# Untangling the semantic web: what does it mean for scholarly publications?

In this presentation Louise Tutton will decode the meaning of the Semantic Web and its ultimate benefits within the publishing industry today. The Internet has revolutionized the way researchers discover and consume content but the full potential of the medium of the web is yet to be realised. The Semantic Web represents the next stage in this evolutionary process and offers exciting opportunities for publishers and their end-users. Louise will discuss real life examples as well as exploring some of the possibilities the Semantic Web can open up for publishers and societies.

Louise Tutton (Senior Vice President, Scholarly Division, Publishing Technology)

As Senior Vice President of Publishing Technology's Scholarly Division, Louise is responsible for the management operations within the division which include Client Management, Library Services, Product Development, Engineering, advertising and scouting new markets. Louise's 12 years in the electronic publishing industry have focused on client relations but also incorporated editorial and project management experience. Positions with Taylor & Francis, ABC-Clio in addition to CatchWord and Ingenta (now Publishing Technology), have given her a breadth of expertise and practical experience. Louise speaks regularly at conferences around the world and sits on several committees focused on scholarly and society publishing.

## INTRODUCING THE INTERNET AND THE SEMANTIC WEB

I think it is important to say that the web—the Internet—has dramatically changed the way people conduct their research and find data. Although research techniques have changed significantly, the web still follows a very traditional print model. Journals follow the journal, issue, and article model and books follow the book and chapter model. The huge amount of information available on the web makes it harder for users to quickly find what they are looking for. Another issue that we face is that computers understand syntax but not meaning. Therefore, they may not always understand synonyms and context and may not be able to provide other relevant material. Search engines are useful, but they often blindly retrieve all possible content related to a search term. A user may be presented 10 or 20 pages of research, making it rather difficult to find what they had initially been looking for. Currently, the web is actually a web of individual websites. Therefore, applications tend to function independently, without much interaction with other websites. They are not as integrated and connected with each other as they could be.

What is the Semantic Web and how can it help with some of these issues? Well, to draw a comparison, the World Wide Web, as we know it, is a web of documents, some of which function in isolation. The Semantic Web, on the other hand, is a web of data; it involves making connections and integrating websites. Approximately 5 years ago, Sir Tim Berners-Lee described the Semantic Web as "data on the web defined and linked in a way that it can be used by machines not just for display purposes but for automation, integration, and reuse of data across various applications."

How can the Semantic Web help address some of the issues mentioned earlier? The Semantic Web should be viewed as an enhancement of the current web, the next step, if you like, in the evolution of its life. It enables us to add more meaning to data. In other words, if an author name is stored as a content object, more information can be added about that author such that a machine as well humans can understand the additional information. By supplementing data, we can improve navigation routes as well as make searches more relevant to the end user. A common format and data tagging allow us to gather data resources from a number of previously independent and separate websites. By doing so, the content becomes more visible. This can drive extra traffic and add more value to content because additional information is available. The content thus becomes more useful to end users, subscribers, and members.

The younger Google generation will probably not remember life before the Internet, the PC, and the BlackBerry, all of which help in research. Therefore, we must make the search for relevant, reliable content easier and assisting the next generation as well. The Semantic Web can certainly make it easier for researchers from different disciplines to collaborate their research and make new discoveries. As Sir Tim Berners-Lee said, “The most exciting thing about the Semantic Web is not what we can imagine doing with it, but what we can’t yet imagine it will do.” Therefore, this is just the first step of many that are not yet known to us. So, we attempt to visualize how the Semantic Web connects and integrates different data resources.

As a real life example, let’s consider DBpedia. You may be familiar with Wikipedia. DBpedia is the Semantic Web version of Wikipedia. DBpedia attempts to more intelligently link Wikipedia’s 7 million articles. It extracts and tags information about people, places, music, albums, and films. For a machine to identify people, places, and so on, it gathers data sets from external websites and external applications. Thus, it enriches the information already available on Wikipedia.

To elucidate its significance in terms of content object, while Wikipedia has 7 million articles, DBpedia describes 2.6 million “things” and an additional 274 million “facts” using RDF, which is Resource Description Framework, and this is a Semantic Web format. It is an area that we at Publishing Technology are experimenting with. We take in XML and PDF formats and convert them into RDF to allow for more interesting prospects on people’s websites.

