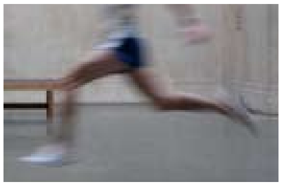
School of Information and Library Science
Summer 2008 e-publishing project:
Open Access in the age of Semantic Web
professor: Tula Giannini, PhD
student: Cedomir Kovacev

the past

Every several years Web is powerfully demonstrating its stubborn ability to reinvent itself. The initial excitement over discrete, but linked text documents flowing through frustratingly slow telephone lines has already acquired a kind of a romantic allure commonly applied to the things and happenings of the distant past.

It is clear now that the Web and the technologies that enabled it have placed an incredible power into our hands, relentlessly feeding us with knowledge, while at the same time, after blasting our brains with mostly meaningless information, allowing us to anonymously choose always convenient option of being ignorant all over again. And what else is to blame but our insatiable desire for infinite entertainment and instant reward so gracefully and generously offered by ever changing global information playground.



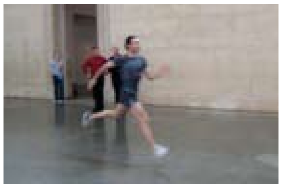
Martin Creed, Work No. 850,
Tate Britain, London 2008

There is a heated and unrelenting debate going on about the meaning and the future of the Web, but the controversy extends to the past as well, into differing interpretations of what actually happened. This is not surprising since the arguments over interpretation or more often reinterpretation of the past are pretty much a staple of any kind of historical discourse.

This paper, given the subject, is inevitably going to be, at least in part, speculative in nature. It would not be fair to say that Semantic Web exists as just an idea. The main components are at the advanced development stage, and are being successfully applied in various domains.

At the beginning of his in many ways impressive presentation at the Second Bloomsbury Conference in London in July 2008 Andrew Walkingshaw from University of Cambridge remarked how Semantic Web concept has over time acquired what could be described as a bad reputation. When more then a decade ago pioneers of the idea talked about the future of the Web they described something that looked like a fairly quick transition to the world pretty much controlled by intelligent agents that roam information cyberspace and in the process make fast and competent decisions related to both vital and mundane aspects of our lives. In their ground braking article “The Semantic Web - A new form of web content that is meaningful to computers will unleash a revolution of new possibilities” Tim Berners-Lee at al. (2001) outlined exactly this vision and almost a decade later described it as premature, unfortunate foray into science-fiction. Technology it turns out could not keep up with overblown expectations boldly and unwisely nurtured by modern preachers of brighter future.

On the other hand the traces of new emerging Web with underlaying semantics as well as the pressing need for meaningfully structured and described web content are more then obvious today. And it is clear that soon enough nothing is going to be the same. Almost all the speakers at our e-publishing program as well as participants at the Second Bloomsbury Conference in London kept making references to emerging Semantic Web. Everybody seems to be aware that the virtual environment is bound to experience dramatic changes and all the institutions concerned, from libraries and repositories to commercial content providers, are taking steps to prepare for that.



Martin Creed, Work No. 850,
Tate Britain, London 2008

This paper argues that Semantic Web is going to greatly influence and possibly transform publishing in general and scientific publishing in particular. Consequently, Open Access controversy is going to acquire a new dimension, the pressing need for free information flow within scientific community will achieve a new level of urgency and possibly inevitability. Creation of new tools for effectively managing research and article content will possibly break the monopoly commercial providers currently have on scientific data and information in general.

Everybody, with a possible and understandable exception of a few commercial providers, agrees that present situation in scholarly publishing is untenable. I will try to demonstrate the possibilities the Semantic Web concept offers without claiming that implied changes are inevitable. Most likely the changes are going to take place over a period of time and precise outcomes are notoriously hard to predict. This should not prevent us though from arguing for better practice as well as an environment that does not place a price tag on access to the information that should be in a public domain in the first place.

the present

In the current debates on the future Web developments terminology sometimes seems to be fluid which is not surprising since the field is just over a decade old. What comes next is usually described as Web 3.0, a legacy of the fact that sometime along the way the current stage was named Web 2.0. Consequently and retrospectively the initial stage was named Web 1.0.

