A map without a legend

The semantic web and knowledge evolution

Jérôme Euzenat
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
E-mail: Jerome.Euzenat@inria.fr

Abstract. The current state of the semantic web is focused on data. This is a worthwhile advance in web content processing and interoperability. However, this does only marginally contribute to knowledge improvement and evolution. Understanding the world, and interpreting data, requires knowledge. Not knowledge cast in stone for ever, but knowledge that can seamlessly evolve; not knowledge from one single authority, but diverse knowledge sources which stimulate confrontation and robustness; not consistent knowledge at web scale, but local theories that can be combined. We discuss two ways in which semantic web technologies can greatly contribute to the advancement of knowledge: semantic eScience and cultural knowledge evolution.

Keywords: semantic web, linked data, big data, open data, machine learning, web of data, reproducible research, eScience, cultural evolution, evolutionary epistemology

1. Life before the semantic web

Many animals are able to learn from their environment. This can be achieved through perceiving the environment and experimenting with it: acting and learning from the results of actions. Some of these animals learn from others. Most of them do it by mere imitations, such as birds learning songs [1]. However, it is known that some monkeys are able to signal danger and even that some other species can learn the signal from others [2].

The resulting behaviour shared among a population is called culture. Culture encompasses know-how, knowledge and beliefs, among other things, and does not have to be explicitly expressed.

Human beings are special among these species because they can express the knowledge that they learn and communicate it to other human beings. Bees can communicate how to reach food sources in an articulated language, but this stunning capability seems to be fixed. Direct knowledge communication is a very powerful mechanism because it leads to transmit knowledge without having to relearn it from experience. This capability grew slowly and led to teaching, monasteries, universities, and conferences.

Besides articulated expression, written expression and communication have allowed to get rid of time and space in knowledge transmission. Through tablets, rolls, books, libraries, or scholar journals it is possible to build on someone else knowledge without even meeting.

This led to a variety of ways to acquire knowledge: by observing and experimenting, by imitating, by communicating (talking, writing, reading). These key features have provided a selective advantage to the species. This has been more eloquently storytold [3].

The worldwide web facilitating cultural exchange is a culminating point in this story, so far. One initial achievement of the web has been to make knowledge more easily and readily accessible across the planet. Knowledge was expressed in a not-very different way than before: through natural language texts and pictures, later through movies, interpretable by human beings. Those who have experienced the transition of the world before and after the web can only be grateful. The ease of use of the web pushed knowledgeable people all over the world to provide worldwide access to their knowledge. In this respect, wikipedia in which millions of individuals care about articulating what they know for others, is just one but a wonderful and precious success.

Hence, a semantic web [4, 5], allowing machines to grasp this knowledge, is a tremendous idea.

2. The data advent: a little semantics went a long way... but may remain stuck here

The semantic web could be characterised by one of its early slogans: a web for machines. The web was already processed by machine, but its content was aimed at being interpreted by human beings. More precisely it meant that the *content* of the web would be more machine processable.

However, there may be various understanding of 'content'. It could be metadata: a set of attributes that help web search and classification. It could be precise data that is interpreted by specific programs, e.g. calendars or software dependencies. It could be relatively general knowledge representations of the content, such as causal relations or universal statements. Technologies developed by the W3C Semantic web activity covered these three aspects.

The semantic web did not immediately took off. This is not particularly surprising, but in our world in which everything has to change fast, this is a major shortcoming. In particular, though enthusiastic scholars have been eager to develop sophisticated representations or ontologies, these were instantiated with little data. Hence of little *practical* use.

It was obviously necessary to provide data to this semantic web because no one was interested by a bunch of theorem provers. Fortunately, by that time, the web was not short of data. Linked data [6], building on semantic web technologies, was welcome to benefit from these data and semantic web technologies. The linked data trend was spot on as it met other trends such as openness, big data, and later, data flow from the internet of things. Governments supported the process to connect data sources of administrations. Companies competed to acquire giant meshes of data on which they could run cleverly engineered pieces of software to learn what was there.

Moreover, data is the fuel for the knowledge mill. The availability of data, and computing power, unleashed the amazing capabilities of data analytics and machine learning from these data. This made some sense on top of raw data, that which can lead to impressive applications. Hence, it obviously contributes to make the web more intelligible to machines [7].

