the easy access points are going to help drive it home. Even though some folks might not understand the full impact of the "how" you get it done, everyone is on board with "what" we get from it, the benefits that we get from it.

I think that we are migrating towards a fuller awareness of the power of the Semantic Web.

Paul: An organization like Dow Jones has traditionally been in the knowledge management business. A lot of organizations of that type have tended to monetize the access to the basic data; you've paid a subscription to get access to the basic stuff.

I think a lot of media and knowledge companies are going through a shift at the moment, where they're providing easier access to the raw data and they're monetizing the value added around it, with a range of services of various kinds. Is that something that Dow Jones is doing?

Clearly, we're seeing it with increasingly free access to things like the Wall Street Journal on the consumer side; but what about on the enterprise side?

Christine: In terms of monetizing access to the raw data, sure. We do sell raw data. We provide content, we aggregate content, and now we're also providing analysis of content.

There is a shift going on, because it is no longer just about providing snippets of facts. People want to understand the relationships among those facts.

This week we announced the acquisition of Generate, Incorporated, and the formation of a new business unit inside of Dow Jones called Business and Relationship Intelligence, BRI.

What that is going to do is provide those value-added services to the raw data, which will analyze and visualize, for our customers, the relationships among data.

While it's easy to start with companies and executives and those kinds of relationships, like we see in the social Web, it actually is a very powerful technology that can take it into the abstract layer, into concepts as well.

We are there. We think it's going to be game changing, especially for us.

It's a paradigm shift for media companies. So we see the Reuters/ ClearForest integration doing this as well.

We definitely believe there are numerous opportunities to provide additional value on the basis of the raw data that we have in our archives.

Paul: Sounds good. In that you mentioned Generate, Inc., which I think you were announcing the acquisition of yesterday...

Christine: Yesterday, yes.

Paul: What do you get with them?

Christine: We get a number of capabilities. In fact, around identifying the relationships among various entities, companies, as I said, is one way. People are another. But the real

value is combining their abstraction and disambiguation capabilities from a technological standpoint with our vast repositories of data.

It is taking that content. Right now Generate indexes content out on the Web, which is going to be a valuable addition to our technological portfolio here at Dow Jones. But combine that with decades' worth of historical data and we will be able to provide our customers with a full-spectrum view of what's happening in a given industry, or with a given company, or with specific people.

That's one of the key benefits to this new business unit that we're forming with the Generate acquisition. It's really an exciting time. You're going to see some of the initial capabilities displayed in our sales works business.

Salesworks is a means for sales professionals to identify new opportunities. They can track information, news about their target companies; about people within those companies.

Now we've just exponentially increased their ability to identify those news stories, the information about the companies, the people within those companies. We're really excited. It's been a really busy week here, as you can imagine.

Paul: Yes, I can imagine. So whereabouts on the continuum do you see these products lying between manual or machine application of taxonomy-type structures on the one hand, and the natural language processing on the other?

You know, reading through a news story and extracting, magically, the appropriate key terms and things and linking them that way. Presumably there's a need for both.

Christine: There is, yes. We will continue. We have always and will continue to have a hybrid approach to this. We've got teams of very smart people who understand their subject areas who will continue to do both coding and data-quality work on our corpus of information. We have teams of folks who are experts in metadata and taxonomies who will continue to help deliver the most advanced capabilities in terms of how the schema's are organized; the taxonomy values that are applied to those schema.

We will continue to use best of breed technologies in terms of automating as much of those processes as possible; gaining those consistency and speed benefits that you get from the technologies.

We will continue to have various natural language processing, probabilistic models of the mathematical... the algorithms. We will use all of those to power the best and richest database of facts that we can provide to our customers for their own analysis. We can provide analysis up front, but really what you want to do is provide the customer the means to do the analysis for how it works for them.

It's one thing to provide a chart. That's great, it's useful and it's fast. But we might not necessarily be analysing the data in a way that makes sense for our customer. If we can let them do that analysis, that's the real win.

Paul: That presumably means providing very intuitive tools around the underlying power, to allow people to really get to grips with the data.

Christine: Right, so exposing these various best-of breed technologies to the right place at the right time, in a usable format as you said. Be it various visualization tools or easy-to-use search and discovery interfaces, to easy-to-use, almost drag-and-drop functionality in terms of performing ad-hoc analysis. Those are the things that that's where we're moving. That's where we have to go.

It's looking at the entire research process, not just one small part of it.

