Wikidata/Wikibase Vision

High-level overview

> “The Web as I envisaged it, we have not seen it yet.
> The future is still so much bigger than the past.”
> *-- Sir Tim Berners-Lee*

> “I'd like to know what the Internet is going to look like in 2050.
> Thinking about it makes me wish I were eight years old.”
> *-- Vinton Cerf*

Authors:
Lydia Pintscher, Lea Voget, Melanie Koeppen, Elena Aleynikova

Contributors:
Leszek Manicki, Jan Dittrich, Raz Shuty, Jens Ohlig, Birgit Müller, Amanda Bittaker, Ramsey Isler, Josh Minor, Ben Vershbow

August 2019

Background **3**
What are Wikidata and Wikibase? 3
What happened so far and why is now a point for strategy? 5
Vision: Giving more people more access to more knowledge **10**
More People 10
Overcoming barriers 10
Bringing in new types of contributors 11
More Access 11
Allowing users to access knowledge in whichever way they want 11
Making Wikimedia's knowledge usable everywhere 12
More Knowledge 13
Increasing the sum of accessible knowledge 13
Ensuring the integrity of Wikimedia’s knowledge 13
Main roadblocks **14**

Background

What are Wikidata and Wikibase?

Wikidata: a 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.

- Free: The data in Wikidata is published under CC-0, allowing the use of the data in many different scenarios.
- Collaborative: Data is entered and maintained by people and their tools. The Wikidata editors decide on the rules of content creation and management. Humans and machines work together hand-in-hand.
- Multilingual:[1] The Wikidata data model is not based on any specific language, but rather expects translations of so-called Q-numbers into all languages. Once labels are translated, their content can be seen in any language. For example, if the new name of a newly elected president is added, this information can be consumed immediately in all available languages.
- Verifiability: Wikidata records not just statements, but also their sources. This helps us reflect the diversity of knowledge and supports the notion of verifiability.
- Collecting structured data: Imposing a high degree of structured organization allows for the data to be easily used in the context of Wikimedia and non-Wikimedia projects, and enables machines to process and "understand" it.
- Support for Wikimedia wikis: Wikidata assists Wikimedia projects among other things with more easily maintainable infoboxes and links to other languages, thus reducing editing workload while improving quality. Updates in one language are made available to all other languages. It also builds the base for new projects and initiatives in areas like ORES and GLAM collaborations.
- Anyone in the world: Not just Wikimedia projects, but anyone can use Wikidata to build applications, websites, visualisations and more.

[1] "According to the engco model of language forecasting, based on first-language speaker numbers, the most widely spoken language in 2050 will still be Mandarin. The model predicts that Spanish will become the second most spoken language, followed by English with Hindi moving to fourth and Arabic retaining its position as the fifth most spoken language in the world." - Considering 2030: Demographic Shifts – How might Wikimedia extend its reach by 2030?

"The global population is expected to reach 8.4 billion by 2030, a 15 percent jump from 2015. Low-income regions have a projected 35 percent growth by 2030, with Africa growing fastest at 40 percent. Wikimedia must focus on serving people in high-growth regions to respond to these trends. Africa currently accounts for a small portion of Wikimedia traffic, but the region presents a crucial opportunity for growth as its mobile and fixed IP traffic are predicted to increase substantially." - Strategy 2030 Wikimedia's role in shaping the future of the information commons, page 11