Nowadays, web users are not expected to provide knowledge, nor to access it. It seems that they are mere data provider, mostly through their actions, e.g. click, buy, like. These data are machine processable, but not open. They are kept secret, in silos, to the exclusive exploitation of a single organisation. They are processed by corporations which eventually learn knowledge from that data. But this knowledge, in turn, is not shared nor even prone to be communicated because not necessarily expressed in an articulated language. Instead, it is directly actioned. Hence, knowledge does not improve.

3. For knowledge!

After twenty years, the semantic web field is mostly focused on data, even when it is made of so-called knowledge graphs.

Of course, schemata and vocabularies are used. But, they are most of the time used for helping machines to parse data. Most of them simply express the structure and eventual relations between entities, but not what we know about them. Hence, the interpretation of data lies within the system after parsing. Knowledge is not made explicit nor shared; it is not available for scrutiny, reuse or evolution.

A semantic web without knowledge is for machines like a map without a legend for humans. The lack of a legend does not prevent to distinguish a road from a river, a bridge from a boat shuttle. But it does not help understanding how these can be used to travel. It does not enable the explorer to plan her trip, by reasoning on elevation or humidity.

Hence, this seems like a regression which brings us down the knowledge evolution ladder: humanity may eventually be fed with data, but any knowledge will have to be learnt again and again.

This is why the grand goal of formally expressing knowledge on the web must be rehabilitated. One ambition of the semantic web was to make knowledge available to machines, so that at least they can help us work with this knowledge, not only to find it out but to process it, to improve it and to communicate it to the world [7], as the next knowledge medium [8]. This could have led to a further knowledge ecosystem in which knowledge is elaborated while it is used for providing services. Like the web can be seen as a data washing machine [9], all the knowledge of the web, may be used as a knowledge washing machine. Reconsidering knowledge sharing and evolution at web scale would empower humans and machines alike with knowledge that can be both used to provide services and jointly refined and elaborated.

The goal is not to build the all-encompassing ontology on which everyone will agree with. Knowledge does not have to be centralised: diversity is source of disputation and robustness. Knowledge is not cast in stone for ever, but it has to seamlessly evolve. Knowledge need not be consistent at web scale, it can come in local theories that can be combined.

The semantic web initiative has already provided a good basis for expressing and sharing knowledge. First, universal and well-defined languages such as RDF, RDFS, OWL, or SPARQL have been designed. Then many specialised ontologies using these technologies have been developed from the most abstract, e.g. work, provenance, to the more concrete, e.g. proteins, scientific articles.

In the remainder, two different aspects are considered: how explicitly sharing formalised knowledge contributes to improve scientific practices, hence our shared knowledge; how knowledge, shared or not, may be seamlessly evolved and how this can be studied in an effective way. Of course, these directions are not without connection.

4. eScience: an example of web of knowledge

Let us take a typical example of collaborative knowledge elaboration, sharing and evolution: scientific research. eScience is a paramount example of knowledge expression and evolution.

Science has not been left aside of the professional computerisation: it has been very active from this standpoint. Use of statistical packages, plotting and editing notebooks are now commonplace and *in silico* simulations are accepted. This facilitates the production, analysis and dissemination of knowledge by and for humans helped by machines. But little is made for helping machines to have a grasp at the content.

Semantic eScience is a further step in that direction [10]. It exploits semantic web technologies to provide an interoperable and machine-interpretable infrastructure for scientific inquiry. Within the past 10 years, it has been a continuing source of attention.

IRIs provide a natural way to identify entities such as researchers (ORCID) or documents (DOI). Support has been provided to express bibliographic data in RDF [11]. This allows to express metadata about scientific literature.

On the content side, i.e. the objects of scientific statements, there are already many resources to express these with semantic web technologies. In life and health sciences, Gene ontology, OBO Foundry, bioportal, and Bio2RDF were already there 10 years ago and continuously improved [12]. On the mathematical side, efforts have been made to provide ways to offer a semantic encoding [13].

Beside bibliographic data, it is possible to deal with the methodological aspect of research. Some work have proposed the expression of hypotheses [14]. On the experimental side: researchobject provided a way to describe protocols [15]. Attempts are made to describe evaluation methods [16]. Finally, efforts have been made to address open science requirements to publish data sets. Google dataset search [17] offers search among data sets described with schema.org dataset and DCAT. Relations between such elements —which process produced which data, what hypothesis is it supposed to support, in which paper this has been published— can be kept track of through provenance assertions [18].