Paul: Yes. End-to-end, cradle-to-grave; semantic technology.

Christine: Yes.

[laughter]

Christine: Right, about discovery of a concept, about gathering and analysing the data that you've found and then really distilling it and delivering it. How do we support all of those areas?

Paul: Right. That's presumably an open question, actually. OK, let's change tack a little bit.

Next month we're both going to be at the Semantic Technology Conference in San Jose. You're talking about semantic coding at 120 miles per hour; pardon?

[laughter]

Christine: One of the fascinating things that I learned when I got to Dow Jones is that we have requirements; design requirements that ask us to provide content in milliseconds. If you think about a good size of our business being financial information and the rapidity with which these professionals move it's unbelievable.

They need information the split second it comes out so that they can get ahead of the market and drive the maximum value for their portfolios.

One really interesting example of this happened a couple months back, where Dow Jones was able to announce that AIG was rethinking some of its derivatives. We beat our competitors to the punch by, I think, eight minutes in terms of delivering this content; which actually knocked them off a $15 billion move in their market cap.

This kind of speed changes the game. So, when we have to deliver content to our customers before our competitors to stay competitive, it comes down to milliseconds. The complications come in, in terms of "What do you get out there? What kind of processing, how fast can it go?"

It does come down to selecting the right cables, which fortunately I don't have to do. But it also is important to think about what kind of entities are we extracting?

How are we encoding this data? How can we do this as quickly as possible?

Is it possible to take the human out of the equation in that first delivery of a headline, for example, and automate as much of that as possible? Having pre-written categorization rules, classification rules, entity extraction rules. Then modify the story after the initial publication of a headline, for example, to include the fuller set of data.

That's what I'll be talking about. How do you go about doing that? How do you scale that? What kind of trade-offs do you need to make to make that happen? That's the gist of the talk. So hopefully people will enjoy it.

Paul: Sounds good. I'll be there. [laughs] It does sound interesting.

It sort of ties into some of the trends that I've been looking at elsewhere; the difference between semantic technology as applied in a research context, and how you apply those same technologies to meet a business requirement.

As you say, time counts in the kind of things you're doing, in a way perhaps, it doesn't if it's your PHD project.

Christine: Right, right. I think there's some really fascinating stuff going on in the academic realms. We know from a recent, several months back Library of Congress report that they still prefer to do pre-coordinated data versus post-coordinated tagging.

But, the thing is, it doesn't scale necessarily inside of a business. You have to do both. While I would love to say "Here's my schema; here are all of my elements. I want to tag everything in every single article that comes through this place."

It's just not realistic in the real world to do that. And so you have to decide which are the key elements. And then amongst those which are the ones we have to tag before publication; before it goes out over the feed and what can wait until transaction time when somebody's actually searched for the data. Definitely gets into the whole long-tail aspect of what's being used and what's of value to the customer.

Paul: Yeah, absolutely. I think we often find that the things we labour hardest on to get just right, perhaps don't matter that much when it comes down to it. But, I guess one of the ways we find out is by trying it. Which is what you're doing.

This leads on quite nicely to my next question. I recently had a conversation with Tim Berners-Lee. And in that, he said that essentially the Semantic Web technologies that W3C had been working on are ready for mainstream adoption.

Now, you're adopting them in the mainstream for real. Was he right? Do they work?

Christine: I think they do work. I think that parts of them are easier to embed than others. For a lot of companies that were early adopters of taxonomies and XML and all of those various components that make up the Semantic Web, there were a lot of hacks that went on. So there are going to be shifts to move towards the standards that are now recommendations. And requirements, as I believe we'll see soon with XBRL.

Really dead simple tools are yet to be seen

But I think that things like micro formats or RDFa, those are easy to adopt and definitely ready.

I think that things like SPARQL and GRDDL still need some baking time inside the enterprise. They are definitely valuable. But to change massive architectures in terms of moving from a relational database to, say, a triple store is a huge undertaking.

I think that there are also some questions about "Do we want to do a triple store? Do we want to stay where we are?" Those kinds of things. "Do we want a quad store?" I've heard that come up a few times now. The importance of the time element for each of these extracted facts, as it were; these triples. I think it's still early days, but I definitely think it has started in terms of adoption, absolutely.

Paul: Is there anything major missing from that public set of specifications that make up that W3C toolkit?

Christine: From the specifications, no. I don't think there's anything missing. I think that using those specifications to deliver things for the average user needs some work. I think that rather than saying "Hey, we have these markup languages, go create your own grammars and vocabularies" is a problem.

