

# **Improving an Academic Organisation's Presence on the Web Through Semantic Data**

**Bachelor thesis written by:**

**Veronika BARTA**

**Representative of mandating institution:**

**Jens VIGEN (CERN)**

**Academic advisor:**

**Julien GOBEILL, HES lecturer**

**Geneva, 15 July 2020**

**Information Documentaire**

**Haute École de Gestion de Genève (HEG-GE)**

## Declaration

This Bachelor thesis is written for the final examination given by the Haute École de Gestion de Genève, required in order to be awarded the degree of Bachelor of Science in Information Science.

The student certifies that her work has been subject to verification by plagiarism detection software (Urkund).

The student accepts the terms of the confidentiality clause, if relevant. Making use of the conclusions and recommendations contained in this Bachelor thesis, independently of their value, carries no liability for the author, the academic advisor for this Bachelor thesis, the independent juror or the HEG.

“I certify that I have carried out the present work alone, without having made use of any sources other than those listed in the bibliography.”

Geneva, 15 July 2020

Veronika BARTA

## Acknowledgements

First, I would like to thank my advisor, Julien Gobeill, especially for his patience with my many, many, **many** questions.

I would also like to thank Jens Vigen, whose enthusiasm is always a stimulating force, and Pablo Iriarte for agreeing to be a juror on this Bachelor project.

Thank you to Alex Kohls, whose feedback and kindness I find very valuable.

To my editors, Jelena Brankovic and Morgane Kozuchowski: you deserve, respectively, a giant donut, and a mountain of unicorns.

Then, somehow, I have managed to rack up debt all over the place: at CERN, thank you to Micha Moskovic, Stella Christodoulaki, Pamfilos Fokianos, Anne Gentil-Beccot and Salomé Rohr, as well as the entire RCS-SIS group, for being very welcoming; at the HEG, thank you to Michel Gorin for always encouraging students and having answers to every question one could possibly have; at Wikidata, thank you to Martin Poulter for his patience and advice; at Google, thank you to Matthew Prosser, for actually answering my email; and at Scholia, thank you to Finn Årup Nielsen, for creating such an awesome tool.

I would also like to thank my family, friends and classmates (especially Stéphanie, who shared my pain) just for being there.

And last, but certainly not least, I would like to thank my father, who has put up with having his hip a bit crowded for the last 30 years, and without whom this thesis (and this human) would not be possible.

# Abstract

CERN, the European Organisation for Nuclear Research, located in Geneva, Switzerland, was founded in 1954. Even before its first experiments started operating, the Organisation was very clear on wanting to be as open as possible with the data and research it would be producing. In 1993, CERN became a pioneer in the field of open source and Open Access by putting the software behind the World Wide Web into the Public Domain. Today, it is fighting on multiple fronts for these ideals, whether developing new software that will help share data, analyses and research openly, or being the driving force behind historic agreements, between publishers of High-Energy Physics (HEP) literature and institutions who publish in that field, aiming to make every single HEP article Open Access.

The internet, and more specifically the Web, is the perfect substrate for open and collaborative projects. For the last 20 years, one method for this openness has become increasingly ubiquitous: the semantic Web. An idea launched by the inventor of the Web himself – Sir Tim Berners-Lee – together with James Hendler and Ora Lassila, its intention was to go beyond the original principle of the first Web (Web 1.0) that had human-readable documents (Web pages) on separate servers, which a human had to “visit” to read. The semantic Web adds semantics (here, machine-understandable meaning) to data, and makes it findable and interlinkable.

This Bachelor project posits that collaborative Web projects such as the ones under the Wikimedia Foundation umbrella, and more specifically Wikidata which uses semantic technologies, are the perfect medium to help CERN be even more open and discoverable on the Web. It starts by looking for data about CERN that is interesting and harvestable (from INSPIRE – a HEP repository), then goes on to test pushing that data to Wikidata, and finishes by trying to assess the effect of the changes. This assessment being close to impossible to carry out, the thesis concludes with a few aspects that show how this kind of work can nevertheless have a great deal of value.

The two deliverables of this project were programs harvesting data from Inspire and transforming it into batches of semantic statements ready to be uploaded to Wikidata. The first, dealing with links between CERN people and their publications, created close to 6300 statements about 270 authors. The second, dealing with links between people and CERN, created 68 statements, representing an increase of over 12% of such links already on Wikidata.

Keywords: semantic web, web of data, open data, high-energy physics, HEP, CERN, Wikidata

# Table of Contents

<b>Declaration.....</b>	<b>i</b>
<b>Acknowledgements.....</b>	<b>ii</b>
<b>Abstract.....</b>	<b>iii</b>
<b>List of Figures.....</b>	<b>vi</b>
<b>List of Abbreviations.....</b>	<b>vii</b>
<b>1. Introduction.....</b>	<b>1</b>
<b>2. CERN.....</b>	<b>2</b>
<b>2.1 History and RCS-SIS .....</b>	<b>2</b>
<b>2.2 Open Philosophy .....</b>	<b>2</b>
2.2.1 Yellow Reports .....	3
2.2.2 SCOAP <sup>3</sup> .....	3
2.2.3 CERN Open Data .....	4
2.2.4 HEPData .....	4
2.2.5 CAP .....	4
2.2.6 PIDs .....	4
<b>2.3 Repositories and Catalogues.....</b>	<b>4</b>
2.3.1 CERN Document Server (CDS) .....	4
2.3.2 INSPIRE .....	4
<b>3. The Semantic Web and Wikidata.....</b>	<b>6</b>
<b>3.1 The Semantic Web.....</b>	<b>6</b>
<b>3.2 Wikimedia Foundation and Wikidata.....</b>	<b>9</b>
<b>4. Improving Discoverability .....</b>	<b>13</b>
<b>4.1 Harvesting Data from INSPIRE .....</b>	<b>13</b>
4.1.1 Methodology .....	13
4.1.1.1 Data source .....	13
4.1.1.2 First deliverable .....	16
4.1.1.3 Second deliverable .....	18
4.1.2 Results .....	21
4.1.2.1 First deliverable .....	21
4.1.2.2 Second deliverable .....	22
4.1.3 Limitations .....	22
<b>4.2 Improving Wikidata Data .....</b>	<b>23</b>
4.2.1 Methodology .....	23
4.2.1.1 Target website .....	23
4.2.1.2 First deliverable .....	23
4.2.1.3 Second deliverable .....	26
4.2.2 Results .....	26
4.2.3 Limitations .....	26
<b>5. Discussion.....</b>	<b>28</b>
<b>5.1 Assessment .....</b>	<b>28</b>

<b>5.2 Value.....</b>	<b>31</b>
5.2.1 Google.....	31
5.2.2 Scholia.....	32
5.2.3 Wikipedia.....	37
<b>5.3 Future Work .....</b>	<b>39</b>
5.3.1 Continuation and communication.....	39
5.3.2 Crossref.....	40
5.3.3 Wikimedia Commons.....	41
<b>6. Conclusion .....</b>	<b>42</b>
<b>Bibliography .....</b>	<b>43</b>

## List of Figures

Figure 1: The Semantic Web Technology Stack (not a piece of cake...)	7
Figure 2: Example of a knowledge graph	8
Figure 3: Sample of XML bibliographic record	9
Figure 4: Sample of RDF/XML bibliographic record	9
Figure 5: Datamodel in Wikidata	11
Figure 6: Example of statement in Quickstatements tool	11
Figure 7: Example of a SPARQL query on the Wikidata Query Service	12
Figure 8: INSPIRE search bar	14
Figure 9: SPIRES search syntax	14
Figure 10: Simple search result	15
Figure 11: API search result	15
Figure 12: Example of requests and json modules	16
Figure 13: Affiliation and <code>bai</code> in JSON	17
Figure 14: Example of Wikidata URL query	17
Figure 15: Example of <code>qwikidata</code> query	17
Figure 16: Example of DOI <code>qwikidata</code> query	18
Figure 17: INSPIRE authors	19
Figure 18: Example of <code>bai</code> or ORCID <code>qwikidata</code> query	20
Figure 19: Example of INSPIRE author profile	20
Figure 20: Example of INSPIRE author profile: affiliations	21
Figure 21: Example of batch ready for upload	22
Figure 22: Example of a researcher's profile on Reasonator	24
Figure 23: Example of From related items section on researcher's Reasonator profile	25
Figure 24: Example of Wikidata manual P2930 statement	25
Figure 25: Comparison of Virdee articles before and after a batch upload	26
Figure 26: Daily views on Virdee's Wikipedia page	28
Figure 27: Daily views on Virdee's Wikidata item page	29
Figure 28: Daily total views of Scholia tool	29
Figure 29: Daily total views of Scholia tool between July 2019 and June 2020	30
Figure 30: Daily total views of Wikidata	30
Figure 31: Example of a Google Knowledge Graph Card	31
Figure 32: Example of a Google Knowledge Graph Carousel	32
Figure 33: Top of CERN's Scholia profile	33
Figure 34: Employees and affiliates on CERN's Scholia profile	33
Figure 35: Topics employees and affiliates have published on on CERN's Scholia profile	34
Figure 36: Topics employees and affiliates have published on in a bubble chart	34
Figure 37: Awards on CERN's Scholia profile	35
Figure 38: Page production on CERN's Scholia profile	35
Figure 39: Detail of John Richard Ellis's co-author graph	36
Figure 40: Gender statistics on CERN's Scholia profile	36
Figure 41: SPARQL query to get images of people linked to CERN	37
Figure 42: Results of the SPARQL query to get images of people linked to CERN	37
Figure 43: CERN's infobox on English Wikipedia	38
Figure 44: Example of a tweet about a Wikidata query	39
Figure 45: Example of an article on Crossref	40
Figure 46: Example of an article on INSPIRE	40
Figure 47: Detail of the media in the CERN Wikimedia Commons Category	41

## List of Abbreviations

APC	Article Processing Charge
API	Application Programming Interface
arXiv	pre-print Open Access archive for physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics
ATLAS	A Toroidal LHC ApparatuS, an experiment which is part of the LHC
bai	INSPIRE author identifier
CAP	CERN Analysis Preservation
CC-BY	Creative Commons licence
CDS	CERN Document Server
CERN	European Organisation for Nuclear Research
CSV	Comma Separated Values
DESY	Deutsches Elektronen-Synchrotron (German Electron Synchrotron)
DOI	Digital Object Identifier
Fermilab	Fermi National Accelerator Laboratory (USA)
HEP	High-Energy Physics
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IHEP	Institute of High-Energy Physics (China)
IN2P3	Institut national de physique nucléaire et de physique des particules (French National Institute of Nuclear and Particle Physics)
IP	Internet Protocol
ISBN	International Standard Book Number
JINR	Joint Institute for Nuclear Research (Russia)
JSON	JavaScript Object Notation
LHC	Large Hadron Collider
LIS	Library and Information Science
OA	Open Access
ORCID	Open Researcher and Contributor ID



PID	Persistent Identifier
RCS-SIS	Research and Computing Sector - Scientific Information Service
RDF	Resource Description Framework
REST	REpresentational State Transfer
SCOAP <sup>3</sup>	Sponsoring Consortium for Open Access Publishing in Particle Physics
SEO	Search Engine Optimisation
SLAC	National Accelerator Laboratory (USA)
SPARQL	SPARQL Protocol and RDF Query Language
SPIRES	Stanford Physics Information Retrieval System
UNESCO	United Nations Educational, Scientific and Cultural Organisation
URI	Uniform Resource Identifier
XML	Extensible Markup Language
YR	CERN Reports, also known as Yellow Reports

# 1. Introduction

Most Library and Information Science (LIS) professionals agree on the reason they do what they do: to make knowledge as freely and as widely accessible as possible. That very noble, if somewhat vague, goal is omnipresent in librarians', archivists' and documentalists' daily work. But the ways these professionals go about trying to achieve that goal are plentiful and varied. They range from public libraries having only free services, to projects like the European Organisation for Nuclear Research (CERN)'s SCOAP<sup>3</sup> project (see section 2.2.2). Crowdsourced projects such as Wikipedia The Free Encyclopaedia have the same guiding principles. But although the LIS world has been offering free services for hundreds of years, and talking about Open Access for over 30, its interest in crowdsourced online projects is only in its infancy.

I was properly introduced to this world of free crowdsourced knowledge bases during my first internship with CERN's Scientific Information Service (RCS-SIS). I was tasked to improve the information on the English and German Wikipedias regarding CERN's physicists. More specifically, I was to add the tag *People Associated with CERN* (or *Person (CERN)* in German) and the justification for the tag (for example sources) on pages of CERN people who did not yet have it. This seemingly simple (dare I say menial) task is actually quite important. It adds a semantic layer to a website that is famously hard to harvest data from, mostly because most of its information is exclusively human-readable text. So, in this example, that means there is a list page that exists which groups together all the people who are associated with CERN, and that can only exist because the individual pages have that tag.

Now, one could ask why it is important to work on Wikipedia and similar websites, if the information is already online, on CERN's websites (CERN 2020 a, CERN Scientific Information Service no date a, CERN Document Server 2020, INSPIREHEP 2020, etc.). If it is true that all the information is already online, all sources are not equal: INSPIRE does not get anywhere near the same number of visitors as the English Wikipedia; of course, that is completely understandable, as Wikipedia tries to cover all areas of human knowledge, as opposed to INSPIRE, which concentrates exclusively on High-Energy Physics (HEP). The fact remains that information on websites such as Wikipedia has a much better chance of being seen than information on CERN websites. This means that it is extremely important for CERN to put the necessary resources into monitoring the information available outside of its own websites, adding to it, improving it, and in general making sure it is available and accurate.

So it came about that I proposed, as part of my Bachelor's project, to work on this subject for CERN's RCS-SIS. Thus, this project tries to answer the question: would putting time and effort into projects such as Wikidata, or more generally projects under the Wikimedia Foundation umbrella, help CERN be even more open on the Web than it currently is?

This thesis will be structured as follows. The first section describes the institution (CERN); the second explains what the Semantic Web is and deals briefly with the Wikimedia Foundation and its many projects, including Wikidata. What, in practice, has been done for this project follows, and I will finish by explaining what value there is to this kind of work, and how it can be taken further.

## 2. CERN

### 2.1 History and RCS-SIS

At the end of the Second World War, in an effort to try and contain the Brain Drain (large-scale emigration of trained individuals) to the United States, a few eminent scientists put forward a proposal first to the European Cultural Conference in 1949 and later to the fifth UNESCO General Conference in 1950, of creating a European atomic physics laboratory. The first resolution for a *Conseil Européen pour la Recherche Nucléaire* (CERN) was adopted in 1951 in Paris, at an intergovernmental meeting of UNESCO. In 1952, Geneva was chosen as the site for the Laboratory, and in 1953, the CERN convention was completed, stating among other things that “[...] the results of its experimental and theoretical work shall be published or otherwise made generally available”<sup>1</sup>. (CERN 2020 b)

In 1957 CERN's first accelerator, the Synchrocyclotron, was built; in 2015, CERN reported “the discovery of a class of particles known as pentaquarks” (CERN 2015). In between, CERN and its staff have been responsible for inventions such as the World Wide Web, pioneers of the touch-screen, discoverers of new particles (including the famous Higgs boson), and, in general, at the forefront of High-Energy Physics research for the last 60 years.

Today, CERN has 23 Member States: Austria, Belgium, Bulgaria, Denmark, Czech Republic, Finland, France, the Federal Republic of Germany, Greece, Hungary, Israel, Italy, the Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovak Republic, Spain, Sweden, Switzerland, and the United Kingdom. It also has two Associate Member States in the pre-stage to Membership, Cyprus and Slovenia (a third, Estonia, will be added soon), and six Associate Member States, Croatia, India, Lithuania, Pakistan, Turkey and Ukraine, as well as six entities with Observer status: Japan, the Russian Federation, the United States of America, the European Union, JINR and UNESCO. (CERN 2020 c)

CERN has four main goals: Discovery through science; Technological innovation; Diversity and bringing nations together; and Inspiration and education (CERN 2020 d).

CERN's Scientific Information Service is comprised of the Library, the Archive, the INSPIRE section (see section 2.3.2) and the Open Science section (see section 2.2). Beyond its classical library and archive missions (acquire and manage information resources for the worldwide high-energy physics community, safeguard documents of interest to CERN or for historical research, provide information about the archives, as well as access to them, and advise other CERN groups on archival and records management issues), RCS-SIS is also responsible for the distribution of CERN publications, and last, but certainly not least:

“It should advise CERN divisions on the best ways of ensuring that all relevant documents are made publicly available.” (CERN Scientific Information Service no date b)

### 2.2 Open Philosophy

“On 30 April 1993 CERN issued a public statement stating that the three components of Web software [...] were put in the Public Domain [...]” (Smith and Flückiger no date)

---

<sup>1</sup> (European Council for Nuclear Research Drafting Committee 1953)

If from the beginning CERN was always clear on wanting to make its results publicly available, putting the Web's software in the Public Domain was the start of its status as a pioneer in that field (CERN 2020 e). This section will outline a few projects led or co-led by the RCS-SIS team that fall under the umbrella of open science.

### **2.2.1 Yellow Reports**

As mentioned in section 2.1, CERN's SIS is responsible for, among other things, CERN's own publications, particularly CERN Reports, also commonly known as Yellow Reports (YR). These reports are collations of some of CERN's scientific publications, although some articles might be published in other journals as well. The Yellow Reports are made up of three sub-series: Monographs, School Proceedings and Conference Proceedings. SIS makes these reports as widely available as possible. Thus, unless another publisher owns the rights to an article, the reports are Open Access (OA). Today, it is easy to make a new Yellow Report OA, partly because of its "digital native" nature, but the YR started being published in 1955, so RCS-SIS is making a concentrated effort to make the whole collection OA, as a majority of the more than 1200 reports were published only in paper form. That work includes tasking an intern to do a bibliographic analysis of the cataloguing practices and overall discoverability of the YR series and tasking an administrative student to improve the digitisations of older YRs. There were several digitization campaigns over the years, but unfortunately the quality varies greatly, so the student is now currently working on improving the overall quality and readability of these digitisations.

