FUJI Xerox

White Paper
Semantic technologies and the future of document management

May, 2008

Author:

Anni Rowland-Campbell

Contents

Context

"I don't think we kind of understand just how profoundly Google has changed the context of how we work, day in and day out."
(Former Director of PARC, John Seely Brown, 2007)

Imagine you are searching for a document that you know relates to a person or project but you can't remember where you put it, or have no idea how to find it. Your first thought is to "Google" some key words, utilise "Boolean search" and wade through as many pages as you can tolerate. Alternatively, you call or email someone you think might be able to point you in the right direction and use your social network to elicit their tacit knowledge and wisdom.

However, all of this is a "work around". The average executive spends some eight hours per week just looking for documents. The business of "search" has driven much (frenzied?) development in information technologies over the past decade as organisations seek to manage knowledge and information in a more productive, efficient and effective way.

There may be a new solution in sight, and one that, if it eventuates, could revolutionise the Web and bring the concept of true Knowledge Management a step closer to reality.

The evolving World Wide Web

Sir Tim Berners-Lee originally designed the World Wide Web as an "information space" for both human-to-human communication and for that between machines. The "Semantic Web" is Berners-Lee's vision for a "networked intelligence" which gives machines the capability of seeing data and information contextually regardless of whether it resides in existing corporate databases, managed document bases, multi-media resources and other information sources.

"Put simply the Semantic Web would make it possible to treat the entire Web as if it were a database. In the same way that a developer can query data in a standard database and build applications that use that data, people would be able to query data from across the entire web and build as-needed applications that pulled related but diverse data from multiple sources."
(Sir Tim Berners-Lee)

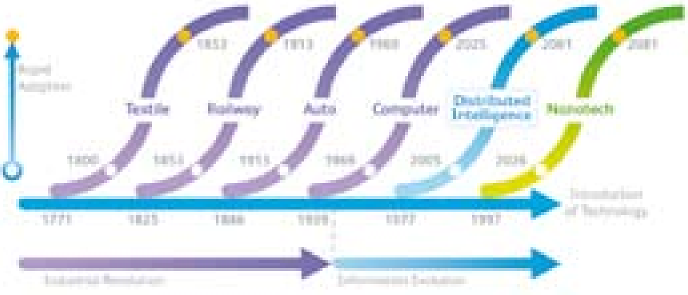
Therefore as the Web evolves from an essentially static medium to one which is immersive, and seamlessly brings together the real and virtual worlds, the challenge is of how to represent not just information but also the context and meaning of that information into data, information and knowledge bases so that these can be harnessed more effectively and productively. The Semantic Web is a "webby way to look at data, that is all" (Dave Beckett, Open Source Software Developer).

Since it's inception the World Wide Web has now grown to gargantuan proportions containing some 19.2 billion web documents, 1.6 billion images, and over 50 million audio and video files (http://www.ysearchblog.com/archives/000172.html, 2005 data and estimated to be 29.7 billion pages in February 2007 - http://www.boutell.com/newfaq/misc/sizeofweb.html). Each website has been create by a human being, or team of human beings, who have a certain logic relating to their organisation, purpose or mindset, which dictates how these web pages and websites are constructed, what language and terminology is used, and how the pages within link to each other and to the outside world. As with the search for documents the navigation of a website can often be a confusing and frustrating experience, particularly when there does not seem to be a "user friendly" interface. So, if it is difficult for humans to navigate the web, what must it be like for machines which operate on a logic based entirely on what has been programmed into them, with no pre-history, social network or memory with which to liaise?

All of this means more problems for business, and more opportunities for companies whose business it is to solve them, and many are seeing the solution in "The Semantic Wave".

The Semantic Wave

In the 1940's Austrian economist Joseph Schumpeter noticed that, since the Industrial Revolution, there had been a number of "long waves" of industrial activity which ebbed and flowed every 50 – 60 years and each of which brought a "new economy" which led to investment, excessiveness in the market, and then a massive rationalisation, all of which ultimately moved the world forward in terms of development. Norman Poire, an economist at Merrill Lynch, has modelled Schumpeter's theory projecting it out to the 21st century, and highlighting the fact that that the present time is at the intersection of the "computer" and "distributed intelligence" waves, seen by many as the "semantic" wave (Davis 2004).