I think the approach taken, for example, by the Dublin Core group makes it much simpler to adopt those kinds of technologies than other more advanced, deeper... well, things like XBRL.

I think that's where the gap is. It's not necessarily in the specifications, but in the change management around getting people to use them. Really dead simple tools are yet to be seen; again, with the exception of things like RDFa and micro formats.

Those are really, really simple. And those are fantastic. But I'm not sure that the average Internet citizen understands how to make use of things like a triple store of SPARQL as a query language. Or even how to apply simple semantic codes to their content.

So it will be great. The Semantic Web Education and Outreach group was fantastic. I think there needs to be more of that. But I think it needs to move out of the high tech and academic realm and into the How-does-the-average-Joe use it sort of "Semantic Web for Dummies" kind of realm.

Paul: Yes, we've got a little way to go to get there haven't we?

You were talking about easy to use tools. Within Dow Jones, are you able to go and pull tools off the shelf commercially, or do you have to build them all in house?

Christine: We do both. We do have some proprietary tools that we've built over the years to solve pressing business problems. We do have a taxonomy and metadata management tool that we actually sell to customers; which is Synaptica.

We are continually working on making that easier to use. In fact, the last revision that we launched in March, V7, we did major overhauls in terms of the user interface; which I was really proud of.

But we also do use commercial, off-the-shelf technologies to try to improve our products. So everything from... now, if you go to Factiva, if you're a subscriber to Factiva and have an article come back in your search results you can listen to it now, rather than read it.

So we are actively pursuing means by which to give the user content in the format that they prefer; either at their desktop, via a mobile connection, that sort of thing.

We're not snobs in terms of how we get this. If we can build it ourselves efficiently, we do. If we can acquire it, either by licensing or another means, we do. So that's Dow Jones' philosophy on the matter.

Paul: OK, sounds good. So, where does Dow Jones go next with Semantic Technologies then?

Christine: I think that given the pending requirements from the SEC, I think we'll be focusing a great deal on XBRL in the near future for the financial information side of our world.

From the content side, I think what you're going to see is a move away from the more commodotized facts and taxonomies and vocabulary lists to really rich databases of relationships; of connections... more than just the six degrees kinds of connections.

Allowing it to be viewed from our data stores, connected inside the organization so that customers can personalize it for themselves and for their organization. I think that's one of the benefits of moving towards open standards, is the ability to integrate more rapidly with the customer data sets.

And I think that's where we're going to go. It really opens up more opportunities for products and services that better align with our customer needs then detracts from it.

Paul: Yes. Presumably, you talked there about integration with customers needs and their products. This ties in quite nicely to some of the length data activity that's going on more widely in the Semantic Web community.

It really opens up more opportunities for products and services that better align with our customer needs then detracts from it.

To what extent do you deliver, or could you deliver data and capabilities that plug in to what a customer is doing within your work flow? To what extent do they come and deal with one of your products and get the answer within that product?

Christine: We actually do license our data as feeds already. So you can get an XML feed of a subset of our content. Or even, for example, our company. We have a company feed with which it is provided in an XML feed, and our subscribers to that feed can then integrate it into whatever internal system they choose to. So it could be a customer relationship system. It could be a supply chain system. It could be a document or content management system.

We already do that. What we're going to do is make those feeds richer in terms of the amount of data, the quality of the data, and the types of data. So the relationship types the disambiguation of the data. That's where we're going to change the game.

Paul: That sounds like it could be interesting; sounds good.

Christine: It'll be a fascinating journey; it's going to be great.

Paul: Oh, absolutely yes. It's been great Christine, thank you very much. Before we finish off, do you have any final things that you would like to say that I forgot to ask you about?

Christine: [laughs] No, I can't think of anything. It's been an interesting conversation, I appreciate the opportunity.

Paul: OK, thank you very much Christine. In that case, I'll say thank you now and I'll look forward to seeing your presentation in San Jose. Thank you very much.

Christine: Thanks, I look forward to meeting you.

*See **tinyurl.com/67fwg4** to hear the podcast itself and for links to resources mentioned during the conversation.*

This podcast is part of a series in which leading proponents of the Semantic Web share their thoughts.

See www.talis.com/platform/podcasts for the full catalogue.

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/

Barak Pridor Talks about ClearForest, Calais, Reuters and the Semantic Web

Barak Pridor of Thomson Reuters' ClearForest subsidiary talks to Paul Miller in this transcript of a recent podcast. See tinyurl.com/68akq2 to hear the podcast and for links to resources mentioned during the conversation.