### **2.2.2 SCOAP<sup>3</sup>**

A pioneer project led by the SIS team is the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP<sup>3</sup>). The main goal of SCOAP<sup>3</sup> is to make all HEP publications Open Access. To that end, the project, initiated by CERN, proposed an agreement between HEP publishers (that means journals that publish only, or partly, HEP articles) and countries (or more precisely institutions representing countries or parts of countries). This agreement includes not only CERN and its member states, but all countries that publish or read HEP literature. The basic premise is that the Consortium pays a set amount of money to the publishers every year and, in exchange, HEP researchers publish in these journals without having to pay Article Processing Charges (APCs), and the articles, if not the journals, are OA. If a journal has articles that are not HEP, so do not fall under the agreement, and the publisher does not want the whole journal to be OA, libraries receive a discount on their subscription fees. Each participating institution representing a country, or institutions within that country, that publishes or reads HEP literature pays a small part of the total amount being paid to publishers, with the principle that countries with more money pay 10% more, so that countries which cannot afford to pay may still publish in these journals without paying APCs.

This project, of course, is not perfect. The contracts have to be renegotiated every three years, and CERN is liable for the payments to the publishers, so if a country or institution is late with its part, CERN temporarily (or permanently, if the payment never comes) has higher costs. And of course, the agreements are limited to the very narrow scope of High-Energy Physics. But despite the relatively small scale and the imperfect model, projects like this, or very recently the read & publish agreement between swissuniversities and Elsevier (Université de Genève 2020), are extremely important, in fact essential steps in the movement for Open Access.

### **2.2.3 CERN Open Data**

CERN is also developing software and databases with Open Access or Open Data in mind. One of these projects, led again by the RCS-SIS team in collaboration with CERN's IT department, is the CERN Open Data Portal. The purpose of this portal is to ensure access to the data produced by CERN's research and experiments. The project adheres to established standards in data preservation and Open Science: open licences and Uniform Resource Identifiers (URIs – in this case, DOIs), to make the data (including accompanying software and documentation, needed to understand and analyse the data) reusable and citable. (Open Data CERN 2020)

### **2.2.4 HEPData**

In a similar vein, CERN's SIS helps Durham University operate a database called HEPData, an Open Access repository for data from experimental particle physics, more specifically, "data points from plots and tables related to several thousand publications including those from the Large Hadron Collider (LHC)". (HEPData 2020)

### **2.2.5 CAP**

CERN Analysis Preservation (CAP) is another of CERN's SIS projects, and its purpose is to help not only the raw data coming out of the experiments and the articles to be preserved, but also the steps in between (i.e. the analysis used to reach the conclusions in the published articles). This includes datasets, software and documentation, and CAP aims to help and encourage researchers to preserve and document the stages leading up to conclusions or publications. The main purpose of this preservation is to ensure future reproducibility, reusability, findability, accessibility and understandability. (CERN 2020 f)

### **2.2.6 PIDs**

Finally, CERN SIS is also working on encouraging the creation and use of persistent identifiers (PIDs), with projects such as Freya. Freya is funded by the European Commission under the Horizon 2020 programme. Its purpose is to encourage and extend the use of PIDs, which are an integral part of an open science model. Extending the infrastructure for PIDs will "improve discovery, navigation, retrieval, and access to research resources". (Freya no date)

## **2.3 Repositories and Catalogues**

### **2.3.1 CERN Document Server (CDS)**

The CERN Document Server (CDS) serves as the library and bookshop catalogue, as well as the institutional repository. The institutional repository contains all CERN publications (CERN Courier, CERN Bulletin, Yellow Reports, Annual Reports, theses, etc.). For now, catalogue and repository are in one system; however, the SIS team is working on a new system, where the library catalogue and the institutional repository would be separated, since they are two vastly different things. A library catalogue references all the books the library has in its possession and, in CERN SIS's case, also the books available in the bookshop. The institutional repository is significantly more interesting, since many of its resources may not be referenced anywhere else on the web.

### **2.3.2 INSPIRE**

INSPIRE is a disciplinary repository curated by CERN, together with DESY, Fermilab, IHEP, IN2P3 and SLAC. It interlinks several databases, among others for publications (including

books, scientific articles, science journalism articles, theses, etc.), conferences, jobs, experiments, etc. “[It] has been serving the scientific community for almost 50 years. Previously known as SPIRES, it was the first website outside Europe and the first database on the web” (Christodoulaki 2020). INSPIRE, like most other CERN projects, attempts to be as open as possible. That means, among other things, that if a document has a CC-BY licence, the repository will have the full text available, or at least link to an external source for it; it also gives free access to its data on citation, publication, co-authorship, etc., as well as free and totally open access to its API. INSPIRE makes its data available in a format called JSON (JavaScript Object Notation), which is a data interchange format that resembles a combination of XML and a Python dictionary.

## 3. The Semantic Web and Wikidata

### 3.1 The Semantic Web

There are several stages in the history of the Web. It was invented in 1989 at CERN by Sir Tim Berners-Lee. Frustrated by the different machines and programs that one had to deal with at CERN (because everyone brought their own computer and own data format from their countries of origin) and because it would take some time for someone to learn enough to get the information/document they wanted, Berners-Lee came up with an idea that made all the documents in different incompatible formats part of one big virtual documentation system on the internet, using hypertext technologies (HTTP and HTML). This first web (Web 1.0) was very static: it consisted simply of pages (i.e. documents) with some formatting (HTML) and hyperlinks. (TED 2009)

The second iteration of the Web is often called Web 2.0 or Participative Web. The websites encourage participatory culture, user-generated content, and generally a more social mode of behaviour. This includes social media sites, blogs, video or photo sharing sites, etc., that put the emphasis on commenting and tagging. (Web 2.0 2020)

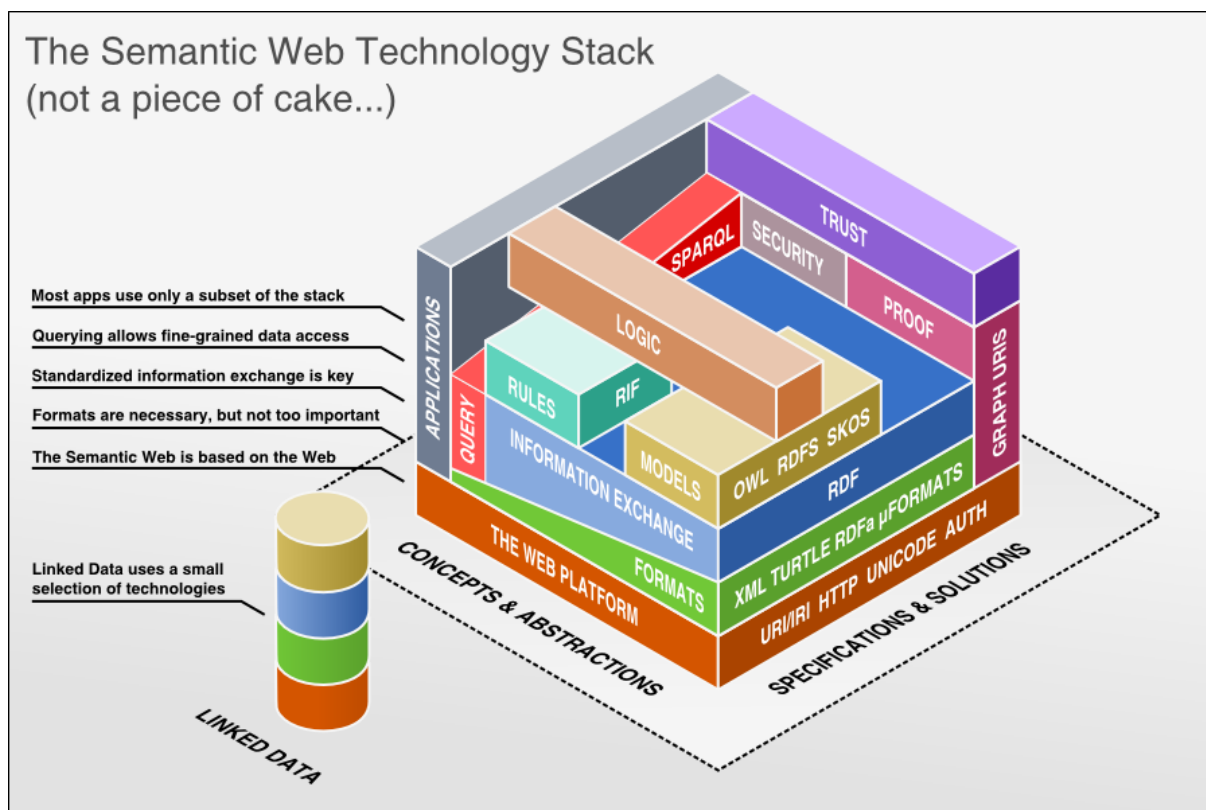
With the exponential growth of the Web<sup>2</sup> came the problem of sharing, accessing, retrieving and consuming the information. Indeed, keyword-based searching could not even touch data stored anywhere else than in HTML (in relational databases, for example), and did not take into account the meaning of the content, as web pages had mostly human-readable and understandable content. That is where Web 3.0 (or the Semantic Web, term coined by Berners-Lee, Hendler and Lassila in 2001) comes in. The Semantic Web offers ways of describing web resources that tackle this problem. The goal is to enrich existing information by adding semantics<sup>3</sup>, to make the resources understandable both by humans **and** machines. “This will make information easier to discover, classify and sort [...]” (Konstantinou and Spanos 2015).

---

<sup>2</sup> 50 Web servers worldwide in January 1993, 1500 in June 1994 (Raggett, Lam and Alexander 1996), 26 million pages indexed by Google in 1998, reaching the one billion mark by 2000 (Alpert and Hajaj 2008), and as of July 2020, the Indexed Web contains at least 5.52 billion pages. (De Kunder 2020)

<sup>3</sup> In this paper, “semantics” is used to mean “machine-understandable meaning”.

Figure 1: The Semantic Web Technology Stack (not a piece of cake...)



(Nowack 2009)

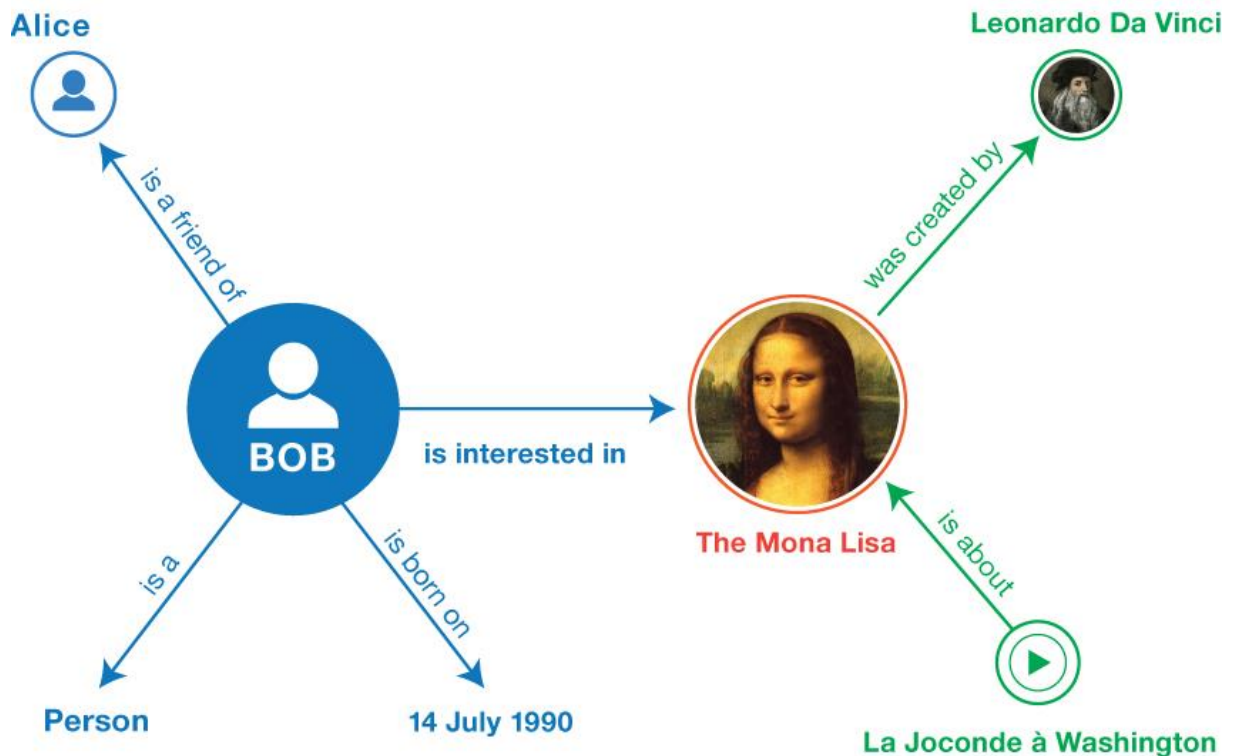
The Semantic Web is very complex, with many layers (see Figure 1). Without trying to be exhaustive, this part will attempt to explain a few of the core concepts. These are illustrated in Figure 3 and Figure 4 below.

**URIs**, Uniform Resource Identifiers, are strings of characters that unambiguously represent resources on the Web. Their main objective is to provide a unique name for the resource, but they may also give more information when they are dereferenced. Each part of a dereferenced URI is a locator for the resource: for example, <https://cds.cern.ch/record/1988646> gives the protocol (https), the server name (cds.cern.ch), a directory on that server (record), as well as a specific document (1988646). URIs are usually URLs (Uniform Resource Locators) as in the above example, but they may also more rarely be URNs (Uniform Resource Names). (Curé and Blin 2014)

**RDF**, Resource Description Framework, is a data model to describe and organise Web resources and their related metadata in a logical way, for the purpose of automatic handling (Curé and Blin 2014). The structure is made of triples, composed of a subject, a predicate and an object. For example, in *<the book 1988646> <has title> <RDF database systems: triples storage and SPARQL query processing>*, *<the book 1988646>* is the subject, *<has title>* the predicate and *<RDF database systems: triples storage and SPARQL query processing>* the object. Semantic data being particularly suitable for representation as a graph, knowledge described with RDF or any other such semantic model is often referred to as a “knowledge graph” (see Figure 2: in this case, one of the triples has *<Bob>* as the subject, *<is interested in>* as the predicate and *<The Mona Lisa>* as the object, but also *<The Mona Lisa>* as the subject in the triple *<The Mona Lisa><was created by><Leonardo Da Vinci>*).



Figure 2: Example of a knowledge graph



(W3C 2014)

**Ontology:** “[...] an ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.” (Ontology (information science) 2020)

In the example triple above, <has title> in the Dublin Core vocabulary<sup>4</sup> is a property called “title”, and its description is “A name given to the resource”. In a typical RDF/XML<sup>5</sup> document, the <has title> predicate would look like this: <dc:title>. So a machine would not know what the <has title> tag meant but, since the Dublin Core vocabulary is very precisely defined at <http://purl.org/dc/terms/> (Dublin Core Metadata Initiative 2020) and linked to other vocabularies<sup>6</sup>, the machine will “know” that the string <RDF database systems: triples storage and SPARQL query processing> is a title.

**SPARQL**, in full *SPARQL Protocol and RDF Query Language*, is a query language used for databases of RDF data.

<sup>4</sup> Although ontologies and vocabularies are slightly different, in this work the terms are used interchangeably.

<sup>5</sup> A syntax to express an RDF data structure as an XML document.

<sup>6</sup> (Linked Open Vocabularies 2020)

Figure 3: Sample of XML bibliographic record

```

1 <?xml version="1.0"?>
2 <catalog>
3   <book id="bk101">
4     <title>RDF database systems: triples storage and SPARQL query processing</title>
5     <author>Olivier Curé</author>
6   </book>
7 </catalog>

```

Figure 4: Sample of RDF/XML bibliographic record

```

1 <?xml version="1.0" encoding="utf-8"?>
2 <rdf:RDF xmlns:dct="http://purl.org/dc/terms/" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
3   <rdf:Description rdf:about="https://cds.cern.ch/record/1988646">
4     <rdf:type rdf:resource="http://purl.org/ontology/bibo/Book"/>
5   </rdf:Description>
6   <rdf:Description rdf:about="https://cds.cern.ch/record/1988646">
7     <dct:title>RDF database systems: triples storage and SPARQL query processing</dct:title>
8   </rdf:Description>
9   <rdf:Description rdf:about="https://cds.cern.ch/record/1988646">
10    <dct:creator rdf:resource="https://dblp.uni-trier.de/pers/c/Curacute=:Olivier.html"/>
11  </rdf:Description>
12 </rdf:RDF>

```

## 3.2 Wikimedia Foundation and Wikidata

*"Imagine a world in which every single human being can freely share in the sum of all knowledge."*<sup>7</sup>

Wikipedia The Free Encyclopedia was born in January 2001, founded by Jimmy Wales and Larry Sanger. The first edit was made on 15 January 2001, the six-millionth article was added to the English Wikipedia on 23 January 2020 (History of Wikipedia 2020), and as of July 2020, there are over 54 million cumulative articles over 310 language editions of Wikipedia. (Wikipedia 2020) The five pillars (i.e. fundamental principles) of Wikipedia are as follows:

1. Wikipedia is an encyclopaedia;
2. Wikipedia is written from a neutral point of view;
3. Wikipedia is free content that anyone can use, edit, and distribute;
4. Wikipedia's editors should treat each other with respect and civility;
5. Wikipedia has no firm rules (only policies and guidelines). (Wikipedia:Five pillars 2020 and, for more, see Wales 2020)

According to Alexa (Alexa Internet 2020), Wikipedia is, as of July 2020, ranked no. 13 in global internet engagement.

Not long after the start of Wikipedia, in 2003, co-founder Jimmy Wales decided to start the Wikimedia Foundation, a non-profit and charitable organisation to help fund Wikipedia and its sister projects. (Wikimedia Foundation 2020) The Foundation's main goal is to help the volunteers editing and adding knowledge to the projects. It also upholds the values that support free knowledge. The Wikimedia Foundation hosts Wikipedia and 11 other projects: Wikibooks, Wikiversity, Wikinews, Wiktionary, Wikisource, Wikiquote, Wikivoyage, Wikimedia Commons, Wikispecies, MediaWiki and Wikidata (Wikimedia Foundation 2020 a).