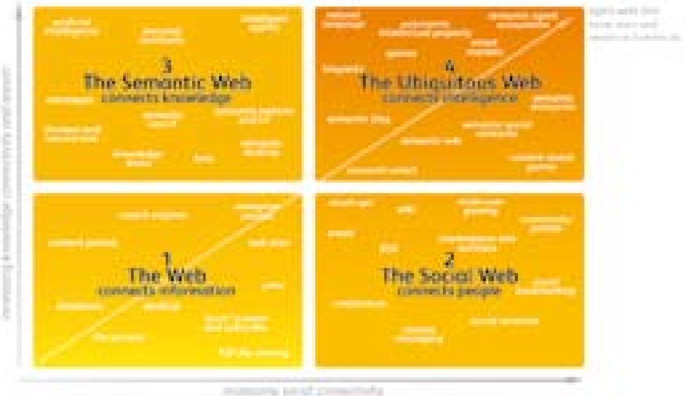


Conceptual advances occur about twice a century and lead to wealth of nations

The Semantic Wave embraces the four stages of internet growth including Web 1.0 which connected information through access to the internet; Web 2.0 which connects people; Web 3.0 which is about representing meanings, connecting knowledge and combining these to make the experience of the internet more relevant, useful and enjoyable; and Web 4.0 which connects intelligences in a ubiquitous Web "where both people and things reason and communicate together" (Davis 2008, p 3).

As the railway, automobile and computer industries have revolutionised the worlds in which they were invented so the "distributed intelligence" industry is now changing the very fabric of modern life by enabling such things as increasing globalisation, changes in demographics and work practices, and emerging business models which challenge the industrial age concept of industrial production by posing alternative solutions for virtually all aspects of doing business.

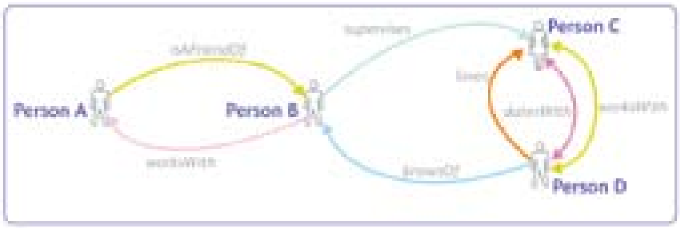
All are now underpinned by the internet, and dependent on the way that it evolves.



Semantics in context

Semantics are shared meanings, associations and know how about the uses of things (Davis 2004). They relate to context and meaning, and whilst computers "think" in terms of data - essentially digital ones and zeros - humans think in terms of documents, where data has meaning and is referred to in context. So, when we enter someone's name in a database we know the context within which we are entering it - they may be a customer or a supplier or a friend. To the system itself, unless we clearly state who we are and what this person's relationship is to us, that name is just data. If we leave the company then there is no record of that relationship, either tacit or explicit.

One of the things about semantics when utilized by computers is the ability to describe "things" in relation to other "things" so that we can find them via a range of associations rather than just keyword searches.



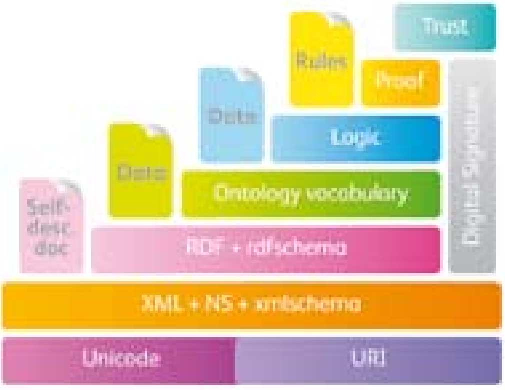
What the Semantic Web, and associated Semantic Technologies, can do is to articulate these relationships within the data and information itself, so that virtually all unstructured data can become structured. It does this by tagging data in a specific way that captures the relationships and embeds this information within the data being described. Thus, regardless of where that data is actually stored, the relationships remain and can be drawn upon at any later stage.

