euroCRIS conference St. Andrews, Scotland, UK

Title

Can machines understand what researchers look for? Conceptualizing the research world

Author: Anna Guillaumet, aGuillaumet, SIGMA AIE

Abstract

Research information is a key topic for the researchers. Throughout history, researchers need to find what they want mainly through paper publications or books of previous researchers. Due the advance of the Internet, a large number of possibilities of data interaction appeared, making impossible to process and track all the research information and this could be a disadvantage. For this reason, semantics searches could help researchers to find and discover information in a reliable way. We must conceptualize the research word, through ontologies and the semantic data modeling techniques such as Resource Description Framework (RDF) and Web Ontology Language (OWL), to create a virtual scenario that a machine can "understand", in this way, when a researcher search or seek something, the machine provides results ordered by categories and discards the results that are not relevant. Also it can make recommendations: helping researchers find colleagues, affinities with groups, best projects for them, and so on. To make this possible, we must define a good interface (using user experience techniques) and use a powerful semantic search engine (using i.e. machine learning, data mining techniques). The results must show as clear as possible, maybe with data visualization techniques.

Keywords

Semantics, computer, researcher, interaction, research information, research data, machine learning, data mining, graphs, discover.

Paper

It's clear that Research information is a key topic for the researchers. Throughout history, researchers need to find what they want mainly through paper publications or books of previous researchers.

With the advance of the Internet, researchers must face a new paradigm, the research scenario has change radically, and new ways and forms to find information appeared.

Sometimes, a researcher could not know all the possibilities that the Internet offers and this could be a clear disadvantage for them. Digital publications in international databases, platforms, social networks, specialized collaboration networks, scientific webs, open access repositories, open data repositories and a large number of possibilities to access and obtain the information, appeared.

Nowadays it's impossible to process and track all of the available information, making the researcher's work more difficult and tedious, and perhaps not as optimum as it could be. Besides, the information is not in the same format, sometimes unstructured and difficult to access and reuse. Large collections of data that the human brain cannot process easily, so the computers need to help researchers achieve this. We are clearly talking about "big data", amounts and amounts of unstructured information "surfing on the web", in repositories, platforms and social media (Twitter, Facebook, Researchgate, and so on), information that grows exponentially as the minutes pass by... As David G. Bolaños said in an article, about Business management software platform Domo, released in February of 2016 [13]: **Data never sleeps.** About 3.2 billion people, sharing and uploading information to the Internet: many hours of videos, photos, and so on, a real mess.

Fortunately we can count on technology that has evolved as this new scenario occurred.

Different ways to process information appeared, like the **Data Curation**. Management / preservation of data (Data Curation)[1] is related to a very common problem in research: how to organize and preserve the data and materials used in the creation of knowledge, once the results (Article / essay / paper) have been published. According to the classic definition refers to data management across their life cycle, including maintenance and facilitating recovery and reuse.

Although management (or "Data Curation") is not related to their open access to itself, it has been linked to two concepts: Data Sharing (the practice of making data used for scholarly research available to other investigators) and Open Linked Data (about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods). Open Linked Data provides a step further, promoting this research data, can be part of data repositories. In this context, researchers could access them, reuse them and quote them. And also, they could gain recognition and credit to their own datasets.

For a time, 'Data Curation' has been a good alternative and remains, but it's no longer enough. So, what is the alternative? What can technology do for the research world? …A lot.

The way researchers interact with the information can and must be improved, and this is the main goal of SIGMA roadmap. SIGMA AIE, is a non-profit higher education consortium created and led by 10 of the most important Spanish universities, to optimize Academic and Research with information technology. This model allows the Group to minimize costs and time, implementing the best practices of university management. The above aspects are reflected in SIGMA's tools and services, given the changing environment of the universities and higher education institutions.

SIGMA RESEARCH is a complete research & innovation management system to promote collaboration, obtain better impact & visibility, enable decision making and optimize administrative procedures for higher education and research institutions, which develop and maintains SIGMA. SIGMA aware of the research scenario, works to improve the researchers interaction with the information, so, we are talking of changing the way researchers interact with the information.

We will explain a new way researchers can access and discover information, focussing our attention in the word "**discover**". SIGMA has a lot of experience working together with higher education institutions in the research field (more than 20 years). Thus, our

aim is to provide them the best innovative tools to solve main barriers the researchers must face.