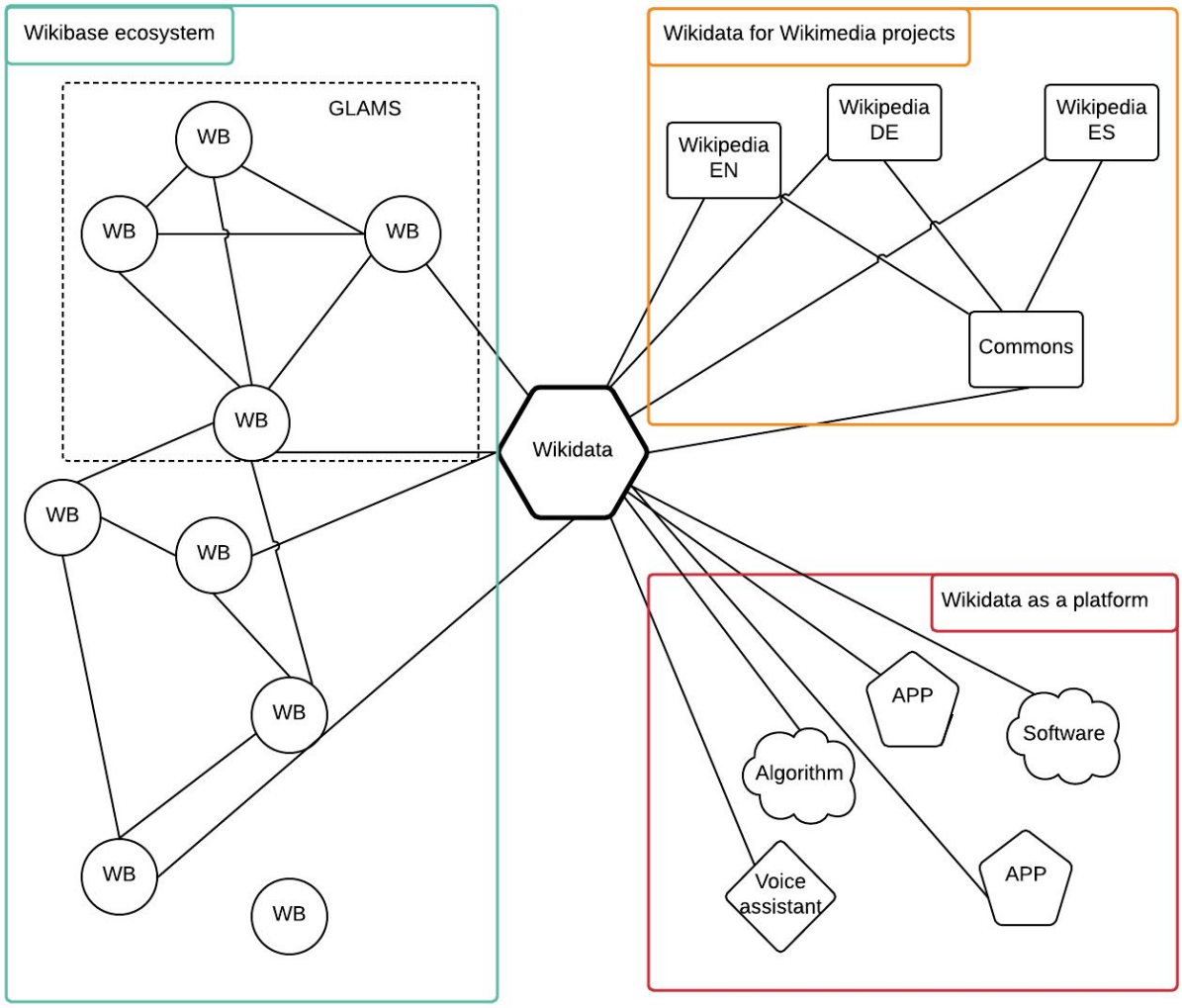
Wikidata provides data, an ontology and links to other databases.

- Data: Wikidata collects general purpose data about the world, starting with the kind of data found in Wikipedia's infoboxes and expanding from there.
- Ontology: Wikidata provides an ontology of the world through the concepts it describes and the relations between them (e.g. Berlin is a city, Berlin is the capital of Germany). By uniquely identifying concepts and the relations between them, Wikidata acts as a general purpose vocabulary that can be used for language independent tagging, search improvements and more.
- Links to other databases: Wikidata acts as a hub in the linked data web. Based on a Wikidata identifier one can easily find matching entries with additional data in other databases. It also makes it possible to understand that two entries in unrelated databases are describing the same concept (e.g. a machine can now understand that Douglas Adams' entry in Project Gutenberg matches a particular entry in IMDb).

Wikibase: an open source software suite for running a collaborative knowledge base. One installation of it is Wikidata. Wikibase enables humans and machines to work together productively in the same shared space. Wikibase is a collaborative knowledge base software that allows users to tie in to a vast network of linked open data, easily adding to and ingesting knowledge from Wikidata and other instances of Wikibase, creating an ecosystem of easily shared, accessible knowledge. As its central unit Wikibase has Items and Properties that are used to describe concepts and can be extended to other domains like language and media files. It is focusing on medium-scale data as opposed to big data. It is focusing on storing heterogenous data rather than data that fits easily in a dataset. (E.g. mayor, number of inhabitants, country, etc. statements about a city rather than temperature measurements for a city in 1-hour intervals for 100 years).

