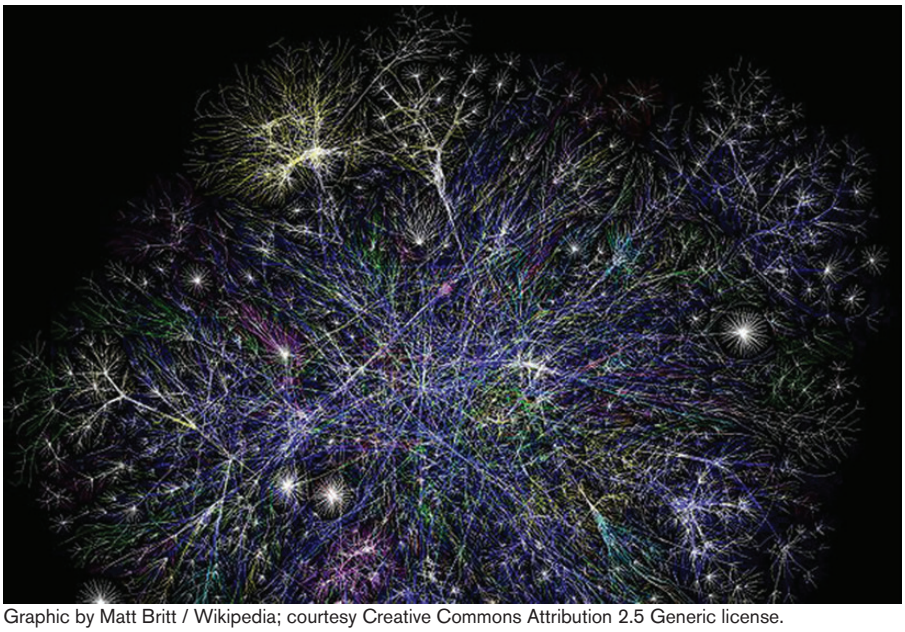
IPTC Mirror
No 156 / Feb 10
INFORMATION TECHNOLOGY FOR NEWS

Can news media use linked data for a stronger future?

Linking information from scattered public sites may be desirable for many, but is it right for profit-oriented news media?



Graphic by Matt Britt / Wikipedia; courtesy Creative Commons Attribution 2.5 Generic license.

A Yahoo! Answers respondent estimated in 2009 that by 2010 the size of the Internet would equal 12 piles of pages, each one twice as long as the distance between the Sun and Pluto. Another estimated the indexed World Wide Web soon would contain at least 19.71 billion pages. However, while documents and Web sites are linked, the information in them is not.

Call it "the taming of the Web". Call it fuel for the Semantic Web. Call it the "Next Big Thing".

Whatever you call it, do not call Linked Data a "passing fad". Recent evidence proves that, while still in its infancy, Linked Data is becoming integral to the Web's future. Activities on both sides of the Atlantic confirm this.

Governments are linking data

In the UK, the government has launched a new venture to increase public access to official data. It is known by its URL: data.gov.uk. The effort is being led by none other than Sir Tim Berners-Lee, inventor of the World Wide Web, and Professor Nigel Shadbolt of Southampton University. Many are already looking at how the vast amounts of public data can be harnessed to serve citizens.

One application makes the locations of England's almost 11,000 General Practitioners available from iPhones and iPods. Other developers postulate that multiple types of information about locales may help people decide where the best schools are or point to the most crime-free neighborhoods.

In the US, the Library of Congress — often considered the largest library in the world — now is implementing the Linked Data movement's approach of interconnecting data on the Web via dereferenceable URIs. In doing so, it is also establishing rules for other government agencies to follow. This means making dissimilar vocabularies of different agencies communicate with each other.

Library of Congress vocabularies that will become open source include the Library of Congress Subject Headings (LCSH), a thesaurus of Geographic Materials, geographic area codes and language codes. The Library's efforts support the Obama's administration's goal of making government information — already in the public domain — more transparent and accessible.

What are news media doing?