All this allows for expressing scientific knowledge on the web, not only the results of scientific enquiry but the whole process that led to establish such results.

Scientific knowledge expressed with semantic web technologies would clearly facilitate searching results [19]. However, more can be obtained from it with the help of machines. Knowledge expressed formally on the web could lead to formal scientific collaboratories [20]. For instance, data analysis workflows can be exploited for continuous reevaluation of hypotheses by updating data analyses when new data is available [21]. It may be checked, beforehand, that an experiment is prone to refute a claim. Papers whose results contradict one another could be identified. On the mathematical side, proof checking, now performed at the scale of one theory may be attempted at larger scales. Experiments testing a particular hypothesis, or a more specific one, can be found to avoid duplication. A protocol may be analysed to pinpoint what should be changed to affect the results of experiments. Literature can be exploited to predict averse effects [22]. More simply, machine learning could help cleaning and mining result as well as suggesting interesting tracks. These applications require knowledge in addition to data.

In addition to contributing to open research, semantic expression of research processes may help addressing reproducibility issues. Providing accurate descriptions of computer-based experiments can allow a computer to reproduce them. Tools may be developed to help (debugging and) peer-reviewing experiment preregistrations. It may also help to identify missing information in descriptions and so even facilitate off-line reproducibility

This is where full knowledge can play its role, by enabling machines to help us improving our knowledge by confronting it to existing data, by finding contradiction in other pieces of knowledge, by learning from data and knowledge. Machines could then join the course of knowledge evolution adding value to knowledge and massively confronting it.

5. Evolving knowledge

Knowledge that does not evolve, and systems based on it, are at risk of obsolescence. We need to ensure that knowledge representations evolve seamlessly and continuously.

Human knowledge evolves independently from scientific research, which is a relatively recent way to deal with knowledge. Contrary to science, knowledge evolution does not aim at elaborating knowledge for itself, but to evolve it through its use. However, both processes contribute to ‘improve’ knowledge and evolutionary interpretations of science are not new [23].

Natural selection can be thought of as a simple control mechanism based on variation, selection and transmission [24]. This can be implemented in computers as had already been done for genetic programming.

In the 20th century, anthropologists [25, 26] provided evidence of cultural evolution, habits transmitted by humans, so that it became an accepted discipline [27]. This was considered at a 'macro' level inspired from population biology: observing the evolution of knowledge of whole populations without figuring out the specific mechanisms implementing this cultural evolution.

From another standpoint, biologists attempted to generalise Darwinism, i.e. to provide an abstract description of evolution mechanisms [1]. They introduced the replicator-interactor pattern as a generalisation of the genotype-phenotype articulation: the replicator generates the interactor, which being in contact with the environment receives the selective pressure and induces differential reproduction. The replicator-interactor pattern is a 'micro' level view of evolution.

Finally, from the discipline of epistemology, some created along these lines an evolutionary epistemology which seeks to apply generalised Darwinism to knowledge elaboration [28]. In this perspective, evolution is described as a way to gain knowledge. Like genetics, this approach aims at providing the operational details.

In computer science, there are at least two non-exclusive directions inspired by natural selection: evolutionary computation and at its extreme genetic programming [29] which tries to be as close as possible to genetic principles, and experimental cultural evolution as applied to natural language [30].

Although both approaches follow different paths, they are meant to computer implementation and thus belong to the 'micro' approaches: they provide computational ways to evolve programs and culture. It is perfectly reasonable to apply such an approach to knowledge. Knowledge generates individual behaviour, which is subject of selective pressure from the environement and thus spreads differentially. Knowledge evolution can indeed be implemented as a mechanism which makes knowledge evolve seamlessly while it is used.

Experiments have been performed with knowledge-carrying agents applying minimal changes over their knowledge when it reveals inadequate. Like in cultural evolution, knowledge transmission does not necessarily happen through inheritance. On the contrary agents may transmit knowledge by cooperating or by directly exchanging it. Such minimal distributed social mechanisms for evolving and shaping knowledge are desirable features for social machines [31].