There are 3 major themes related to Wikidata and Wikibase:

- **Wikidata for Wikimedia projects**: Wikidata supports the Wikimedia projects, most noteworthy Wikipedia and Wikimedia Commons, as well as efforts like GLAM-Wiki, ORES and structured citations. It does so by centralizing data in one place so that every Wikimedia project can benefit from it independent of project type and language.
- **Wikidata as a platform**: Wikidata is a project that is useful independent of its usefulness for the other Wikimedia projects. Many external applications, visualisations and more are being built using Wikidata's data.
- **Wikibase Ecosystem**: Wikibase can be used to run other knowledge bases, not just Wikidata. These knowledge bases can then be tightly connected to Wikidata and form an ecosystem of linked data projects around Wikidata. It provides a lever for Wikimedia to increase the sum of accessible knowledge, and to tightly collaborate with other institutions. Right now, we already have interest from or ongoing projects with many different institutions like the MET and the French, German and Swedish National Library, research initiatives like FactGrid and community projects like LinguaLibre who use Wikibase to open up their data.



What happened so far and why is now a point for strategy?

In 2012 we started out with the goal of building a project to centrally maintain infobox data for Wikipedia to enable the work of our volunteers to scale better and have more reach. Around 2013/14 we started to also focus on providing data to users outside Wikimedia because the data we collected is valuable far beyond Wikimedia. Around the same time we also put more focus on the other Wikimedia projects like Wikimedia Commons. In 2017/18 we added the idea of an ecosystem around Wikidata as an important next step because we believe that this is a key step to spreading more knowledge. Many new Wikibase installations are a core part of this linked data ecosystem. The idea of a Wikibase Ecosystem has been transformed into one of the strategic threads of the multi-year strategy of WMDE[2].

[2] Results of the “Zukunftsprozess” of Wikimedia Deutschland and the resulting strategic threads

Since the start in 2012 Wikidata has gained a lot of traction both inside and outside of Wikimedia, winning prestigious awards along the way[3]. The number of contributors has been growing steadily[4]. The amount of available content is increasing rapidly[5]. More and more articles in Wikimedia projects make use of Wikidata's data (e.g. almost 90% of articles on Basque and Catalan Wikipedia having infoboxes powered by Wikidata)[6]. More and more applications and visualisations are built on top of Wikidata's data and ontology by third parties[7]. And interest is growing rapidly outside Wikimedia in using Wikibase also for powering other knowledge bases.

At the same time we are reaching the limits of our technical, organisational and social infrastructure. We cannot accommodate all the data and interest[8] with the level of resources we are currently investing.

- We need to be able to channel and address the huge demand for Wikidata and Wikibase from external institutions and partners to increase the sum of accessible knowledge. This entails establishing close personal relationships with key partners to catalyze the ecosystem, especially now that there is a potential of Wikibase fundamentally shaping the open linked data environment.
- The Wikibase Ecosystem and its primary users have needs and requirements towards the underlying software that differ from the users inside the Wikimedia movement, such as connecting to other Wikibase instances for sharing data, easily setting up and filling Wikibase instances, well working APIs or creating multiple levels of data accessibility. We cannot continue developing features for Wikidata expecting that they also address the most important needs from the Wikibase community. We will also need specialised tools and interface options for the different sectors for the Wikibase Ecosystem to thrive.
- Wikidata needs to continue growing and improving its data set. The more complete the data set of Wikidata, the more questions can be answered by it. For new applications to utilize Wikidata, trust needs to be established by the quality and completeness of the data.
- There is still a high potential for Wikidata and Wikibase supporting the editing communities of the Wikimedia projects. Structured Data on Commons shows promising results, and there is much more that can be improved through Wikidata and Wikibase,