Some media publicly support the

*Continued, "**News Linked Data**" page 2*

SUMMARY

1 Industry weighs value of linking archival news data to the Semantic Web

4 Extraordinary General Meeting called for 10 March in Paris; Standards recently updated

5 European Broadcasting Union implements NewsML-G2

***News Linked Data**, from page 1*

Linked Data movement. The UK's British Broadcasting Corporation (BBC) — the largest broadcasting network in the world — is bringing information to its Web sites by linking to outside databases that have made their vocabularies open source. Essentially, BBC imports their data and combines it with its own to create richer programs.

Its lauded online program, BBC Wildlife Finder, is an example of the type of enriched information that can be constructed when one's own efforts are combined with the efforts of others. The BBC Wildlife Finder's richness results from linking to datasets at the World Wildlife Federation, the University of Michigan and DBpedia that are maintained by those entities. The linking is complex and technically sophisticated, but the costs of maintaining the material is distributed across the originating organisations.

There is a lot of discussion of "native to the Web vocabularies", i.e., to vocabularies that are published, conform to Linked Data standards and are open-sourced. Some would say they are being given away.

In 2009, the New York Times (NYT) announced its intent to break with 100 years of tradition and publish its thesaurus — more than one million terms in five vocabularies that are used to tag NYT articles — under a Creative Commons BY license that lets others use it and contribute to it. As of mid-January 2010, the first 10,000 subject headings already had been published. The NYT plans first to release tags back to 1980 and later back to 1851.

Open Linked vs News Linked Data

It is one thing to open access via linked data to information that was paid for by tax payers. News may be another matter.

In January, The Guardian, the BBC, and the Media Standards Trust sponsored a one-day "News Linked Data Summit" in London and invited everyone in the news industry interested in discussing the potential of Linked Data for news organisations to attend.

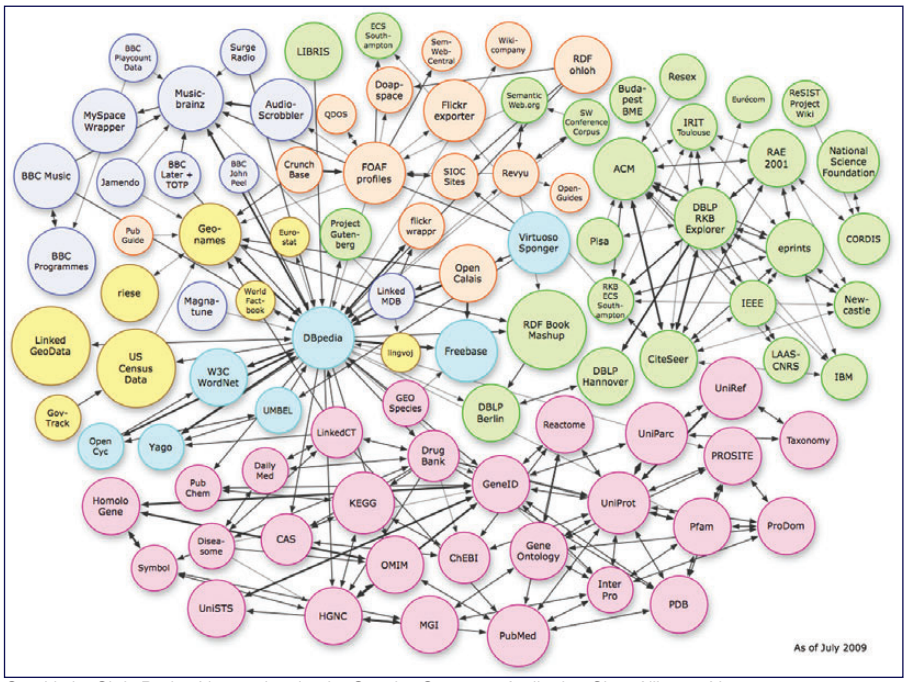
Martin Moore of the Media Standards Trust noted that the meeting was not so much a forum for decision-making, as for exploration.

"We did talk about a bunch of potential Linked Data experiments", said Moore, "that would enable people to stick a toe in the water and get a better sense of how much work Linked Data involves and what its value might be. Most people there wanted to get involved in some way", he added. The differences lay in how much or how little involvement participants were prepared to give.