Why does DBpedia function like this? Well, the community wants to enrich user experience, and with all this additional information, they can support more sophisticated natural language user queries. For example, DBpedia would be able to answer the question, “Give me all cities in New Jersey with more than 10,000 inhabitants?” because the machine knows that New Jersey is a city and understands other demographic information as well. Another benefit is that the machine understands the query and can therefore gather related information.

A second example, which you may have heard of in the context of authors, is the “Friend of a Friend” or the FOAF Project. This project describes people, their activities, and their relationships with other people. It includes information such as the institution that a person is affiliated to, the project being worked on, who they know, who are their co-authors, which societies are they members of, and so on. In other words, it allows them to store such information.

I earlier mentioned the RDF record in the Semantic Web format. When this site is open, it begins integrating and connecting with other sites. FOAF also integrates with Facebook, MySpace, Blogs, LinkedIn, and Flickr (for photographs). At Publishing Technology, we extract additional information about authors so that it can be included within a publisher’s website or not included in the article information that is available. However, Semantic Web technologies can draw relevant information from external sites, for example, FOAF.

I find the third real life example rather interesting because it is Japanese and has just been launched this month by Toshiba’s research center. It is called the Word-Of-Mouth Scouter. This Word-Of-Mouth Scouter aims to improve the shopping experience. It is now being tested in bookshops and electronic stores. Essentially, what happens is a shopper visits a shop, takes a photograph of a barcode, and later compares reviews on different blogs, which will provide the shopper with a simple positive or negative response to help him or her decide whether to buy that product.

## THE SEMANTIC WEB FOR PUBLISHERS

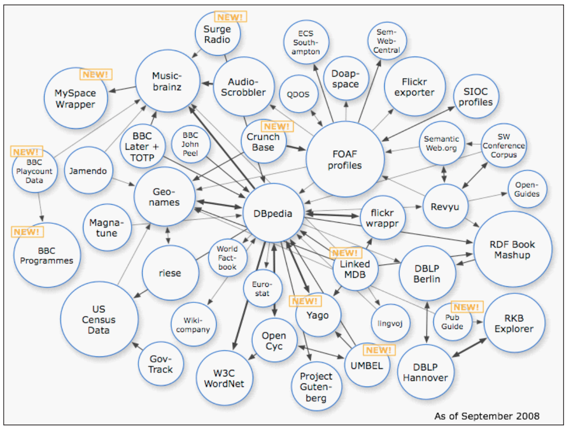
How is the Semantic Web relevant for publishers and what can you do to prepare your data Semantic Web? First, the more tags you add to new content coming through the web flow, the better. This is particularly with regard to subject-related information, whether to do with the taxonomy or classification system. For backfile data, which may be available in PDF or XML, more likely in PDF, a number of natural language processing techniques and data mining techniques can be used. Our publishing technology uses subject-specific open source tools to extract information from journal articles, book chapters, etc. Extra tags are added to data so that it can be used on websites.

## EXAMPLES OF DATA MINING TOOLS

Some data mining tools available within the community have been developed by the European Bioinformatics Institute. For example, Thomson Reuters' Calais looks at events, people, and places. There are a number of tools available, and it is important that we refrain from attempting to reinvent the wheel and become specialists in every subject when in fact, the researchers in the field do so, and these techniques can be applied to publishers' data.

A large number of available open datasets already use Semantic Web technologies within certain subject areas. For example, a project from the W3C encourages researchers to use this format. At the moment, there are already 17 billion entries that are connected by 3 million links, and this is only the initial stage. Therefore, you can imagine how quickly possibilities become a reality.

As of September, some of the datasets that we found or have been working with include DBpedia and FOAF, which I mentioned earlier, as well as Bio2RDF, which is within the biosciences area. Moreover, there are various information resources from the BBC, US Census data, etc. Therefore, there is a lot of information available to enrich the end user's resources (Figure 1).

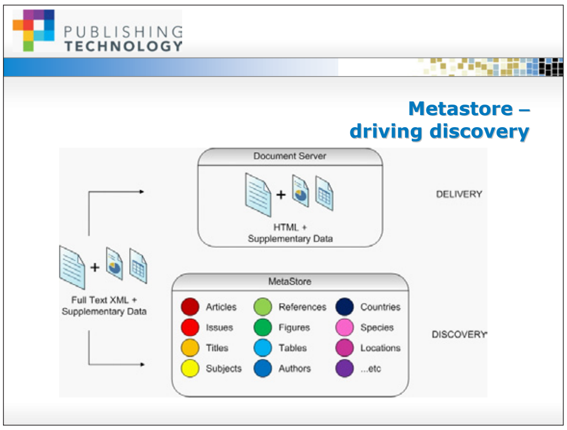


(Figure 1)
Image Credit: Chris Bizer. Reproduced under the Creative Common's License