---

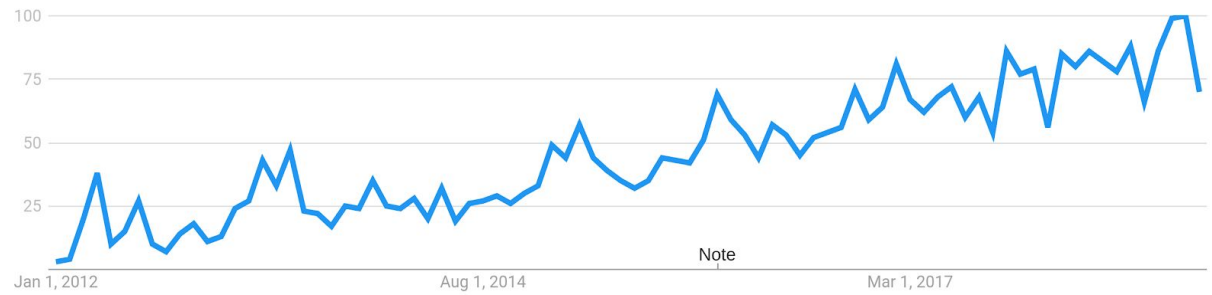
[3] For example, Wikidata has been awarded with the first ever ODI Open Data Award by Sir Tim Berners-Lee and Sir Nigel Shadbold. It has also been recognized as one of 100 innovative places in Germany by the German government ("Land der Ideen").
[4] See graph below
[5] See graph below
[6] See graph below and the Wikidata Usage Dashboard
[7] For some examples see the current usage highlights in the strategy document for Wikidata as a platform. Furthermore Wikidata already plays a vital role in the GLAM sector, e.g. Wikidata being named the 5th most-used source of linked data in a 2018 OCLC survey.
[8] Wikidata is demonstrating a steady growth for Items, Properties, and Lexemes. 2019-20 predicted growth for items is 10 million - 20 million, resulting in approx. 73 million items; 2019-20 predicted grown for properties is 1500 - 3000 property increase, resulting in approx. 9000 properties; Lexemes were only released to the world in 2018, but for the last 9 months (to March 2019) have seen an increase from 3509 to 43500. (Wikidata growth estimates)

from new forms of contributing with dedicated campaigns to improved tools for translation of content across languages and projects.
- Apart from that, we also need to make sure that the technical foundation of Wikibase and Wikidata are put on more solid footing concerning scalability, robustness, performance, availability and can keep up with demand.

We are now at a strategic decision point and need to consciously decide where to take Wikidata and Wikibase in the future and what their roles are in relation to Wikipedia and the other Wikimedia projects. We have amazing opportunities in front of us but in order to fulfill them we need to scale our processes and organisational set-up.

Relevant statistics:

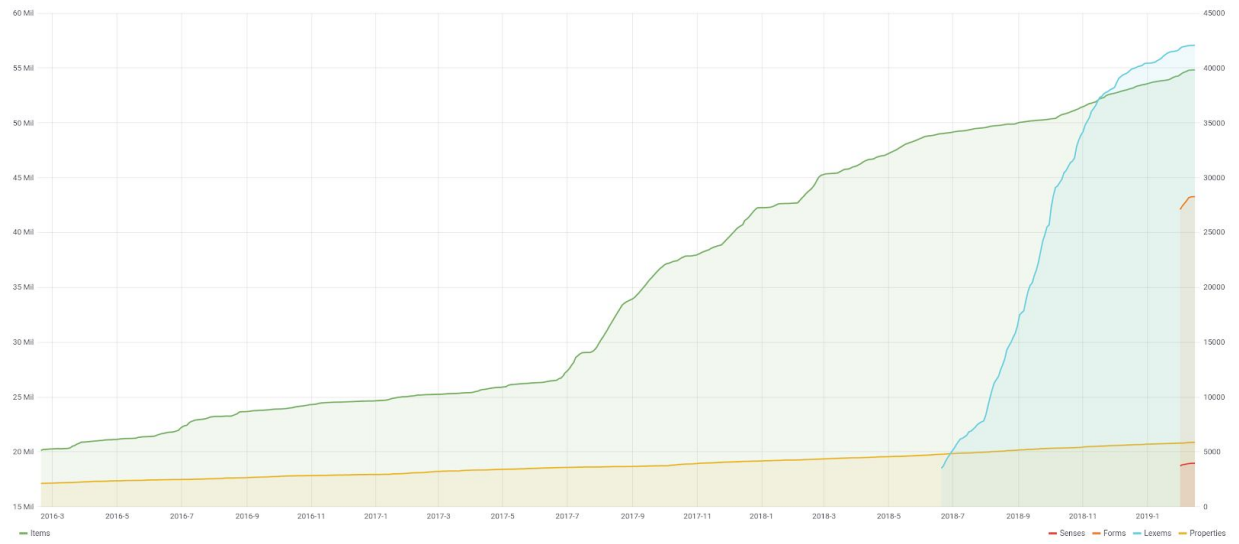
Google Trends graph showing growing general interest in Wikidata:



Active editor[9] statistics showing Wikidata's growing community:

Entity graph showing growing amount of content on Wikidata:

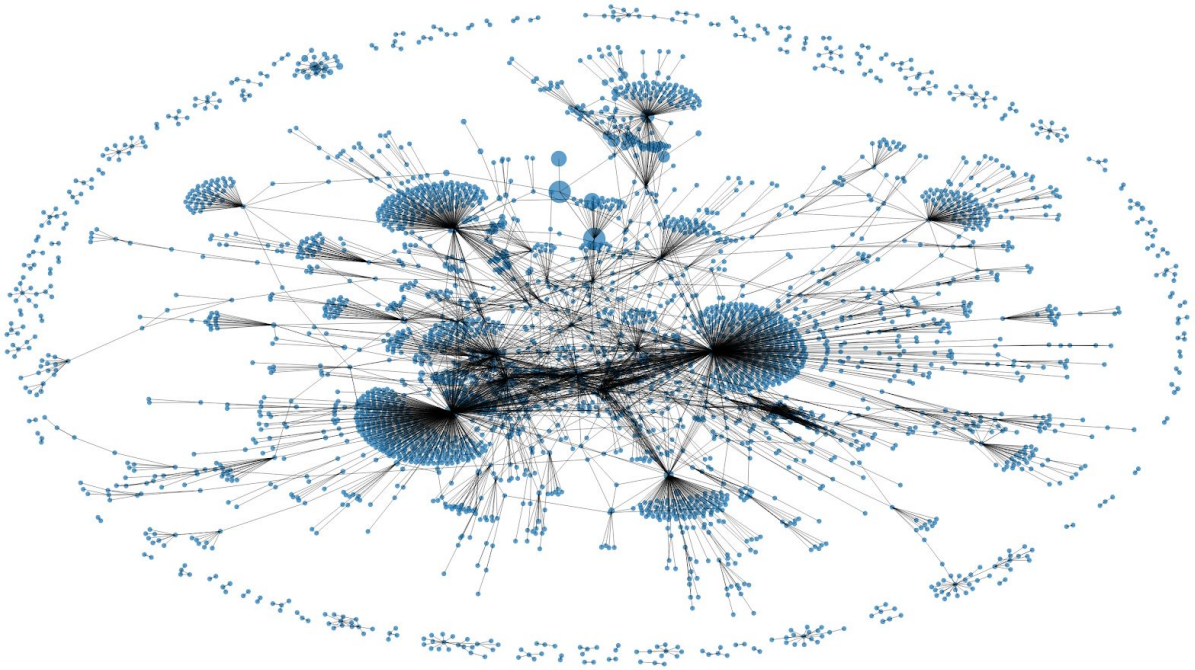
[9] 5 or more edits in the last 30 days



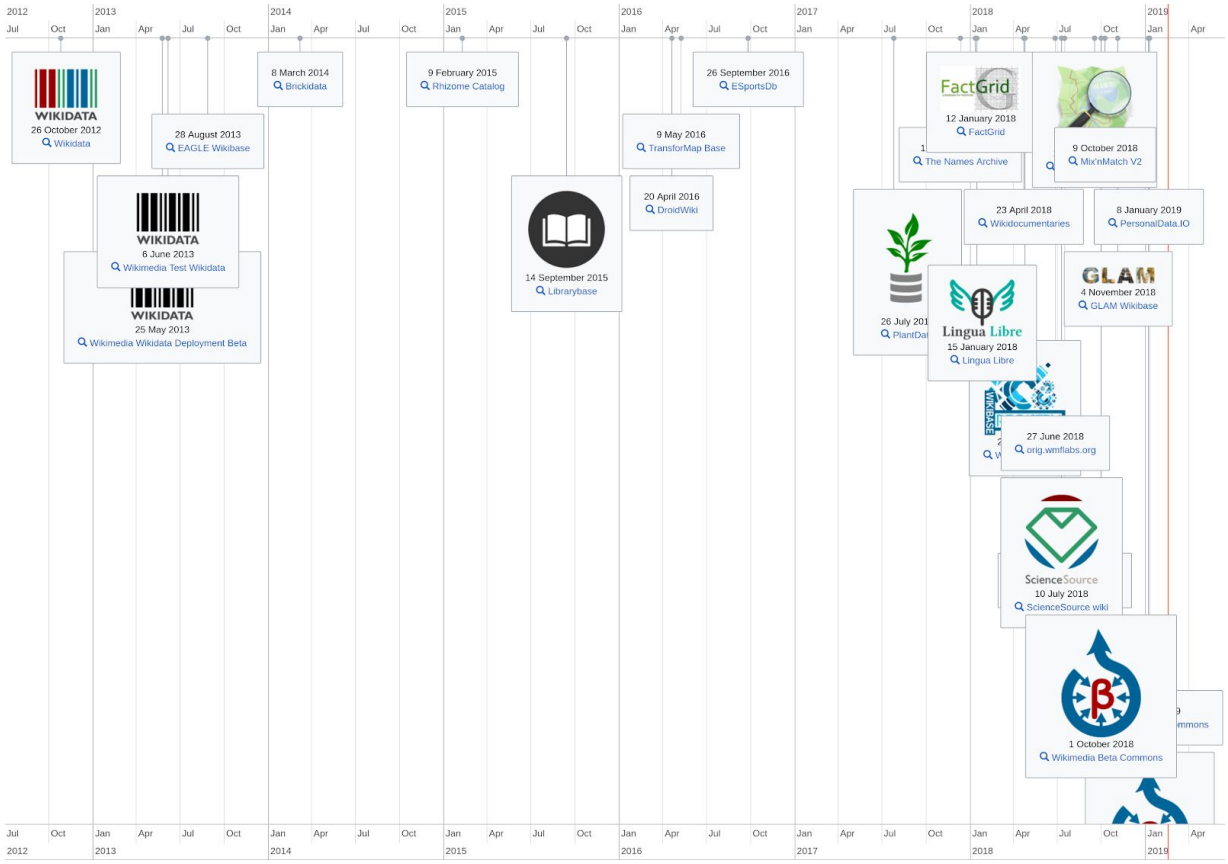
Graph showing increasing use of data from Wikidata in the Wikimedia projects[10]:

Network of over 3800 external identifiers on Wikidata and how their usage overlaps, showing the usefulness of Wikidata as an identifier hub:

[10] The large dip in 2019 is an error in the tracking data, not a dip in actual use.



Timeline of Wikibase installations showing increasing installation numbers:



Vision: Giving more people more access to more knowledge

The vision is to give more people more access to more knowledge. There are three components to this: More people, more access and more knowledge, and each of them will be explained below.

Contributing data to Wikidata and the Wikibase Ecosystem means that data can be understood by humans and machines and thereby spread far and wide, overcoming barriers of language, culture and technology.

> **Access:** When talking about access we usually mean an empowered and active form of access, that goes beyond just consuming content and includes contributing.

More People

Wikidata gives people access to information regardless of the language they speak or the culture they come from. The Wikibase software also provides a new opportunity for contribution: people who like structuring and sorting information as well as institutions, researchers and other groups who have corpora to contribute and that they'd like to structure and maintain.

Overcoming barriers

To overcome the inequalities in contribution[11] to and consumption of knowledge we need to empower people to:

1. make contributions that have an impact far beyond their language and
2. consume knowledge no matter where and in which language it originated.