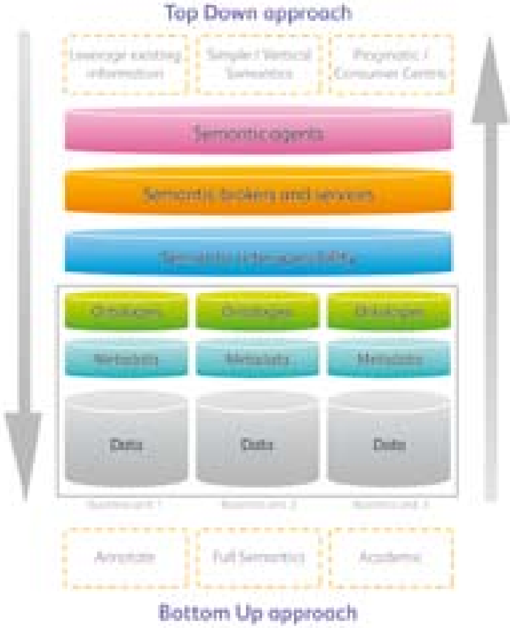
Differing approaches to the Semantic Web

The original vision of the semantic web is as a layer on top of the current web, which is annotated in a way that computers can understand.

One of the key challenges for the Semantic Web is how to go from today's unstructured web to a web rich with semantic information. There are two main approaches being developed.

The first, termed the "bottom up approach", is where people, or automated systems, annotate data and information and gradually build semantically enabled systems. The W3C Consortium, on which Berners-Lee sits, is authoring specifications for RDF (Resource Description Framework) and OWL (Web Ontology Languages), which enable the collective capture and description of information, along with the ontology (or dictionary) and the relationships with other pieces of information, in a rigorous mathematical way. The challenge of standards here is fundamental because each organisation and often systems within organisations have their own descriptors of data, none of which can talk to each other. Another attempt, called Microformats, is to embed the basic semantics right into HTML pages, and, whilst this is less powerful, this approach is becoming popular (Iskold 2006).





All of these approaches involve people annotating pages and, for many organisations, the cost of doing this has never been justified, the tools required have not existed and the challenge of teaching computers so-called "natural language" is a significant one.

The alternative is the "top down" approach: to develop smart software agents which can extract information from existing web content format. This involves leveraging existing, unstructured information to build vertical, semantic services, and then deliver the information back to the consumer via a consumer-centric application. The results are not as reliable or accurate as those that would be produced by the bottom-up approach because the semantics wouldn't be as accurate and the recognition is still algorithmic rather than being based on an underlying RDF representation. However, it could be a good start.

The Promises of Semantics – Key steps

The main thrust of The Semantic Web is to help people enjoy their on-line experiences and reduce the frustration of dealing with voluminous amounts of information that is incoherently organised and often irrelevant to a particular person's need. In other words, to be able to create systems which can truly deliver what a human being needs when and how they need it.

The key steps to building the Semantic Web are firstly to annotate all data and information, and then to develop persistent personal preferences with which semantic agents and search engines can match the data and information to the needs and desires of people at any particular time. As the world moves towards having a "services" focus this customisation is becoming more and more valuable, and the focus of services, to learn about the individual likes and preferences of customers, is something that is now being built into intelligent systems through iterative learning. In theory once this occurs all data will be annotated and then easily identified and retrievable.

Given these challenges the pain points of business and the complementary benefits to organisations are going to have to be significant for organisations to invest.

The main business pain points associated with the use of the Web for information is the sheer amount of time that employees, at all levels and in all roles, spend searching, looking at things that are irrelevant to their task or need, and explaining to computers exactly what they want. This, combined with the shrinking tolerance for wasted time and resources as organisations seek to maximise productivity and efficiency means that the economics of investing in semantics is beginning to make sense.

