Information Services & Use 35 (2015) 175–179
DOI 10.3233/ISU-150779
IOS Press


# Putting data at the heart of the Open Web

Phil Archer
*World Wide Web Consortium (W3C)*

The World Wide Web Consortium (W3C) was founded in 1994 at MIT by Tim Berners-Lee with support from the European Commission and DARPA (Defense Advanced Research Projects Agency, which pioneered the Internet). The mission of W3C is achieving agreement among key industry players to ensure compatibility among the elements of the infrastructure that are essential to developing the World Wide Web. W3C does not write any standard – rather, its members use the process, infrastructure and intellectual property regime offered by W3C to develop documents that represent the consensus of the relevant stakeholders. The success of the Web derives in part from the fact that only two standards MUST be followed by everyone: URIs as identifiers, and HTTP(s) as the transport protocol. Everything else is optional, which is why you can access HTML pages, images, videos, PDFs, and proprietary document formats over the Web.

W3C standards are developed by working groups comprising representatives of its member organizations. Almost all work is conducted in public with the majority of working groups using a publicly accessible GitHub repository as a communal workspace. During development, standards go through a number of steps as follows:

(1) Working Draft – documents for review by the community (i.e. this is what we are working on, what do you think?);
(2) Candidate Recommendation with standards set for review and testing of routes to implementation (i.e. prove that it works!);
(3) Proposed Recommendation for final approval by the W3C advisory council (i.e. yes, it works, the work is done, subject to the wider membership's approval);
(4) W3C Recommendation.

## 1. Origins of the Open Web

The Web is not just a way of putting a document online. It is not just about making datasets available. It is a web of connected things, for example a connection between a person, their work, and the data that backs it up. Consider Fig. 1, part of the original proposal for 'Mesh' at CERN that became the World Wide Web.

From its origins, the Web was meant to connect things. The way it has evolved since 1989 has made it a massive interactive platform, that is also an open platform. The Web is highly functional, interactive, and has very advanced graphics capabilities. Ask a young person to read a traditional book. I tell you that a likely response is: "It is broken, you cannot click it, zoom it, and do interesting things with it. A book is dead material, it is not interactive."