All this starts with the data and information that underpins knowledge. One of Wikibase's core features is structuring data in a language agnostic way: It connects concepts, and every concept can then be translated into different languages. For example, it stores "Q2" which in English is translated to "earth" and in German to "Erde". This allows people to collaborate across language and culture barriers on a shared knowledge base. Wikibase does this without forcing an agreement where legitimate different world-views exist, for example around border disputes.

Wikidata helps us support the smaller Wikimedia projects better by helping them benefit from all the work done in the bigger projects. But it also - equally important - helps people, who would

[11] "Most of Wikimedia's contributors and users are not located in parts of world where the fastest growing populations are projected between now and 2030." - Strategy 2030 Wikimedia's role in shaping the future of the information commons

otherwise not be able to share their knowledge with so many people, have a much bigger impact. In addition, Wikidata helps us better understand where we are lagging and which users we aren't serving. It helps put a mirror in front of us and see where our content is biased, unbalanced, and not representative. How many articles do we have on a given Wikimedia project about women? How are these articles distributed across professions? And how about time periods? What does the geographic distribution of images on Wikimedia Commons look like? Which writers' works do we make available on Wikisource? These and more are questions we can answer with the help of Wikidata. And the answers can help us make a difference in the knowledge we cover. This is how Wikidata contributes directly to the Knowledge Equity pillar of Wikimedia's strategic direction.