We talk about new forms to access the information and we, as technologists, are making a lot of efforts to simplify the access to it. We can develop webs and platforms with sophisticated search engines and so on, but we are always using the same technology. We must change the paradigm and we must provide alternatives to use the computer capabilities for helping researchers. In this way, we need to "talk with the computers" to tell them what we want or what researchers want. How can we make it? through Semantics.

Semantics will be the next revolution of the information society. But, first of all, we must understand what does semantics means.

Semantics [2] (from Ancient Greek, means "significant") is the study of meaning. It focuses on the relationship between *signifiers*—like words, phrases, signs, and symbols—and what they stand for, their denotation. Linguistic semantics is the study of meaning that is used for understanding human expression through language. Other forms of semantics include the semantics of programming languages, formal logics, and semiotics.

The word *semantics* [2] was first used by Michel Bréal, a French philologist. It denotes a range of ideas—from the popular to the highly technical. It is often used in ordinary language for denoting a problem of understanding that comes down to word selection or connotation. This problem of understanding has been the subject of many formal enquiries, over a long period of time, especially in the field of formal semantics. In linguistics, it is the study of the interpretation of signs or symbols used in agents or communities within particular circumstances and contexts. Within this view, sounds, facial expressions, body language, and proxemics have semantic (meaningful) content, and each comprises several branches of study. In written language, things like paragraph structure and punctuation bear semantic content; other forms of language bear other semantic content.

In computer science,[2] the term *semantics* refers to the meaning of languages, as opposed to their form (syntax). According to Euzenat, semantics "*provides the rules for interpreting the syntax which do not provide the meaning directly but constrains the possible interpretations of what is declared*." In other words, **semantics is about interpretation of an expression.** Additionally, the term is applied to certain types of data structures specifically designed and used for representing information content.

Terms such as *semantic network* and *semantic data model* [2] are used to describe particular types of data model characterized by the use of directed graphs in which the

vertices denote concepts or entities in the world, and the arcs denote relationships between them.

So, the Semantic Web refers to the extension of the World Wide Web via embedding added semantic metadata, using semantic data modelling techniques such as Resource Description Framework (**RDF**) and Web Ontology Language (**OWL**).

The semantic web is based on ontologies. **Ontologies** [9] **are formalized vocabularies of terms**, often covering a specific domain and shared by a community of users. They specify the definitions of terms by describing their relationships with other terms in the ontology.

RDF [3] is a **standard model for data interchange on the Web**. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This *graph view* is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations.

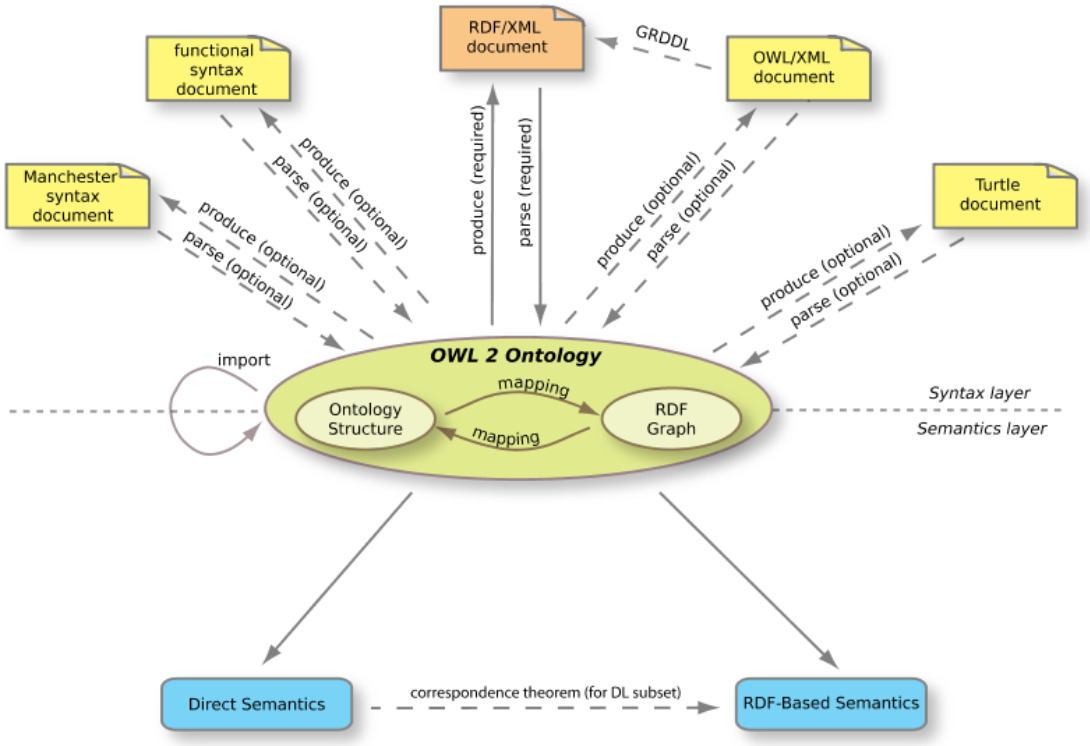
The RDF 1.1 specification consists of a suite of W3C Recommendations and Working Group Notes, published in 2014. The World Wide Web Consortium (W3C) [5] is an international community where Member organizations, a full-timestaff, and the public work together to develop Web standards. Led by Web inventor Tim Berners-Lee and CEO Jeffrey Jaffe, W3C's mission is to lead the Web to its full potential.