## PUBLISHING TECHNOLOGY AND SEMANTIC WEB TECHNOLOGIES

I would like to share with you some of Publishing Technology's work with Semantic Web technologies. We have been developing a database to manage metadata behind our IngentaConnect and pub2web products. You may have heard of the IngentaConnect product. Pub2web is a standalone publisher branded web product. Across these two products, we are working with over 300 publishers and managing a significant amount of content. It is important for us to ensure that the database—the foundation of the products—is robust, future proof, and can scale very quickly. After 2 years of researching which technology should be used for the database, we decided on RDF and the Semantic Web format.

Metastore (Figure 2) is the database, the foundation between these two services. It equips end users with useful data by providing additional ways of finding content. It increases the visibility of content by offering cross-promotion of products and allows for easy integration of datasets such as the examples mentioned earlier with all the open datasets that are available.



PUBLISHING TECHNOLOGY

Metastore – driving discovery

Document Server

HTML + Supplementary Data

DELIVERY

Full Text XML + Supplementary Data

MetaStore

| | | |
|---|---|---|
| ● Articles | ● References | ● Countries |
| ● Issues | ● Figures | ● Species |
| ● Titles | ● Tables | ● Locations |
| ● Subjects | ● Authors | ● ...etc |

DISCOVERY

(Figure 2)

Then I will now introduce you a prototype that we have been working on and how some of these technologies can work. To offer some background, we incorporated some XML Open Access data from BioMed Central into a very basic pub2web site.

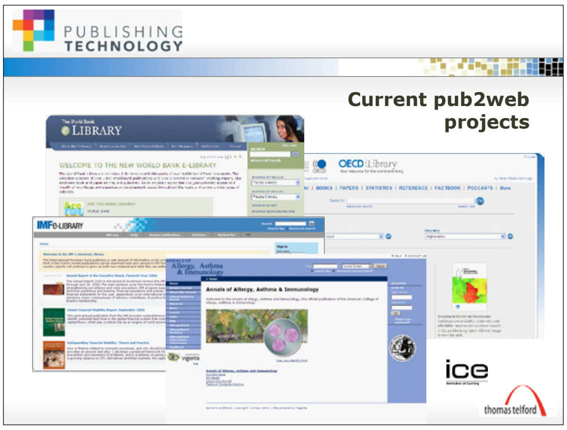
What we have done is apply the What Is It data mining techniques from the European Bioinformatics Institute to XML data and pull out the species names and gene terms. Thus, we have a list of species that appears within that journal and the number of papers associated with each species. We should also have a species map with all those terms but the more significant names or where there is more available content associated with that species name. Therefore, when we click on a species name, because we have extracted the species information from the text, we can store that as an individual content object and add more information. Thus, from the data that we have from the publisher, we can work out how many other papers are available within this website and within this dataset and use the Bio2RDF dataset to collect extra information. For example, it could be the rank, common name, taxonomy ID, and also an image. All this additional information comes from external datasets, making the publisher's content richer and more useful to the end user.

We also extract gene terms and split them into the biological processes, cellular components and molecular components, so that we have different ways of navigating the content. Thus, if a researcher and/or end user is searching for something very specific, all the information relevant to the search is very easily accessible. Similarly, we can have a search map for gene terms with all the gene terms as well as a listing of all the available gene terms. If I select one of the gene terms, I am then navigated to the homepage for that gene term. In such a scenario, we can extract data from within our datasets or say what other available content talks about this gene term. So we used the Bio2RDF dataset to find a description of the gene term, its ID, as well as the synonyms associated with it. Once this is keyed into the search engine, it makes the results a bit more relevant for the researcher. Although, this is an example from within the biosciences, I hope you can imagine how this could work in other subject areas as well.

We now follow the link to another paper that mentions, for example, the same gene term. So if we follow the link to an author name, we use data that we already have within the XML so that we can create an author homepage, include the e-mail addresses of the authors, their affiliations, and other articles written by them within this dataset and by their co-authors as well. However, for improving the potential of page, we could add photographs of the authors or perhaps connect to Flickr for the same. We could also connect to FOAF for information on the projects that they are currently working on and the conferences that they are attending or presenting at. Thus, a lot more information could be added.