Therefore, for business the benefits of semantic technologies can be categorised into three major areas:

- Efficiency – cost savings – reducing repetition, resources and time
- Effectiveness – return on assets – productivity gain, leveraging intellectual capital & knowledge
- Edge – return on investment – developing new ways of doing things

In addition semantic technologies, once developed, should add value through:

- Accessing, connecting and developing legacy systems
- Changing business processes and IT infrastructure
- Mergers, acquisitions and divestments
- Co-ordinated access to external systems
- Global approach to customers and resources
- Tailoring of systems to local conditions – markets, regulations
- Large amounts of information
- Managing multiple databases simultaneously
- Accessing information sources in context
- Tailoring content for differing media, differing audiences, different languages

Impediments and road blocks

Whilst the promise of semantics has long been the dream of artificial intelligence (AI), semantics shows enormous potential in making software more efficient, adaptive, and intelligent. While the technology exists, the human element remains elusive and it is less a challenge of technology than it is a challenge of psychology, politics, linguistics and philosophy. From our perspective it is "semantic technologies" which offer greater promise than the fairly narrow prescriptive "Semantic Web".

For those who are closely involved it seems that whilst the technologies themselves are almost developed there are three key "people" challenges that must be overcome:

1. Senior people need to be educated about the potential offered by these technologies, and encouraged to invest both in their development and in readying their organisations. This means approaching data, information and knowledge from the top down and insisting that ontologies be developed and data be accessed and shared between departments and divisions.
2. That any pilots that are conducted are done so properly, well scoped and not rushed or undertaken half-heartedly. Those in organisations will resonate with this and the example of projects which are begun with insufficient resources and planning and then either cannot be begun (let alone completed!) or fail to deliver due to weak foundations.
3. This then links to the third challenge, that of expectations. Too often ICT projects are undertaken with unrealistic expectations and unachievable goals. Much of the hype around Knowledge Management exemplifies this, and, nine times out of ten, it was the "people" aspects that proved to be the weak link in the chain. Peoples' willingness to both contribute to and use KM systems has more to do with politics and power than technology.

The reality of Semantics

Whilst the Semantic Web was first talked about in 1998 the reality has taken time to develop and, for many organisations, the sheer cost of laboriously going through and tagging data and information simply hasn't been worth it (Iskold 2007). With the development of Web 2.0, increasing globalisation and the changing dynamics of how people are using information, there is a growing need to address this problem because:

- many organisations absolutely must link globally to customers and networks, both internal and external, which means opening up and sharing databases, developing common languages, and ensuring smoother information flows.
- the ability to provide true customer service relies on being able to "mash" data between applications (a Web 2.0 phenomenon) which also leads to greater productivity and more effective relationship management.
- the promise of eGovernment is totally dependent on the notion that a constituent needs to "Just Ask Once" and, instead of negotiating the myriad of government departments, can go through a single portal.
- for constituents and consumers the ability to obtain accurate and trusted information from a range of sources at any point in time means less frustration, greater utilisation of resources and, in some cases, perhaps lives saved.

This may sound futuristic and, like many IT promises, full of rhetoric and low on reality but our research is now beginning to show that this "semantic" world is coming and coming fast.

- Google and Yahoo have both financed projects built using semantic technologies

- Citigroup is heading an initiative using semantic technologies to organize and correlate content from various financial data feeds to help identify capital-market investment opportunities
- Oracle is working with a range of governments to develop technologies which can interrogate legacy databases and find more relevant information through semantic search
- HP is providing an open source toolkit to create semantic web applications
- Radar Networks is developing a semantic search engine
- Massachusetts General Hospital and Harvard University are undertaking a project to annotate clinical data using semantic tools to make it more easily and accurately searchable and sharable among researchers.
- The German government is developing "Theseus" working with SAP and Siemens among others to develop a "semantic web" around government information and services

and, finally, perhaps the most ambitious project we have seen to date:

- The government of Finland has developed FinnONTO (http://www.seco.tkk.fi/projects/finnonto/), which aims to "lay a foundation for a national metadata, ontology, and ontology service framework in Finland, and demonstrate its usefulness in practical applications. In our vision, a conceptual semantic infrastructure is needed for the semantic web in the same way as roads are needed for traffic and transportation, power plants and electrical networks are needed for energy supply, or GSM standards and networks are needed for mobile phones and wireless communication."