What topics were discussed? According to Moore, "Basics were discussed, like 'What does Linked Data mean?' and 'How might it have value?' There were technical questions about how you link together. If you link to one dataset, then you're linked to masses of them. Where would the junction points be, the canonical data points? There will be masses of such points. There needs to be agreement about how we will find them".

As an example, he said, "Let's say you have a news event like the Haiti earthquake. Once it's defined somewhere, like in Wikipedia, then that becomes a junction point about that event."

There are other issues than the strictly technical or structural, however. "For an event like the World Cup, for example", Moore said, "many organisations already hold rights to different aspects of the event, so data linking to that information could run into li-



Graphic by Chris Bezier. Licensed under the Creative Commons Attribution-ShareAlike 3.0 License.

This diagram, updated regularly, represents instance linkages within the Linking Open Data Community Project's datasets. The diagram grows more complex as new datasets are added.

The Linked Data Principles

1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI, provide useful RDF information.

4. Include RDF statements that link to other URIs so that they can discover related things.

Tim Berners-Lee, 2006, on the underlying rules for structuring linked data for the Semantic Web.

*Continued, "**News Linked Data**" page 3*

***News Linked Data**, from page 2*

censing difficulties." Resolving such legal issues adds an additional layer of complexity to making Linked Data work, especially for news organisations.

In other words, the legal, business and concept roadblocks that must be overcome regarding Linked Data are considerable and may be more difficult to resolve than technical issues.

Putting it to the test

Although it was not an objective, a consensus grew out of the meeting that the national elections in the UK, which will take place in May or June of this year, could be a good test of the value of news linked data and offer technical implementation experience as well.

Whether news media in the UK end up using this event as a Linked Data feasibility test is not known or is not being shared. Both the January News Linked Data Summit and a late February meeting that was more technically focused were held under "Cheltenham Rule". This means everyone agreed not to talk about who attended or who voiced which opinion. The confidentiality rule would ensure that issues could be discussed and brainstormed without fear that opinions expressed would find their way to competitors or unsympathetic executives. The objective was to secure the free flow of ideas.

Martin Belam, The Guardian's information architect, is a respected blogger on news and technology issues. He organized, spoke at and participated in several linked data meet-ups in the past five months. In his estimation, even though UK election coverage plans might not include experiments with news linked data, "It has been really positive so far to get people from across varying competing news organisations into the same room to talk about solving some common problems that we share. I'm cautiously optimistic that we'll see further developments in using Linked Data for news this year."

The business case

Answering the question "Can Linked Data work?" is just the beginning. "Is there a business case for it?" is the rest of the question. Some are taking a wait-and-see approach; others are, in Moore's words, "sticking a toe in the water" with projects that vary in scope.

While in a different form, Thomson-Reuters embraces Linked Data via its Open Calais product. The goal of Open Calais, as noted on the company's Web site, is "to make all the world's content more accessible, interoperable and valuable".

Dave Compton with Reuters in the UK said, "Tim Berners-Lee has previously stated that 'The Semantic Web isn't just about putting data on the Web. It is about making links, so that a person or machine can explore the web of data. With Linked Data, when you have some of it, you can find other, related data'. Applying this thinking, news providers can significantly improve the content discovery process by focussing on providing qualified links between the data they publish. This interconnected data web creates