It is interesting that the stages in Web development were given names that commonly denote new software application release. Cynics argue that the term was coined at the Conference on the Future of the Web, following the disastrous market crash from 2001 that nearly wiped out the tech industry. The new name was supposed to reinvigorate the interest in the tech industry and retrospectively looking this branding strategy proved to be extremely successful.

South Bank Graffiti, Wateloo, London 2008

While working at the prestigious Swiss CERN Laboratory Tim Berners-Lee developed in his spare time a complex intranet system that ultimately set off a World Wide Web frenzy. It seems strange today that the project was initially rejected by the Laboratory although it is obviously easy to make judgments from this perspective. In any case, the newborn Web 1.0 was about linked discrete documents and at a time that was a huge development. The content and the meaning of the document was impenetrable but the linking feature opened up a whole new world of communication.

The story about Web 2.0, the stage we are apparently currently in, and Web 3.0 that should follow is a lot more complicated. Nova Spivack is a well known proponent of new technologies that should transform the Web. The current stage of web development in his view is signified by creation of social networks as well as by user's ability to create desirable access points:

"Web 3.0, in my opinion is best defined as the third-decade of the Web (2010 - 2020), during which time several key technologies will become widely used. Chief among them will be RDF and the technologies of the emerging Semantic Web. While Web 3.0 is not synonymous with the Semantic Web (there will be several other important technology shifts in that period), it will be largely characterized by semantics in general.

South Bank Graffiti, Wateloo, London 2008

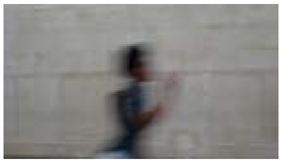
Web 3.0 is an era in which we will upgrade the back-end of the Web, after a decade of focus on the front-end (Web 2.0 has mainly been about AJAX, tagging, and other front-end user-experience innovations.) Web 3.0 is already starting to emerge in startups such as my own Radar Networks, but will really become mainstream around 2010." (Nova Spivack, 2007)

Leaving aside the unfortunate fact that he is not that subtly promoting the services of his company, Spivack's comments raised the issue of what Web 2.0 actually is. Critics emphasized the dubious claim that Web 2.0 is mainly about front-end technologies, the claim that does not relate to reality.

Tim O'Reilly, another prominent name in tech circles, reacted with the article that tried to clarify what Web 2.0 actually is:

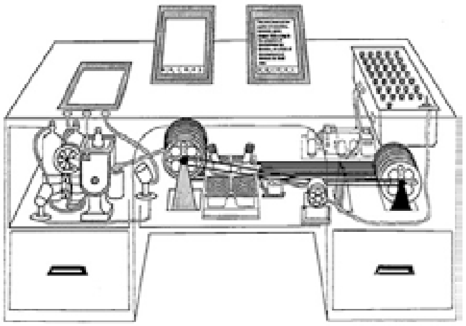
"I find the Web 3.0 arguments as clear evidence that the proponents don't understand Web 2.0 at all. Web 2.0 is not about front-end technologies. It's precisely about back-end, and it's about meaning and intelligence in the back end.

The real difference between Web 2.0 and the semantic web is that the Semantic Web seems to think we need to add new kinds of markup to data in order to make it more meaningful to computers, while Web 2.0 seeks to identify areas where the meaning is already encoded, albeit in hidden ways." (Tim O'Reilly, 2007)



Martin Creed, Work No. 850,
Tate Britain, London 2008

One of the most interesting and important precursors to the Semantic Web concept dates back almost seventy years. In the 1940s Vannevar Bush, a prominent scientist better known for his later involvement in the Manhattan Project, developed a detailed plan for MEMEX, machine that was supposed to make a creative use of technologies of its time. MEMEX was supposed to have the ability to handle different mediums - print, sound, microfilm, and to store vast amounts of data. But what would make it fundamentally different was its ability to record a trail of associations a scientist is following while working on a project. Bush considered the creative process in analyzing data extremely important and tried to capture it. MEMEX machines would be connected into a network and that way exchange of scientific information would be enhanced and creativity fostered. Ultimately, the machine was never built, but the powerful idea behind it remained influential.

