Weaving webs for the Web:
Linking data and their vocabularies
for a machine-friendly Web

Fabien Gandon,
director of research, INRIA

Abstract:
The Web has more than three billion direct users. For several years now, it has been used not only by people but also by machines. How have a Data Web and a Semantic Web been woven to formalize descriptions and enable machines worldwide to exchange structured data in all fields and to share rules and vocabularies that make these data meaningful, useable and reusable?

I speak any language. I have more than three billion direct users. I might be private or public. I am spreading onto all networks around the world. I am lurking underneath your reservations for vacations, your telephone, your e-books, your exchanges with your insurance company or… I am… I do… I have… the Web, the Net (or *ouaibe* for our cousins on the other shore of the Atlantic). We have always looked to history for an efficient means to collect and access the masses of information we are creating. This is the main reason for libraries. It was also the motivation of Tim Berners-Lee when, in 1989, he proposed setting up a global hypertext system at CERN in order to better share information on a campus with thousands of persons distributed among many specialties and equipped with various devices (BERNERS-LEE 1990).[1]

It is still noteworthy that the Web is both very familiar and poorly known, evidence of this being the much too persistent confusion between uses of the words "Web" and "Internet". Even though their inventors received two different Turing Awards, respectively in 2004 and 2016 for two quite distinct inventions, "Internet" and "Web" are still too often used interchangeably. It is worth repeating: the Internet allows for interconnections between networks of computers and connected devices in general. It is the infrastructure for communications where, at a higher level, several applications can be run for e-mail, telephones, videos, etc. In contrast, the Web is a distributed hypermedia that has become the majority software architecture for applications on the Internet. Less well known about the Web is that, since the end of the 1990s, it is no longer consulted and used by us, human beings, alone but also by machines, in particular via the Data Web and Semantic Web.

[1] This article has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references. All websites were consulted in December 2020.

A hypermedia network of resources

As of 1996, Tim Berners-Lee reviewed the architecture of the Web while insisting on three core concepts: addresses (Web addresses, URLs and URIs formatted as http://…), the data transfer protocol (HTTP), and the procedures for negotiating data types (contents). The latter is part of the HTTP protocol that enables a Web server to provide the same URL (uniform resource locator) with different representations ("*in terms of language and data format*") of a single resource, depending on what is known about the party interrogating it. If, for example, the server knows the languages spoken by the party accessing the address, it will be able to respond in one of those languages. The data type is negotiated every time we access the Web without our even knowing about it.

The possibilities arising out of the negotiation of data types extend farther and, in a way, downgrade the importance of "web page" since one of these possibilities is to negotiate with a Web server the types of formats for its response. So, the Web is not just a web of documents. Indeed, it offers the possibility of serving and linking any- and everything. Since the URI (uniform resource identifier) can identify any type of resource (a page, image, person, product, molecule, etc.) and not just Web contents, the Web and its languages can be used to describe and link whatever can be identified in the world. The Web thus declares its independence from a model or data structure; and the HTML language of web pages rebecomes a mere prerequisite for using a browser (BERNERS-LEE *et al*. 1994). Initially, HTML made it possible to propose a uniform format of hypertextual documents and to "document" the network of resources that was becoming the Web. The Web was ready to exchange many other things than pages.

A machine-processible Web

In 1996, the PICS language was being used to standardize the filtering of inappropriate contents, in particular for children. PICS brought along, incidentally, the idea of labeling contents with data for machines. The Web thus adopted the concept of metadata, whence the trend toward a Web of documents and structured data.

As part of this trend, the language of cascading style sheets (CSS) represented a major step toward separating substance (or content) from form (or its presentation) on the Web. By separating a document's form from its structure, style sheets opened the possibility of using the same form for several documents, or, on the contrary, of varying the form of a single document. Soon afterwards, the Web underwent further change owing to the XML standard for creating and managing its own structures for documents and data. In line with all this and with his articles in 1994 (BERNERS-LEE 1994, BERNERS-LEE *et al*. 1994), Tim Berners-Lee published in 1998 a road map for what he called the "Semantic Web". He foresaw a change in Web objects (most of which were, at the time, documents for human beings) toward resources with a semantics oriented toward machines: "*the Semantic Web approach instead develops languages for expressing information in a machine processable form*". This 1998 roadmap opened the way for work on the Data and Semantic webs and their standards (RDF, RDFS, SPARQL, OWL, etc.).