*Continued, "**News Linked Data**" page 4*

LINKED DATA GLOSSARY

DBpedia — a project at two German universities to pull structured information from information created as part of the Wikipedia project and make it available using standardised formats.
Dereferenceable URI — a mechanism defining how to use a URI for retrieving a copy or representation of the resource it identifies. Different Internet protocols may be used; the most widely used is HTTP.
FOAF — acronym for "friend of a friend"; a machine-readable ontology describing persons, their activities and their relation to other people and objects.
hNews — a microformat for adding metadata to news content in a Web document; includes fields that describe journalistic principles upheld by the journalist.
LCSH — Library of Congress Subject Headings (US); government project to define tags to be used with public information; they are made available by standardised formats compatible with the Semantic Web.
Linked Data — a method of identifying and establishing formal relationships between data which exist in (Web) documents using dereferenceable URIs; a W3C project that is a subset of the Semantic Web.
News Linked Data — material from news media archives, prepared according to Linked Data principles and distributed on the Web.
Open Linked Data — information that is publicly available, prepared according to Linked Data principles and distributed on the Web.
OWL — Web Ontology Language; a RDF-based technology which is meant to be used when the knowledge represented by the content of a (Web) document, i.e. persons, locations, topics, dates etc., needs to be formally notated.
RDF – Resource Description Framework; a standard model for the formal notation of assertions about resources; a.k.a., a formal notation of metadata.
RDFa — a standard that adds a set of attribute level extensions to XHTML for embedding rich metadata within Web documents
Semantic Web — the extension of the World Wide Web that enables the linking of data contained in Web pages.
SEO — Search Engine Optimization; the process of improving Web site ranking in search engine results.
SPARQL — a query language for RDF that is considered a key technology of the Semantic Web.
URI — Uniform Resource Identifier: a string of characters used to identify a resource. URIs are split into two groups: URNs as abstract names and URLs as locators of resources.

***News Linked Data**, from page 3*

a much richer content resource for both internal editorial usage and for the consumer. If content is well linked, it will be found more easily. The power of being able to discover and join data sets provides new and exciting possibilities for monetising derived data".

A more restrained approach

The Associated Press, which was represented at the News Linked Data Summit, is still in an investigative mode. According to AP's Stuart Myles, "AP is interested in Linked Data, particularly in learning how it can benefit news organisations (both as publishers and consumers of Linked Data), and our members and clients. We don't have any firm plans in this area", he said. "We have spoken to other news organisations informally about the possibilities (both at the Summit and elsewhere).

"[Those attending] the Summit decided to focus on a very UK-centric initial project — the forthcoming UK general election — so we have agreed to observe, but not take an active role in that initial pilot, since we can't offer as much specific expertise in that area," Myles said. "We hope to take a more active role in future efforts."

Does IPTC have a role to play?

"The technical side of Linked Data is quite mature", said Myles. "There are several tools, both commercial and open source, for creating, storing, publishing and consuming linked data. There is quite a lot of work about how to convert existing information into Linked Data." In Myles' opinion, IPTC might have a valuable role to play in creating controlled vocabularies.

"IPTC could publish its controlled vocabularies using linked data technologies", said Myles. "It could also map between the different vocabularies of others".

Getting numerous controlled vocabularies to work together is an important aspect of Linked Data within the Semantic Web. IPTC's role could be central to doing it right, said Myles. IPTC has vast experience to offer.

Resources & Readings

Linked Data: Connected Distributed Data Across the Web

Media Standard Trust: Martin Moore

Martin Belam, blog/The Guardian: News Linked Data Summit

RDF: Resource Description Framework

Semantic Web Community

Thomson-Reuters: Open Calais

URI: Uniform Resource Identifier

U.S. Library of Congress: LCSH

Extraordinary General Meeting in Paris

An extraordinary General Meeting of the IPTC will take place 10 March 2010 in Paris at the Millenium Hotel Paris Opera, 12 Boulevard Haussman, Paris 75009 at 1500 hrs., local time. A membership vote on changes to the IPTC Articles of Association is required by law.

Reminder: IPTC standards updated

The IPTC released updates of a few of its standards in December:

- NewsML-G2, the standard for exchanging multimedia news and packages thereof, was updated to developer version 2.4, which provides only updated schemas.
- EventsML-G2, the standard for exchanging event and coverage planning data, was updated to developer version 1.3, which provides only updated schemas.
- NITF, the standard for article markup, was updated to version 3.5
- SportsML-G2, the standard for sports data, was already updated to version 2.1 in November.