Paul Miller: Hello, and welcome to this podcast with your host Paul Miller. Today, I talk with Barak Pridor, Chief Executive Officer of ClearForest. ClearForest is a division of Reuters, and is working actively in the Semantic Web space. During our conversation, we talk about Reuter's interest in ClearForest, some of the things that ClearForest have been trying to achieve with their technology, and take a look forward to some of the ways in which their open platform Calais allows other developers to benefit as well.

Barak, thank you very much for joining me today. You're doing some very interesting things at ClearForest, and at the moment, you seem to be attracting a lot of interest for that, which is definitely great to see. We'll dig into ClearForest and Calais and things in a moment. But before that, can you tell listeners a little bit about yourself and where you came from to get to where you are today?

Barak Pridor: Sure. First of all, thanks for taking the time Paul. It's my pleasure. My name is Barak Pridor, I have been the CEO of a company called ClearForest for the past eight years, since 2000. ClearForest was a company that was started by two Israeli entrepreneurs. One of them spent his entire academic career in an area that later changed its name to text analysis, or text analytics, which is really about harnessing technology in order to feed textual information into machines for processing.

If you think about data versus content, or text, data largely gets processed by machines. Very few of us actually read spreadsheets for pleasure, but most of us do read text and very few

machines actually process text. Text analytics was set out to change that it actually feed textual content into machines for business information, business intelligence sort of processes. The company itself grew from nothing in 2000, and expanded into spaces like business intelligence in financial services, obviously, quite a number of defence intelligence types of applications and publishers, business publishers.

Some of our biggest customers still today include folks like Reuters, which we're a part of today, as well as Reuter's biggest competitors. In the middle of last year, as part of raising additional financing for the company, we started conversations with Reuters where it became clear that Reuters' strategic outlook of where the world is going has changed.

Text analytics was set out to change that it actually feed textual content into machines for business information, business intelligence sort of processes

They were looking for some key pieces in order to implement that strategy. It turned out that what ClearForest did in our experience and expertise were a key part of that strategy and we ended up joining forces.

Paul: OK. So, you talked about those early days at ClearForest and the text analytics space. What was the problem, specifically, that ClearForest was set up to address?

Barak: Initially, the business problem itself was not well defined. Initially, it was just a very interesting technology discipline that could allow people to slice and dice through huge collections of textual information. The first two business applications that emerged, one of them was with the publishing community - people that create content for a living and have been facing pretty major margin pressures over the past few years.

All of us have been willing to pay less and less for content. They were struggling to find ways, they still are to a certain extent, find ways by which they can re-purpose all those content assets for new, innovative, premium revenue generating products. This was a platform, ClearForest platform that enabled them to do that - to re-purpose their content, to harmonize it, to feed it into a higher price-point types of applications.

The other one was the defence intelligence one. Obviously, specifically after September 11th, the need to sift through huge amounts of content, most of it unstructured in nature, and look for information around people, organizations, how they linked together, network evolution and stuff like that, that emerged to be a second, very important

application.

Paul: OK. Under the hood, how much of it is secret source and how much of it is using existing technology components from, for example, the Semantic Web stack.

Barak: In terms of numbers, I'd probably go for 50/50. Fifty percent of what we do is proprietary. When I talk to people, and people say "Well, we've heard about quite a number of organizations or companies or individuals that try to do these things," my point is that there's not a lot of rocket science here in terms of technology. Most of it is covered in academic research. But, there is a lot of know-how in bridging the gap between the academic aspect and real life.

Superimposing the analysis of text on real life problems with people's tolerance for precisions with dirty content versus clean content, and so on and so forth - something that just takes time and experience - and that's largely what we've developed at ClearForest; superimposing this on real life.

Paul: And real life is often dirty and messy. And real life data's always dirty and messy.

Barak: More than we'd care to admit.

Paul: Yes, absolutely. That actually raises an interesting point. To what extent does the ClearForest approach require either existing structure in the data or the application of structure to the data? Or is it doing an autonomy style read across the existing content and simply inferring the structure without anyone having to do anything to edit?

Barak: It is the latter. The vast majority of what the platform does is go through textual information with absolutely no structure and derive structure out of that content. I will say that a very important aspect, albeit not a big one of the product is the ability to look at existing structure and make intelligent assumptions based on that pre-existing structure.