RDFa [6] is an extension to HTML5 that helps markup things like People, Places, Events, Recipes and Reviews. Search engines and Web services use this markup to generate better search listings and give better visibility on the Web, so people can find certain information more easily.

W3C[4] standards define an **Open Web Platform** for application development that has the unprecedented potential to enable developers to build rich interactive experiences, powered by vast data stores, that are available on any device. Although the boundaries of the platform continue to evolve, industry leaders speak nearly in unison about how HTML5 will be the cornerstone for this platform. But the full strength of the platform relies on many more technologies that W3C and its partners are creating, including CSS, SVG, WOFF, the Semantic Web stack, XML, and a variety of APIs.

In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. **The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network.** The term **"Semantic Web" refers to W3C's vision of the Web of linked data.** Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

OWL 2 [7][8] is an extension and revision of the OWL Web Ontology Language developed by the W3C Web Ontology Working Group and published in 2004 (referred to hereafter as "OWL 1"). OWL 2 is being developed (and this document was written) by a follow-on group, the W3C OWL Working Group. Like OWL 1, OWL 2 is designed to facilitate ontology development and sharing via the Web, with the ultimate goal of making Web content more accessible to machines.



The conceptual structure of OWL 2 ontologies is defined in the OWL 2 Structural Specification document [*OWL 2 Structural Specification*]. This document uses UML [*UML*] to define the structural elements available in OWL 2, explaining their roles and functionalities in abstract terms and without reference to any particular syntax. It also defines the functional-style syntax, which closely follows the structural specification and allows OWL 2 ontologies to be written in a compact form

There are many Ontology repositories and utilities that can be used[10]: BioPortal, Ontology Design Patterns, Prefix.cc, Schemapedia, http://www.opensemanticsearch.org/

So, we can conclude that [11] Semantics is the study of meaning. It focuses on the relation between words and phrases and due to the relationship of what they mean in context. **Semantics is a game changer** when it comes to the management of large volumes of structured and unstructured data. Having the ability to understand words in context using **lexical or linguistic semantics** will drive linked data initiatives and ultimately, as many believe, will make Web 3.0 a reality. With this in mind, possessing a **"semantic platform"** is the new marketing target for many software companies and they are quickly moving to describe themselves as a semantic platform in hopes of aligning themselves with the semantic snowball as it continues to gather speed.

The Semantic Web is a Web of data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. RDF provides the foundation for publishing and linking your data. Various technologies allow you to embed data in documents (RDFa, GRDDL) or expose what you have in SQL databases, or make it available as RDF files. Using OWL (to build vocabularies, or "ontologies") and SKOS (for designing knowledge organization systems) it is possible to enrich data with additional meaning, which allows more people (and more machines) to do more with the data.

Having all this technology, we can develop new and innovative ways to discover, track and work with the research information.

Researchers not only want to find information, they want to find a lot of more possibilities, related to networking, financial, projects, and so on, so **we must provide an environment where the researcher can find all they need.**

Imagine that we can conceptualize the research world, through this technology, defining ontologies and so on, **so that computers can understand what the researchers want** when they look for information, and not only what they want but what they seek.