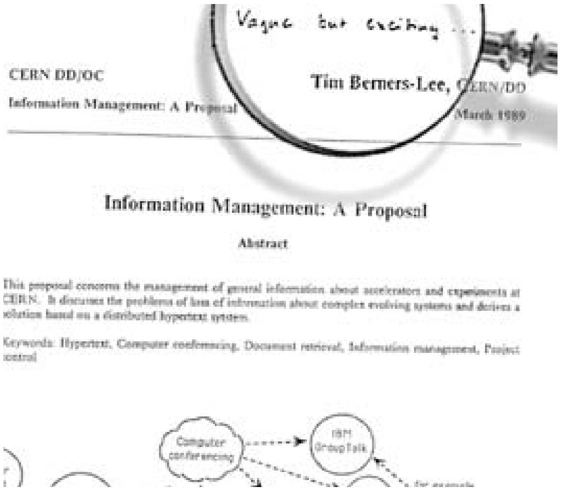


Memex: enabled associative indexing of information;
particular trail of association is saved for later use

The Semantic Web concept was developed by Tim Berners-Lee, who about three decades ago laid foun-

the future

dation for what eventually became known as the World Wide Web. The goal he defined in his ground breaking Scientific American article includes the environment in which machines can independently penetrate semantics of information which would enable them to make something close to what we call rational and logical decisions. That would entail completely new way of structuring information on the Web, the process that slowly, but inevitably is taking place right now.



Vague but exciting ...

CERN DD/OC

Information Management: A Proposal

Tim Berners-Lee, CERN/DD

March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

“Vague but exciting...” were the words Tim Berners Lee’s boss wrote on Lee’s proposal for an information management system that ultimately led to the World Wide Web. He allowed him to continue with the project.

Martin Creed, Work No. 850, Tate Britain, London 2008

“The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. Such an agent coming to the clinic’s Web page will know not just that the page has keywords such as ‘treatment, medicine, physical, therapy’ (as might be encoded today) but also that Dr. Hartman works at this clinic on Mondays, Wednesdays and Fridays and that the script takes a date range in yyyy-mm-dd format and returns appointment times. And it will ‘know’ all this without needing artificial intelligence on the scale of 2001’s Hal or Star Wars’s C-3PO. Instead these semantics were encoded into the Web page when the clinic’s office manager (who never took Computer Science 101) massaged it into shape using off-the-shelf software for writing Semantic Web pages along with resources listed on the Physical Therapy Association’s site.

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and ‘understand’ the data that they merely display at present.” (Tim Berners-Lee at al. 2001)

South Bank Graffiti, Wateloo, London 2008

Semantic Web relies heavily on Artificial Intelligence research. In its infancy Web was connecting documents as separate entities. Web 2.0 brought the ability to penetrate the document and search its content with keywords. Such search is more often then not indiscriminate. Hence the thousands of results we are frequently faced with. Semantic Web is supposed to make a big leap forward by structuring information in such a way so that its meaning can be extracted by machine. The assumption is that a program can make free associations between seemingly unrelated data.

The main components of Semantic Web were defined almost a decade ago by Tim Berners-Lee. It is interesting that his vision in this regard was fairly accurate. Three technologies or components he considers to be central to the future development of Semantic Web are XML, RDF and Onthologies.

XML (eXtensible Markup Language) allows for creation of custom tags. Authors are free to structure documents using their own tags that describe sections of a page. That way XML is open for creation of distinct arbitrary structures of information and its only common XML denominator is the fact that the document has a form of an ordered labeled tree. As Michel Klein (2001) says “this generality is both XML’s strength and its weakness. You can encode all kinds of data structures in an unambiguous syntax, but XML does not specify the data’s use and semantics”. In other words XML document may be well structured but if its vocabulary is arbitrary and during data exchange its meaning is going to remain obscure. Therefore it is necessary for those involved in data exchange to agree on common vocabulary. XML schemas offer such a solution to a certain extent. One of the main requirements for an effective XML schema is the level of its interoperability. Still, XML schemas do not specify the meaning of a document but its structure.