Experimental cultural evolution has been adapted to evolve abstract cultures [32], natural language features [30], and, closer to the semantic web, ontology alignments. In particular, it can be used to repair alignments better than blind logical repair [33], to create alignments based on entity descriptions [34], to learn alignments from dialogues framed in interaction protocols [35, 36], or to correct alignments until no error remains and to start with no alignment [37, 38]. Each study provides new insight and open directions.

There remain various challenges to provide general principles of cultural knowledge evolution. For instance, the formation and evolution of common knowledge, i.e. proper culture, must be studied. Another relevant issue is the co-evolution of knowledge learnt from the environment and knowledge acquired through communication.

6. Conclusions

The semantic web effort has so far provided impressive outcomes in terms of linked data that can be interpreted by machines as well as ontology languages and ontologies.

This is only half of the path towards a semantic web contributing to the knowledge of humanity. A solid basis is available that should be pushed further.

Human beings have progressively shaped their culture and knowledge through evolution. The current emphasis of the semantic web towards data sharing, undermines this enterprise. Knowledge learnt from data is not made explicit nor communicated. Hence, it cannot properly evolve, but has to be relearnt.

For the semantic web to take its full part in knowledge advancement, it has to be complemented by explicit knowledge expression and sharing. This would unleash the capability to properly evolve knowledge as illustrated by work on two directions. Scientific knowledge elaboration processes may be improved by expressing them semantically. Knowledge on the web may be evolved smoothly through evolutionary techniques.

References

[1] R. Dawkins, *The selfish gene*, Oxford University Press, Oxford (UK), 1976.

[2] M. Hauser, *The evolution of communication*, The MIT Press, Cambridge (MA US), 1997.

[3] Y.N. Harari, *Sapiens: a brief history of humankind*, Penguin Random House, London (UK), 2011.

[4] T. Berners-Lee, What the semantic web can represent, *Design issues*, W3C, 1998. https://www.w3.org/DesignIssues/RDFnot.html.

[5] T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web, *Scientific American* **284**(5) (2001), 34–43.

[6] T. Berners-Lee, Linked data, *Design issues*, W3C, 2006. https://www.w3.org/DesignIssues/LinkedData.html.

[7] K. Janowicz, F. van Harmelen, J. Hendler and P. Hitzler, Why the data train needs semantic rails, *AI Magazine* **36**(1) (2015), 5–14.

[8] M. Stefik, The next knowledge medium, *AI Magazine* **7**(1) (1986), 34–46.

[9] S. Auer and J. Lehmann, Creating knowledge out of interlinked data, *Semantic Web* **1**(1–2) (2010), 97–104.

[10] B. Brodaric and M. Gahegan, Ontology use for semantic e-Science, *Semantic Web* **1**(1–2) (2010), 149–153.

[11] P. Ciccarese, D. Shotton, S. Peroni and T. Clark, CiTO + SWAN: The web semantics of bibliographic records, citations, evidence and discourse relationships, *Semantic Web* **5**(4) (2014), 295–311.

[12] M. Dumontier, Building an effective Semantic Web for health care and the life sciences, *Semantic Web* **1**(1–2) (2010), 131–135.

[13] M. Kohlhase and F. Rabe, Semantics of OpenMath and MathML3, *Mathematics in Computer Science* **6**(3) (2012), 235–260.

[14] D. Garijo, Y. Gil and V. Ratnakar, The DISK Hypothesis Ontology: Capturing Hypothesis Evolution for Automated Discovery, in: *Proc. K-Cap Workshop on Capturing Scientific Knowledge, Austin (TX US)*, 2017.

[15] K. Hettne, H. Dharuri, J. Zhao, K. Wolstencroft, K. Belhajjame, S. Soiland-Reyes, E. Mina, M. Thompson, D. Cruickshank, L. Verdes Montenegro, J. Garrido, D. de Roure, O. Corcho, G. Klyne, R. van Schouwen, P. ‘t Hoen, S. Bechhofer, C. Goble and M. Roos, Structuring research methods and data with the research object model: genomics workflows as a case study, *Journal of biomedical semantics* **5**(1) (2014), 41.

[16] M. Stocker, M. Prinz, F. Rostami and T. Kempf, Towards research infrastructures that curate scientific information: A use case in life sciences, in: *Proc. 13th international conference on Data integration in the life sciences (DILS), Hannover (DE)*, Lecture Notes in Computer Science, 11371, 2018, pp. 61–74.