*"The community-based, somewhat chaotic consensus-driven approach of Wikidata can be frustrating ("well, if you'd asked ME, I wouldn't have done it that way"), but I think it's time to accept that this is simply the nature of the beast, and marvel at the notion that we have a globally accessible and editable knowledge graph. We can stay in our domain-*

<sup>7</sup> (Wikimedia Foundation 2020 a)

*specific silos, where we can control things but remain bereft of both users and contributors. However if we are willing to let go of that control, and accept that things won't always be done the way we think would be optimal, there is a lot of freedom to be gained by deferring to Wikidata's community decisions and simply getting on with building the bibliography of life.*<sup>8</sup>

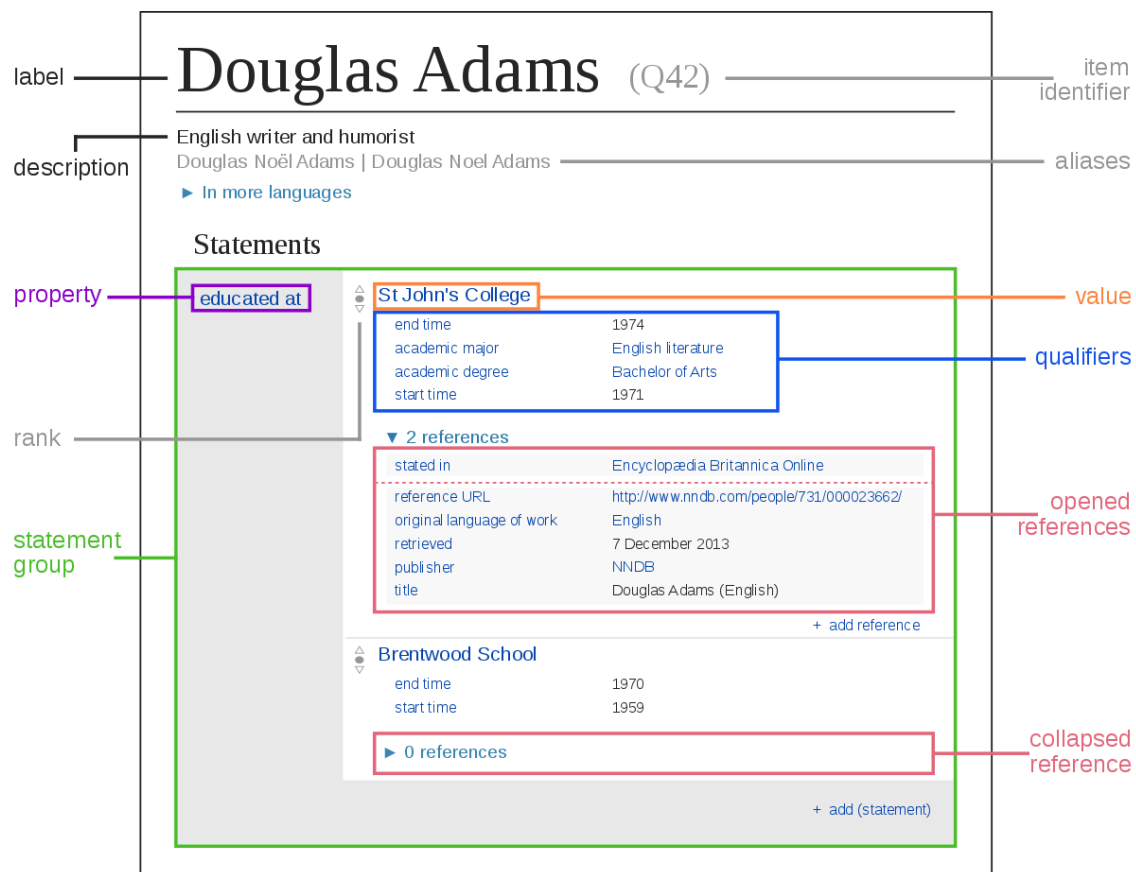
Wikidata, a project driven by Wikimedia Deutschland, was launched in 2012. The original goal, or more specifically the original problem that needed a solution, was the links between different language Wikipedias. Indeed, on Wikipedia, “[...] each of the 207 articles on Rome included a list of 206 links to all other articles on Rome—a total of 42,642 lines of text. By the end of 2012, 66 of the 287 language editions of Wikipedia included more text for language links than for actual article content.” (Vrandečić and Krötzsch 2014) To help solve that problem, as well as the challenge of getting to the data spread across millions of articles in hundreds of languages, Denny Vrandečić and a handful of other German Wikimedians proposed to start a “collaboratively edited multilingual knowledge graph” (Wikidata 2020), which became Wikidata. The project’s key principles are the following: open editing, community control, plurality (it is nigh-on impossible to agree on a universal truth, so Wikidata allows conflicting “truths” to coexist), secondary data (Wikidata collects data from primary sources, with references to said sources), multilingual data (there is only one Wikidata, with data not dependent on language, and the user chooses the language in which to view the data, or the language Wikipedia on which to use the data), easy access, and continuous evolution. (Vrandečić and Krötzsch 2014)

Wikidata is a knowledge graph using the RDF data model. A triple is called a statement, and is composed of an item (subject), a property (predicate), and a value (object). Wikidata has one page per item (see Figure 5), as well as qualifiers (i.e. references) and rank (to rank coexisting “truths”) for values. It uses its own URI system, with numbers preceded by a Q for items (Q42 in the example of Figure 5), or by a P for properties. However, Wikidata also values and encourages the addition of external identifiers, for example ORCIDs for researchers, or ISBNs for books.

---

<sup>8</sup> (Page 2020)

Figure 5: Datamodel in Wikidata



(Kritschmar 2016)

Statements can be added individually by editing the item page directly (like for a Wikipedia page), but batches of statements can also be uploaded directly to Wikidata through a tool called Quickstatements (Wikimedia Toolforge no date a). Statements have a very simple structure, resembling a CSV line, but with tab separations:

Q12787312 P1416 Q42944 S854 https://inspirehep.net/authors/982313

Figure 6 shows what this statement then looks like in Quickstatements before being uploaded to Wikidata.

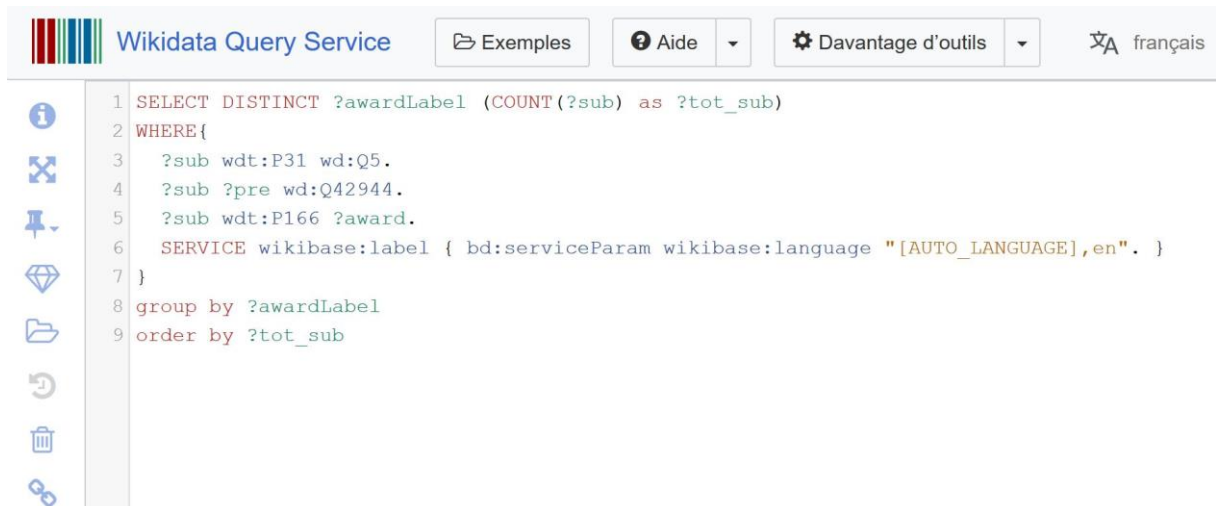
Figure 6: Example of statement in Quickstatements tool

1	init	Danilo Zavrtanik [Q12787312]	<div>ADD</div>	Statement	affiliation [P1416] : CERN [Q42944]
2	init	Danilo Zavrtanik [Q12787312]	<div>ADD</div>	Sources	affiliation : CERN to [P1416] [Q42944] [P854] reference URL : { "type": "unknown", "text": "https://inspirehep.net/authors/982313" }

(Wikimedia Toolforge no date a)

To query the data on Wikidata, there is a SPARQL endpoint available at <https://query.wikidata.org/> (Wikidata no date a). An example query is presented in Figure 7.

Figure 7: Example of a SPARQL query on the Wikidata Query Service



(Wikidata no date a)

## 4. Improving Discoverability

How does one go about improving an academic institution's presence on the Web through the semantic web? This project posits that investing resources into contributing to collaborative and open projects such as Wikidata can help achieve better discoverability and openness. The part below describes what data was chosen for this purpose, how it was harvested (4.1 Harvesting data from INSPIRE) and then pushed to Wikidata (4.2 Improving Wikidata data) to test the theory that this kind of work would or could help CERN be more visible online.

### 4.1 Harvesting Data from INSPIRE

#### 4.1.1 Methodology

The method used for this part of the project was quite straightforward. The first stage had to be the choice of the source of the data, as well as the kind of data that would be useful to work with. The second stage had to do with finding out how to harvest the data (understanding the tools, the formats, etc.). Finally, the data had to be transformed to be ready to be pushed to another part of the Web.

##### 4.1.1.1 Data source

The search for a source of data was done externally as well as internally. Externally, Crossref<sup>9</sup> was considered, before realizing that, as it lacks correct affiliation information, it might be more suitable as a target rather than a source. Wikipedia was also examined, but one of the reasons that Wikidata exists at all is that data is famously difficult to extract from Wikipedia. Internally, CDS was considered but, as mentioned in section 2.3.1, both library catalogue and institutional repository are in one single system. One possibility was to make arrangements for when the new systems would be ready, but the library catalogue was being developed first, so there was not enough information on the future repository to create something useful. INSPIRE was thus left as the main possibility.

As mentioned previously, INSPIRE is a disciplinary repository curated by CERN and other HEP institutions. It is composed of several interlinked databases on HEP literature, authors, jobs, seminars, conferences and institutions. The normal user interface can be seen in Figure 8. Just to the left of the search bar, where there is a button with a drop-down list, the user can choose between literature, authors, jobs, seminars, conferences and institutions, and type free text in the search bar. There is also some information below the search bar (see Figure 9) if the user wants to use SPIRES syntax operators.

---

<sup>9</sup> Crossref is a not-for-profit organisation whose main goal is to make scholarly communications better (Crossref no date a).

Figure 8: INSPIRE search bar



(INSPIREHEP 2020)

Figure 9: SPIRES search syntax

# How to Search

INSPIRE supports the most popular SPIRES syntax operators and free text searches for searching papers.

SPIRES

free text

Search by	Use operators	Example
Author name	a, au, author, name	a witten
Title	t, title, ti	t A First Course in String Theory
Collaboration	cn, collaboration	cn babar
Number of authors	ac, authorcount	ac 1->10
Citation number	topcite, topcit, cited	topcite 1000+

Learn more

(INSPIREHEP 2020)

INSPIRE has a RESTful<sup>10</sup> API that is very easy to use. If the URL of a search for articles with at least one CERN affiliation runs: `https://inspirehep.net/literature?sort=mostrecent&size=25&page=1&q=aff%20cern`, it is enough to add `/api/` in the URL to get the information back in a JSON

---

<sup>10</sup> REST is a software architecture style for creating Web services (Representational state transfer 2020).

---

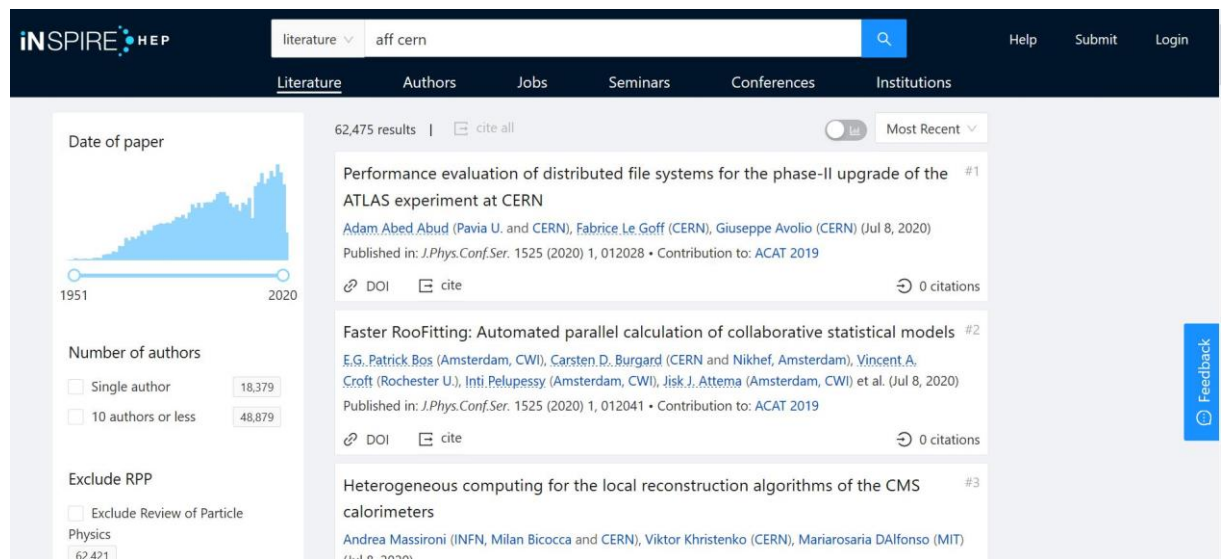
Improving an Academic Organisation's Presence on the Web Through Semantic Data  
Veronika BARTA



format:

<https://inspirehep.net/api/literature?sort=mostrecent&size=25&page=1&q=aff%20cern>. Figures 10 and 11 show the results for these two requests in a browser.

Figure 10: Simple search result



(INSPIREHEP 2020)

Figure 11: API search result

```
▼ hits:
  ▼ hits:
    ▼ 0:
      ▼ metadata:
        ▼ publication_info:
          0: {}
          citation_count_without_self_citations: 0
          citation_count: 0
          author_count: 3
        ▼ facet_author_name:
          0: "1067302_Giuseppe Avolio"
          1: "NOREC_Adam Abed Abud"
          2: "NOREC_Fabrice Le Goff"
        ▼ first_author:
          full_name: "Abed Abud, Adam"
          ids: [...]
          first_name: "Adam"
          last_name: "Abed Abud"
          earliest_date: "2020-07-08"
        ▼ authors:
          ▼ 0:
            raw_affiliations: [...]
            full_name: "Abed Abud, Adam"
            ids: [...]
            affiliations: [...]
            signature_block: "ABADa"
```

Most of the information needed to use the API is available on its GitHub page (Michamos and Jacquerie 2020). Among other information, it explains how to obtain single-record responses



as well as more general searching. One can also find information on rate limiting (maximum 50 requests from one IP address, then maximum 2 requests per second), as well as internal and external (DOIs, ORCIDs and arXiv) identifiers used.

Having understood how to build a URL to use the API, it was necessary to work out how to obtain the data with Python<sup>11</sup>. This can be done with two modules: `requests` and `json`. First, the URL must be fetched (`requests.get()`), then `json.loads()` used to get the text of the data (`.text`). This can be done in three simple lines (see Figure 12).

Figure 12: Example of requests and json modules

```
1 import requests, json
2 r=requests.get("https://inspirehep.net/api/literature?sort=mostrecent&size=25&page=1&q=aff%20cern")
3 data=json.loads(r.text)
```

It was important to understand the JSON format, which proved to be relatively simple using knowledge of how a Python dictionary works and some observation and testing. However, to determine which part of the data was interesting or important was much more difficult. For that, the first necessity was to find out to which website the transformed data would be pushed (see Section 4.2.1).

#### 4.1.1.2 First deliverable

Once Wikidata had been chosen, and it had been decided that one of the main aspects to be improved was the links between CERN people and their articles (that program would become the first deliverable), the kind of information which should be available in the data derived from INSPIRE could be determined. The first step would be to find a reliable way of identifying people. The `full_name` found in the JSON data would not work, because it is not a reliable way to find someone on Wikidata. The best would be to use a unique identifier. The ORCID came to mind first, but unfortunately, not every researcher has one yet, and it is not systematically referenced on Wikidata. There was an identifier called `bai` in the JSON data, sometimes also referred to as an INSPIRE ID (Clegg 2020) or INSPIRE-HEP author ID on Wikidata (INSPIRE-HEP author ID (P2930) 2019), which seemed to be the most prevalent on both sites.

Consequently the data for all articles with at least one CERN affiliation was taken (with an iteration to obtain all ~62'000 articles, 25 at a time to avoid overloading the system), securing the `bai` identifiers for all the authors with at least one CERN affiliation. The method was as follows:

```
data["hits"]["hits"][m]["metadata"]["authors"][o]["affiliations"][q]
["value"]=="CERN" checks the affiliation, and if it is CERN,
```

```
data["hits"]["hits"][m]["metadata"]["authors"][o]["bai"] obtains the bai.
```

`m` is the rank of the article among the 25 per page, `o` is the rank of the author out of the total number of authors for that article and `q` is the rank of the author's affiliation out of five, as some researchers publish with several affiliations at a time. Figure 13 shows the JSON as presented in a web browser.

---

<sup>11</sup> Python is a programming language frequently taught and used in the LIS world.

Figure 13: Affiliation and bai in JSON

```
▼ authors:
  ▼ 0:
    ▶ raw_affiliations: [...]
    full_name: "Abed Abud, Adam"
    ▶ ids: [...]
    ▼ affiliations:
      ▼ 0:
        value: "Pavia U."
        ▶ record: {...}
      ▼ 1:
        value: "CERN"
        ▶ record: {...}
    signature_block: "ABADa"
    uuid: "5ade5eb7-ccfb-45f0-a0e1-b44103b5af88"
    first_name: "Adam"
    last_name: "Abed Abud"
    bai: "A.Abed.Abud.1"
    ▶ name_suggest: {...}
    full_name_unicode_normalized: "abed abud, adam"
```

The next step was to use the bais to check which researcher had an item page on Wikidata. One of the possibilities would have been to use the Wikidata SPARQL query service simply in the URL (see Figure 14). However, a Python module called qwikidata (Kensho Technologies LLC 2019) exists that allowed the same queries in a slightly easier way (see Figure 15).