South Bank Graffiti, Wateloo, London 2008

While XML is a way of encoding data, Resource Description Framework (RDF) is a mechanism to tell something about data. RDF happens to be a simple and effective way of describing metadata about web resources. RDF interprets simple relations in terms of an object, its attribute and its value. But it does not go far enough in terms of semantics of a document.

This is where Ontologies come into play. The term “ontology” comes from philosophy, but it was at some point adopted by artificial intelligence and web researchers. For them ontology means a document that describes relations between terms. Ontologies also contain rules about relations between terms and that way supply powerful tool for defining semantics of a document.

James Handler (2001) defines ontology as a “set of knowledge terms, including the vocabulary, the semantic interconnections, and some simple rules of inference and logic for some particular topic”. Handler does not see the Web of the future as the Web that relies on one overreaching ontology. This is, by the way one of the core arguments regularly put forward when supposed impracticality of Semantic Web concept is analyzed. Handler sees the Web of the future as consisting of a large number of interrelated ontological components, mostly created by users. In other words, semantic markup is going to be a regular by-product of computer use.

Mathieu d’Aquin at al. (2008) describe next generation of Semantic Web applications that are capable of dealing with distinct sets of interrelated ontologies. Traditional knowledge systems, they say, were built in centralized manner and were therefore self-contained, closed and inflexible. Semantic Web applications must be designed to deal with heterogeneous data. And that means the ability to combine ontologies and knowledge resources. Several Semantic Web applications are been developed. Their common characteristic is the ability to combine information from several distinct ontologies related to search query expressed in natural language.

It is important to note that all these projects are still in experimental phase, but they are powerful demonstration of the underlaying concept. As Tim Berners-Lee noted, a set of interrelated ontologies has not just a potential of greatly improved search precision but can also evolve into a powerful tool for making multiple relations between distinct pieces of information.

semantic web and open access

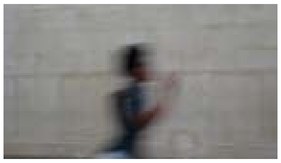
Contrary to common misconceptions Semantic Web is not a closed system with arbitrary, imposed rules. Most of the criticism tends to concentrate on supposed idea of some central authority that dispenses prescribed semantics throughout World Wide Web. It is actually possible to approach the concept, and at this point not just the concept, from a completely different perspective.

Martin Creed, Work No. 850,
Tate Britain, London 2008

Peter Murray-Rust from Unilever Cambridge Centre for Molecular Informatics is one of the leading advocates for Open Access and arguably one of the most prominent practitioners of applied Semantic Web. With Henry Rzepa he developed Chemical Markup Language (CML) that over the years became a mainstream scientific XML language. And with a group of his students he developed OSCAR, application that crawls the Web and searches for latest scientific data on crystallography. The application is then able to extract the data and provide visual representation of new structures.

Peter Murray-Rust believes that Semantic Web is going to rise from ground up:

“In our opinions the first generation of the semantic web will be built on open systems on the public Internet. For rapid development the resources must be open. (Current robots give up when asked to register for a site, add their names and emails, etc.). In biosciences this has been spectacular and a robot can access and re-use a wide range of high quality and comprehensive data (genomes, sequences, structures, etc.) Much of the full-text literature is now being made Openly available and bioscience publishers are looking at new publication models. This would allow the primary literature to become a primary knowledge base for the semantic web.” (Peter Murray-Rust, 2004)



Martin Creed, Work No. 850,
Tate Britain, London 2008

The development of Semantic Web concepts has a potential for breaking existing information barriers. Muray-Rust and his colleagues demonstrated this with the development and implementation of OSCAR, the application that’s entirely based on open source software.