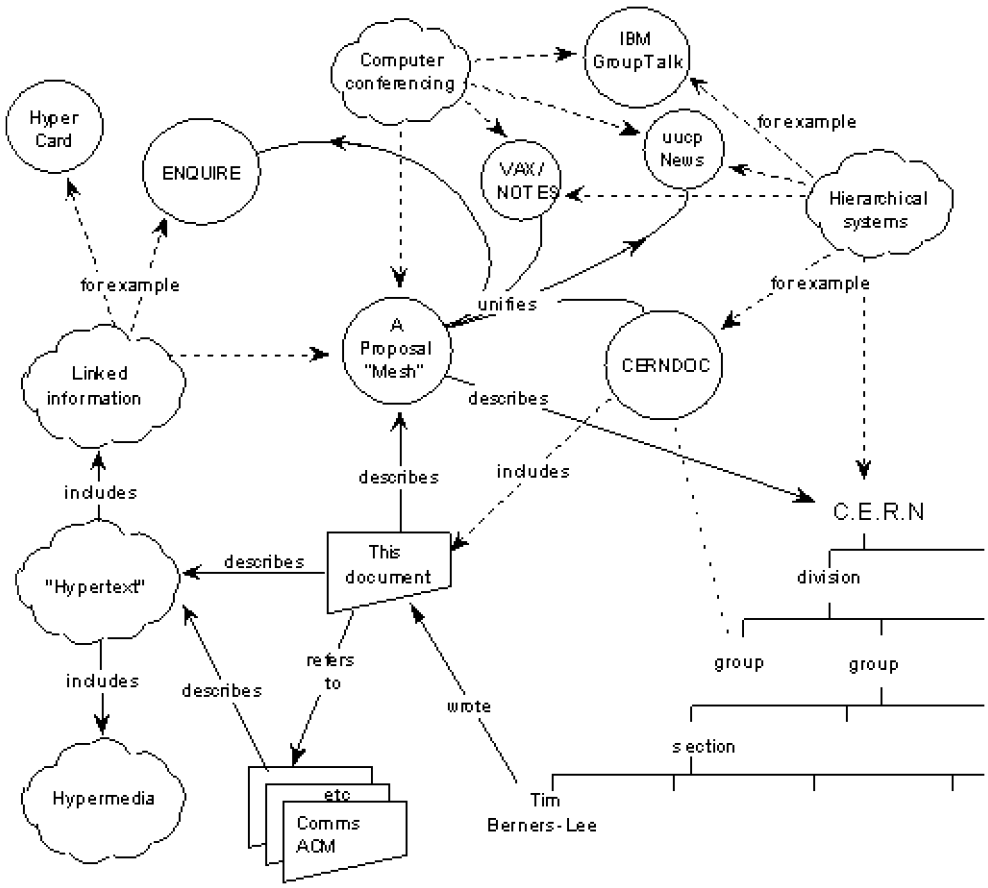


Fig. 1. CERN Mesh: the World Wide Web.

On the Web you can do many interesting things. We call them applications. The Web has become a powerful application platform that essentially is free to use. The basis of this freedom, as designed at CERN originally, is the link.

## 2. Basic agreements needed for further development

In order to use the Web efficiently, and to its full potential as a global platform, there a few simple things we need to agree on.

(1) We must develop a more explicit semantics of identifiers: We must be very clear what identifiers describe, otherwise machines cannot relate objects, i.e. make the link. For example, if you dereference an ORCID number (the 'identifier'), it is very easy to assume that the same number means many things, i.e. a profile document, an online account, a person, a list of publications, and so on. ORCID is a great tool to disambiguate people (researchers), and we are working with ORCID to fix it. The issue is that ORCIDs largely assume that a person is reading, but that is not good enough, for the service to be efficient and scalable, the identifier(s) need to be machine-

readable. We need variations on an ORCID number so that the machines can connect the person to the documents, to the account and so on.[1]

(2) We need a better understanding of core concepts like dataset and distribution, i.e. we must acknowledge the importance of metadata. Platforms for academic information typically have little or bad metadata. The best practice I could find of useful and well-crafted metadata is CIARD RING, a web-based service for accessing sources of agricultural research and development.[2] More importantly, none of the platforms I surveyed have common metadata standards when describing their datasets. Without standards, interoperability cannot be achieved. To make progress, W3C recommends DCAT, the Data Catalog Vocabulary.[3] Another way of understanding the importance of metadata is to consider the visibility of your offerings, i.e. marketing. All the major search engines collaborate on schema.org to promote structured data. If you are not using schema.org vocabulary, it will be so much harder for users to find your offerings.

(3) We must set standards and achieve best practice models: RFC 2119 is the most cited standard, defining what 'Must', 'Should', and 'May' mean in standards.[4] It is an example of an 'Internet Best Current Practice', and we need more such standards that we all abide by.

In sum, I strongly encourage you to move along this path towards more standardization – that is, consensus among the relevant stakeholders – so that we get more benefits from the Web as an open platform. For publishers, it adds value to the papers they sell; for researchers, it increases the likelihood of citation; for funders it allows you to access more easily the supported work; for humanity it makes knowledge available, and connections visible, on the greatest data infrastructure platform yet built.

## 3. Efforts at W3C

Among the activities of the World Wide Web Consortium, the following activities are particularly relevant to academic information and scholarly publishing:

- The Share-PSI 2.0 Thematic Network[5] focuses on the European Public Sector Information Directive (PSI) to transform public data into sharable open data. The difficulties encountered with PSI are the same as with academic information. There is a large potential for adoption and implementation.
- Data on the Web Best Practice Working Group[6] to achieve interoperability between open data, data under less permissive license terms, and enterprise data.
- Standardized metadata for CSV (tabular data) on the Web.[7] CSV is a prevalent data format and we are developing a standardized format for identifying column and row headings, and data types etc.