Bringing in new types of contributors

Wikimedia's vision of "a world in which every single human being can freely share in the sum of all knowledge" can only be reached if we allow many different ways to contribute - people should be able to play to their strengths. Wikidata appeals to people who are interested in structure, in checking things off, and in contributing many small pieces rather than a longer, well phrased text. It thus represents yet another option for people to contribute to the Wikimedia projects and opens up contribution opportunities for a different type of person (e.g. micro contributions on mobile devices or through special contribution experiences like the Wikidata Game). This also allows the Wikimedia Movement to scale more as it strives to become the essential infrastructure for free knowledge.

By investing in a Wikibase Ecosystem, we also reach institutional contributors and people who identify with other open knowledge institutions: Wikibase instances with specific purposes (such as storing the data about books and authors of the German National Library) attract contributors interested in these specific purposes. Since people can now share their knowledge from where they already feel comfortable, one can expect high quality contributions: contributors to these Wikibase installations are specialists.

More Access

Wikidata gives people access to information in many new and different ways (e.g. Wikipedia, voice assistants, major web services and special-purpose websites). Thereby we increase the accessibility of our content for everyone.

Allowing users to access knowledge in whichever way they want

Today, and even more so in the future, knowledge will be consumed through machine-powered intermediaries such as search engines, apps or Wikipedia. Not only are people consuming content on a variety of platforms, such as mobile phones, tablets or gaming consoles, they also prefer different kinds of media formats: from YouTube videos, podcasts, short messages to virtual reality applications or interactive games. They consume information while commuting,

drinking coffee with friends or while being bored in meetings, just to name a few. Wikidata is built in a platform and media agnostic way. It is ready for a world in which applications as diverse as voice activated devices like Alexa or classic Wikipedia-like applications like Reasonator are accessing the very same data and turning it into vastly different ways of consumption[12].

Machine learning, voice recognition and virtual and augmented reality are trends that will fundamentally change the way people interact with information, and collaborate with each other. Currently these trends are not driven by free and open values but enclosure in proprietary platforms. Wikimedia can change this. Today however, we have only scratched the surface of the potential of new types of interaction with information, but already now, it is certain that for machines to process data, it needs to be machine-readable. Wikidata provides just that.

By enabling new ways of accessing knowledge, we not only give people more choice, we also lower barriers of entry for people who have not been well supported in the past. Impaired or lost vision, for example, should not be a reason for not finding answers to complicated questions online.

Making Wikimedia's knowledge usable everywhere

As a movement, Wikimedia wants to foster open knowledge, both within and outside its own movement. In the words of the Knowledge as a Service pillar of Wikimedia's strategic direction: "To serve our users, we will become a platform that serves open knowledge to the world across interfaces and communities." In order to fulfill that mission, we need to enable the easy reuse of content, and a crucial step for reaching that goal is structuring content so it is bite-sized, machine-readable and connected to many other knowledge repositories. This is what Wikidata provides. This will lead to a wider reach of Wikimedia's free and open knowledge, as well as a new opportunity for anyone to use our data for creating new innovative projects. Because our data is language independent, nuanced and context-rich, we also open the doors for building better applications on top of that data that can represent more of the richness of our world instead of ignoring it for simplicity's sake. And because our data is the base for any kind of application, it can be used as fitting to the current need - be it as part of an encyclopedic article, or as the base for aural explanations.

[12] "Fragmentation of platforms and user habits is slated to accelerate. In two widely cited reports on technology innovation and usage, Mary Meeker and Amy Webb lay out the most likely new content types and platforms to mature between now and 2030—several of which, not incidentally, use mobile devices as a base for attracting users and shifting media consumption behaviors. These include bots, interactive interfaces, voice-driven personal assistants, and toys that are powered by artificial intelligence (AI); virtual reality (VR) and augmented reality (AR) for news, education and gaming; wearables; and ubiquitous interactive screens." - page 15 in Strategy 2030: Wikimedia's role in shaping the future of the information commons

More Knowledge

Wikidata and the Wikibase Ecosystem increase the amount of available knowledge: they make Wikimedia's knowledge available everywhere, and increase the overall amount of free and open knowledge in the network of open knowledge bases.