[17] N. Noy, Making it easier to discover datasets, *Google blog*, 2018. https://www.blog.google/products/search/making-it-easier-discover-datasets/.

[18] T. Lebo, S. Sahoo and D. McGuinness (eds.), PROV-O: The PROV Ontology, Recommendation, W3C, 2013. https://www.w3.org/TR/prov-o/.

[19] W. Hu, H. Qiu, J. Huang and M. Dumontier, BioSearch: a semantic search engine for Bio2RDF, *Database* **2017** (2017), 059.

[20] J. Euzenat, Building consensual knowledge bases: context and architecture, in: *Towards very large knowledge bases*, N. Mars, ed., IOS press, Amsterdam (NL), 1995, pp. 143–155.

[21] Y. Gil, D. Garijo, V. Ratnakar, R. Mayani, R. Adusumilli, H. Boyce, A. Srivastava and P. Mallick, Towards Continuous Scientific Data Analysis and Hypothesis Evolution, in: *Proc. 31st AAAI Conference on artificial intelligence, San Francisco (CA US)*, 2017, pp. 4406–4414.

[22] A. Noor, A. Assiri, S. Ayvaz, C. Clark and M. Dumontier, Drug-drug interaction discovery and demystification using Semantic Web technologies, *Journal of the American Medical Informatics Association* **24**(3) (2017), 556–564.

[23] D. Hull, *Science as a process: an evolutionary account of the social and conceptual development of science*, University of Chicago Press, Chicago (IL US), 1988.

[24] C. Darwin, *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, John Murray, London (UK), 6th edition: 1872, 1859.

[25] L.L. Cavalli-Sforza and M. Feldman, *Cultural transmission and evolution: a quantitative approach*, Princeton University Press, Princeton (NJ US), 1981.

[26] R. Boyd and P. Richerson, *Culture and the evolutionary process*, University of Chicago Press, Chicago (IL US), 1985.

[27] A. Mesoudi, A. Whiten and K. Laland, Towards a unified science of cultural evolution, *Behavioral and brain sciences* **29**(4) (2006), 329–383.

[28] H. Plotkin, *Darwin machines and the nature of knowledge*, Harvard university press, Cambridge (MA US), 1993.

[29] A. Eiben and J. Smith, From evolutionary computation to the evolution of things, *Nature* **521** (2015), 476–482.

[30] L. Steels (ed.), *Experiments in cultural language evolution*, John Benjamins, Amsterdam (NL), 2012.

[31] J. Hendler and T. Berners-Lee, From the Semantic Web to social machines: A research challenge for AI on the World Wide Web, *Artificial intelligence* **174**(2) (2010), 156–161.

[32] R. Axelrod, The dissemination of culture: a model with local convergence and global polarization, *Journal of conflict resolution* **41**(2) (1997), 203–226.

[33] J. Euzenat, First experiments in cultural alignment repair (extended version), in: *Proc. ESWC 2014 satellite events revised selected papers*, Lecture notes in computer science, 2014, pp. 115–130.

[34] M. Anslow and M. Rovatsos, Aligning experientially grounded ontologies using language games, in: *Proc. 4th international workshop on graph structure for knowledge representation, Buenos Aires (AR)*, 2015, pp. 15–31.

[35] M. Atencia and M. Schorlemmer, An interaction-based approach to semantic alignment, *Journal of Web Semantics* **13**(1) (2012), 131–147.

[36] P. Chocron and M. Schorlemmer, Vocabulary alignment in openly specified interactions, in: *Proc. 16th International conference on autonomous agents and multi-agent systems (AAMAS), Saõ Paolo (BR)*, 2017, pp. 1064–1072.

[37] P. Chocron and M. Schorlemmer, Attuning ontology alignments to semantically heterogeneous multi-agent interactions, in: *Proc. 22nd European conference on artificial intelligence (ECAI), The Hague (NL)*, 2016, pp. 871–879.

[38] J. Euzenat, Interaction-based ontology alignment repair with expansion and relaxation, in: *Proc. 26th International Joint Conference on Artificial Intelligence (IJCAI), Melbourne (VIC AU)*, 2017, pp. 185–191.