Figure 14: Example of Wikidata URL query

```
1 import json, requests
2 url_wiki="https://query.wikidata.org/sparql"
3 url_query="SELECT ?sub WHERE {?sub wdt:P2930 'A.Abed.Abud.1' .}"
4 dico={"query":url_query,
5 "format":"json"}
6 r=requests.post(url=url_wiki, data=dico)
7 data=json.loads(r.text)
```

Figure 15: Example of qwikidata query

```
1 import json
2 from qwikidata.sparql import return_sparql_query_results
3 sparql_query = """
4 SELECT ?sub
5 WHERE{
6   ?sub wdt:P2930 'A.Abed.Abud.1' .
7 }"""
8 res = return_sparql_query_results(sparql_query)
```

This last step had to come before going back to INSPIRE to get the DOIs of all the articles written by the researchers, because some CERN physicists working on big collaborations such as ATLAS, for example, have “written” over 1400 articles. Otherwise a large amount of data would have been fetched, when only a fraction of the researchers have Wikidata item pages. So once there was a list with all the researchers with at least one CERN affiliation on INSPIRE, and who also have an item page on Wikidata, their DOIs could be obtained from INSPIRE. However, if only 25 (or even fewer) article DOIs could be fetched at a time, and supposing there were upwards of 100 researchers to check, some with up to 1400 articles, it would take a very long time to get all the data needed. With the help of a colleague from the INSPIRE team, an argument was placed in the header of the `get` request (`headers={"accept": "application/vnd+inspire.record.ui+json"}`) to obtain less data for each article (but still including the DOI, the only information which was actually interesting), which allowed fetching 200 articles at a time. The output of this step was one text document per researcher that had a Wikidata item page, containing all the DOIs available on INSPIRE of articles that the researcher had authored.

Wikidata was accessed for a second time to check every DOI in every document against Wikidata’s data, to see which ones also had an item page. The SPARQL query was now a little more complex, because articles that already had the connection to the author were not needed. Indeed, academic articles on Wikidata often already have some information on their author(s). Very often, the metadata for these articles was uploaded somewhat automatically, so there were no checks to see if the author already was on Wikidata. In these cases, the statement uses the property Author name string (P2093), with the value, in this case the name of the author, as a simple string (for example “Abed Abud, Adam”), instead of a URI (Q123456789, for example). However, some articles have correct author statements, with an author (P50) property and a URI as a value. To make sure only to get articles that do not already have a correct P50 author statement, a line has to be added in the SPARQL query to screen for those cases (see Figure 16).

Figure 16: Example of DOI qwikidata query

```

1  import json
2  from qwikidata.sparql import return_sparql_query_results
3  sparql_query = """
4  SELECT ?obj
5  WHERE{
6      ?obj wdt:P356 '10.1103/PhysRevLett.124.082002'.
7      MINUS{?obj wdt:P50 wd:Q67220367.}
8  }"""
9  res = return_sparql_query_results(sparql_query)

```

#### 4.1.1.3 Second deliverable

Another part of the data on Wikidata that could be improved was the links between people and CERN (this would become the second deliverable). The assumption was that there were researchers with Wikidata item pages that did not yet have such a link to CERN. The first step to remedy that was deciding which property would be best to use (see section 4.2.1.3). The second step was to find what criteria to use to determine if a person should be considered as affiliated with CERN. Is one affiliation (in a single article) enough? It is a difficult question, and

after consultation with Jens Vigen (the representative of the mandating institution), the criterion used in the past when adding tags to Wikipedia pages was chosen: a person is considered linked to CERN when that person has published three articles or more with a CERN affiliation.

With these decisions made, data from INSPIRE could now be harvested. This time all the authors on INSPIRE (about 127'000) were examined, 25 at a time, and the ORCID IDs as well as the `bais` were fetched, in order to increase the probability of finding the Wikidata item pages of the researchers (i.e. if the INSPIRE-HEP author ID was not referenced on the item page, but the ORCID was). The unique ID (made up of numbers, as opposed to the `bai`) as well as the `preferred_name` was also needed, both of these for the subsequent INSPIRE step. Thus a document was created with researchers' INSPIRE ID and name, as well as `bai` or ORCID (or both - see Figure 17, line 2 or 4, for an example).

Figure 17: INSPIRE authors

1	H.H.Chavez.Sanchez.1	1774790	Helder Chávez
2	M.Tomii.1	0000-0003-0118-7703	1275248 Masaaki Tomii
3	M.Zamaklar.1	982388	Marija Zamaklar
4	W.J.Zakrzewski.1	0000-0002-0629-2812	982405 Wojtek J. Zakrzewski
5	T.Zannias.1	982353	Thomas Zannias
6	L.Zanini.1	982355	Luca Zanini
7	R.Zaliznyak.1	982391	Renata Zaliznyak
8	A.F.Zakharov.1	982413	Alexander Fyodorovich Zakharov
9	J.Zak.2	982415	John W. Zak
10	V.S.Zamiralov.1	982380	Valery S. Zamiralov
11	F.Zamani.1	982387	Farid Zamani
12	A.M.Zaitsev.1	0000-0002-4961-8368	982420 Alexandre M. Zaitsev
13	M.Zanolli.1	982349	Manfred Zanolli
14	R.Zannoni.1	982351	Roberto Zannoni
15	A.Zalite.1	982393	Andrey Zalite
16	S.Zakrzewski.1	982406	Stanislaw Zakrzewski
17	R.Zalaletdinov.1	982402	Roustam M. Zalaletdinov
18	J.M.Zanotti.1	0000-0002-3936-1597	982346 James M. Zanotti
19	P.Zalewski.1	982395	Piotr Zalewski
20	P.Z.l.n.1	982401	Peter Zalan
21	I.Zahed.1	982428	Ismail Zahed
22	M.Zanolini.1	982350	Margherita Zanolini
23	J.Zanelli.1	982363	Jorge Zanelli
24	M.Zanabria.1	982369	Manuel Eugenio Zanabria
25	L.E.Zamora.1	982375	Luis Benito Zamora
26	A.Zaitchenko.1	982421	Alexandre Zaitchenko
27	B.G.Zakharov.1	982411	Bronsilav G. Zakharov
28	C.Zambon.1	982382	Cristina Zambon
29	N.Zamiatin.1	982381	Nikolai Zamiatin
30	H.W.Zaglauer.1	982430	Helmut W. Zaglauer
31	D.Zanon.1	982348	Daniela Zanon
32	N.Zaganidis.1	982435	Nicolas Zaganidis
33	J.Zajac.1	982418	Juliusz Zajac
34	M.Zanin.1	982357	Marco Zanin
35	A.Zannoni.1	982352	Alberto Zannoni
36	A.Zallo.1	0000-0002-2212-2821	982390 Adriano Zallo
37	W.A.Zajc.1	0000-0002-9871-6511	982417 William A. Zajc
38	O.Zanotti.1	982345	Olindo Zanotti
39	R.L.Zako.1	982409	Robert L. Zako
40	B.V.Zagareev.1	982429	Boris V. Zagareev
41	I.Zakout.1	982408	Ismail Zakout

With the data in this document, Wikidata was queried with the `qwikidata` module used previously, to get all the item pages for the researchers in the document. This query checked for `bais` or ORCID IDs, as well as made sure that the researcher did not yet have a link to CERN (see query in Figure 18).

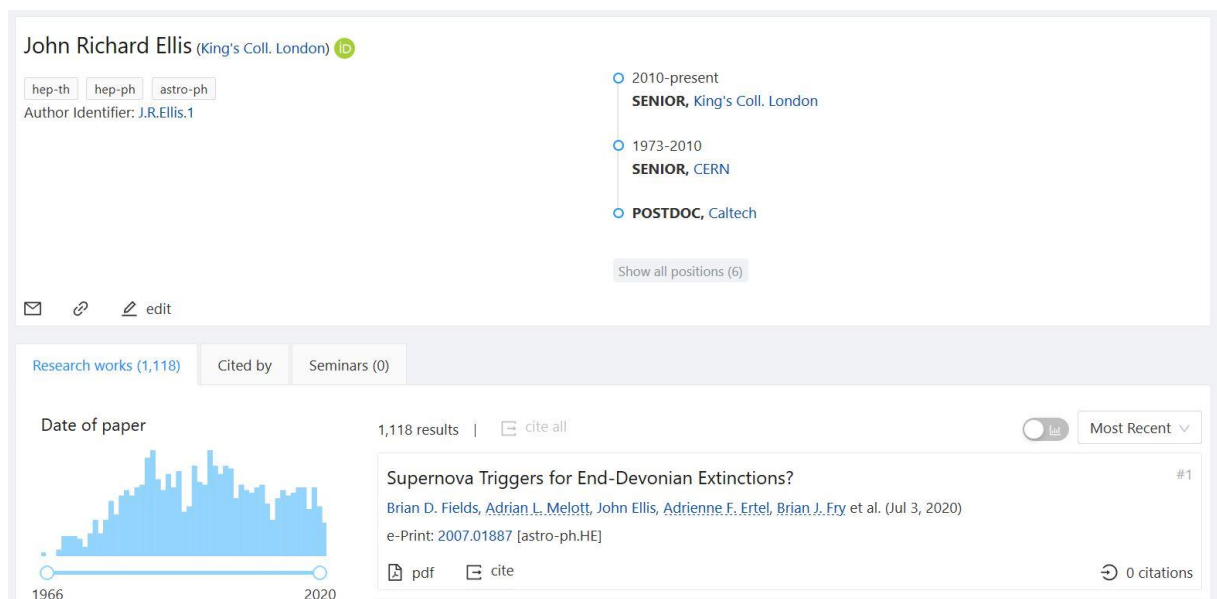


Figure 18: Example of bai or ORCID qwikidata query

```
1 import json
2 from qwikidata.sparql import return_sparql_query_results
3 sparql_query = """
4 SELECT DISTINCT ?sub ?obj
5 WHERE{
6     {?sub wdt:P2930 'M.Tomii.1'}
7     UNION
8     {?sub wdt:P496 '0000-0003-0118-7703'}
9     MINUS{?sub wdt:P108 wd:Q42944}
10 }"""
```

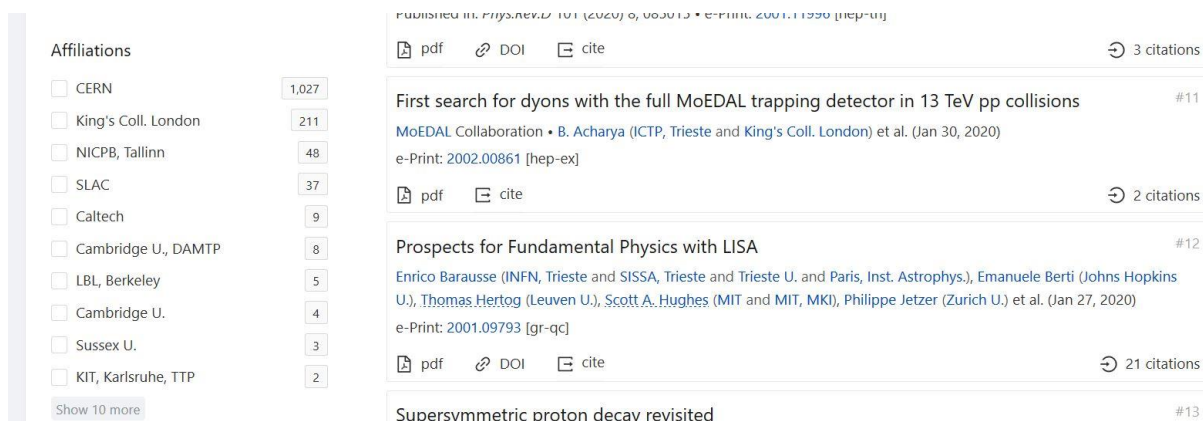
Now with a list of researchers who have a Wikidata item page, INSPIRE was used again to get information on the researchers' affiliations. Here, the goal was to find researchers with three or more articles authored with a CERN affiliation. INSPIRE, luckily, already had the information needed. Indeed, it has profiles for authors (see Figure 19) that not only give the articles authored, but also statistics on date of publication, citation and, crucially, affiliations (see Figure 20).

Figure 19: Example of INSPIRE author profile



(INSPIREHEP 2020)

Figure 20: Example of INSPIRE author profile: affiliations



(INSPIREHEP 2020)


This affiliation facet is available through the API, with a URL using the unique ID and the name of the researcher, for example: [https://inspirehep.net/api/literature/facets?sort=mostrecent&size=25&page=1&author=1275248\\_Masaaki\\_Tomii&facet\\_name=hep-author-publication&author\\_recid=1275248\\_Masaaki\\_Tomii](https://inspirehep.net/api/literature/facets?sort=mostrecent&size=25&page=1&author=1275248_Masaaki_Tomii&facet_name=hep-author-publication&author_recid=1275248_Masaaki_Tomii).

## 4.1.2 Results

### 4.1.2.1 First deliverable

The first deliverable's program, when run, went through about 62'000 articles with at least one author with a CERN affiliation, which resulted in 7381 authors with at least one CERN affiliation. Of those, 270 had a Wikidata item page, and so 270 individual batches were created (after obtaining sometimes upwards of 1500 DOIs per researcher from INSPIRE), as in Figure 21. In total, 6297 statements were created, ready to be uploaded (every batch individually - unfortunately, there is no faster way to do that) to Quickstatements (see section 3.2, p.10). The biggest batch had 227 statements, and there were also 87 empty batches, either because those researchers' articles are not on Wikidata yet, or because the semantic links between article and author are already there.

Figure 21: Example of batch ready for upload

 P.S.Wells.1\_clean\_batch.txt - Bloc-notes

Fichier	Edition	Format	Affichage	Aide
Q64451273		P50	Q57612653	
Q63440618		P50	Q57612653	
Q63440610		P50	Q57612653	
Q64451193		P50	Q57612653	
Q63440612		P50	Q57612653	
Q61501214		P50	Q57612653	
Q59456410		P50	Q57612653	
Q61501167		P50	Q57612653	
Q58957622		P50	Q57612653	
Q60193268		P50	Q57612653	
Q57451440		P50	Q57612653	
Q58820779		P50	Q57612653	
Q57451436		P50	Q57612653	
Q56906815		P50	Q57612653	

As there are currently (as of July 2020) almost 129'000 statements between people linked to CERN and their articles already on Wikidata, adding 6297 statements would represent an increase of almost 5% of such statements.

#### 4.1.2.2 Second deliverable

The second deliverable's program yielded 68 statements (adding affiliation links between researchers and CERN), based on over 125'000 authors on INSPIRE, of whom 274 had item pages on Wikidata. As there are currently (as of July 2020) 549 statements linking people with CERN on Wikidata, adding 68 statements would increase this total by over 12%.

#### 4.1.3 Limitations

There were unfortunately quite a few obstacles that came up while trying to harvest data from INSPIRE and prepare it to be pushed to Wikidata. The first major hurdle was that the data on INSPIRE, as structured as it is, is not semantic. That means that although it uses its own vocabulary, it is not semantically defined nor is it linked to any other external vocabulary or ontology, and although it uses persistent identifiers, the majority of the data is not in the form of URIs. Were INSPIRE data to be closer to an actual knowledge graph, it would be possible (if perhaps a little beyond the scope of a project like this one) to try to map whole sections of the repository's databases to prepare them to be transferred to Wikidata in their entirety (see Chardonens's work with the Montreux Jazz Digital Project, 2019, although he was also working with a non-semantic database).

Another hurdle had to do with the data proper: although INSPIRE is a very good repository (as good as it can possibly get, actually), it is not complete, as that is entirely impossible to do for a field constantly growing and changing. So despite the team working very hard to maintain and improve it continually, not all author profiles have a `bai` or ORCID, not all articles have a DOI, etc. That fact, together with the limits on the data on Wikidata (as it is maintained by volunteers, and especially as it is not focused solely on HEP, it is much further from being complete in HEP than INSPIRE is), makes it impossible for any program to harvest **all** the data that would need to be added to Wikidata.

The last big hurdle for this part of the project was the technical limits. Some of these were the rate limitations of INSPIRE mentioned in section 4.1.1.1, which meant that a second had to be added between each request (done with the `sleep()` method of Python's `time` module), to make sure not to overload the system. Not only did the rate of the requests need to be closely controlled, but there was the added difficulty that the API itself was easily overwhelmed, often answering with a “502 Bad Gateway” when there were too many articles or authors per requested page, so that the number of these had to be scaled down as well. Lastly, Wikidata also seemed to have some problems. Although it was never discovered where the problem lay, when queried, the site did not always give an answer. The program therefore had to include an additional iteration to make sure that Wikidata was queried until there was an answer for each individual query.

## 4.2 Improving Wikidata Data

### 4.2.1 Methodology

This part of the project dealt with finding a target for the data to be pushed to. The first stage was to find this target. The subsequent stages were finding the best ways (both technical and theoretical) to push the data, and testing the method.

#### 4.2.1.1 Target website

If Wikidata was always planned as a target website for data to be pushed to, some other targets were considered before being abandoned in favour of concentrating solely on Wikidata. As mentioned above, Crossref was briefly considered, as it has some information missing, especially regarding affiliations. Unfortunately, it is complicated to push data to the website, as an agreement would have to be signed between Crossref and the institution wanting to upload or correct some data, and so it would only be worth trying if one was interested in uploading whole datasets or knowledge graphs to the site. As this project is rather concentrating on improving data already on the web, this would not really fit within its scope. Wikipedia and Wikimedia Commons<sup>12</sup> were also considered, but as neither is actually semantic, and both are almost (or completely) impossible to contribute to in an automated way, they were both dropped.

#### 4.2.1.2 First deliverable

After deciding to concentrate the first deliverable on improving the links between CERN researchers and their articles on Wikidata, the choice had to be made as to the property that would be the most useful. Having started the data harvesting with information on authors, the first idea was to use the List of Works (P1455) property, which would allow the information to be visible directly on the author's item page. Unfortunately, there were several obstacles: the value of a statement using that property should be a bibliography<sup>13</sup>, which implies that one would need to create a new item for the bibliography, which could then contain all the articles; the property is also not very commonly used (not a single case of a statement using that property for a person linked to CERN was found), and in fact it is quite often misused. In the

---

<sup>12</sup> A collaborative database of freely reusable media files which is a part of the Wikimedia Foundation (Wikimedia Commons 2020).

<sup>13</sup> See <https://www.wikidata.org/wiki/Q1631107> (bibliography (Q1631107) 2020) for a description.



end, a decision was made to use the property author (P50)<sup>14</sup>. It is very widely used, and although the information would not appear directly on the author's item page, there are tools that allow an easy, human-readable way of grouping this data, without having to know or use SPARQL with a query endpoint. For example, the tool called Reasonator (Wikimedia Toolforge no date b) allows users to view the profile of a researcher built with data from Wikidata (see Figure 22), including statements that have the author as the value instead of the item (see From related items section on reasonator profile – Figure 23).

Figure 22: Example of a researcher's profile on Reasonator

Tejinder Virdee (Q1033005)

Tejinder Singh Virdee | Virdee | Jim Virdee | تَجیندر ویردی | ਤੇਜਿੰਦਰ ਵਿਰਦੀ | 特金德·福狄 | テイジンダー・ヴィルディー | تَجیندیر قردی

British physicist

Tejinder Virdee is a [British physicist](#) and [university teacher](#).  
He was born on [October 13, 1952](#) in [Nyeri](#).  
He studied at [Queen Mary University of London](#). He is/was a member of [Royal Society](#). He worked for [Imperial College London](#) and for [CERN](#).