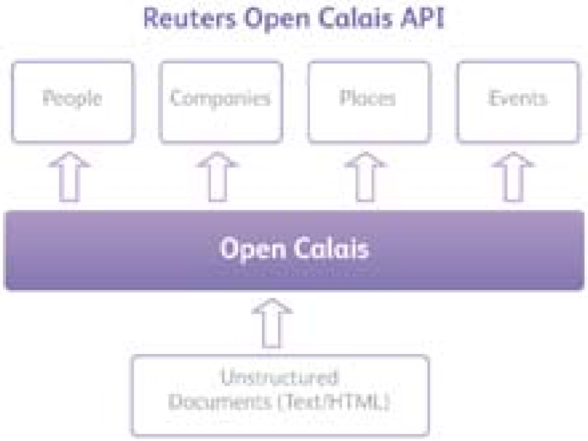
Real world examples

Within Australia a number of organisations and, in particular, specific units within larger enterprises, are well along the semantic journey having organised their data and are now awaiting the availability of next generation technologies. They have managed to overcome many of the political problems either through a gradual development of internal standards and ontologies, or by the imposition of them from above. Either way, they are preparing themselves for the benefits that they see semantics can bring. Some are taking it a step further and offering what they have to the world, free of charge.

Open Calais

One real example that anyone can experience is "Calais" (http://www.opencalais.com).

Thompson Reuters announced recently that it intends to "open its content to the world" and to facilitate this has acquired a company called Clear Forest which has built "Calais", a semantically driven engine which semantically enriches content and information on any website or data submitted to it. Calais works by analysing the text submitted and then returning that text to the user tagged in terms of key terms which identify elements within the text, such as "person", "company", "city", "industry term" and a range of others as defined by Reuters' vocabulary. This can either be done by submitting text directly to the Calais site or by downloading a plug-in for the Firefox browser called "Gnosis" (http://sws.clearforest.com/Blog/?page_id=32/ or https://addons.mozilla.org/en-US/firefox/addon/3999) which will, when enabled from the Tools menu, automatically submit the text of any web page being browsed to Calais and return it semantically tagged.



Calais' power is in the fact that it brings together three key elements and makes them publicly available to anyone with an internet connection and a web-browser:

1. Calais is open source and free to use
2. Calais is web based
3. Calais utilizes RDF and semantic technologies to tag the data

What does this mean? It means that you or anyone can submit a query to the engine and receive semantically enriched data which could potentially then be used to drive more effective search or then even generate customised documents.

Reuters are offering Calais in order to enable people to experience the reality of semantics. Possibly, by getting as many people to use it as possible and submit their content, it will help them build ontologies to semantically enrich their own data. Reuters are taking a bet that the future of documents will be "semantic" and they are utilising open source software and cultures to enrich their own data as a way to differentiate their service in the marketplace. It is a move to combat the threat of users no longer valuing the mere delivery of information. It is the content generation, search and contextualization – the very content and form in which data is delivered that now adds value.

Xerox applies Semantics in its Research

Within the Xerox world there is a great deal of development around semantic technologies, from the work done at PARC (Palo Alto Research Centre) through to the research on Parsing and Semantics being undertaken at Xerox Research Centre Europe, Grenoble. This work "utilises theoretical models of communication, language, dialogue, computation, and inference which take into account the context in which these activities are occurring". Two of the key products within this are XIP, the Xerox Incremental Parser, and "Factspotter", a semantically based search engine, which together enable semantic search and "information discovery". Xerox states that FactSpotter is capable of combing through almost any document regardless of the language, location, format or type; take advantage of the way humans think, speak and ask questions; and discriminate the results highlighting just a handful of relevant answers instead of returning thousands of unrelated responses.

In addition, Powerset, which licensed PARC Natural Language technologies and has recently been acquired by Microsoft", has recently announced the public beta of its semantic search engine (http://www.powerset.com/) which can "read and extract meaning from a user's query; and from the resources that they are searching" (http://blogs.zdnet.com/semantic-web/?p=141).

Research Pilots

As part of our research we are conducting a number of pilots, a number of which are running text and documents of various types through the Calais engine to see what is returned, and to work through a number of processes whereby this could be used within an organisation to more effectively organise and manage data.