Publisher's Statement

The *IPTC Mirror* is published five times per year by the International Press Telecommunications Council (IPTC). The IPTC, based in London, U.K., is a consortium of the world's major news agencies, news publishers and news industry vendors. Founded in 1965, the IPTC develops and maintains the technical standards for improved news exchange that are used by virtually every major news organization in the world. Membership is open to organisations and companies concerned with news collection, distribution and publishing, as well as to system vendors supporting the news industry. The IPTC keeps the industry apprised of issues and developments through the IPTC Mirror and the IPTC Web site: www.iptc.org.

IPTC Managing Director: Michael W. Steidl (mdirector@iptc.org) ▪ Editor: Sue Sherrard Fine (editor@iptc.org) ▪ 20 Garrick Street, London WC2E 9BT, United Kingdom ▪ Tel: +44 (20) 3178 4922 ▪ Fax: +44 (20) 7664 7878.

Graphic representations of the Internet that appear on pages 1-2 and, as icons, elsewhere were created by Matt Britt and Chris Bezier, published at Wikimedia.org and are used under Creative Commons licenses.

The IPTC is registered in England as "Comité International des Télécommunications de Presse" at 10 Portland Business Centre, Datchet, Slough, Berks, SL3 9EG, Registration No. 1010968.

Photo courtesy European Broadcasting Union. Used with permission.

The EBU has long been a leader in research and development for new media. Its commitment to identifying the best solutions for its members led to the decision to join IPTC and drove its participation in developing the NewsML-G2 specification.

European Broadcasting Union: an early adopter of the NewsML-G2 standard

By Jean-Pierre Evain and Benoît Sergent, European Broadcasting Union

The EBU (European Broadcasting Union), founded in 1950, is the world's largest professional association of national broadcasters. Its members reach a weekly audience of 650 million people in over 56 countries.

In a nutshell

The organization provides services to the broadcasting community at large, along with expertise — specific to members — on legal, technical and programming issues. It also conducts economic and market analyses and offers targeted training programmes.

One of the EBU's main activities is exchanging content. The EBU coordinates daily transfers of programmes, music, sports events and news between members and other media operators, and transmits this content via the Eurovision network (for video) since 1954, and Euroradio network (for audio) since 1989.

EBU Headquarters are in Geneva with offices in Beijing, Brussels, London, Madrid, Moscow, New York, Singapore, and Washington DC.

News and sports exchange

Major Eurovision activities include the coordination of news, sports and entertainment transmissions.

EBU News Exchange is a platform based in Geneva. The EBU News Exchange's desk manages content provided by Eurovision Members and by agencies (APTN, CBS and Reuters). This content, identified as "items", can be news packages or live events. Today, more than 40,000 news items are exchanged every year.

Contributors that wish to share material can exchange material over the Eurovision satellite network once that material has been accepted by Eurovision editorial staff, thereby allowing all recipients to receive the material in real time. News Producers handle the daily exchange of News and Sports items. They also provide information (metadata) about the items available for exchange.

*Continued, "**EBU + NewsML-G2**" page 6*

***EBU + NewsML-G2**, from page 5*

Technical innovation and standardisation

EBU TECHNICAL is at the forefront of research and development in new media and has helped develop many radio and TV systems, including Radio Data System (RDS), Digital Audio Broadcasting (DAB), TV-Anytime (TVA), Digital Video Broadcasting (DVB), and high-definition TV (HDTV).

It is in this framework of identifying and selecting best solutions for its members that EBU TECHNICAL (the EBU Technical department) decided to join the IPTC and actively participate in defining the NewsML-G2 standard.

Eurovision's NewsML-G2 implementation

At the beginning, contribution and distribution of material used to be simultaneous over satellite links.