The Semantic Web: When links make sense

Although the concept of a Semantic Web arose in 1998, the best known article for the public was published three years later (BERNERS-LEE *et al*. 2001). Now, twenty years afterwards, we can explain what has happened in two major stages; the idea of "linked data" and the Data Web, and the idea of linked schemata and the Semantic Web.

Linked data and the Data Web

The first principle is to create links between data just as we create links between pages and, by extension, to create links between databases just as we create links between websites.

The first stage involved putting the Web's standardized identifiers to use for identifying the subject and relations between data. For example, I can make a Web address (URI) to identify a sedan in my company's parking lot (*e.g*. http://www.my-company.fr/car/sedan-n3). It is clear that this car cannot be accessed via the Web at this address but that this address can represent it in the data describing it. Since this same identifier can be reused in several sources of data, it can be used for links between these data and between their sources. This is what is called "linked data". Likewise, I can make a Web address (URI) to identify a relation between a person and a car, (*e.g.*, http://www.my-company.fr/car/isDriverof). Once again, this identifier can represent all the occurrences of the relation "driver" between a person and the car. Any data set may reuse it, as needs be; and this enhances the interoperability between the applications that produce or use data.

In these two examples, the URIs use the HTTP protocol, *i.e*., they begin with *http://*. This is called an HTTP URI. They are not called URLs because, unlike an address (*e.g.*, http://fabien.info), they do not correspond to a resource (page) available on the Web. Nonetheless, using the HTTP protocol provides a default method for discovering what these HTTP URIs identify: we go to the address they indicate to obtain the data on them. This is called "redirection" or "forwarding".

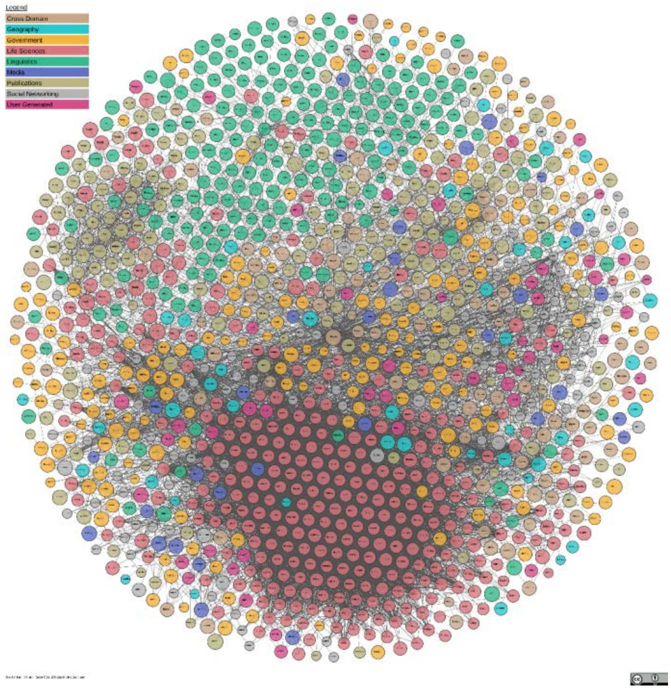
The Data Web thus puts in relation sources of data of variable sizes while relying on, and extending, the Web's conventional architecture. Since not just pages are being linked but arbitrary resource identifiers, the question crops up: what is obtained when we use these identifiers? When an identifier is consulted, the server responds with data that describe the resource — a car, animal species, protein or author not necessarily on a web page — and adjusts to the query about the data. For a single identifier and thanks to the procedure of data-type negotiation, the software making the query will receive data for its own database whereas users with browsers will receive web pages in HTML to read.