Other properties

employer

Imperial College London public research university located in London, United Kingdom

CERN international organization which operates the world's largest particle physics laboratory

award received

High Energy and Particle Physics Prize EPS award  
point in time : 2013

Glazebrook Medal Awarded annually by the Institute of Physics  
point in time : 2015

James Chadwick Medal and Prize the Institute of Physics (IOP) biennial award made for distinguished research in particle physics  
point in time : 2009

Special Breakthrough Prize in Fundamental Physics  
point in time : 2013


ITEM NOT LOADED : Q891037  
point in time : 2017

doctoral advisor

ITEM NOT LOADED : Q7173711

place of birth

Nyeri town in the Central Highlands of Kenya.



External sources

Fellow of the Royal Society

tejinder-virdee-12466

GND

1119264448

INSPIRE-HEP author

T.S.Virdee.1

ISNI

0000 0001 3548 5544

ORCID

0000-0001-7429-2198

Scholia

Scholia

VIAF

7505147967372084200004

(Wikimedia Toolforge no date b)

<sup>14</sup> This implies that the researcher becomes the object of the triple that has the article as the subject, so the information would only appear on the article's, and not the author's, item page.

Improving an Academic Organisation's Presence on the Web Through Semantic Data

Veronika BARTA

24

Figure 23: Example of From related items section on researcher's Reasonator profile

<b>residence</b>	Geneva city in Switzerland and capital of its canton
<b>date of birth</b>	1952-10-13
<b>educated at</b>	Queen Mary University of London public research university in London, United Kingdom; constituent college of the federal University of London
<b>family name</b>	Virdee family name
<b>noble title</b>	Knight Bachelor part of the British honours system

From related items

**author**

Search for anomalous  $t$  production in the highly-boosted all-hadronic final state wetenschappelijk artikel  
series ordinal : 1449  
stated as : T. Virdee

Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV artikel  
series ordinal : 1527  
stated as : T. Virdee

Search for supersymmetry in the vector-boson fusion topology in proton-proton collisions at  $\sqrt{s}=8$  TeV im November 2015 veröffentlichter wissenschaftlicher Artikel  
series ordinal : 1551  
stated as : T. Virdee


Search for neutral MSSM Higgs bosons decaying into a pair of bottom quarks im November 2015 veröffentlichter wissenschaftlicher Artikel  
series ordinal : 1545  
stated as : T. Virdee

Search for a Higgs boson in the mass range from 145 to 1000 GeV decaying to a pair of W or Z bosons im Oktober 2015 veröffentlichter wissenschaftlicher Artikel  
series ordinal : 1542  
stated as : T. Virdee

Search for neutral color-octet weak-triplet scalar particles in proton-proton collisions at  $\sqrt{s}=8$  TeV im September 2015 veröffentlichter wissenschaftlicher Artikel

Current language Wikipedias

en Tejinder Virdee  
Big Wikipedias  
de Tejinder Virdee  
fr Tejinder Virdee  
Wikiquote  
en Tejinder Virdee  
Other Wikipedias  
Concept cloud



(Wikimedia Toolforge no date b)

The next step was to test out adding Wikidata statements. Over 500 manual edits<sup>15</sup> were made, mostly adding author (P50) statements, or removing author name string (P2093) statements, as well as adding employer (P108), INSPIRE-HEP author ID (P2930) and ORCID iD (P496) statements. Manual edits are made directly on an item page (see Figure 24).

Figure 24: Example of Wikidata manual P2930 statement

INSPIRE-HEP author ID

T.S. Virdee 1

✓ publish ✕ cancel ?

+ add qualifier

▼ 1 reference

reference URL

https://inspirehep.net/authors/984609

remove

+ add

+ add reference

(Wikidata no date b)

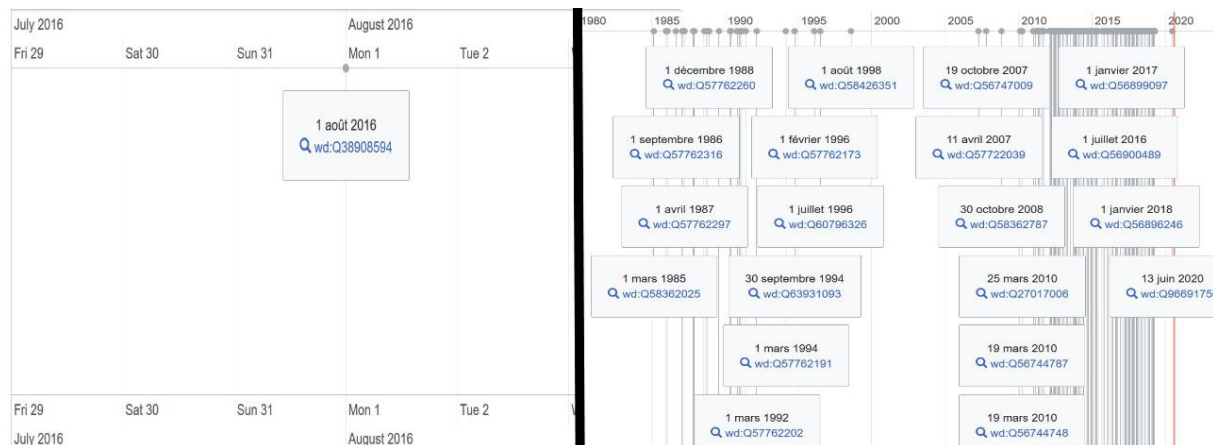
Regarding INSPIRE-HEP author ID (P2930) statements, these made up almost half of the 500 manual edits, and were done because over half the researchers linked to CERN on Wikidata did not have their `bai` referenced there, and that would mean the program not finding them at all.

Finally, a test batch was also uploaded to Quickstatements. The batch, created with a test run of the first deliverable, added 210 author (P50) statements to Wikidata about Tejinder S.

<sup>15</sup> See list at <https://www.wikidata.org/wiki/Special:Contributions/Vbarta> (Wikidata no date b).

Virdee. A comparison of timelines of articles on Wikidata known to be written by Virdee, before the upload (only one article, in 2016), and after the upload (211 articles, between 1985 and 2020), can be seen in Figure 25.

Figure 25: Comparison of Virdee articles before and after a batch upload



(Wikidata no date a)

#### 4.2.1.3 Second deliverable

The main element that needed to be determined for the second deliverable was the property that was to be used. There are five properties used on Wikidata to describe the relationship between a person and CERN: four that are barely ever used (member of (P463); work location (P937); educated at (P69) and affiliation (P1416)), and one that is used in most cases: employer (P108). As of July 2020, there are 549 people with Wikidata item pages who are linked to CERN with one (or several, in the case of Sir Tim Berners-Lee (Tim Berners-Lee (Q80) 2020) of these five properties. Although P108 (employer) is the one used in most cases on Wikidata, that might not be the best choice. Indeed, people working at CERN and publishing with a CERN affiliation are often not considered CERN employees. As it is extremely difficult to find out if a person is actually an employee, it was decided that it would be slightly more precise to use the affiliation (P1416) property.

#### 4.2.2 Results

A total of 853 edits (including the test batch uploads to Quickstatements) were done over the course of five months. As both deliverables are merely proofs of concept from the point of view of uploading and improving data on Wikidata, most batches created by the programs have yet to be uploaded (see sections 4.2.3 Limitations and 5.1 Assessment for more information).

#### 4.2.3 Limitations

Beyond the limitations already mentioned in section 4.1.3, there are a few additional ones linked to the very nature of proof of concept deliverables. Indeed, the first limitation on any concrete result for this part of the work is the fact that the batches created have not yet been uploaded. That step would be quite time consuming, as it would mean copying lines of text from 271 individual documents to Quickstatements, and the tool itself takes some time (between a few minutes and a few hours, depending on the size) to process each batch.

Another problem might arise during the upload. Although authorship of an article is almost never questioned, the same cannot be said of the affiliation to CERN. This is undoubtedly one of the big hurdles to adding the *People associated with CERN* tag on Wikipedia, as it often leads to heated debates. If Wikidata rarely matches Wikipedia in the violence of its debates, uploading almost 70 statements linking people to CERN might become the origin of a new debate, so one would have to be prepared to justify the choice.

The last and greatest hurdle, of course, regards assessment, but that point is discussed in the next section, 5.1 Assessment.

## 5. Discussion

### 5.1 Assessment

It is extremely difficult to assess the kind of impact such a project can have. The main reason for this is that the data on Wikidata is rarely viewed directly, but rather used on other websites, or in other projects, typically Wikipedia. This can be easily seen if one compares the Wikipedia page views and Wikidata item page views of T. S. Virdee. Sir Tejinder S. Virdee FRS has authored over 1000 articles, is a Fellow of both the Royal Society as well as the Institute of Physics (IOP), has worked at CERN for several decades, and is generally a very recognizable name in the HEP field. However, the statistics clearly show that if his Wikipedia page has an average of around 23 views a day (see Figure 26 for daily views between July 2019 and June 2020), his Wikidata item page averages less than 1 view per day for the last year (see Figure 27 for daily views between July 2019 and June 2020), even after uploading the test batch on 19 April 2020.

Figure 26: Daily views on Virdee’s Wikipedia page

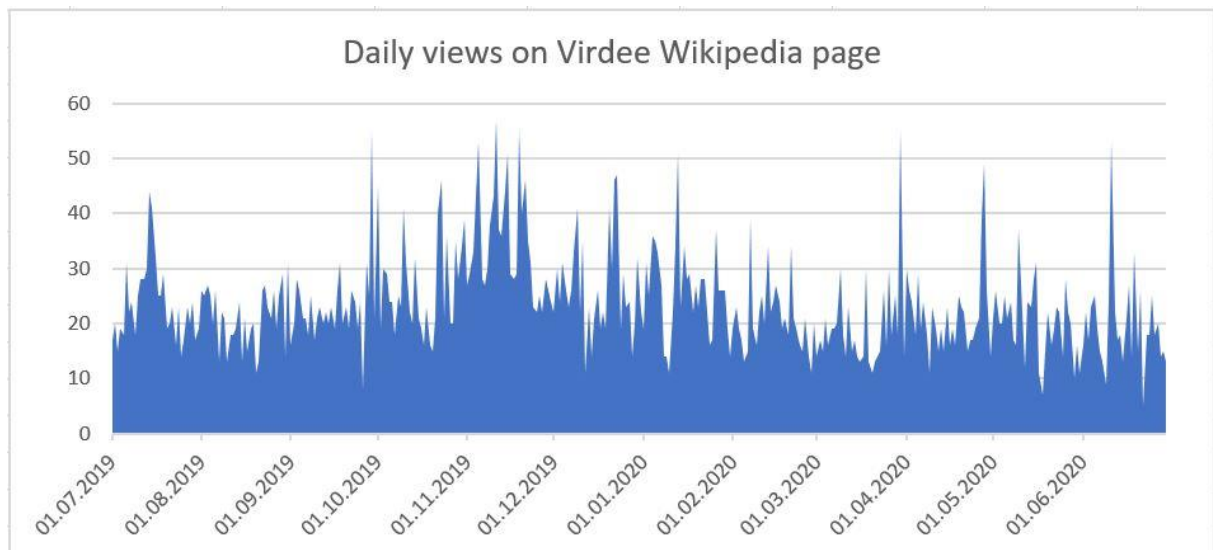
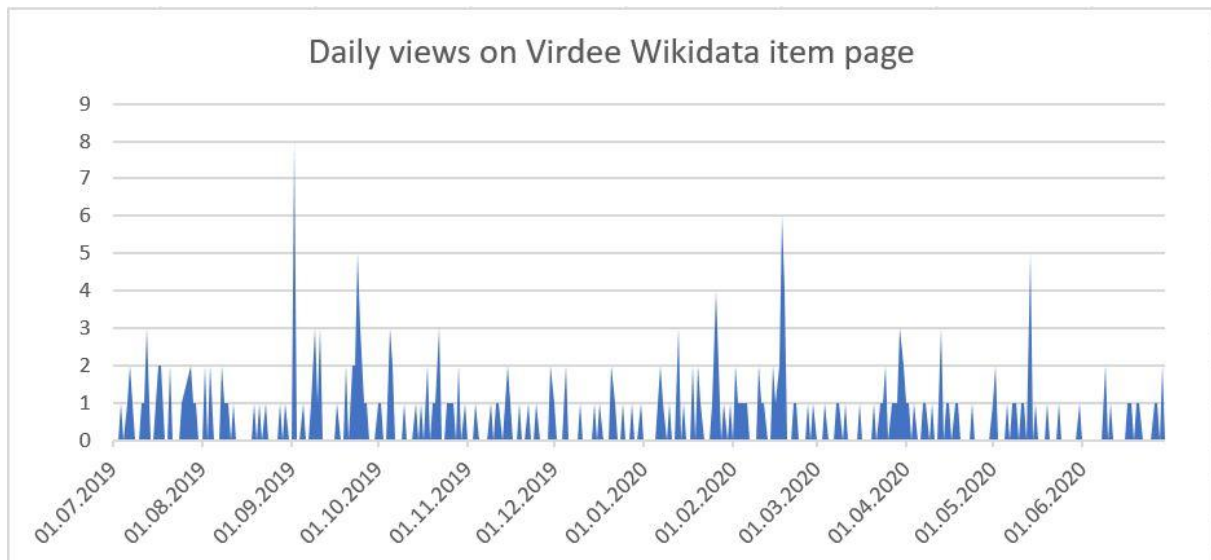


Figure 27: Daily views on Virdee's Wikidata item page



If it is close to impossible to assess the effect that adding specific amounts of data to Wikidata could have, one can try different other angles to approach the answer as closely as possible. One part of the answer is looking at the general value of the data on Wikidata, and especially how it can be used by other projects or websites. That subject is covered in section 5.2 Value. Another way to approach the problem is looking at general usage statistics of INSPIRE and comparing them to Wikidata in general as well as Scholia, a tool that uses data from Wikidata. If these different sites are not directly comparable, as INSPIRE is solely dedicated to the HEP field, whereas Scholia deals more generally with academia, and Wikidata, like most other Wikimedia projects, tries to cover all areas of human knowledge, it can still be interesting to look at some usage and involvement statistics. In 2019, INSPIRE had an average of 22K daily views, when Scholia had 11.7K (see Figure 28, although the tool has had more views recently, with daily views averaging more than 15K between July 2019 and June 2020 – see Figure 29) and Wikidata 6M (see Figure 30). As for unique visitors, INSPIRE had a monthly average of 180K, where Wikidata had a monthly average of 2M unique devices, with no data from Scholia.

