Strategy for Wikidata for Wikimedia projects

> “Knowledge has to be improved, challenged, and increased constantly, or it vanishes.”
> *-- Peter Drucker*

Authors:
Ramsey Isler, Lydia Pintscher, Lea Voget

Contributors:
Josh Minor, Amanda Bittaker, Danny Horn, Melanie Koeppen, Raz Shuty, Adam Shorland, Elena Aleynikova, Ben Vershbow

August 2019

Abstract

There are new opportunities available now that Wikidata has matured and communities across multiple projects are embracing it. When Wikidata is better integrated into Wikipedia, we can add advanced content generation capabilities like suggesting articles that should exist but do not yet, surfacing new sources that can be added as citations in articles, and more.

Beyond content generation, Wikidata on Wikimedia projects helps narrow the knowledge equity gap. By utilizing the multitude of language labels and descriptions already on Wikidata, projects will become easier to use for non-English speakers. Additionally, data-driven features like automated work lists would help contributors easily find a meaningful task based on automatic recommendations that take into account their expertise and interest, facilitating the recruitment of new contributors to improve neglected areas of content.

Finally, Wikidata can provide the backbone for a system that accommodates diverse ways of consuming knowledge. This system could reconfigure content or present it in native formats/contexts on other platforms (messaging, social media, virtual assistants, etc), thereby increasing accessibility and discovery across the internet as a whole.

Background

The Wikidata project started in 2012 and has had a dramatic effect on the open data movement. Wikidata is Wikimedia's free, collaborative, multilingual, knowledge base focused on verifiability, collecting structured data to provide support for Wikipedia, the other wikis of the Wikimedia movement, and to anyone in the world who needs general purpose structured data.

Wikidata also provides a robust machine-readable system that enables for other software. Via Wikibase (the software suite that drives Wikidata and other structured data repositories), Wikidata provides structure and consistency in ways that free text can't. This is true even on the world's largest user of wikitext, Wikipedia.

Almost every Wikipedia article has a corresponding "Item" on Wikidata. These Items include links to pages on the same topic in other language Wikipedias and other Wikimedia projects, as well as labels and descriptions in different languages, and structured statements about the topic. Additionally, Wikidata has much broader inclusion criteria than Wikipedia, and contains many Items that do not currently have Wikipedia articles. Wikidata provides most of the interwiki links to other language Wikipedia articles (although a small number of custom interwiki links remain locally, e.g., to sections of articles on other languages). We also use Wikidata information in templates such as {{Authority control}}, which provides links to catalog entries on the article subject.

Wikidata can be used to display all of the content in infoboxes (the data summary boxes on the right side of articles,) just some of it, or none of it. Images, captions, coordinates, locations, links, dates, units, websites, and maps can all be shown using Wikidata values.

WMF and WMDE are currently piloting more powerful ways to integrate Wikidata with the other Wikimedia projects. The first step is the complex and ambitious Structured Data on Commons project (SDC), which aims to bring multilingual structured metadata to the media repository that all Wikimedia projects rely upon. There are intriguing new opportunities available now that Wikidata has matured and communities across multiple projects are embracing it. We can begin bold projects on Wikipedia that are similar to SDC in scope and potential impact. The next evolution of the Wikimedia projects is in sight. Now we just have to embrace it.

Strategy: Make Wikimedia projects ready for the future

The data within Wikidata needs to be strengthened and expanded in order to meet the needs of other Wikimedia projects. More established Wikipedias have concerns over uncited data, control over data, ontological structures, and data comprehensiveness. Data quality and ontological predictability need to be improved if search algorithms are to rely on it. These concerns can be addressed through new tools, features, policies, and integration of community feedback.

Audience

Users increasingly consume content through new devices and services, so Wikimedia content will need to become more atomized and suitable for re-use. Wikidata is a crucial component in this transition, as the backbone for how Wikipedia can structure data that can be used in chat bots or voice assistants like Siri and Alexa, and making multimedia available for media bots and screen-enabled virtual assistants. The Wikimedia Foundation, affiliates, and other developers can also build rich interactive tools like timelines and other visualizations to help contextualize information. Thus our audience includes both *people* and *platforms*.

Goals

This strategy addresses four primary goals:

- Further knowledge equity
- Maintain Wikimedia's growing content
- Enable new ways of participating in knowledge sharing
- Ensure the integrity of our content

The global internet space is rapidly changing, and our opportunities are narrowing in regions where awareness of our projects is low. With Wikidata as a more mature platform, we have the chance to act rapidly. We can build new ways to automate content creation for underserved languages and regions. We can empower users with new and easier ways to contribute. But the internet waits for no one, and if we don't do this now, we'll miss countless opportunities and miss an opportunity to be a leader in local language content in places where open source knowledge is needed most.