A new way of interaction must be defined…

SIGMA is working with these new kinds of interactions, to offer researchers the information they want in the way they want and also helping them with new approaches. In this way, we define a very usable **interface** between the researcher and the computer (using the best user experience techniques), so the computer can interpret what the researcher is trying to find. It's not only to search information and general recommendations, like Google, are to **discover things**, and maybe things that the researcher **is not trying to find initially**. When a researcher writes something in the interface, the system must be capable to **interpret this information and translate it in**

a semantic search. With the well-defined ontologies, the system provides results ordered by categories and discards the results that are not relevant. All this information must be provided in a clear format, maybe in graphic models (with visualization techniques) to understand better the results. The semantic searches, not only provide information related to the search information, but provide relationships at some levels of the searched information. Relations between researchers, research projects and research groups, to work with recommendations, interesting calls related with searches and so on.

It is so important to define correct ontologies as the scope of information we access, in this way, we must define a "universe of data" that the semantic platform must track to provide the best semantic information.

There are some European initiatives about this topic and we are interested in collaborating in these new and innovative scenarios. We use the open innovation paradigm to find the best search semantic engine to achieve this objective. Now we are working on a new and innovative semantic engine that can provide these results. The next step is to build the best interface for the researcher.

There are nowadays some examples of these kinds of platforms, but we must improve and offer researchers the best platform with more semantic services. We can talk about platforms like VIVO. VIVO [12] is an open community, an open information model, and an open source semantic web application, supporting the advancement of scholarship by integrating and sharing information about scholars, their activities and their accomplishments.

VIVO [14] was developed by the Cornell University Library beginning in 2003 to meet individual and institutional research discovery needs and already addresses many areas of importance to researchers related to the health sciences. VIVO supports ontology as well as content editing, and is also a simple content management system that enables the representation of the resulting structured information in web pages. It uses the standard Resource Description Framework (RDF) Semantic Web data model and Web Ontology Language (OWL) schemas that identify distinct types of data and defines properties to connect these data with consistent, bi-directional relationships. For example, the profile of a person includes simple text attributes including name, title and statements of research or teaching interests, but extends much farther to include affiliation, activity and outcome relationships to departments, research grants, publications, talks, courses, research areas and geographic areas. Each of these entities is a defined type in its own right that may in turn have its own relationships to other people or to a funding agency, event sponsor, research center or topic.

VIVO project want to extend from a single institutional installation to a multi-institutional distributed model, VIVOWeb.

The current initiatives are like VIVO. A web where we can find all the research information of the institution ordered by categories and it can be extended to a multi-

institutional model, and so on. SIGMA is trying to go beyond and provide a new form of interaction with the information and, in this way, we are evaluating some alternatives.

We think that there must be a **disruptive change** in the way researchers interact with the information, the information must "explain" things, and not only show or discover information to researchers, and we work with some partners to achieve it, through innovative mining and machine learning techniques, with non-assisted platforms and no-dictionary definitions, letting the machine propose its own alternatives, self-learning more and more each time.

Also, is not just only a platform for researchers, but also it could be for students, administrative and service officials, prospective faculty and students, donors, funding agencies, and the public and to empower them to contribute – each in their own way in the research scope to improve funding, the university-business relationship, to find experts, to create international relationships and so on.

This is an exciting and innovative challenge that SIGMA faces, with the security that is the right and necessary way.

References:

[1] http://www2.ual.es/apoyoinv/preservacion-de-datos/que-es/

[2] https://en.wikipedia.org/wiki/semantics

[3] https://www.w3.org/RDF/

[4] https://www.w3.org/standards/

[5] https://www.w3.org/Consortium/

[6] https://rdfa.info/

[7] https://www.w3.org/2001/sw/wiki/OWL

[8] https://www.w3.org/TR/2012/REC-owl2-overview-20121211/#Documentation_Roadmap

[9] https://en.wikipedia.org/wiki/Ontology_(information_science)

[10] https://www.w3.org/wiki/Ontology_repositories

[11] http://www.dataversity.net/what-is-a-semantic-platform/

[12] http://www.vivoweb.org/

[13] http://cincodias.com/cincodias/2016/02/03/lifestyle/1454522046_347468.html

[14] NationalInstituteofHealth.pdf