Figure 28: Daily total views of Scholia tool

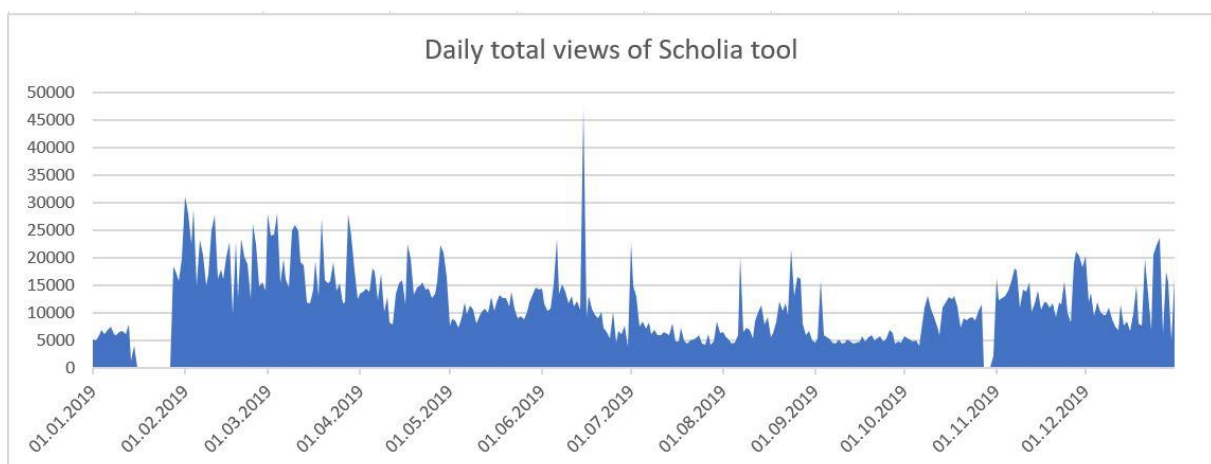




Figure 29: Daily total views of Scholia tool between July 2019 and June 2020

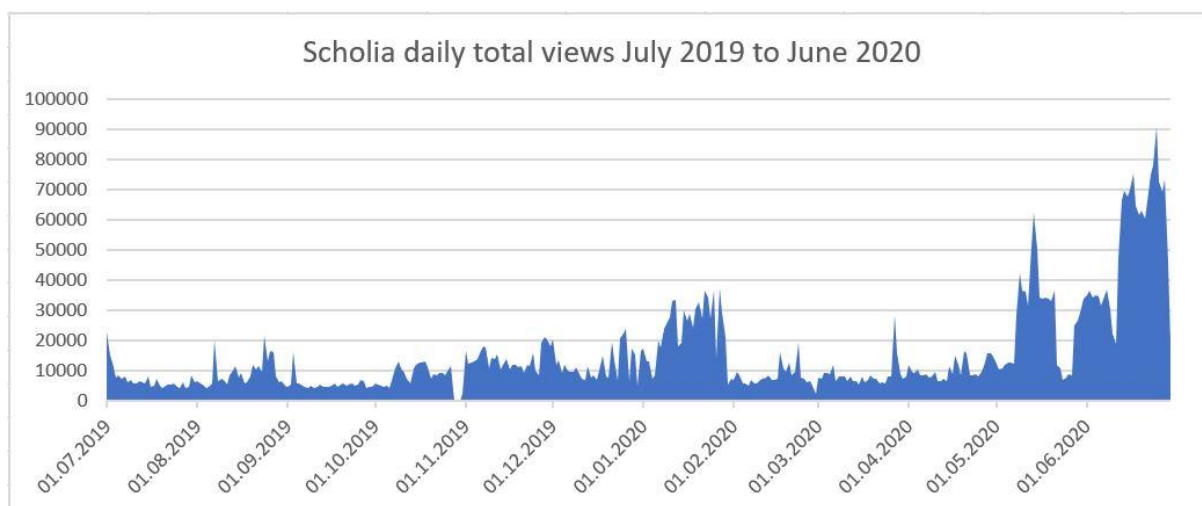
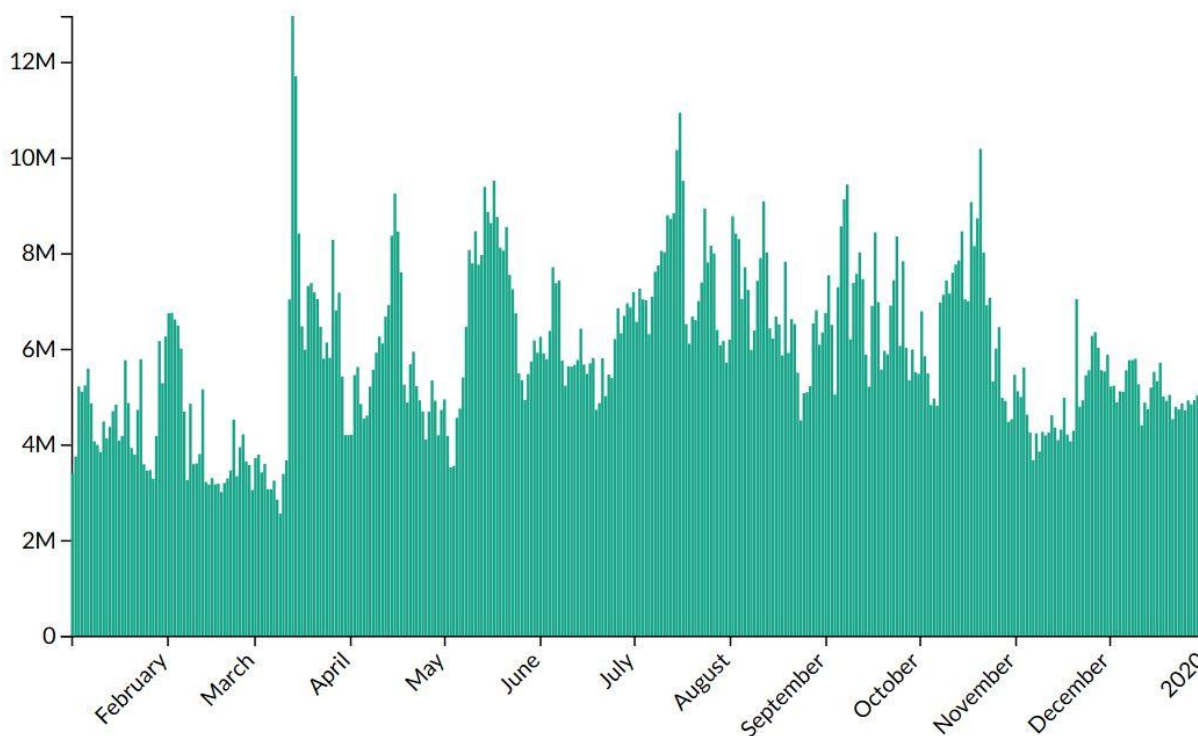


Figure 30: Daily total views of Wikidata



(Wikimedia Foundation 2020 b)

These numbers, together with the fact that “[...] site views are a tiny measure of the usage of Wikidata’s content, for a start because it is used to generate more than 2.5 million infoboxes for Commons, so the views on those profiles need to be counted as well”,<sup>16</sup> clearly show that Wikidata (and by extension Scholia as well, as it uses Wikidata’s data) has a much wider audience than INSPIRE. Even if that has a lot to do with the subjects covered, it is worth mentioning, as the data this project deals with could be interesting to more than just the HEP field.

<sup>16</sup> (Poulter 2020 a)

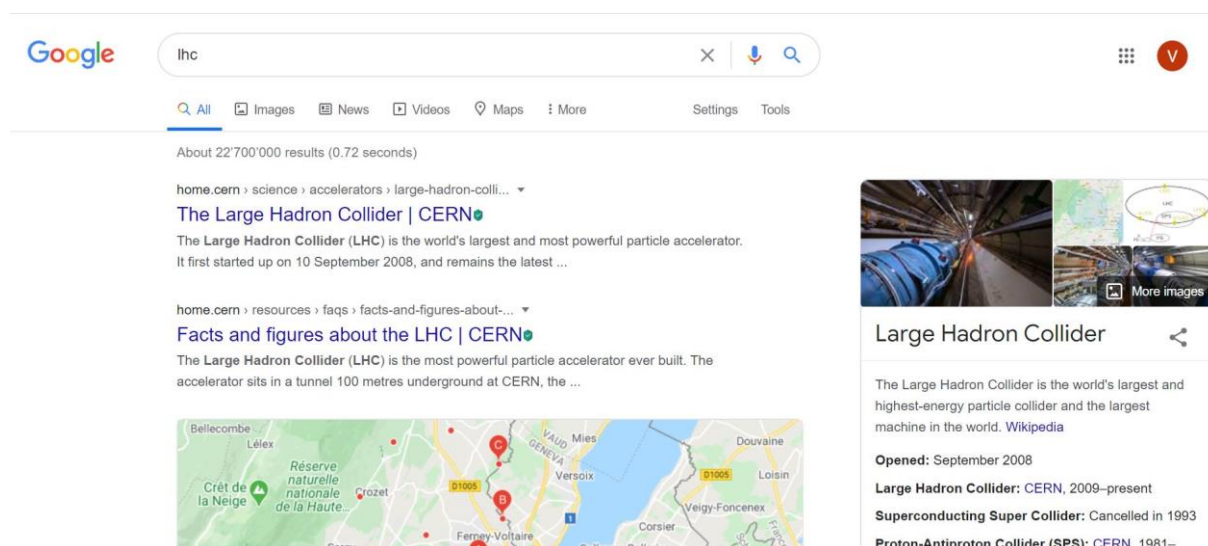
But of course, beyond the numbers and statistics, the value of this kind of work lies in less quantifiable elements, described in the next section.

## 5.2 Value

### 5.2.1 Google

One of the main ways that Wikidata's data is used is by Google. There are probably several ways they do this (they are famously secretive about their search algorithms), but one of the ways is with their Knowledge Graphs. Introduced in 2012, Google Knowledge Graphs offer what Google think is the most pertinent result or results (WikiNative 2019) either in the form of a Card (to the right of the top results – see Figure 31) or in the form of a Carousel (which appears above the top results – see Figure 32). These Graphs use, among others, public sources of data, such as Freebase<sup>17</sup> and Wikipedia (Singhal 2012).

Figure 31: Example of a Google Knowledge Graph Card



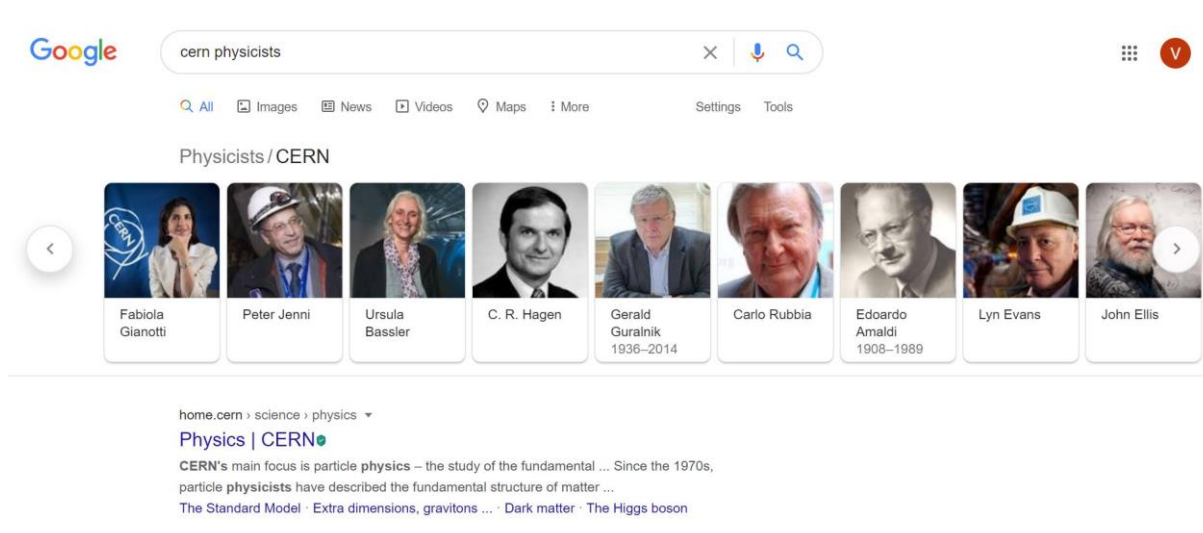
(Google 2020)

---

<sup>17</sup> Freebase was a collaborative knowledge base, launched in 2007, bought by Google in 2010 and shut down in 2016 (Freebase (database) 2020).



Figure 32: Example of a Google Knowledge Graph Carousel



(Google 2020)

*"Google had their own kind of linked open data project that they actually killed in 2016 because they saw what Wikidata was doing and said You folks are doing it faster, better, and on a scale we can't replicate, so that shows you how influential Wikidata has been. So Google uses the Wikidata graph database for a lot of what it does with the knowledge graph that returns content to you in the search form."* (Andrew Lih<sup>18</sup>)

*"\*Officially\* [Google] say they don't harvest from Wikidata, but some of us who add data to Wikidata and see rapid changes in the knowledge graph boxes know there must be some harvesting going on. [...] the guy who created Wikidata now works for Google and Google sponsor a lot of the development of Wikibase and associated events."*<sup>19</sup>

Google itself almost admits as much: "The best is to follow best practice in the editing of source material (Wikipedia, etc.) and this will eventually influence the Knowledge Graph organically."<sup>20</sup>

It seems that having data in Wikidata also helps with Search Engine Optimisation (SEO) generally speaking (WikiNative 2019).

### 5.2.2 Scholia

Another significant way in which Wikidata's data is used is by tools such as Scholia or Reasonator. Reasonator offers a service to display Wikidata item pages in a more reader-friendly way, as well as related items (i.e. items that are the subject in a statement where the original item is the object) (Wikimedia Toolforge no date b). Figures 22 and 23 are an example of such a visualisation by the Reasonator tool.

Scholia is a service that uses data from Wikidata to build profiles for researchers, institutions, funders and other scholarly topics (Wikimedia Toolforge no date c). CERN's profile on Scholia is very interesting to look at, for example. At the top of the page, Scholia takes the definition of CERN from Wikipedia (in this case, the English Wikipedia, but the language depends on the reader's preferences) and displays it at the top, as well as some related profiles on Scholia (see Figure 33).

<sup>18</sup> (OCLC 2018)

<sup>19</sup> (Poulter 2020 b)

<sup>20</sup> (Prosser 2020)

Figure 33: Top of CERN's Scholia profile

SCHOLIA

AuthorWorkOrganizationLocationEventProjectAwardTopicToolsHelp

Search...

authororganizationpublisherlocationsponsortopic

CERN (Q42944)

The European Organization for Nuclear Research (French: Organisation européenne pour la recherche nucléaire), known as CERN (; French pronunciation: [sɛʁn]; derived from the name Conseil européen pour la recherche nucléaire), is a European research organization that operates the largest particle physics laboratory in the world. Established in 1954, the organization is based in a northwest suburb of Geneva on the Franco-Swiss border and has 23 member states. Israel is the only non-European country granted full membership. ... (from the English Wikipedia)

Related: World Health Organization · UNESCO · European Space Agency · National Association for the Advancement of Colored People · NATO · Interpol · Académie des Sciences Morales et Politiques · Royal Academy of Dutch language and literature · Council on Foreign Relations · World Meteorological Organization

(Wikimedia Toolforge no date c)

After this top part, organisation profiles have different tables, graphs and other visualisations of the data linked to these organisations. In the case of CERN, the first table is of its employees and affiliates (see Figure 34), which is of course mostly based on the employer (P108) and affiliation (P1416) properties used to link people to CERN. At the bottom of every table or graph, Scholia links to the SPARQL query used to get the data. So for example, for the table that lists the topics that employees and affiliates have published on (see Figure 35), one can follow the link to Wikidata's query endpoint and get the results to the query with a different visualisation, such as a bubble chart (see Figure 36).

Figure 34: Employees and affiliates on CERN's Scholia profile

Employees and affiliated

Past and present employees, affiliated, and members

Show 10 entries

Search:

Works	Wikis	Researcher	Researcher description	Orcid
1474	0	<a href="#">Giovanni Marchiori</a>	chercheur et physicien	
1391	0	<a href="#">Guido Volpi</a>	researcher	<a href="#">0000-0003-1058-8883</a>
1300	0	<a href="#">Maurizio Pierini</a>	particle physicist at CERN	<a href="#">0000-0003-1939-4268</a>
1232	0	<a href="#">Domenico della Volpe</a>	researcher ORCID ID = 0000-0001-8530-7447	<a href="#">0000-0001-8530-7447</a>
1217	1	<a href="#">Paul Lecoq</a>	physicist	<a href="#">0000-0002-3198-0115</a>
1163	0	<a href="#">Francesco Cerutti</a>		
1145	0	<a href="#">Concezio Bozzi</a>	researcher	<a href="#">0000-0001-6782-3982</a>
1141	0	<a href="#">Frans Meijers</a>	onderzoeker	
1129	0	<a href="#">Olivier Schneider</a>	Swiss physicist, professor at the Swiss Federal Institute of Technology Lausanne (EPFL)	<a href="#">0000-0002-6014-7552</a>
1125	0	<a href="#">Philippe Farthouat</a>	ingénieur au CERN	<a href="#">0000-0002-4779-5432</a>

[Edit on query:Wikidata.org](#)

(Wikimedia Toolforge no date c)



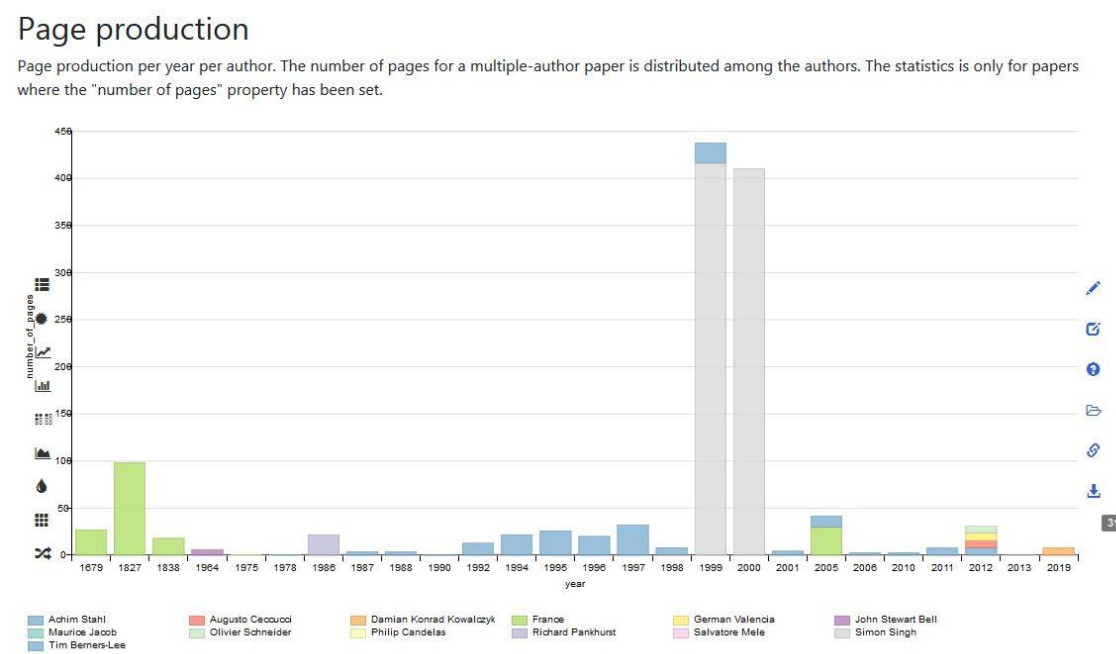
Scholia organisation profiles can also have awards tables (see Figure 37), page production bar charts (see Figure 38), as well as co-author graphs (see Figure 39 – co-author graph of John Richard Ellis<sup>21</sup>) or even gender statistics (see Figure 40).