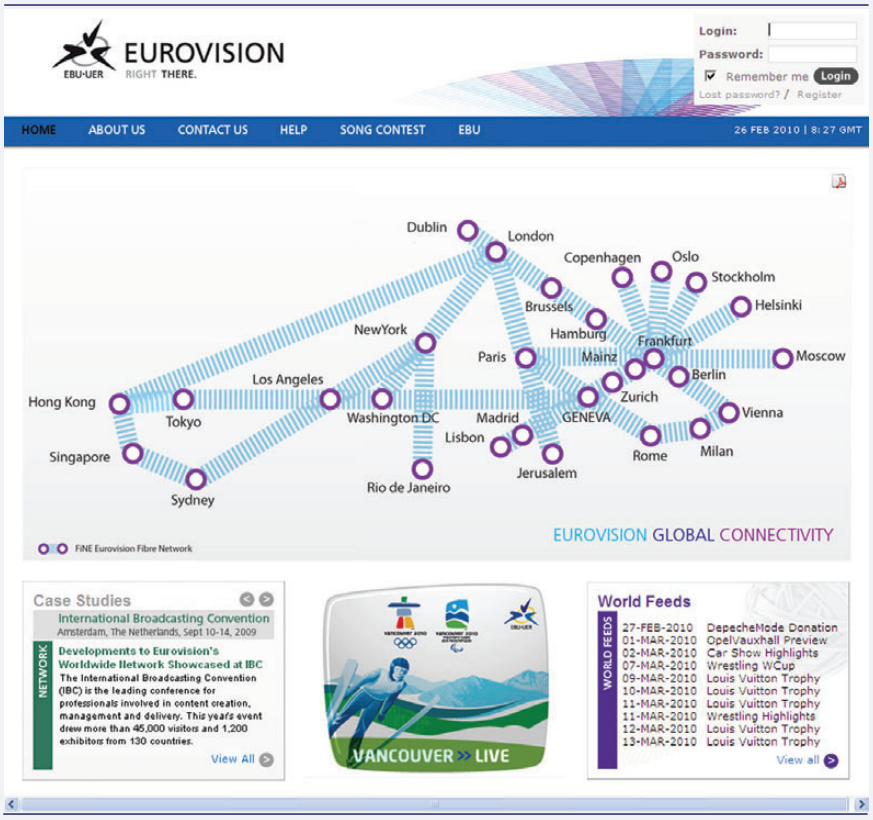
This was followed after several years by the distribution of MPEG stream files to remote recorder stations named SuperPOP. Finally, file-based contribution has been implemented; this soon will be followed by full file exchange (uplink and downlink). The NewsML-G2 standard plays a key role in this evolution of the distribution network and newsroom management system.

EBU's NewsML-G2 implementation was deployed on 1 January 2009. Metadata files are distributed in this format over EBU's private satellite network. NewsML-G2 replaces the former EBU format called "NMS XML". As a distinctive additional feature, NewsML-G2 allows us to provide references to keyframes, compressed video and video sources.

EBU News Exchange's MAM or Newsroom Management System (NMS) is based on a relational database, which is the heart of the system. Newsroom users, such as news editors, can access the database and amend the description of an item, such as references to news or sport content (shotlist), and its associated metadata (e.g. dopesheet), as they obtain additional information about the event. The system tracks changes, which are then forwarded to all of EBU's external platforms composed of Web sites (XTRANET, the Eurovision Web site — www.eurovision.net) and members' stations named POP and SuperPOP. This applies to metadata (NMS XML in a transitory period and NewsML- G2), keyframes, and low resolution video or high resolution video.

Mapping EBU NMS attributes to NewsML-G2 proved to be easy after a limited number of iterations. The generation of NewsML-G2 metadata descriptions is made from the batch processing and transformation of data from the NMS Database. Eurovision uses its own QCodes (e.g. for roles, station names). As mentioned above, an essential advantage is the ability to link data to metadata description using partMeta, which former EBU NMS XML did not allow. EBU Technical has made key contributions for the development of this part of the specification.

Collaborating with IPTC for the development of NewsML-G2 proved to be a good choice. EBU encourages other news providers to make the effort to learn and use NewsML-G2.



EBU's Eurovision and Euroradio networks daily exchange news, sports and music programmes.

Graphic courtesy European Broadcasting Union. Used with permission.

Jean-Pierre Evain, is a Project Manager with the European Broadcasting Union and works in EBU Technical.

Benoît Sergent, Ph.D., is a Systems Architect for the European Broadcasting Union and works in IT Services.