I guess, the bottom line is, which is part of what I said earlier in terms of really solving the problem as opposed to trying to be neat and unique, the platform will take anything it could get in order to establish structure, whether it exists or not. Any clue is more than welcome.

Paul: The more clues the better, presumably. This, the inferred structure - this is my last question in this area. The inferred structure that the application adds to the content - Can you see that or is it sort of black box type stuff where you don't actually know where the inferences came from?

Barak: You mean when you look at the output?

Paul: When you look at the output, yeah.

Barak: The information is there. Obviously, there is sort of a crumb trail that would suggest why the system decided to apply or superimpose this type of structure on this particular type of element. Whether that crumb trail is being made completely available or not depends on the particular application, on the particular platform, so on and so forth. But, the information is there.

the platform will take anything it could get in order to establish structure

Paul: OK. That's fine. I remember some of the early sort of neural network type applications, which I realize were different, but there was a lot of black box magic happening inside and you were never quite sure why it had drawn some of the conclusions that it had.

Barak: I think, that's a very, very important point that you're making. Part of what we discovered, again superimposing this on real life is that when you try to scale something like this. When it's in a lab, and it's a black box, you can always find a few people who would just get pleasure out of trying to break down the black box and understand why this little movement to the right actually causes this big anomaly on the left.

But, when you try to scale this to Web scale types of usage, both in terms of number of users and amount of content, black box approaches have no future. I say that almost with full confidence. You need to be able, in a pretty straightforward fashion, to see cause and effect types of phenomena and not be a rocket scientist in order to do that.

Paul: I can agree with that, yes, definitely. Reuters, as you said in the beginning, you were acquired by Reuters last year because they saw a clear alignment between the direction they wanted to go in and the kind of things you're doing. Can you say a little bit more about that? What was it about you that they thought "Ooh, this is so good that we're going to buy the company"?

Barak: It's hard for me to say good things about ourselves, but I'll give it a shot. Reuters has been in a pretty interesting position for quite a number of years. Few people know, I didn't know anything about the history of Reuters before I became a part of Reuters, but it is a company that has innovated a lot over its 150 year existence - starting from figuring out that using pigeons would get information from stock exchange to stock exchange much quicker than sending people on horses, then through the first commercial satellite network for delivering news, and so on and so forth.

So, it is an innovative organization. It's been seen over the past few years that the battlefield has changed a little bit. First, the continuing erosion in the value that people are willing to attribute to just raw content. That's one trend that's happening. The other is there's a big war that's been going on for a couple of centuries now around latency. How quickly we can get the data to people and to systems in order to act on that data, and that battle is coming to an end because we're approaching zero.

I mean, all information - pick information, stock information, even news - the time it takes to appear on our terminal or inside our algorithmic trading system is coming close to zero. That battle is

almost over, and when Reuters looks at where it sees the growth opportunity, one of the major areas that scream out is the shifting of the value that people attribute from just the content to content with context - something that will enable them to look at the content through their own particular set of spectacles, whether it's for investment purposes, entertainment or what have you.

Being able to provide people with these customized spectacles is becoming where a lot of the effort is being diverted and ClearForest obviously is a fundamental part of that strategy because of its ability to create and provide these customized spectacles or lenses for looking at content.

Paul: I probably should know the answer to this question. Does Reuters still charge for the basic content or is that now freely available?

Barak: Reuters does still charge for the basic content, although parts of it are being made available for public usage on Reuters.com. Largely, for commercial purposes Reuters does charge for its content.

Paul: OK. And presumably part of the argument then is that as the value of that is perceived to fall, whether the value has actually fallen or not is another matter, but as the value is perceived to fall, you need to begin to add these additional contextual services that are exactly the kind of thing you do at ClearForest.

Barak: I'm not sure that the trigger is the falling of the value. The trigger is the growth prospect. There is less and less markets for just delivering raw information, but there is very fertile ground and a lot of opportunity for delivering context on top of that content. So, it's really the growth opportunity, and less I think the falling of prices at this point.

Paul: OK. I read an interesting blog post from Tim O'Reilly reporting on their MoneyTech conference in New York a couple of weeks ago. He had an interview with your incoming CEO, Devin Wenig, who said some very interesting things. I commented at the time at our blog that he really seems to get it. Does he get it or was he well briefed?

Barak: Oh, he gets it. Devin gets it. I don't think it would be possible for Reuters to just talk the talk without walking the walk. In order to command on a shift like this - and we'll be talking about the Calais initiative a little bit later on - everybody needs to get it in order for the company to do something quite as bold as Calais, I believe. Devin totally gets it.