It is possible to argue that the future development of similar applications will establish strong and wide network of openly available scientific information. That in turn may render closely guarded scientific information to some extent obsolete since advance in science heavily relies on collaboration:

“To appreciate the need for better data integration, compare the enormous volume of
experimental data produced in commercial and academic drug discovery laboratories around the
world, as against the stagnant pace of drug discovery. While market and regulatory factors play a
role here, life science researchers are coming to the conclusion that in many cases no single lab,
no single library, no single genomic data repository contains the information necessary to

discover new drugs. Rather, the information necessary to understand the complex interactions
between diseases, biological processes in the human body, and the vast array of chemical agents
is spread out across the world in a myriad of databases, spreadsheets, and documents. (Tim Berners-Lee, 2007)

Semantic Web concept and emerging reality provides a wide array of possibilities for Open Access initiative. Apart from obvious ethical issues related to commercial exploitation of at least partly publicly funded scientific research the wider issue of necessity for unrestricted access to fundamental scientific research within the new semantically powered digital environment may play an important role in braking the commercial hold on something that should be the "property" of humankind.

sources

Berners-Lee, Tim, James Hendler, and Ora Lasilla. "The Semantic Web." Scientific American. 17 May 2001. Scientific American. 11 Apr. 2008 <http://www.sciam.com/article.cfm?id=the-semantic-web>.

O'Reilly, Tim. "Today's Web 3.0 Nonsense Blogstorm." O'Reilly Radar. 4 Oct. 2007. 4 Apr. 2008 <http://radar.oreilly.com/archives/2007/10/todays-web-30-nonsense-blogsto.html>.

Spivack, Nova. "Web 3.0 -- The Best Official Definition Imaginable." Minding the Planet. 4 Oct. 2007. 8 Apr.

2008 <http://novaspivack.typepad.com/nova_spivacks_weblog/2007/10/web-30----the-a.html>.

Packer, R, and K Jordan, eds. Multimedia: from Wagner to virtual reality. London: W.W. Norton & Company, Ltd, 2001.

Berners-Lee, Timothy. “Testimony of Sir Timothy Berners-Lee Befor the United States House of Representatives.” . 2007. 6 Aug. 2008 <net.educause.edu/ir/library/pdf/EPO0719.pdf>.

Antoniou, Grigoris, and Frank van Harmelen. Semantic Web Primer. 2nd ed. Cambridge: MIT Press, 2008.

Murray-Rust, Peter. “Chemistry and the Semantic Web.” 2004. 6 Jan. 2009 <http://markmail.org/message/g4r-e5xvc242ostqx>.

Hendler, James, “A New Portrait of the Semantic Web in Action,” Intelligent Systems, IEEE , vol.23, no.3, pp.2-3, May-June 2008
URL: http://ieeexplore.ieee.org/iel5/9670/4525132/04525135.pdf?isnumber=4525132∏=JNL&arnumber=4525135&arnumber=4525135&arSt=2&ared=3&arAuthor=Hendler%2C+James

d’Aquin, Mathieu; Motta, Enrico; Sabou, Marta; Angeletou, Sofia; Gridinoc, Laurian; Lopez, Vanessa; Guidi, Davide, “Toward a New Generation of Semantic Web Applications,” Intelligent Systems, IEEE , vol.23, no.3, pp.20-28, May-June 2008
URL: http://ieeexplore.ieee.org/iel5/9670/4525132/04525139.pdf?isnumber=4525132∏=JNL&arnumber=4525139&arnumber=4525139&arSt=20&ared=28&arAuthor=d%27Aquin%2C+Mathieu%3B+Motta%2C+Enrico%3B+Sabou%2C+Marta%3B+Angeletou%2C+Sofia%3B+Gridinoc%2C+Laurian%3B+Lopez%2C+Vanessa%3B+Guidi%2C+Davide

Klein, M., “XML, RDF, and relatives,” Intelligent Systems, IEEE , vol.16, no.2, pp. 26-28, Mar-Apr 2001

URL: http://ieeexplore.ieee.org/iel5/9670/19905/00920596.pdf?isnumber=19905∏=STD&arnumber=920596&arnumber=920596&arSt=+26&ared=+28&arAuthor=Klein%2C+M.

Hendler, J., “Agents and the Semantic Web,” Intelligent Systems, IEEE , vol.16, no.2, pp. 30-37, Mar-Apr 2001 URL: http://ieeexplore.ieee.org/iel5/9670/19905/00920597.pdf?isnumber=19905∏=STD&arnumber=920597&arnumber=920597&arSt=+30&ared=+37&arAuthor=Hendler%2C+J.

Berners-Lee, Tim, and James Handler. “Scientific publishing on the semantic web.” . Nature. 12 Nov. 2008 <http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm#au>.