Figure 37: Awards on CERN's Scholia profile

Awards		
Show	10	entries
		Search: <input type="text"/>
Count	Award	Recipients
8	<a href="#">Fellow of the Royal Society</a>	Tim Berners-Lee, Gavin Salam, John Adams, Erwin Gabathuler, John Dowell, Philip Candelas, John Ellis, John Stewart Bell
6	<a href="#">Enrico Fermi Prize</a>	Paolo Giubellino, Gabriele Veneziano, Dimitri Nanopoulos, Fabiola Gianotti, Ettore Fiorini, Sergio Ferrara
5	<a href="#">Nobel Prize in Physics</a>	Simon van der Meer, Georges Charpak, Jack Steinberger, Gerard 't Hooft, Samuel C. C. Ting
5	<a href="#">Fellow of the American Physical Society</a>	Victor Weisskopf, Dimitri Nanopoulos, Yannis K. Semertzidis, Maurice Jacob, Ann Nelson
4	<a href="#">Officer of the Legion of Honour</a>	Robert Gabillard, Robert Klapisch, Georges Charpak, Gerard 't Hooft
4	<a href="#">Dannie Heineman Prize for Mathematical Physics</a>	Gabriele Veneziano, Gerard 't Hooft, Sergio Ferrara, John Stewart Bell
4	<a href="#">Rutherford Medal and Prize</a>	Ken Peach, David E. Plane, Erwin Gabathuler, John Dowell
3	<a href="#">High Energy and Particle Physics Prize</a>	Tejinder Virdee, Georges Charpak, Gerard 't Hooft
3	<a href="#">Oskar Klein Medal</a>	Gabriele Veneziano, Valentine Telegdi, Gerard 't Hooft
3	<a href="#">John Simon Guggenheim Memorial Foundation Fellowship</a>	Victor Weisskopf, Ann Nelson, Jack Steinberger

(Wikimedia Toolforge no date c)

Figure 38: Page production on CERN's Scholia profile

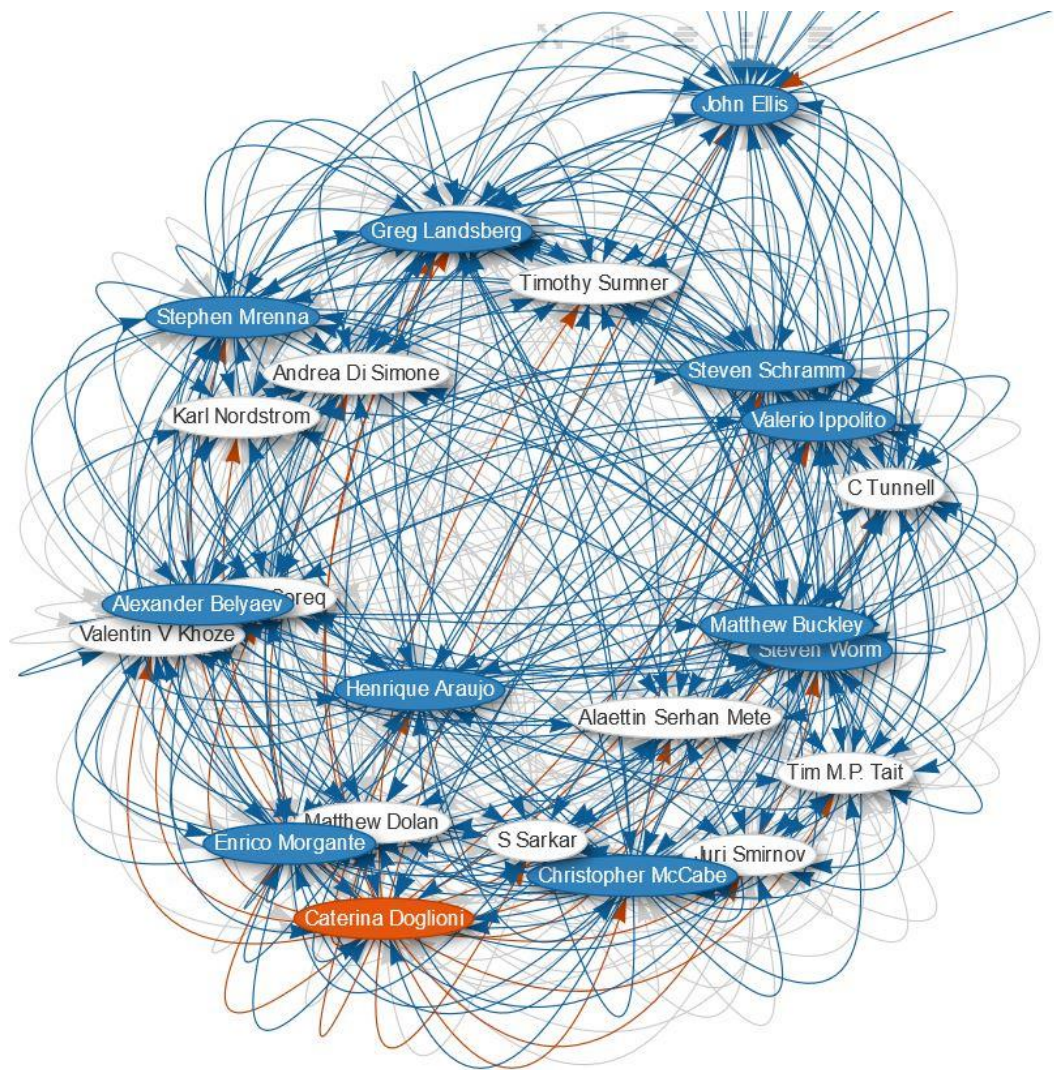


(Wikimedia Toolforge no date c)

<sup>21</sup> The example is from an author's profile, as the co-author graph for all of CERN employees times out because of the sheer amount of data the query tries to process.

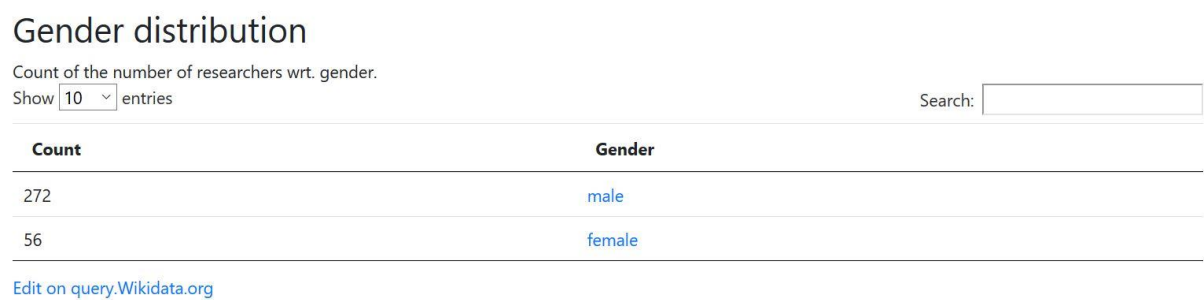


Figure 39: Detail of John Richard Ellis's co-author graph



(Wikimedia Toolforge no date c)

Figure 40: Gender statistics on CERN's Scholia profile

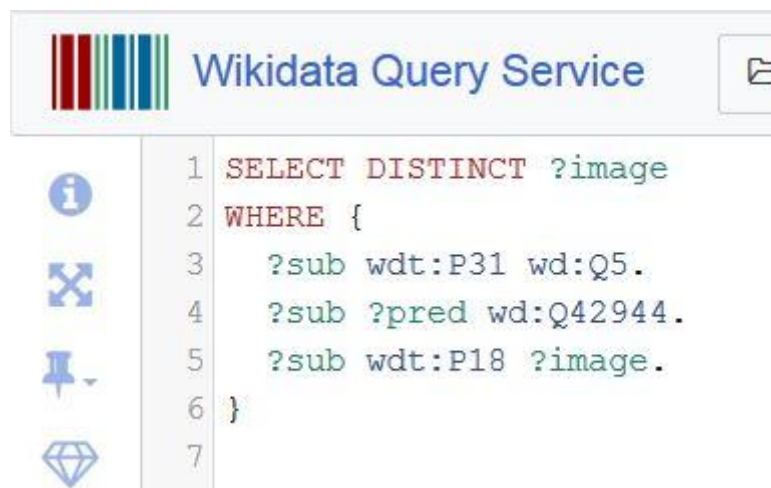


(Wikimedia Toolforge no date c)

These visualisations are the perfect representation of the value of the kind of work this project does. Without data about CERN and its people on Wikidata, these graphs and tables could not be generated by Scholia. If some of these visualisations could be created with INSPIRE's data with some (or probably a great deal of) effort, the process would probably be more complex, as the data is not in a semantic form, and the processes would have to be tailored to INSPIRE's

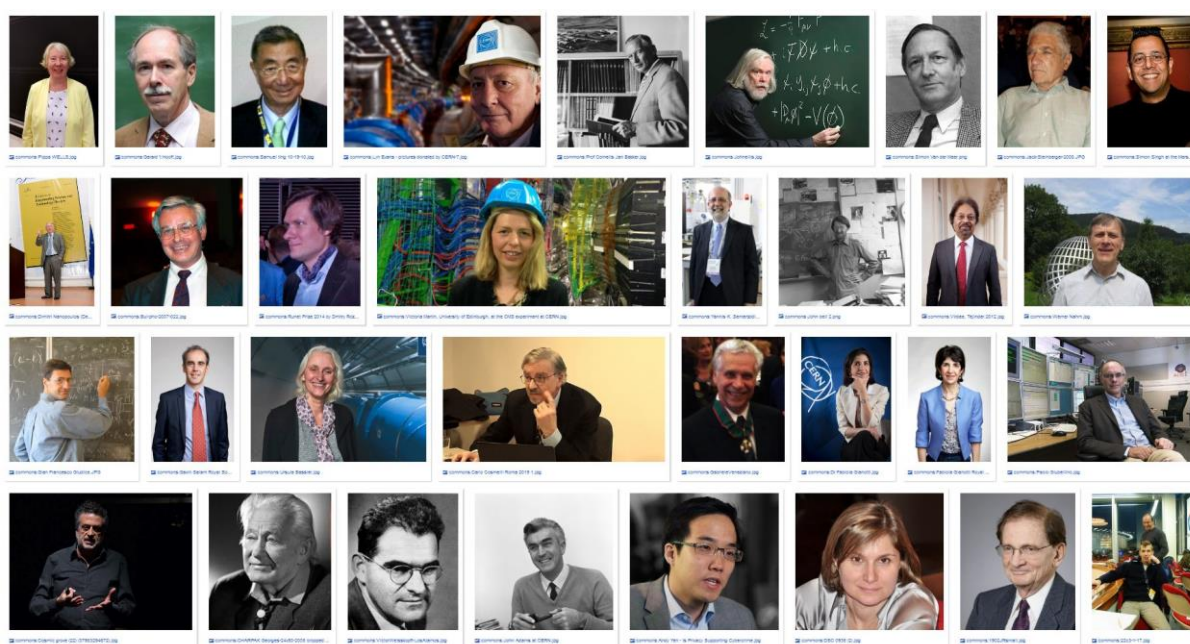
data specifically. Beyond these technical difficulties, INSPIRE also simply does not have all the data that Scholia uses, such as the gender of its researchers. And beyond just Scholia, the semantic links that exist on Wikidata allow one, with three simple lines of SPARQL (see Figure 41), to generate a collection of images of CERN people (see Figure 42).

Figure 41: SPARQL query to get images of people linked to CERN



(Wikidata no date a)

Figure 42: Results of the SPARQL query to get images of people linked to CERN



(Wikidata no date a)

### 5.2.3 Wikipedia

Finally, and this might seem obvious, Wikidata's data is used by Wikipedia a great deal. There is of course the straightforward use of the same data on multiple language Wikipedias, making sure that the current population of Rome is not different on every single Wikipedia, for example (to use the example favoured by Vrandečić and Krötzsch in their 2014 article *Wikidata: A Free Collaborative Knowledgebase*). But another advantage of pushing data or improving the data



already on Wikidata is that it is easier for smaller language Wikipedias (i.e. not English or German, but rather languages that have five or at most 10 editors contributing in all) to create many more articles than they could manage manually, by pulling data from Wikidata to create infoboxes (see Figure 43).

Figure 43: CERN's infobox on English Wikipedia

European Organization  
for Nuclear Research  
*Organisation européenne  
pour la recherche nucléaire*





CERN's main site, from [Switzerland](#) looking  
towards [France](#)



Member states

Formation

September 29, 1954;  
65 years ago<sup>[1]</sup>

Headquarters

[Meyrin](#), [Canton of Geneva](#),  
[Switzerland](#)

Membership

23 countries [\[show\]](#)

Official  
languages

[English](#) and [French](#)

Council  
President

[Ursula Bässler](#)<sup>[2]</sup>

[Director  
General](#)

[Fabiola Gianotti](#)

Website

[home.cern](#) 

(CERN 2020)

There is also a proposal that has just been approved (Abstract Wikipedia 2020) that proposes the creation of a Wikimedia Foundation sister project called Abstract Wikipedia (working title): this project would be similar to Wikidata, in being independent of any one language, but for whole text articles and not just the data underlying the articles. This idea has just been proposed, so the actual realisation is still quite far off, but it is something that should be watched closely, as the potential of this project is immense (Vrandečić 2020).

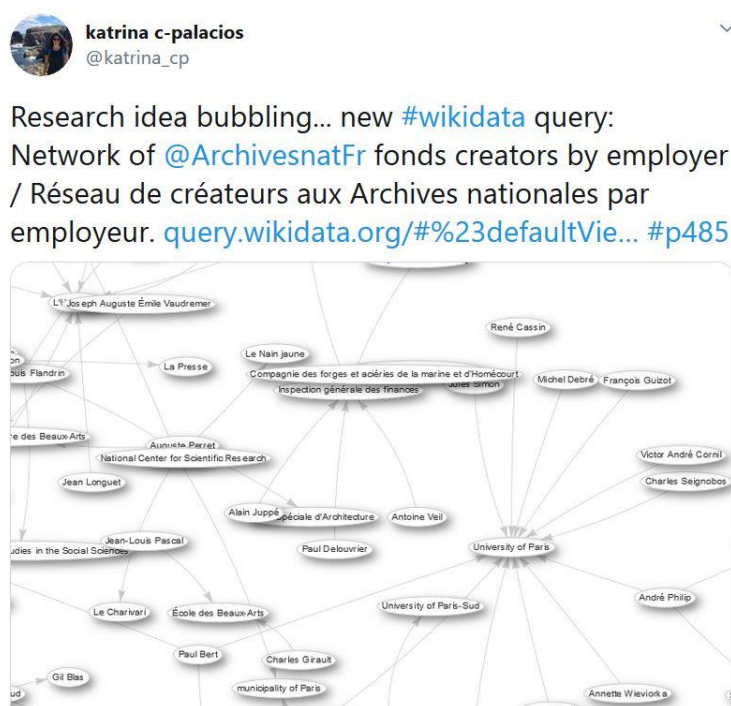
## 5.3 Future Work

During the course of this project, a few ideas were considered that had to be put aside for lack of time or competence. These ideas, as well as simply the possible continuation of the work already carried out, will be detailed here, with the hope that they might be taken up in the future by RCS-SIS.

### 5.3.1 Continuation and communication

As a continuation to this project, the first step should of course be to upload the batches created by the two programs to Wikidata with Quickstatements. But once that is done, one important step not covered in this project would be to communicate about it. Indeed, as Wikidata is not often viewed directly, communicating about a new (or more complete, in this case) dataset would let Wikipedians know that there is new data they can pull to their own language Wikipedias. Martin Poulter suggests writing a SPARQL query for the new dataset, and tweeting the results of the query, making sure that the Wikidata people see the information, and can include it in a newsletter that goes out to the whole Wikimedia community. In this instance, that could mean tweeting the query result for people linked to CERN, mentioning that the dataset has just been increased by 12% (see an example of a tweet of a query in Figure 44).

Figure 44: Example of a tweet about a Wikidata query



(Palacios 2019)



It would also be extremely helpful not only to add statements to already existing items (as the work described here concentrated on) but also creating new items. This means not only items representing people or articles (although articles are one very important dataset that would greatly benefit from additions), since it would be interesting to look into other datasets, such as CERN experiments or CERN discoveries, for example.

### 5.3.2 Crossref

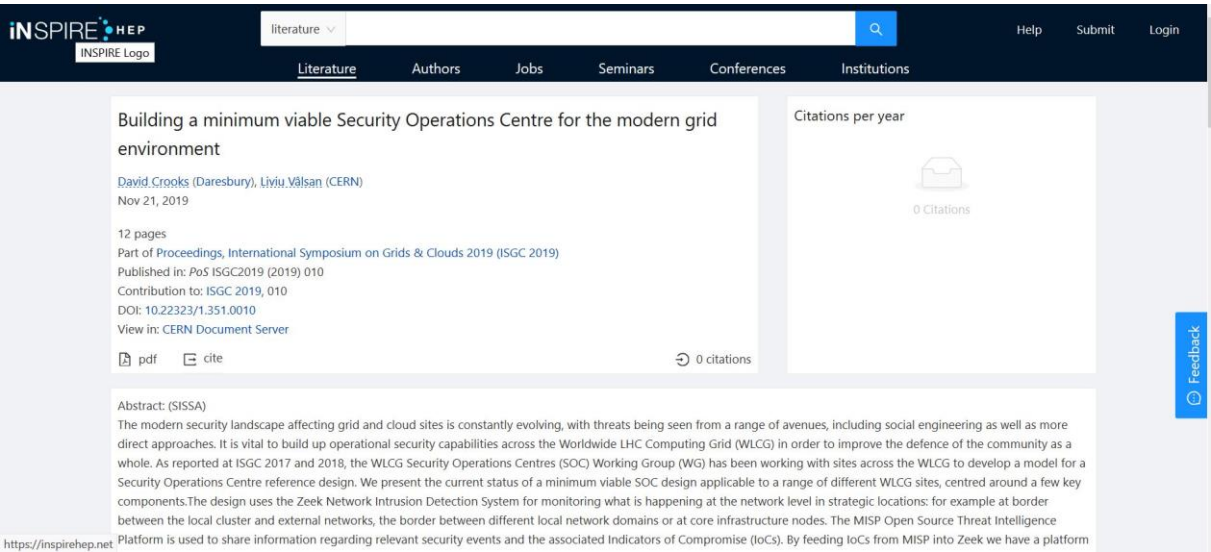
As mentioned previously, Crossref is a not-for-profit organisation whose main goal is to make scholarly communications better. Although Crossref runs an open infrastructure and involves the academic community in the work as much as possible, it is not as easy to contribute data to its database as for collaborative projects such as Wikidata. However, it would be greatly beneficial to find a way for CERN's SIS to do just that, as Crossref's metadata on HEP literature is not as complete as INSPIRE's. For example, for the same article, Figure 45 clearly shows that Crossref does not have the affiliation metadata nor an abstract, whereas INSPIRE (see Figure 46) even has the references of the article, as well as a link to the full text on CDS.

