Teaching Wikipedia to Write Itself

By Rebecca J. Rosen
Rebecca J. Rosen is an associate editor at The Atlantic. She was previously an associate editor at The Wilson Quarterly, where she spearheaded the magazine's In Essence section.

Atlantic; Apr 2 2012, 3:06 PM ET

A new project called Wikidata aims to automate some aspects of the collaborative encyclopedia. Could Wikipedia's infoboxes get a boost from a centralized database?

Say next week there is an election in France (okay it's not for a few more weeks, but just suppose) and suppose that the challenger, Francois Hollande, ekes out a victory over Nicolas Sarkozy, the current president. What happens next?

This isn't a question of politics but of information -- how does the world's information sources come to reflect a French changing of the guard? For Wikipedia, one of the web's largest and most up-to-date compendia of facts, the answer is complicated. It's not just the Wikipedia entry for "France" or "President of France" that requires quick updating, but many, many more pages that reference Sarkozy, and not just in English or French but in the more than 280 languages in which Wikipedia appears. Who will update all those references?

Currently, some 90,000 volunteer Wikipedia editors. Those heroes of the collaborative web do the yeoman's work that keeps Wikipedia updated. But could there be an easier way? Could all of those changes happen automatically?

That is the hope of Wikidata, the first new project from the Wikimedia Foundation since 2006. Led by Wikimedia Deutschland, the German branch of Wikimedia, Wikidata aims to be a central fact repository that can feed all articles throughout Wikipedia. All of the data will be published under a Creative Commons license. The project has received financial backing from Google, Inc., the Gordon and Betty Moore Foundation (Gordon Moore co-founded Intel and invented Moore's Law), and the Allen Institute for Artificial Intelligence (as in, Paul Allen, Microsoft's co-founder).

The current plans are for the project to roll out in three phases over the next year. The first phase, set to be completed by this August, is the centralization of all the different points of data in Wikipedia across languages whose updates could be coordinated. The second phase allows for people to begin collaboratively building the database's datasets. They hope to finish that by the end of this year. The final phase "will allow for the automatic creation of lists and charts based on the data in Wikidata." By March of next year Wikimedia Deutschland hopes to turn the database over to the Wikimedia Foundation.

The project's goal may be automation, but automation doesn't mean there aren't humans involved. This is automation by collaboration, a very Wikipedia kind of automation indeed.
The Problem With Wikidata

By MARK GRAHAM
APR 6 2012, 3:25 PM ET 12

Fundamental changes are afoot at Wikipedia. Changes that have worrying connotations for the diversity of knowledge in the world's sixth most popular website.
Wikipedia, with a new initiative called Wikidata, is radically reconfiguring itself to take advantage of the "Semantic Web." Wikidata will create a collaborative database that is both machine readable and human editable and which will underpin a lot of knowledge that is presented in all 284 language versions of Wikipedia.

In other words, the encyclopaedia plans to become part of the movement from a mostly human-readable Web to a Web in which computers and software can better make sense of information. This system becomes especially useful for facts that are embedded in a variety of pages. If Mitt Romney were to become President of the United States, there would be hundreds or thousands of pages in all of the language versions of Wikipedia that would need to be altered to reflect that fact. Wikidata would allow all of those references to be immediately updated after only one change in the central Wikidata repository.

This is a highly significant and hugely important change to the ways that Wikipedia works. Until now, the Wikipedia community has never attempted any sort of consistency across all languages.

Look, for instance, at the Wikipedia pages about the Bronze Statue of Tallinn (a highly controversial moment in Estonia's history that sparked one of the world's first 'cyberwars' between Russia and Estonia). The Estonian and Russian versions of that article present interestingly different versions of the very same place and events. The Arabic and Hebrew articles about Hezbollah offer perhaps an even starker contrast of the ways in which different communities of editors agree on different types of representation and truths.

Research carried out independently by Brent Hecht, myself, and others has found that each language edition of Wikipedia represents encyclopaedic knowledge in highly diverse ways. Not only does each language edition include different sets of topics, but when several editions do cover the same topic, they often put their own, unique spin on the topic. In particular, the ability of each language edition to exist independently has allowed each language community to contextualize knowledge for its audience.

It is important that different communities are able to create and reproduce different truths and worldviews. And while certain truths are universal (Tokyo is described as a capital city in every language version that includes an article about Japan), others are more messy and unclear (e.g. should the population of Israel include occupied and contested territories?).

The reason that Wikidata marks such a significant moment in Wikipedia's history is the fact that it eliminates some of the scope for culturally contingent representations of places, processes, people, and events. However, even more concerning is that fact that this sort of congealed and structured knowledge is unlikely to reflect the opinions and beliefs of traditionally marginalized groups.

We know that Wikipedia is a highly uneven platform. We know that not only is there not a lot of content created from the developing world, but there also isn't a lot of content created about the
developing world. And we also, even within the developed world, a majority of edits are still made by a small core of (largely young, white, male, and well-educated) people. For instance, there are more edits that originate in Hong Kong than all of Africa combined; and there are many times more edits to the English-language article about child birth by men than women.

What does this mean for structured, semantic data in Wikipedia? If we start to rely on a singular source for our truths, it will undoubtedly in most cases make most articles more accurate and current. But it also means that in contested cases, we will likely see an even more vivid reinforcement of existing core/periphery inequalities of knowledge production. A disagreement over facts or data would no longer be confined to a specific article and language, but would most likely have to be conducted in English to an unfamiliar community of editors.

Without social structures and technologies specifically dedicated to maintaining diversity in Wikidata, the almost certain outcome is that the truths and worldviews of the dominant cultures in the Wikipedia community will win out.

The support of knowledge diversity in Wikidata becomes all that much more important when you consider a primary goal of the Wikidata project: to make information in Wikipedia much more understandable to artificial intelligence systems. In other words, Wikidata - if successful - is going to form the "brains" of many future technologies and online platforms. Google, a key funder of the Wikidata project, will undoubtedly be one of the first to incorporate the knowledge in Wikidata.

This means that certain culturally and politically specific truths and worldviews will become ever more central, integral, and powerful in the information ecosystems of the Web.

The point of this is not that Wikidata is necessarily entirely a bad idea. It will undoubtedly provide useful raw informational material to expand many of the 284 Wikipedias that are currently sorely lacking useful content. The project is also still in its formative stages, and its leaders are concerned about developing technologies and practices to support knowledge diversity. My worry, however, is that the core idea behind the project can only serve to congeal the power and voice of those already at the core of processes and practices of knowledge production and reproduction.

The beauty of Wikipedia has always been its ability to allow for nuanced and complex representations of contested information. This is something that need not change as information from Wikidata is increasingly propagated through the platform. We just need to ensure that we aren't seduced into codifying, categorizing, and structuring in cases when we should be describing the inherent messiness of a situation. Tokyo will always be the capital of Japan, but it will probably be a long time until we can all agree on the true population of Israel.

The author would like to thank Brent Hecht for his valuable feedback on this piece.