Paul: OK. As you say, we'll touch on Calais in a moment. You said that everyone at the top needs to get it. I'm with an organization that has gone through quite a change process recently. The top certainly got it, and there was a process then of insuring that the rest of the organization got it too. To what extent do you think Reuters gets it?

Barak: Reuters is a big organization. After we finish merging the Reuters organization with the Thompson organization, the Reuters organization is going to be a pretty big organization - around 50,000 people deployed in 130 countries around the world. It's going to take a while for all parts of the organization to really shift their thinking and understand that we're doing two things in parallel.

To some people, it may actually appear contradictory in nature. So, there's a lot of internal evangelizing taking place. There's a lot of time that we spend in order to try explain the logic to all our colleagues internally. It's going to take a while, but there's no way back.

Paul: No, there isn't. I think, there could be a perception, internally and externally, that some of the things you're doing around Calais, around some of the things that Devon was saying in his interview - you're cannibalizing your business model, aren't you? That isn't meant in a bad way. The existing model is actually being undermined by some of the new things you're doing. That raises obvious opportunities and challenges.

Barak: I almost tend to think that when you are a product or service company, almost any time you release a radically different type of product that's targeted at a different constituency, in some way you're cannibalize yourself. So yes, there is some cannibalization threat. The same would be true by starting to offer way back whence some of the Reuters content on Reuters.com.

Presumably, there are some people who would have been willing to pay for that. What you're looking for is to do the thing that ultimately will pre-empt any damage that cannibalization costs. I think, in the case of really providing context services to the world, the same context services that we're using for internal Reuters content, there's not a lot of math actually that you need to do in order to understand that down the road it's two completely different orders of magnitude in terms of the outside.

the Reuters organization is going to be a pretty big organization - around 50,000 people deployed in 130 countries around the world

Paul: Yes, absolutely. So, everyone on Reuter's senior management team has read "The Innovator's Dilemma" then?

Barak: I don't know, but I presume.

Paul: It'll be interesting to see. It's certainly an exciting time to be turning a large and venerable ship around, so it'll be interesting to see what happens.

We mentioned Calais a few times, and this, I guess, is something that's been attracting quite a lot of interest online in the past couple of weeks. In case anyone hasn't seen the announcements in the news on places like ReadWriteWeb, can you say a little about Calais and what it's trying to do?

Barak: Sure. Calais... by the way, for those of you who might have forgotten, Calais is a city in France. It is the city that Reuters used at the time to lay down the first telegraph line between mainland Europe and the UK - between Calais and Dover.

Paul: Oh, right! I wondered why you picked the name. I didn't know that. I learn something new every day. Great!

Barak: Don't we all? Yeah. So, that's the reason for the name Calais. It is a service that Reuters is offering openly to the world, where we're inviting people to submit any type of content to the service and get it back richly tagged by the Calais service, free of charge. We're also being bold in the sense that we're encouraging people to use the results for non-commercial or commercial purposes. We don't care.

Obviously, there's a big thrust around developing a developer community, around this service, and that community has been developing very rapidly, actually exceeding our expectations by far. There is already a tool, an application development effort that's starting utilizing this Calais service. The same service that Reuters is offering the world through Calais, it is starting to use on its own internal content.

It's important to mention at this point that obviously we see this as a big, positive impact on the Reuters business in the future since Reuters' content will continue to be generated and unique. And by being able to inter-operate that content with any other external content that gets generated around the world (because the underlying mechanism will be the same), Reuters' ability to turn out new, innovative application information consumption paradigms in the future will be greatly enhanced. That's the Reuters idea behind Calais.

Obviously, for any Calais type of user, we are offering capabilities which traditionally have cost people. They've been very, very expensive and not affordable to most people. We're offering them for free. These will allow people to get plugged into the Semantic Web to offer better search and retrieval capabilities on their own particular Web properties. So, it's sort of a win-win situation for everybody, we think.

Paul: Win-win. You're mad! [laughter] So, Reuters saw the value in ClearForest, and Reuters paid quite a lot of money to get ClearForest. And Reuters is now giving that value that they thought was worth paying a lot for. They've put it out on the open Web and said "Rupert Murdoch, come and use it for free." How does that add up?