Further Knowledge Equity

Wikidata offers a more flexible, less contentious space than Wikipedia for the inclusion of marginalized knowledge, for example, data related to indigenous peoples and cultures[1]. Wikidata's multilingual environment and broad scope can serve as a public platform for this data, opening it to the wider web ecosystem, and establishing a footprint for further coverage on Wikipedia (and Wikimedia Commons). Looking ahead toward 2030, as the Wikimedia movement explores how to incorporate other source formats like oral knowledge, Wikidata may play a further role as a descriptive framework for these assets, facilitating their appropriate citation in Wikipedia and across the web.

There's also an extreme disparity of size and quality articles across language wikis, and the individual language versions cover (sometimes vastly) different topics. But **Wikidata helps make content accessible across languages**. On Commons, for instance, the addition of structured fields that state what is depicted in an image, linked to Wikidata concepts, addresses the major problem of information and categorization being hidden behind English-only categories. By utilizing the multitude of language labels and descriptions already on Wikidata, Commons becomes easier to use for non-English speakers. With Wikidata, we can help every editor's contributions have far larger reach (no matter what language they speak), and we enable more readers to access information that has previously been inaccessible for them.

Large Wikipedias have concerns about using Wikidata that prevent them from fully embracing it. We want to build features and practices that allay fears and empower contributors to use Wikidata on large Wikipedias, before their concerns solidify into policies that are more difficult to change.

Maintain and support Wikimedia's growing content

Wikidata provides one of the building blocks for generating more content in our projects. When Wikidata is better integrated into Wikipedia, we can add advanced content generation capabilities like suggesting articles that should exist but do not yet, surfacing new sources that can be added as citations in articles, and reassembling information from one medium (e.g. long-form article) into another (e.g. visual slideshow).

Wikidata can help highlight gaps in our content. We want to draw on and improve the experience of individuals and groups (such as Women in Red and Wiki Loves Monuments) who currently use Wikidata powered work lists. Such lists should be tailored and made actionable so every contributor can easily find a meaningful task based on recommendations that take into account their expertise and interest. This could help us move away from the assumption that Wikipedia is complete, and help recruit new contributors to improve neglected areas of content.

---

[1] See: https://www.wikidata.org/wiki/Wikidata:WikiProject_Indigenous_peoples_of_North_America

On some Wikis, particularly English Wikipedia, a decreasing number of contributors are responsible for a growing amount of content[2]. We need to support and help these contributors have a larger impact with their work handling the content[3]. **With Wikidata as a base we can build better tools for contributing and maintaining content** (e.g. micro contributions, or machine-parsing and surfacing uncited claims for human tagging, or flagging content for translation or expansion across wikis.) We can help contributors be more effective/efficient by giving their work more reach. This improvement to contributor impact also helps us better understand our content and helps contributors focus on the most pressing tasks in their area of work and expertise. Wikidata is a crucial platform for high-impact contributions. A contribution made and shared through Wikidata reaches much farther than one made on any single Wikipedia.

Enable new ways of consuming and contributing knowledge

Knowledge acquisition on the internet is changing. Research shows that information is no longer mainly consumed through direct-access page-reads but filtered through custom-made experiences that blend into overall media consumption habits.[4] The internet is being increasingly accessed through and filtered by a handful of prominent apps and services, especially in markets with little local language content. Wikidata can provide the backbone for a system that accommodates diverse ways of consuming knowledge. This system could reconfigure content or present it in native formats/contexts on other platforms (messaging, social media, virtual assistants, etc), thereby increasing accessibility and discovery across the internet as a whole. Similar capabilities are also possible on the contribution side.

We should continue to invest in structured data so knowledge is easier and faster to find and associate with other content. Using structured data to integrate new forms of contribution and consumption will enable more people in many different languages and cultural contexts to read, share, and edit the wikis. Structured data also provides the added benefit of making the content more flexible and transmutable for form factors and platforms beyond the web or apps. For more ideas in this area, see the Form Factor topic on the Experiences Product Perspectives doc.

If we don't act on these opportunities, other private-sector, profit-driven actors will create their own knowledge bases and people will move on to other places and ways to find knowledge. They will most likely not be compatible with our mission and values. The longer it takes for us to engage new contributors and readers, the harder it will be to bring them on board as we compete with an ever-increasing list of apps, sites, and devices that absorb people's attention and instill particular habits.

---

[2] See for example the wikipage increase on English Wikipedia vs the editor decline on English Wikipedia
[3] Magnus Manske's take on list editing
[4] Experiences Product Perspective

Ensure the integrity of Wikimedia's content

In a disinformation age, Wikipedia is becoming a crucial provider of neutral and cited information that an increasing number of people, companies and institutions rely on. Wikidata should serve as a common way for internet users to identify or reference a source, one that incorporates advanced capabilities such as queries and systems that locate Wikipedia passages that cite redacted papers, discovery tools that find inconsistencies in content, or "smart" systems that determine if a reference actually supports a claim it is used for. The current WikiCite community is the beginning of this, but tools and processes built with Wikidata should improve the quality of content and data to serve the scholarly community, the media, and education. The more individuals and entities that are working with a mutually-reinforced network of verification and citation, the more reliable Internet-based information becomes.