The linked open data rating system proposed up to five stars, corresponding to improvements in the quality of the publication of open data:

★ The data are on the Web under an open license.
★★ ditto + The data are structured.
★★★ ditto + The data are in a nonproprietary format.
★★★★ ditto + HTTP URIs denote subjects, objects and types of relations.
★★★★★ ditto + The data are linked to other data.

The evolution of the Web from documents and pages toward the Data Web relies on these principles and standards, which enable anything to be identified and described on the Web and weaves a worldwide web of data. By applying these principles to what used to be a Web of web pages mostly for human consumption, a "Data Web" was added as of 2006-2007. It links databases of any size and about any subjects, mainly for automatic processing by machines. Machines thus search across this Web, following links in order to find new sources. This Data Web can be browsed and searched for data just as we browse and search through web pages. Its very name insists on opening our bunkers of data, regardless of the size, ranging from our agendas to vast geographical databases, and on exchanging, linking and combining them as a function of our needs.

Figure 1**: The Linked Open Data Cloud of part of the open databases linked on the Web (29 March 2019).**
Source: https://lod-cloud.net/.



To grasp the impact of this change and the volume of data becoming available, let us examine a few examples. One of the circles in Figure 1 represents DBpedia, which publishes on the Web the linked (RDF) data that it extracts from Wikipedia's encyclopedia. At the time of the writing of this article, this circle alone in the depicted cloud of open databases represented 38 million "subjects" described with 3 billion elementary data (attributes and relations of these subjects) taken from the 125 different languages used on Wikipedia and made available as open structured data. Another circle represents Wikidata, which serves to directly enter and publish structured data for, at present, the 75 million subjects described by 23,000 users. In specialized fields like biology, one of the circles represents Uniprot, which provides 179 million elementary data about proteins. This cloud does not depict many other sources of data. For instance, the major search engines behind Schema.org enable us to place structured data in web pages, something millions of websites do in order to be more accurately indexed. Another example: all the sites with a "Like" button from Facebook include in their pages descriptive structured data for this button when it is clicked. The volume of structured data on the Web has exploded in the past fifteen years without ordinary users necessarily realizing this.

To represent and exchange these data on the scale of the Web, it is necessary to set standards for the models, structures, formats and languages to be used. RDF (resource description framework) is to the Data Web what HTML is to the Web of pages. The RDF language serves to represent and link data about resources. It is incorporated in the Web's architecture, in particular through the use of URIs to identify the resources and relations described by its data graphs.

Coincidentally, RDF furnishes a model of data that serves as the basis for other standards. Above RDF, SPARQL is both a language for modifying RDF graphs and a protocol for submitting queries to a distant server. For example, on the site DBpedia (one of the bases of the cloud in Figure 1), a query can be made in SPARQL for all the URIs of resources named "Paris" in French. Once the URIs are received, a new query can be made to obtain additional data, thus passing from linked data to linked data just as we pass from web page to page. SHACL (Shapes Constraint Language), another standard, validates graph-based data by detecting the rules with which the structure of RDF graphs has to comply in order to be validated (*e.g.*, All books have to have a title). This validation is useful for verifying the data exchanged between applications.

Descriptions like this might come from any source on the Web and be merged with others. The phrase “global giant graph” is sometimes used to refer to this worldwide Data Web woven by thousands of descriptions that, distributed over the Web, declare links between nodes identified by URIs.

Linked schemata and the Semantic Web

In this second stage, ever more schemata of data are being published on the Web, *i.e.*, the lexis and rules that govern the values, structures, uses, interpretations of these data — in brief, their sense or semantics. These schemata and their terms also use Web identifiers (*e.g.*, a URI identifying the category “woman”) and links to declare relations between the concepts they define (*e.g.*, A woman is a person), thus giving them a meaning and weaving the Semantic Web.

So, the Semantic Web allows for the formalization, publication and linking of the vocabularies used in RDF descriptions. These vocabularies allow for applications to more efficiently use data from the Web by recognizing the various types of resources and links encountered and by tapping the meaning and reasons attached to them. An application can thus tell the difference between resources named “Charles de Gaulle” that are of different types: a man, street, place, airport, aircraft carrier, etc.