Barak: A couple of minor corrections here. Reuters is not offering the entire value that it has been getting with ClearForest through Calais. Obviously, there are a multitude of internal new products, applications and services that we're working on inside of Reuters. The ClearForest expertise, and also technology that has been developed over the years, is being applied, And it's being applied in a unique and sometimes proprietary sort of fashion.

The important statement is that the treatment that content gets, is being treated with when it enters Calais, is the same one that at least initially the Reuters news content is being treated. That's the powerful statement. Ultimately, and this is something that I believe in with all of my heart, the competition is not going to be on who disseminates the content better, rather on who creates a more competitive application or end user experience on top of that. That is where the competition is going to be, and we're more than happy to go and compete there.

the competition is not going to be on who disseminates the content better, rather on who creates a more competitive application or end user experience on top of that

Paul: OK. That's quite an important point. You're saying there that the value has shifted from the raw data, the raw content, upon which companies like Reuters historically have built their businesses. You're saying it's moved to analysis, it's moved to aggregation. It's moved to a whole range of new things. I think, that's worth emphasizing.

Barak: Yes, it is. It would be presumptuous of me to say that it has shifted, but it is shifting. There is an enormous amount of value in the content and in the timeliness of delivering that content, and the tools and analytics on top. But, there is definitely a shift taking place, and we'd like to be on the early part of that.

Paul: OK. Companies like Amazon - in a slight change of direction - companies like Amazon have been investing quite heavily in the provision of Web services alongside their traditional e-commerce business. I saw that in their most recent set of figures, the network traffic for their Web services has passed the network traffic for their own properties, which is interesting. Do you see yourself going further into that particular game? Could Reuters do an Amazon in terms of that switch from their original core business, moving towards the sort of Web services application business?

Barak: I'd be surprised if anyone knows the answer to that. I certainly don't. It is a hard question because historically, Reuters is still largely known as a news agency. It does have a huge editorial operation that covers news from all over the world, and it will continue to do that. That is an important part of the Reuters business. Forgoing that, I'm not sure it'll happen in my lifetime.

Whether a transition of where Reuters makes most of its money, whether that transition happens, I believe it will. But, how exactly it will look like, I really don't know. I'm not sure there is anyone that knows.

Paul: It's an evolving space.

Barak: It is; definitely.

Paul: That's fine.

Barak: I don't think we know what's going to happen in the space a year from now.

Paul: Don't tell your investors that. We know what we're doing, we've got a road map. We know exactly where we're going. [laughs] But, yeah, I do agree with you. So, how does Calais present itself to the world? There's a website, obviously, and a developer community. But, if I want to use it within one of my existing applications, is it simply a Web services call?

Barak: Yes. It's NTI, it's Web service, it's well documented. You go to the website opencalais.com. Even I can understand what's going on there, which means that probably 90% of the people with very slight technical inclination could go in and just start utilizing it.

Paul: OK. What do I get back?

Barak: A good way to think about it, I think, is you submit a piece of content, the document of some sort. It needs to be textual in nature. You get back an RTFa file of your original content, except it's marked up.

Paul: OK. And what kind of market would I get back? I don't know. Say, it's a blog post that I send you, what is going to come back in addition to that text that I gave you?

Barak: Just as an example, there is a pretty vast variety of elements that would get tagged or marked up for you. For example, it will take your blog post and highlight all of the different companies that are mentioned there, all of the different people that are mentioned there. It would markup different types of events for example.

So, if your blog discusses a management change within a particular organization, it will tell you that this person entity has been identified as working for this company entity in this type of position. So, it is a pretty wide variety of different element types that it would mark up for you, and obviously tell you what their type is, and different event types and their various attributes that construct that particular event or fact.

Paul: Interesting. And you said that this is getting quite a lot of interest and quite a lot of use?

Barak: Yeah.

Paul: Are there any unexpected uses coming in or is it pretty much the kind of things you would expect people to have done with it?

Barak: First of all, these are early days, so we are only starting to be privy to what people are doing with it. We are just seeing the first applications come online. Based on past experience, I would probably say that 70 to 80% of what we are seeing or what we will see are things that we never thought of, which is obviously part of the motivation: Send it out there and see how people take and innovate around this very powerful core capability. So, a lot of what we are seeing is... I could say, yeah we thought of everything, but that would be a lie.

Paul: OK. And presumably we can follow along in the developer community on your site and see the kind of things that at least the people are talking about doing.

it will take your blog post and highlight all of the different companies that are mentioned there

Barak: Yeah, there is already a pretty interesting and pretty intensive set of discussions that is going on there.