Other organisations are now beginning to use Calais to enhance their data and prepare to leverage semantic technologies.

Sydney's Powerhouse Museum has already done this to a large extent by tagging much of it's collection (both manually and using Calais) and

(http://www.powerhousemuseum.com/dmsblog/index.php/2008/03/31/opac20-opencalais-meets-our-museum-collection-auto-tagging-and-semantic-parsing-of-collection-data/; http://www.readwriteweb.com/archives/australian_museum_uses_open_calais.php; http://reuters.mashery.com/forum/read/13214 comment by Tom Tague). We will also investigate how Calais compares with other engines of a similar nature such as those developed by Xerox such as XIP, Factspotter and Powerset.

Implications and Conclusion

One thing that needs to be clearly understood is that this conversation is similar to the one that we were having fifteen years ago when people were talking about Web 1.0 and Netscape.

Nobody is going to own the semantic online world, but once organisations begin to organise their data and information semantically then there is going to be a big difference between semantic and non-semantic databases and websites. Reuters has opened up Calais to the world because it "wants the world to be tagged". Once the world's content is quickly and easily accessible to customers then Reuters can deliver better, faster, more precise and relevant information, which creates an outstanding customer experience. In addition Calais builds a unique set of assets for Reuters including a growing semantic database of people, places, companies and events and an open API is training the system.

Semantic technologies as they evolve will be applied within a number of areas:

1. e-Commerce – in particular through facilitating communication between semantic agents which reside within specific systems and therefore bringing disparate data and information systems together
2. Search Engines – to find semantically similar terms, as has been described above with Powerset
3. Web Services – to provide rich service descriptions.

The Web is in its infancy and already Web 2.0 has been described as "inspirational" because it has shown people what is possible on the web. They see a marked contrast to how technology is used within their organisations on a day to day basis.

Semantic technologies will have a similar effect, and those organisations who begin the journey early will be the ones who benefit the most, both in terms of developing the required knowledge, skills and capabilities with which to more effectively manage data, information and knowledge in their own businesses, but also through being more effective and efficient in their work with customers.

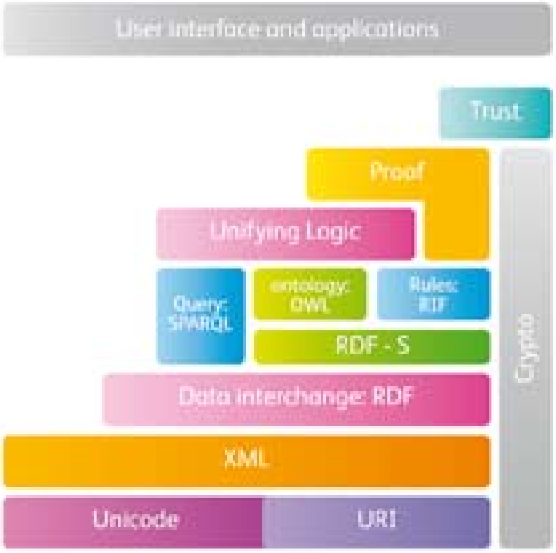
Glossary:

Semantic Web and Semantic Technologies

The semantic web is an “evolving extension” of the World Wide Web and it aims to process information at the level of meaning (semantics). This can enable machines to derive meaning and context from web pages and hence deliver a more meaningful and relevant user experience.

Essentially the Semantic Web is about “inter-operability” between databases and therefore between any systems which store data. It relies on two basic things:

1. common formats for the integration and combination of data drawn from diverse sources, where the original Web mainly concentrated on the interchange of documents.
2. language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.”

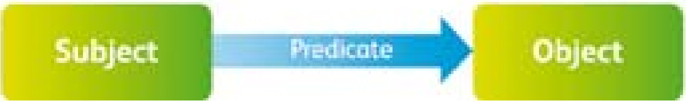


URI – Uniform Resource Identifier

A URI is a unique name that identifies a resource, and that resource can be anything to which we can attach identity, be it an information object (like a document or webpage), or a real world object (like a person or a thing).