[1] See http://www.w3.org/2015/03/orcid-semantics.

[2] http://www.ciard.net – The CIARD Routemap to Information Nodes and Gateways (RING) is a global directory of web-based services that will give access to any kind of information sources pertaining to agricultural research for development (ARD). The RING is the principal tool created through the CIARD movement to allow information providers to register their services in various categories and so facilitate the discovery of sources of agriculture-related information across the world. The RING aims to provide an infrastructure to improve the accessibility of the outputs of agricultural research and of information relevant to ARD management.

[3] http://www.w3.org/TR/vocab-dcat/.

[4] https://www.ietf.org/rfc/rfc2119.txt.

[5] http://www.w3.org/2013/share-psi/.

[6] http://www.w3.org/2013/dwbp/.

[7] http://www.w3.org/2013/csvw/.

This facilitates automatic ingestion and transformation, e.g. to JSON (JavaScript Object Notation) or RDF (Resource Description Framework).

– Digital Publishing Activity[8] builds bridges between the Open Web Platform and the publishing industry, e.g. by extending Web browser functionality to work more readily with electronic books and other packaged publications, and by encouraging greater use of the Web in those publications.

## 4. Text and data mining, and natural language processing

Researchers want their publications and data to be found and cited. There is a boundary where the researchers responsibility for (future) discoverability ends and the publisher takes over. Consider, in the abstract, the following kind of metadata:

> "An investigation in the effect of X in relation to *disease* Y"
> Keywords: X, *disease* Y, Z

Someone searching for keyword X and/or disease Y or Z may or may not find the article and dataset. 'May not' because this kind of metadata really is insufficient. Consider Fig. 2.

By text mining you can pull out (some) simple relationships. That and how the article and dataset describe the relationship between X, Y, and Z should be part of the metadata. It is the relationship that makes the dataset and article interesting. It is why people want to discover and look at it. Hence the metadata should look like this:

> "An investigation in the effect of X in relation to *disease* Y"
> Key finding: Exposure to environmental pollutant X upscales production of hormone Z, which is associated with an increase in prevalence of disease Y.
> Keywords: X, *disease* Y, Z

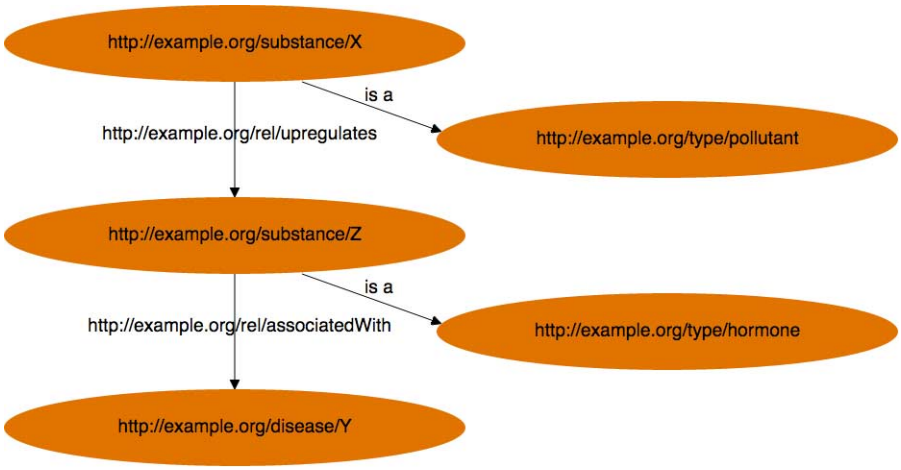


Fig. 2. Text mining for metadata. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-150779.)

[8] http://www.w3.org/dpub/.

Another way of understanding this idea is to see that it is a basic form of search engine optimization. Increasing metadata by a small amount that matters is not a hard thing. As noted above, metadata models should be aligned across the academic publishing industry.