Taking it one step further, we could create a homepage for an institution as well, drawing from dataset papers that have been submitted from this particular institution and any authors that are affiliated to that institution. Further, we could add a logo for the institution along with some facts and figures that DBpedia can provide. A lot of possibilities are being explored through some of the pub2web projects that we are currently working on.

To give you an idea of some of the projects that we are working on at the moment (Figure 3), we are building a new website on our pub2web platform for the OECD, a French statistical publisher. This is a rather interesting project because they have a wide range of content like journals, books, statistical information, and reference works, and they are also starting to experiment with podcasts.



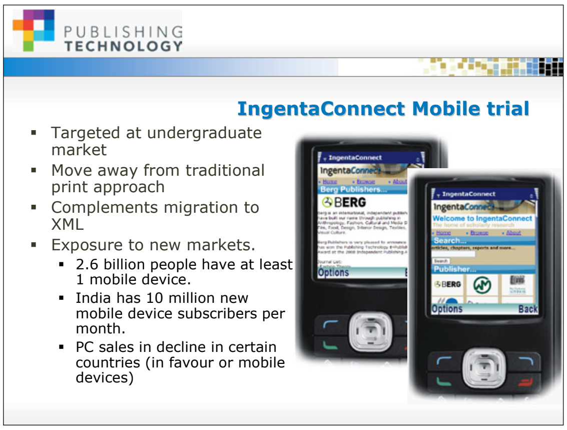
(Figure 3)

We have also recently launched a new website for the World Bank. This website includes all of the World Bank publications which primarily include books, but also journals, reports, and monograms. We have also recently launched a website for all the publications of the International Monetary Fund. At present, this is for their international offices; therefore, they do not have to send printouts to all their offices. Since they want to get into the institutional markets, over the next year, this will be marketed to institutions as well.

In contrast, we are working with an American society—the American College of Allergy, Asthma, and Immunology—to build a journal-focused pub2web site that will be accessible by their members. Some of the features that they are particularly interested in have to do with continual medical education. We are providing them with online quizzes as well as a news and events area for their members. Soon, we will also begin working on a pub2web project for a UK society—Thomas Telford—the publishing home of the Institute of Civil Engineers. Our project will bring together their journals, books, and manuals. It is attracting an institutional and society member audience and will integrate a print bookshop within their electronic products.

As you can see from all the examples, there are a wide range of subject areas and content types that can take advantage of the Semantic Web technologies mentioned earlier.

Technology keeps evolving and changing, and at Publishing Technology, we are continually looking to experiment and innovate such that content is as physical as possible. We also look to cater to different demographics. Two weeks ago, at London Online, we launched an experiment into the delivery of content through mobile devices; it is called IngentaConnect Mobile (Figure 4).



PUBLISHING TECHNOLOGY

**IngentaConnect Mobile trial**

- Targeted at undergraduate market
- Move away from traditional print approach
- Complements migration to XML
- Exposure to new markets.
  - 2.6 billion people have at least 1 mobile device.
  - India has 10 million new mobile device subscribers per month.
  - PC sales in decline in certain countries (in favour or mobile devices)

(Figure 4)

This experiment targets a younger audience. We are looking at the undergraduate student audience and how they use it. Feedback from that user group as well as librarians and publishers will help to determine the next steps for the project.

While deviating from the traditional print format, in this experiment, it is important that we not just focus on websites because users will access content from their mobiles very differently from how they would access content on the Internet. The challenge that we face is how would users want to access content on their mobiles and what should the business model be? Since there are a lot of different answers to these questions, this really is a case of experimenting and finding out what the market wants.

Another benefit of having a mobile platform is exposure to new markets. There are some figures that really stand out. Around the world, 2.6 billion people possess at least one mobile device right now. In India, for example, there are 10 million new mobile device subscribers every month. Therefore, in countries that perhaps do not have the infrastructure for widespread broadband and access to PCs, web mobile sales are increasing dramatically. Now is the time to make content more visible in these markets. We should also consider that the trends show that in some countries, PC sales are declining in favor of mobile devices. Therefore, this is a challenge that we are all going to face in the coming years.

## CONCLUSION

Finally, I think the message for today is that we need to move to an environment that is a web of data and that can be processed by machines—a continual evolution of the World Wide Web.

There are many ways in which data can be made ready for the Semantic Web. It does not require you to discard PDF or XML.

There are a number of benefits for publishers as well, which I have hopefully demonstrated. I would like to conclude by saying that in my opinion, this is the time for experimentation, and experimentation alone will prepare us for the next steps regarding the appearance of our web products.