How to Automate Wikipedia

Michael Upshall, Consult MU – "Enhancing Digital Publishing"

michael@consultmu.co.uk

Nowadays, Wikipedia is so ubiquitous that when you use the Web, you don't even have to look for it – Wikipedia comes to you. It is well known that Google search is deliberately adjusted to ensure that hits from Wikipedia are shown high up in search results, in fact so high that SEO experts wonder how it is achieved. (See, for example "Why Wikipedia is top on Google: the SEO truth no-one wants to hear.") But not so well known is the process that goes on in the background to create Wikipedia content. You may think that this work is largely manual, and there are indeed tens of thousands of enthusiastic and committed volunteers who maintain and compile Wikipedia entries, but there is also an increasing amount of automation providing ways in which the process can be speeded up, and, as you might expect, some disagreement over the best route that automation should take.

Given the scale of Wikipedia, it's not surprising there have been attempts to speed up the process of creating the world's largest encyclopaedia. One short cut was simply to cannibalise older encyclopaedias, for example to cut and paste from the latest available edition of Encyclopaedia Britannica that is in the public domain – that is, the 11th edition, published 1911. Many of Wikipedia's entries for older biographies contain content from this work. The work of transferring content was done entirely manually, as far as I know. In the last few years, much work has been done to try to improve Wikipedia's links to and from other reference resources. This comprises both enabling external resources to link with the relevant Wikipedia entry, as well as automatically delivering Wikipedia content to third parties.

Why do links matter? The problem is that of disambiguation – there are many "John Smith" entries in Wikipedia, certainly more than 125 in the English-language edition alone, and it would be very helpful to make sure which John Smith is which. Wikipedia helpfully classifies all its John Smiths by activity (such as "politician", "criminal" or "writer"), but even that doesn't solve the problem of linking.

One of the visions of the Semantic Web is to be able to link the right John Smith – as simple as that. The vision of Tim Berners-Lee seems to strike a chord in some people's minds, and they immediately embark on a campaign to fix the ambiguous entry problem.

However willing the volunteers, there is a limit to their appetite for low-level linking, and behind the scenes, much work is being done to attempt to automate the linking of Wikipedia content to other resources.

Probably the best-known example of providing content outwards is DBpedia. This is an initiative that came not from the Wikipedia founders themselves, but from universities in Germany (Berlin and Leipzig). Its goal was (and is) quite simply to create out of Wikipedia,

a resource created largely manually, a machine-readable resource. Using automated tools, for example, it's not difficult to write a script that extracts a fact such as "Buenos Aires has a population of 2.89 million" from Wikipedia and to make it available for other devices to use, without any human intervention. The magic is of course that the link is dynamic, rather than static – DBpedia is re-extracted from Wikipedia every 24 hours or so, which means that when the population figures for Buenos Aires are updated, DBpedia would output this change within 24 hours.

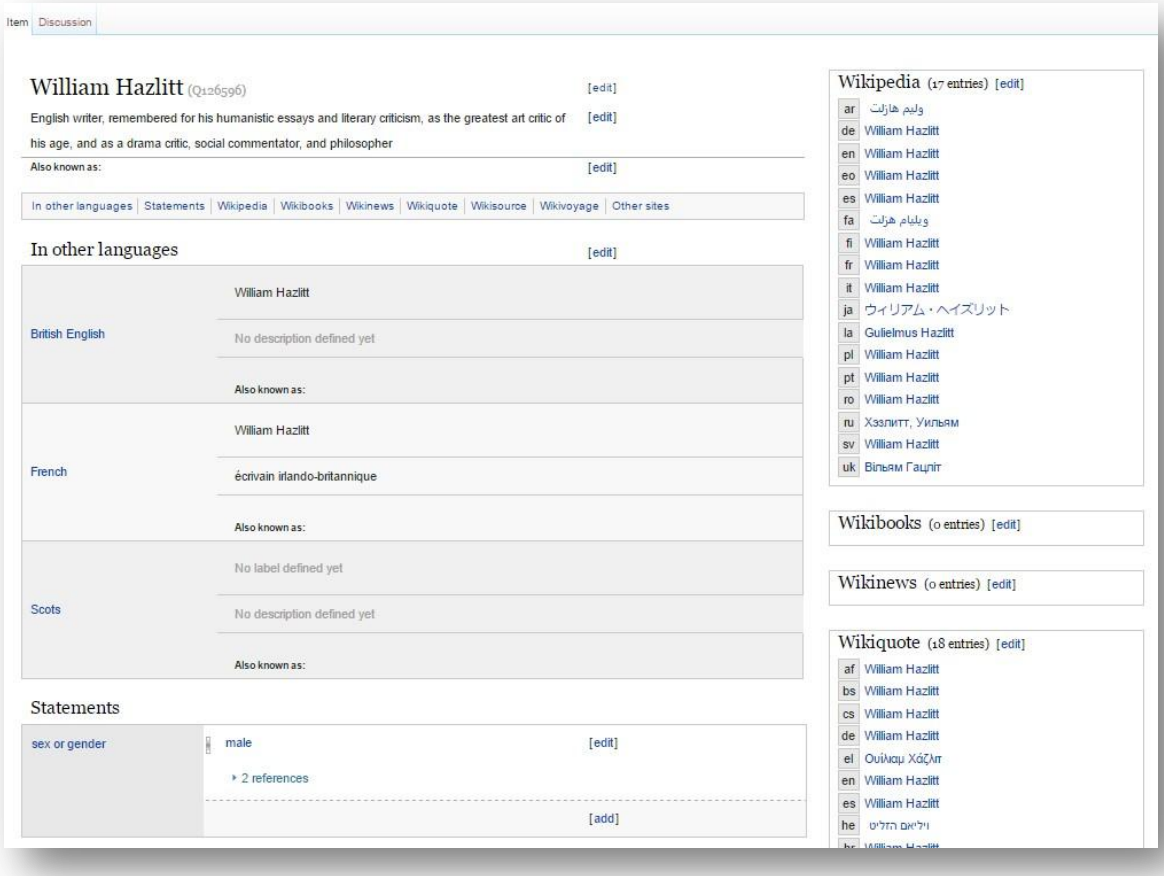
The best way to see DBpedia is to see an example of it being used in practice. There are good (and well-documented) examples of this at the BBC websites for music and for natural history. For example, the page for Paul McCartney in the BBC music site includes the start of the Wikipedia biography. But cleverer than that is the use of linked data to join articles on musicians with facts and figures about them. This work was summarised in an article as long ago as 2009. (Kobilarov, George, et al, "Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections ", in The Semantic Web: Research and Applications (Lecture Notes in Computer Science, vol 5554, 2009, pp 723–737.)