Figure 45: Example of an article on Crossref



(Crossref no date a)

Figure 46: Example of an article on INSPIRE



(INSPIREHEP 2020)

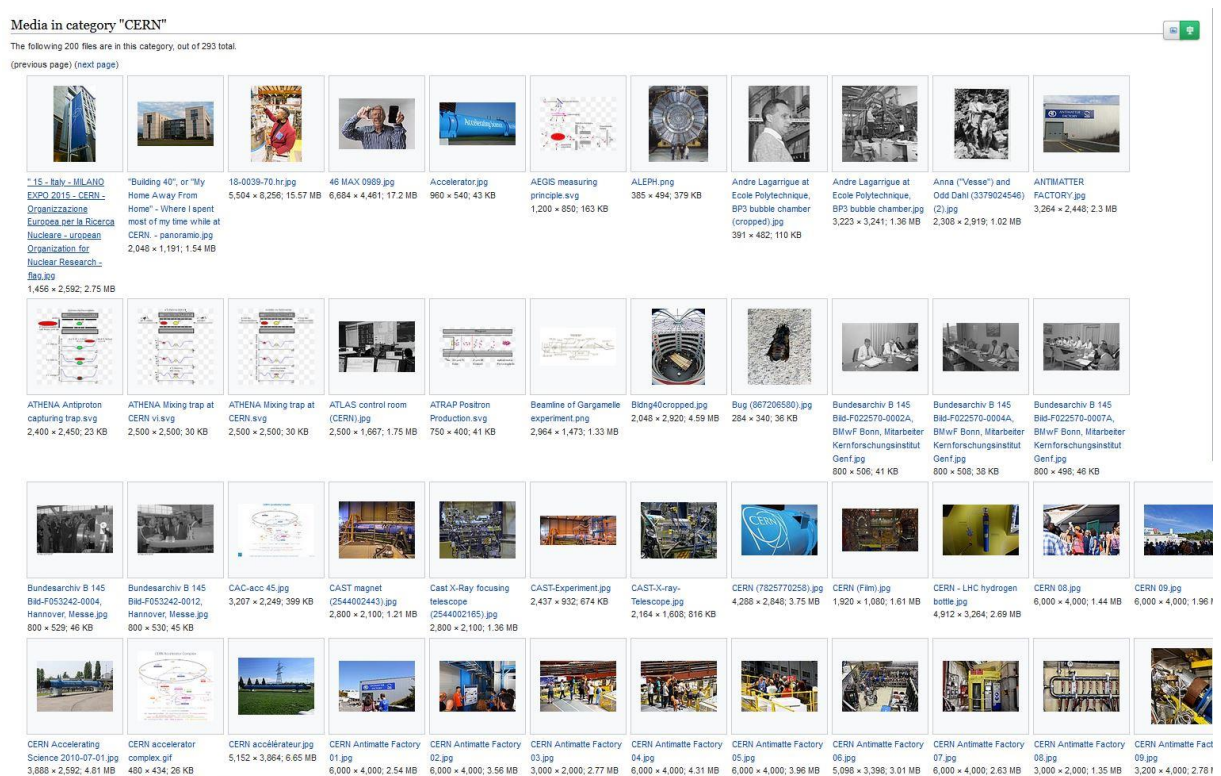
### 5.3.3 Wikimedia Commons

Last, but certainly not least, CERN's SIS should not omit Wikimedia Commons. As mentioned in section 4.2, Wikimedia Commons is a collaborative database of freely reusable media files. On CDS, CERN has over 18K photos of people, experiments, exhibitions, etc. as well as a trove of fascinating archive photos and videos. According to Martin Poulter:

*"I've done work with museums [...] and a common finding is that 50 to 100 times as many people are encountering the museum's collections on Wikipedia as they are on the museum's own website [...] so the main statistic for the reach of these institutions is the number that comes from BaGLAMa<sup>22</sup>."*<sup>23</sup>

So not only would there be intrinsic value in uploading photos and videos to Wikimedia Commons, but it would make it possible to use or link to those photos from Wikipedia, Wikidata and other websites, which would in turn make CERN more discoverable online.

Figure 47: Detail of the media in the CERN Wikimedia Commons Category



(Category:CERN 2020)

<sup>22</sup> BaGLAMa is a tool used to get page view numbers for pages on Wikimedia Foundation projects that contain Wikimedia Commons files from a specific category.

<sup>23</sup> Interview with Martin Poulter, Wikimedian, Skype, 11 March 2020.

## 6. Conclusion

CERN's history clearly shows its desire to be as open as possible. From its founding convention to its latest discovery of a new type of tetraquark (CERN 2020 g), CERN shows this intention again and again. However, an intention is not necessarily enough. Too much information about CERN and its people is still in silos, like CDS or INSPIRE, and that definitely does not help CERN's discoverability online. So what does?

Obviously, developing software that allows openly sharing literature (INSPIRE), data (HEPData), analyses (CAP) etc., as well as setting up ground-breaking agreements to make all publications in the HEP field OA (SCOAP<sup>3</sup>), or even working on better-quality OA (YRs), all contribute towards the goal of openness. But this thesis argues that another way of working on this goal is through collaborative semantic projects, such as Wikidata.

Although it was explained how hard it is to determine the concrete effects of this kind of work, it is very likely that improving the semantic data on Wikidata helps with SEO (Google Knowledge Graphs), visualisation (Scholia) and discoverability in general, as Wikidata and especially Wikipedia have a much wider audience than any CERN project.

I am absolutely convinced that every academic organisation such as CERN, or even smaller ones, needs to invest more resources into these kinds of crowdsourcing projects (such as those of the Wikimedia Foundation), as the semantic Web is only the stepping stone for the next version of the Web, and institutions need to join this process before it leaves them behind completely. Hopefully, this thesis shows a relatively simple way of joining the movement, and perhaps going beyond it.

# Bibliography

Abstract Wikipedia. *Wikimedia* [online]. Last modification of the page on 13 July 2020, at 16:19. [Viewed 15 July 2020]. Available from: [https://meta.wikimedia.org/w/index.php?title=Abstract\\_Wikipedia&oldid=20278161](https://meta.wikimedia.org/w/index.php?title=Abstract_Wikipedia&oldid=20278161)

ALEXA INTERNET, 2020. wikipedia.org: Competitive Analysis, Marketing Mix and Traffic. *Alexa* [online]. [Viewed 15 July 2020]. Available from: <https://www.alexa.com/siteinfo/wikipedia.org>

ALPERT, Jesse and HAJAJ, Nissan, 2008. We knew the web was big.... *Google Official Blog* [online]. 25 July 2008. [Viewed 15 July 2020]. Available from: <https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>

BERNERS-LEE, Tim, HENDLER, James and LASSILA, Ora, 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* [online]. May 2001. [Viewed 15 July 2020]. Available from: [https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American\\_%20Feature%20Article\\_%20The%20Semantic%20Web\\_%20May%202001.pdf](https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf)

bibliography (Q1631107). *Wikidata* [online]. Last modification of the page on 21 June 2020, at 14:22. [Viewed 15 July 2020]. Available from: <https://www.wikidata.org/w/index.php?title=Q1631107&oldid=1213286398>

Category:CERN. *Wikimedia Commons* [online]. Last modification of the page on 18 March 2020, at 07:52. [Viewed 15 July 2020]. Available from: <https://commons.wikimedia.org/w/index.php?title=Category:CERN&oldid=404973164>

CERN. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 11 July 2020, at 10:14. [Viewed 15 July 2020]. Available from: <https://en.wikipedia.org/w/index.php?title=CERN&oldid=967135698>

CERN, 2015. Discovery of pentaquarks. *CERN* [online]. 14 July 2015. [Viewed 15 July 2020]. Available from: <https://timeline.web.cern.ch/discovery-pentaquarks>

CERN, 2020 a. *CERN* [online]. [Viewed 15 July 2020]. Available from: <https://home.cern/>

CERN, 2020 b. Our history. *CERN* [online]. [Viewed 15 July 2020]. Available from: <https://home.cern/about/who-we-are/our-history>

CERN, 2020 c. Our Member States. *CERN* [online]. [Viewed 15 July 2020]. Available from: <https://home.cern/about/who-we-are/our-governance/member-states>

CERN, 2020 d. Our Mission. *CERN* [online]. [Viewed 15 July 2020]. Available from: <https://home.cern/about/who-we-are/our-mission>

CERN, 2020 e. Open source for open science. *CERN* [online]. [Viewed 15 July 2020]. Available from: <https://home.cern/science/computing/open-source-open-science>

CERN, 2020 f. *CERN Analysis Preservation* [online]. [Viewed 15 July 2020]. Available from: <https://analysispreservation.cern.ch/login?next=/>

CERN, 2020 g. LHCb discovers a new type of tetraquark at CERN. *CERN* [online]. 1 July 2020. [Viewed 15 July 2020]. Available from: <https://home.cern/news/news/physics/lhcb-discovers-new-type-tetraquark-cern>

- CERN DOCUMENT SERVER, 2020. *CERN Document Server* [online]. 15 July 2020. [Viewed 15 July 2020]. Available from: <https://cds.cern.ch/>
- CERN Scientific Information Service, [no date a]. CERN Archive. *library.cern* [online]. [Viewed 15 July 2020]. Available from: [http://library.cern/archives/CERN\\_archive](http://library.cern/archives/CERN_archive)
- CERN Scientific Information Service, [no date b]. Our mission. *library.cern* [online]. [Viewed 15 July 2020]. Available from: [http://library.cern/about\\_us/mission](http://library.cern/about_us/mission)
- CHARDONNENS, Alain, 2019. *Valorisation des notices de l'inventaire d'une archive audiovisuelle en Linked Open Data : le cas du Montreux Jazz Digital Project* [online]. Geneva: Haute École de Gestion de Genève. Bachelor thesis. [Viewed 15 July 2020]. Available from: <https://doc.rero.ch/record/327836?ln=fr>
- CHRISTODOULAKI, Stella, 2020. About INSPIRE. *INSPIREHEP* [online]. 4 March 2020. [Viewed 15 July 2020]. Available from: <https://inspirehep.net/help/knowledge-base/about-inspire/>
- CLEGG, Melissa, 2020. INSPIRE Author list tools. *INSPIREHEP* [online]. 30 March 2020. [Viewed 15 July 2020]. Available from: <https://inspirehep.net/help/knowledge-base/inspire-author-list-tools/>
- CROSSREF, [no date a]. *Crossref* [online]. [Viewed 15 July 2020]. Available from: <https://www.crossref.org/>
- CURE, Olivier and BLIN, Guillaume, 2014. *RDF database systems: triples storage and SPARQL query processing*. Waltham, MA : Morgan Kaufmann. 9780128004708
- DE KUNDER, Maurice, 2020. *The size of the World Wide Web (The Internet)* [online]. 15 July 2020. [Viewed 15 July 2020]. Available from: <https://www.worldwidewebsite.com/>
- DUBLIN CORE METADATA INITIATIVE, 2020. DCMI Metadata Terms. *Dublin Core Metadata Initiative* [online]. 20 January 2020. [Viewed 15 July 2020]. Available from: <http://purl.org/dc/terms/>
- EUROPEAN COUNCIL FOR NUCLEAR RESEARCH DRAFTING COMMITTEE, 1953. *Re-Draft of the Draft Convention establishing a European Organization for Nuclear research*. 14 January 1953. Available from: <https://cds.cern.ch/record/25966?ln=fr>
- Freebase (database). *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 6 July 2020, at 02:18. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=Freebase\\_\(database\)&oldid=966261367](https://en.wikipedia.org/w/index.php?title=Freebase_(database)&oldid=966261367)
- FREYA, [no date]. *Project Freya* [online]. [Viewed 15 July 2020]. Available from: <https://www.project-freya.eu/Plone/en>
- GOOGLE, 2020. *Google* [online]. [Viewed 15 July 2020]. Available from: <https://www.google.com/>
- HEPDATA, 2020. *HEPData* [online]. [Viewed 15 July 2020]. Available from: <https://www.hepdata.net/>
- History of Wikipedia. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 7 July 2020, at 06:32. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=History\\_of\\_Wikipedia&oldid=966454388](https://en.wikipedia.org/w/index.php?title=History_of_Wikipedia&oldid=966454388)



INSPIRE-HEP author ID (P2930). *Wikidata* [online]. Last modification of the page on 16 November 2019, at 16:20. [Viewed 15 July 2020]. Available from: <https://www.wikidata.org/w/index.php?title=Property:P2930&oldid=1053898593>

INSPIREHEP, 2020. *INSPIREHEP* [online]. [Viewed 15 July 2020]. Available from: <https://inspirehep.net/>

KENSHO TECHNOLOGIES LLC, 2019. *qwikidata* [online]. [Viewed 15 July 2020]. Available from: <https://qwikidata.readthedocs.io/en/stable/index.html>

KONSTANTINOOU, Nikolaos and SPANOS, Dimitrios-Emmanuel, 2015. *Materializing the web of linked data*. Cham: Springer. 978-3-319-16073-3

KRITSCHMAR, Charlie, 2016. File:Datamodel in Wikidata.svg. *Wikimedia Commons* [online]. 21 June 2016. Updated 25 June 2020. [Viewed 15 July 2020]. Available from: <https://commons.wikimedia.org/w/index.php?curid=49616867>

LINKED OPEN VOCABULARIES, 2020. *Linked Open Vocabularies (LOV)* [online]. [Viewed 15 July 2020]. Available from: <https://lov.linkeddata.es/dataset/lov>

MICHAMOS and JACQUERIE, 2020. inspirehep / rest-api-doc. *GitHub* [online]. 26 June 2020 13:53. [Viewed 15 July 2020]. Available from: <https://github.com/inspirehep/rest-api-doc>

NOWACK, Benjamin, 2009. The Semantic Web - Not a piece of cake... *Bnode* [online]. 8 July 2009, 14:55. [Viewed 15 July 2020]. Available from: <http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake>

OCLC, 2018. Works in Progress Webinar: Introduction to Wikidata for Librarians [video recording]. OCLC [online]. 12 June 2018. [Viewed 15 July 2020]. Available from: <https://www.oclc.org/research/events/2018/06-12.html>

Ontology (information science). *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 14 July 2020, at 17:03. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=Ontology\\_\(information\\_science\)&oldid=967680257](https://en.wikipedia.org/w/index.php?title=Ontology_(information_science)&oldid=967680257)

OPEN DATA CERN, 2020. *CERN Open Data Portal* [online]. [Viewed 15 July 2020]. Available from: <http://opendata.cern.ch/docs/about>

PALACIOS, Katrina, 2019. @katrina\_cp. Research idea bubbling.... *Twitter* [online]. 31 May 2019, 18:04. [Viewed 15 July 2020]. Available from: [https://twitter.com/katrina\\_cp/status/1134490728426549248](https://twitter.com/katrina_cp/status/1134490728426549248)

POULTER, Martin, 2020 a. Re: Wikidata [electronic message]. 17 June 2020.

POULTER, Martin, 2020 b. Re: Wikidata [electronic message]. 4 March 2020.

PROSSER, Matthew, 2020. Re: Improving CERN information on Wikidata and Google [electronic message]. 13 March 2020.

RAGGETT, Dave, LAM, Jenny and ALEXANDER, Ian, 1996. *HTML 3: Electronic Publishing on the World Wide Web*. Harlow: Addison-Wesley. 0201876930

Representational state transfer. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 14 July 2020, at 07:24. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=Representational\\_state\\_transfer&oldid=967606301](https://en.wikipedia.org/w/index.php?title=Representational_state_transfer&oldid=967606301)

SINGHAL, Amit, 2012. Introducing the Knowledge Graph: things, not strings. *Google Official Blog* [online]. 16 May 2012. [Viewed 15 July 2020]. Available from: <https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>

SMITH, Tim and FLÜCKIGER, François, [no date]. Licensing the Web. *CERN* [online]. [Viewed 15 July 2020]. Available from: <https://home.cern/science/computing/birth-web/licensing-web>

TED, 2009. Tim Berners-Lee: The next Web of open, linked data [video recording]. *YouTube* [online]. 13 March 2009. [Viewed 15 July 2020]. Available from: [https://youtu.be/OM6XIIICm\\_qo](https://youtu.be/OM6XIIICm_qo)

Tim Berners-Lee (Q80). *Wikidata* [online]. Last modification of the page on 9 July 2020, at 05:51. [Viewed 15 July 2020]. Available from: <https://www.wikidata.org/w/index.php?title=Q80&oldid=1227674432>

UNIVERSITE DE GENEVE, 2020. «L'argent public doit aller à la recherche, pas aux éditeurs». *Le Journal* [online]. 28 May 2020. [Viewed 15 July 2020]. Available from: <https://www.unige.ch/lejournalejournal/ejournal-08/negociations-editeurs/>

VRANDEČIĆ, Denny and KROETZSCH, Markus, 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* [online]. September 2014. Volume 57, Issue 10. [Viewed 15 July 2020]. Available from: <https://dl.acm.org/doi/10.1145/2629489>

W3C, 2014. *RDF 1.1 Primer* [online]. 24 June 2014. [Viewed 15 July 2020]. Available from: <https://www.w3.org/TR/rdf11-primer/>

WALES, Jimbo, 2020. User:Jimbo Wales/Statement of principles. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 2 July 2020, at 10:15. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=User:Jimbo\\_Wales/Statement\\_of\\_principles&oldid=965608752](https://en.wikipedia.org/w/index.php?title=User:Jimbo_Wales/Statement_of_principles&oldid=965608752)

Web 2.0. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 6 July 2020 at 11:49. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=Web\\_2.0&oldid=966316447](https://en.wikipedia.org/w/index.php?title=Web_2.0&oldid=966316447)

WIKIMEDIA COMMONS, 2020. *Wikimedia Commons* [online]. 17 May 2020. [Viewed 15 July 2020]. Available from: [https://commons.wikimedia.org/w/index.php?title=Main\\_Page&oldid=419734877](https://commons.wikimedia.org/w/index.php?title=Main_Page&oldid=419734877)

Wikimedia Foundation. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 14 July 2020, at 16:18. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=Wikimedia\\_Foundation&oldid=967674325](https://en.wikipedia.org/w/index.php?title=Wikimedia_Foundation&oldid=967674325)

WIKIMEDIA FOUNDATION, 2020 a. *Wikimedia Foundation* [online]. [Viewed 15 July 2020]. Available from: <https://wikimediafoundation.org/>

WIKIMEDIA FOUNDATION, 2020 b. Wikidata: Monthly overview. *Wikimedia Statistics* [online]. [Viewed 15 July 2020]. Available from: <https://stats.wikimedia.org/#/wikidata.org>

WIKIMEDIA TOOLFORGE, [no date a]. *QuickStatements* [online]. [Viewed 15 July 2020]. Available from: <https://quickstatements.toolforge.org/#/>

WIKIMEDIA TOOLFORGE, [no date b]. *Reasonator* [online]. [Viewed 15 July 2020]. Available from: <https://reasonator.toolforge.org/>

WIKIMEDIA TOOLFORGE, [no date c]. *Scholia* [online]. [Viewed 15 July 2020]. Available from: <https://scholia.toolforge.org/>

WIKINATIVE, 2019. Why is Wikidata essential to search engine results? *WikiNative* [online]. 5 June 2019. [Viewed 15 July 2020]. Available from: <https://www.wikinative.com/why-is-wikidata-essential-to-search-engine-results/>

Wikipedia. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 14 July 2020, at 18:56. [Viewed 15 July 2020]. Available from: <https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=967695346>

Wikipedia:Five pillars. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 22 June 2020, at 08:16. [Viewed 15 July 2020]. Available from: [https://en.wikipedia.org/w/index.php?title=Wikipedia:Five\\_pillars&oldid=963870876](https://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=963870876)

Wikidata. *Wikipedia, The Free Encyclopedia* [online]. Last modification of the page on 9 July 2020, at 10:19. [Viewed 15 July 2020]. Available from: <https://en.wikipedia.org/w/index.php?title=Wikidata&oldid=966812640>

WIKIDATA, [no date a]. *Wikidata Query Service* [online]. [Viewed 15 July 2020]. Available from: <https://query.wikidata.org/>

WIKIDATA, [no date b]. *Wikidata* [online]. [Viewed 15 July 2020]. Available from: [https://www.wikidata.org/w/index.php?title=Wikidata:Main\\_Page&oldid=1086709037](https://www.wikidata.org/w/index.php?title=Wikidata:Main_Page&oldid=1086709037)