Various types of models have been designed to provide the vocabularies for describing our world on the Web, whence the talk about “ontologies” and “thesauri”. By querying these models and reasoning with them, it is possible to improve existing features and propose new ones. Above the RDF layer is a stack of schema languages, with increasing meaning and computing costs. The higher in the stack, the better the logical definitions of the vocabulary for describing the structure and meaning of data — but also the more costly in terms of complexity and computing time the procedures thus made possible. The first “light schema” layer is RDF Schema, which allows for declaring and naming classes of resources (such as books, films, persons) and their properties (*e.g.*, author, actor, title) and for organizing these types in hierarchies. These schemata are called light ontologies. Above RDFS is the recommended OWL (Ontology Web Language) for formally representing the definitions of heavier ontologies. It is organized in several fragments of a more or less extensive expressiveness that allow for making more logical deductions but that cost more in computing time.

In line with the Data Web, this Semantic Web emphasizes the possibility of exchanging the schemata of our data and the related semantics. Formalized and published as standards, these models enrich the range of automatic processing that can be done on data. By opening data and their models, the Data Web and the Semantic Web open all uses that can be made of them.

A literally infinite project

The Data and Semantic webs have already been rolled out; many applications have adopted them. Nevertheless, much work in R&D is still being done on several questions, among them: how to improve scalability, how to more efficiently process data, or how to more robustly handle the variety, quality or uncertainty related to data. For the Web, this is one direction among several others.

In the 1990s, Tim Berners-Lee's problem was to imagine a world with the Web before the Web had taken shape.[2] But we are now in the opposite situation: people are forgetting, or can no longer imagine, what the world would be like without the Web (SAVAGE 2017). However it is still very important to defend the Web, its openness and expansion. The Web is universally useful and used, but it is still vulnerable. Its initial ideal might be bygone if we do not constantly keep guard to preserve it, in particular to fend off any form of recentralization (such as the centralization of databases by certain firms). We must always bear in mind that the Web is not a realization to be taken for granted; it is an endless, ongoing project.

References

BERNERS-LEE T. (1990) "Information management: A proposal", the original proposal for the software project at CERN that would become the World Wide Web, available at https://www.w3.org/History/1989/proposal.html.

BERNERS-LEE T., CAILLIAU R., LUOTONEN A., NIELSEN H.F. & SECRET A. (1994) "The World-Wide Web", *Communicatons of the ACM*, 37(8), pp. 76-82, available at https://dl.acm.org/doi/10.1145/179606.179671.

BERNERS-LEE T. (1994) "Plenary at WWW Geneva 94", first WWW Conference, available at https://www.w3.org/Talks/WWW94Tim/.

BERNERS-LEE T. (1996) "The World Wide Web: Past, present and future", August, available on https://www.w3.org/People/Berners-Lee/1996/ppf.html.

BERNERS-LEE T. (1998) "Semantic Web road map", available at https://www.w3.org/DesignIssues/Semantic.html.

BERNERS-LEE T., HENDLER J. & LASSILA O. (2001) "The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities", *Scientific American*, 284(5), pp. 28-37.

GANDON F. (2017) "Pour tout le monde: Tim Berners-Lee, lauréat du prix Turing 2016 pour avoir inventé… le Web", *Bulletin de la Société informatique de France*, 1024, available at https://hal.inria.fr/hal-0162368

GANDON F. (2018) "A survey of the first 20 years of research on the Semantic Web and linked data", *Revue des Sciences et Technologies de l'Information - Série ISI: Ingénierie des Systèmes d'information*, pp. 11-56, available via https://hal.inria.fr/hal-01935898.

SAVAGE N. (2017) "Weaving the Web", *Communications of the ACM*, 60(6), pp. 20-22, available at https://cacm.acm.org/magazines/2017/6/217732-weaving-the-web/fulltext.

[2] For further reading about this adventure, *cf.* Gandon (2017, 2018).