Paul: That is good. Where does it go next then, is it just scaling what you have got or do you have clear plans already that you can talk about to extend some of the capabilities?

Barak: We have a roadmap for Calais where we are planning to make sort of major quarterly releases and upgrades of the service. Obviously, we are going to incorporate additional languages. Currently, we have only opened it up for English type of content. There is going to be additional languages incorporated later on this year.

I think, the major next step that we are looking to do - and it is a pretty drastic step forward, I believe - is we will open up the service not only for people to submit content to, but we will be opening it up for people that want to augment the service with their markup types of capabilities.

So, if tomorrow morning someone thinks that the ability to look at content and markup actors is important - I just made that up - we will be inviting people to come in and add their own markup capabilities to the system and make them available to everyone and then we believe the free market will determine which types of modules or markup services gets adopted by the wider sort of community and which don't make it. It is going to be a Darwinistic type of paradigm.

We don't believe that we will be able to accommodate the world's needs in terms of the different lenses by which people want to look at content, so we want to invite the world to contribute to that as well.

Paul: That is quite sort of Freebase idea, isn't it?

Barak: Yeah.

Paul: That's the kind of thing they have been talking about doing, are you talking to them? You may not be able to tell...

Barak: No, not at this point, at least not that I know.

Paul: Right, OK. So, if they are listening, they know where you are. That is interesting. Presumably, the Web service at the moment: If I send some data, I get it back and it is then available to me to reuse it as I see fit. Are there any ideas to start leveraging the stream of content that is going into Calais to actually begin to add value across those streams? I am not asking this question very well. Do you

see any value in mining the click stream, mining the stream of data to actually begin to draw inferences across those different streams, to add value to third parties?

Barak: Well, I think, there is an enormous... Well, first of all, there definitely is some value there, but I think that there is even more value, it is actually a very fundamental part of the service. When metadata gets matched to a certain article, for that metadata to be available for that piece of content throughout its life time, so we want to make sure that if you submit a blog post to Calais and get it back and that blog post finds its way somewhere else in the ecosystem, we want to make sure that later on through its life cycle, it gets matched with that particular metadata again should the need arise. So, it is, if you will, this footprint or fingerprint of each and every piece of content.

And obviously looking at the aggregate of metadata have a lot of value and will enable us also to navigate the evolution of Calais better. So, I am also not answering it very well, but my bottom-line is that there is a lot of value, a lot of know-how in the accumulating metadata as well as the way in which people deploy it.

Paul: Yeah, absolutely. So, I own my blog post when I give it to Calais and I own it when I get it back again, who owns the metadata?

Barak: Well, that is a question for lawyers, but I will make the point that it is not Reuters that owns the metadata, but we do retain, since we are the people that created the metadata for you, we maintain the right to use it. It is your metadata, you could do whatever you want with it, but we maintain a copy of that metadata and we will make use of it, otherwise it is not very valuable because that piece of content will never be interoperable with anything else. That is the idea behind it. But, we are not in the business of stealing intellectual property from anybody.

Paul: OK. No, I didn't think you were. This is just interesting. This is something we have been grappling with a bit at Talis. We have been doing some work on licensing for data and have funded some work on something called the Open Data Commons to try and clarify where ownership falls and where sort of usage rights fall when it comes to data. So, trying to apply some of the sort of Creative Commons like ideas, but apply them to raw data. It would be interesting to look at how some of that fits with what you are doing. It will be interesting to see. Good.

I think, as we draw to a close, I'm conscious of the time, those were the main areas I wanted to dig into and really to raise awareness of some of the things you are doing and to get people looking at it and thinking about what it might mean for them. Were there any other things you feel I haven't covered that you wanted to say before we wrap up?

Barak: Well, just to encourage anyone that is listening to this and thinks that there is even a remote chance of them being able to do something that is innovative or breakthrough using these capabilities together with other types of capabilities, make the effort. I think that, as I said earlier, we are probably not aware of 80% of what is possible with this and we want to see the world get cranking, so go for it.

Paul: OK, thank you very much for your time Barak and I look forward to seeing what people do with this moving forward. Thank you.

Barak: Same here; thanks very much Paul.

See tinyurl.com/68akq2 to hear the podcast itself and for links to resources mentioned during the conversation.

This podcast is part of a series in which leading proponents of the Semantic Web share their thoughts.

See www.talis.com/platform/podcasts for the full catalogue.

SAVE THE DATES

2009

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

www.semantic-conference.com