Opportunities and risks for Wikimedia

Why should Wikimedia invest in this now?

By building new tools and processes based on more tightly integrated structured data, we can make contributing and consuming knowledge easier for everyone, especially smaller projects. And large Wikipedias have concerns about using Wikidata that prevent them from fully embracing it. We want to build features and practices that allay fears and empower contributors to use Wikidata on large Wikipedias, before their concerns solidify into policies that are more difficult to change. Directly addressing the concerns and social/governance issues is of crucial importance.

What are the risks of not acting now?

The global internet space is rapidly changing, and our opportunities in emerging markets are narrowing: The internet is being increasingly accessed through and filtered by a handful of prominent apps and services, especially in markets with little local language content. The longer it takes for us to engage new contributors and readers, the harder it will be to bring them on board as we compete with an ever-increasing list of apps, sites, and devices that absorb people's attention and instill particular habits. With Wikidata as a more mature platform, we have the chance to act rapidly. We can build new ways to automate content creation for underserved languages and regions. We can empower users with new and easier ways to contribute. But the internet waits for no one, and if we don't do this now, we'll miss countless opportunities and miss an opportunity to be a market leader in local language content in places where open source knowledge is needed most.

Wikidata as a platform needs to be strengthened and expanded in order to meet the needs of other Wikimedia projects: More established Wikipedias have concerns over uncited data, control over data, ontological structures, and data comprehensiveness. Data quality and ontological

predictability need to be improved if search algorithms are to rely on it. These concerns can be addressed through new tools, features, policies, and community engagement. Not addressing these concerns means we cannot build the features that will achieve our goals.

Appendix

Existing usage highlights

Statistics: usage of Wikidata content in Wikimedia projects

Project	**Why it matters**
Wikidata-powered infoboxes on Basque and Catalan Wikipedia (example): almost all infoboxes on these projects are powered by Wikidata	A large amount of work is being taken off the shoulders of these communities so they can concentrate more on the articles they care about while serving their readers. Additionally, Basque and Catalan content is made available to more people by sharing it through Wikidata.
Commons category infoboxes (example): the vast majority of category pages on Commons use Wikidata-powered infoboxes	Category pages on Commons historically did not contain infoboxes. It would have been too tedious to maintain and is not core to the work of the Commons community. Using Wikidata for creating these infoboxes adds value for the reader without burdening the editors.
Editable infoboxes on Russian Wikipedia: Russian Wikipedians developed a gadget to allow editing of Wikidata's data for a limited number of infobox types directly from their Wikipedia	A community is taking the lead in discovering how to better integrate Wikidata editing workflows in their project and remove a barrier to greater acceptance of Wikidata.
WikiShootMe: a tool with a map with all the information on Items without a picture on it. It lets you choose a missing image and upload your picture	We can easily show people places near their current location that are missing a picture in Wikimedia projects. This way we open up a clear path to contribution that is meaningfully enriching our content.
ArticlePlaceholder (example): a tool to show a data sheet when people search for an article that doesn't exist in their language Wikipedia	We provide our readers with at least basic information on a topic they were looking for and we have a way to turn readers into contributors by showing them a call to action to write the missing article. In addition it helps us avoid automated mass-creation of static articles that are never going to be updated.

Wikidata powered worklists such as the Women in Red worklist and Wikiproject Built Heritage worklist	Wikidata-generated work lists are a great way to show how Wikidata helps the broader movement, and also call attention to the fact the creation of such lists (using listeriabot and other grassroots workflows) is inaccessible to all but an "in the know" few. These practices call us to imagine a future in which improved tools, and a skilled-up community, have empowered more actors across the movement to map knowledge gaps and systematically close them through inspiring campaigns and thematic projects.
Wikidata-powered interwiki links	By using Wikidata as the central source connecting the wiki projects, more than 240 million lines of wikitext code have been removed from the Wikipedia projects alone in the effort to replace locally defined interwiki links.

Guiding principles and beliefs

- We do not force any project to adopt Wikidata's data. Integration with Wikidata is only sustainable if it is driven by the communities.
- Each project is unique in what benefits and drawbacks Wikidata adoption has for them. It's ok to have different adoption rates and focuses based on each project's maturity and needs.
- Sometimes local overwrites and exceptions to Wikidata's data are necessary. That's ok but we strive for sharing as much data as possible to the benefit of everyone while allowing local autonomy where needed.
- We do not privilege the large projects over the smaller ones (sometimes we do the opposite).
- Every project has something to contribute. If we all share, then everyone benefits - even the large Wikipedias.

Capabilities map