RDF – Resource Description Framework

RDF is a framework for describing and linking resources on the web and it allows URIs to be organised into directed graphs, which are themselves composed of RDF statements or “Triplets”. These Triplets follow a very simple logic, which is that one has a relationship with the other,



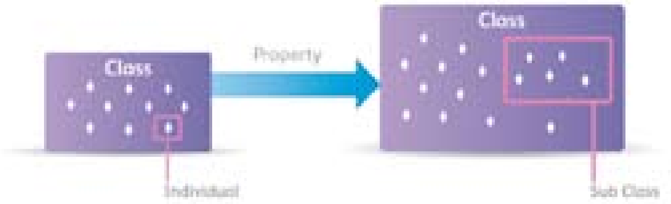
Therefore in RDF we can:

- declare classes like Country, Person, Student and Australian
- state that Student is a subclass of Person
- state that Australia and England are both instances of Country
- declare hasNationality as a property relating the classes Person (its domain) and Country (its range)
- state that hasAge is a property, with Person as its domain and an integer as its range
- state that Peter is an instance of the class Australian and that he hasAge of value 48.

RDF Schema

RDF Schema is a vocabulary description language which:

- ·allows us to define classes and properties
- ·allows us to organise classes into hierarchies
- allows us to connect classes using our own properties
- provides the facilities needed to define and describe classes and properties
- does not provide the classes and properties themselves - we need to create our own or use pre-existing ones



OWL – (Web Ontology Languages)

- ·An ontology is an agreed way of describing “things” within a shared environment, essentially the terms, relations and constraints that are formally used to specify a body of knowledge.

Semantic Agents

Agents are software programmes that can assist users and act on their behalf within the digital world. A Semantic Agent will utilise knowledge about resources, content, media, language, processes, functions, and how to communicate with other agents and they collaborate with other agents across platform(s) to provide services and capabilities.

References:

References:

- Berners-Lee, Tim – “The Semantic Web Road Map”, www.w3.org/DesignIssues/Semantic.html, 1998
- Cregan, Anne – Senior Engineer, National ICT Australia, Metadata Open Forum, Sydney, May 2008
- Davis, Mills - The Business Value of Semantic Technologies, Top Quadrant, September 2004, colab.cim3.net/file/work/SICoP/Semantic_Technology_Conference_2004_0908/BusinessValue_v2.pdf
- Davis, Mills – Project 10X’s Semantic Wave 2008 Report: Industry Roadmap to Web 3.0 & Multibillion Dollar Market Opportunities, February 2008, project10x.com/dispatch.php?task=exsum&promo=sw20081000
- Iskold, Alex - The Road to the Semantic Web, http://www.readwriteweb.com/archives/semantic_web_road.php, November 14, 2006
- Iskold, Alex - Semantic Web: Difficulties with the Classic Approach, September 19, 2007, http://www.readwriteweb.com/archives/semantic_web_difficulties_with_classic_approach.php
- Iskold, Alex – Top-Down: A New Approach to the Semantic Web, http://www.readwriteweb.com/archives/the_top-down_semantic_web.php, September 20, 2007
- Kelly, Kevin - “Scan this book”, NYT, 14 May, 2006
- Poire, Norman - Social Dynamics and the Investment Cycle, http://www.market-innovations.com/social.html
- Pridor, Barak - Interview about Reuters’ “Open Calais”, http://blogs.talis.com/nodalities/2008/03/barak_pridor_talks_about_clear.php
- University of Berkeley – “How Much Information” - http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Web References:

(http://www.ysearchblog.com/archives/000172.html

About Fuji Xerox Australia and Semantic Research

Fuji Xerox Australia and RMIT University have been researching Semantic Technologies for the past three years funded by a Commonwealth Government Australian Research Council Linkage grant. This research investigates the impact these technologies will have on document-centric industries such as printing and publishing, and covers the research being undertaken worldwide as well as the ways organisations are beginning to utilise semantics as part of their strategy to manage data, information and knowledge. The working documents of the project are publicly available through Fuji Xerox Australia and will be presented at a conference in 2009.

For more information please contact:

Anni Rowland-Campbell
Innovation and Research - Semantic Technologies
Email: **Anni.RowlandCampbell@aus.fujixerox.com**