Remarkably, DBpedia turns out not to be the only initiative in this area – there is now a parallel, or possibly competing resource, from the Wikipedia team themselves, an initiative called Wikidata (not really surprising that there is a "wiki" in the title – I get confused at the proliferation of "wiki" initiatives and projects). Why create an alternative system? What's wrong with DBpedia, when in 2009 it was seen as the answer? When I asked this question at a recent Wikimedia Meetup, one curious answer I received was that "DBpedia wasn't created by us – it's not part of our community." It's intriguing that anyone could feel themselves part of Wikipedia, but this is clearly a very powerful motivation for Wikipedia (and Wikidata) editors and (to use their preferred term) collaborators.

Wikidata was set up to address what they perceive to be one of the biggest limitations of DBpedia: the fact that Wikipedia is created by humans, and in a very non-structured way. In fact, extracting anything from Wikipedia apart from facts and dates is very difficult. Instead, Wikidata switched from a dissatisfaction with DBpedia, by starting completely afresh. Well, almost afresh - Wikidata uses Wikipedia data, but the team state the data is derived from, but not using directly, Wikipedia content, even though the same team in many cases compiles and edits both products. The term "team" is used very loosely, because Wikipedia is compiled by a very loose collection of freelancers, each doing as much or as little on the project as they would like. There is currently very little co-operation between each of the national Wikipedia products. This approach leads to some unexpected consequences, as will be described later.

The Wikidata project started in 2012. Funded by external donations, it is an initiative created by volunteers. It describes itself as creating a "free knowledge base about the world that can be read and edited by humans and machines alike". That definition is not very enlightening. The goal is to do what DBpedia does, which is to provide a machine-readable link so that computers can pull information from a Wikipedia-type resource automatically and dynamically.

While DBpedia is not readable by humans, Wikidata is, after a fashion. You can look subjects up in Wikidata, just as you can on Wikipedia. What you are shown is rather more elementary, since it is compiled entirely automatically. It makes some sense, but it is more of a checklist than a full encyclopaedia entry. Take a typical biographical entry, that for William Hazlitt, the English writer:



Wikidata example entry: William Hazlitt

What is shown on the screen is the result of or the opportunity for a considerable amount of linking. Firstly, the Wikidata software has attempted to unite the various entries from the different language editions of Wikipedia (on the right) and has done the same for Wikiquotes and for the other wiki publications. Unfortunately, you can't take for granted that the "William Hazlitt" you write about in the English edition is the same William Hazlitt who exists in the other language editions of Wikipedia. Hence the list on the right-hand side of the screen. Wikipedia editors can click on individual articles in the other editions to confirm that the two William Hazlitts are indeed the same man. This is where a tool called "The Reasonator" comes in: it is a clever tool that provides some automation of the work of linking Wikipedia entries with other sources, by lining them all alongside each other; all the compiler has to do is to accept (or reject) the machine-proposed link.

Using The Reasonator, and human comparison, Wikidata creates a set of "statements" about the entry that have been validated. In the case above, the statement that William Hazlitt is male has two references.

The problem that Wikidata tries to tackle is the problem of authority. How do we know on what authority Wikipedia states a fact? Even a simple fact such as "William Hazlitt was a male"? Well, the authority given by Wikidata (and some might claim this is a rather circular authority) is Wikipedia itself. There are references to Hazlitt being male in two editions of Wikipedia, the Swedish and Italian versions (although not, apparently, the English edition). You could ask what authority they have, but that would be a topic for another article.

As a result of all this linking, a "category" of William Hazlitt is created, which links all the various media created by the different editions of Wikipedia relating to Hazlitt. In other words, the Wikidata team are doing an ORCID-style disambiguation for everyone in Wikipedia.

In this way Wikidata is one enormous linking engine, seeking to disambiguate references. The same process of disambiguation is being done with other reference sources as well. For example, Charles Matthews, one of the Wikidata contributors, described to me at a Wikidata Meetup in Cambridge how he had linked every biography in the Oxford Dictionary of National Biography to the relevant Wikipedia entries via Wikidata. Thus, William Hazlitt is listed in Wikidata as having link number "101012805" to the ODNB.

While any human could read this, the intention is of course to provide an automated link.

What will be the outcome of this? I'm sure we will all benefit from the work being done to improve links and to make Wikipedia a slicker tool. But whether it resolves the fundamental problem of knowledge organisation, familiar to every librarian, is another matter. Essentially, anything created by humans is usually difficult for machines to process correctly without errors, but what a machine creates is not usually very readable (or interesting) for a human. So far Wikidata is, well, just a lot of links.

Michael Upshall has been providing consultancy for publishers on digital content and delivery since 2002. He managed the team that produced the UK's first online encyclopaedia, The Hutchinson Encyclopaedia, in 1999. He writes a blog about reference and encyclopaedia publishing.