Increasing the sum of accessible knowledge

As is stated in Wikimedia's strategic direction, we want to provide the essential infrastructure of free knowledge that "enable[s] us and others to collect and use different forms of free, trusted knowledge." One software for creating such an infrastructure is Wikibase. Our goal within "more knowledge" is to increase the sum of accessible knowledge by creating a network of connected knowledge bases. Wikidata and the Wikimedia projects are part of the network, and so are institutions like GLAMs, science organisations and other open data projects. This semantic web makes it possible to store all kinds of data, even data that doesn't fit directly into the content of the existing Wikimedia projects. Furthermore, the broad variety of data allows connections to all kinds of information. Previously data about artists and their artwork was locked up in one database and data about artists and their countries of origins in another; now, the network of knowledge bases connects them, so that one can also connect that information. And not only that, even truly interdisciplinary connections can be created - leading to new information and knowledge.

Ensuring the integrity of Wikimedia's knowledge

Wikipedia is under increasing pressure to be a place where people go to find reliable information because the rest of the web is losing the trust of the people[13]. This is a big burden we have to live up to. But with an increasing amount of content[14] having to be taken care of by fewer contributors[15] and increasing efforts to try to influence the content in partisan ways we have to provide our contributors with the tools they need to live up to those expectations. Many of those tools will at their core need to understand provenance and other context about our knowledge in a machine-readable way. This is what Wikidata provides both with the references for individual statements as well as metadata about sources that is collected through WikiCite and similar initiatives.

Main roadblocks

There are several roadblocks that are keeping us from realizing the full potential of Wikidata and the Wikibase Ecosystem:

---

[13] "YouTube will flag conspiracy theory videos with additional information from Wikipedia in an effort to tackle the spread of disinformation on its platform." - The Guardian
[14] statistics for English Wikipedia
[15] statistics for English Wikipedia

- Acceptance inside the Wikimedia projects: Right now the movement consists of multiple, very separated projects. This leads to fractured communities. To truly fulfill its potential, Wikidata requires these separate projects to work together to some extent. This causes friction that needs to be addressed.
- Data quality: Wikidata's data quality needs to be high if we want to have the trust of Wikimedia contributors and not lose the trust of the rest of the world. However Wikidata does not yet live up to the expectations of the bigger Wikipedias about references for the data. Additionally data quality may decrease because of challenges during data import, as well as through vandalism and the struggle to keep up to date with an ever-changing world. A lot of effort has already been put into detecting issues, for example by using ORES to find bad edits more easily or adding constraint checks that make finding inconsistencies and mistakes in the data easier. With increasing amounts of data, both improving the verification of data quality and improving data quality itself are important areas to tackle over the next years. One important step is getting to an agreed upon measure for data quality as that currently does not exist.
- Data access usability: Getting data out of and truly benefiting from Wikidata still requires a lot of effort and technical knowledge. For example, building an infobox powered by Wikidata requires skills in template and Lua programming that especially smaller projects often lack. Writing a query on the Wikidata Query Service requires knowledge of SPARQL, which is too technical for many of our intended users. Our APIs are not living up to the expectations of today's programmers of what a good API should be.
- Shared understanding of vision and priorities: Everyone in the movement should be aligned on the vision and priorities for Wikidata and the Wikibase Ecosystem in order to execute on a shared vision instead of pulling in different directions.
- Community Structures: The concept of knowledge is shaped by the culture and the context it exists within. This means that the structures that shape Wikidata and the knowledge that is created with this tool are modeled after our "western" way. If we want to live up to our ambition of creating equity we need to create structures within our movement that allow communities across the globe to shape our tools and knowledge base according to their needs. Only then we will become truly equal.
- Software Development: Today the core of our software, be it MediaWiki or Wikibase, is created by the two biggest organisations in our movement - the WMF and WMDE. The same way the Wikibase Ecosystem becomes a decentralized network of nodes that exchange data with each other, the software we use in this ecosystem should also be built of software components fulfilling needs of different organisations, individuals, institutions. The responsibility of developing the software should be shared by different actors in the ecosystem, allowing for the software to capture the needs of specific contexts advancing the whole movement. Only this way knowledge about our tools can flow across geographies, advance the whole movement and make us more resilient in the process.
- Resources: This vision and the subsequent strategy documents are painting an ambitious picture of the next 3 to 5 years. It will create tremendous opportunities to

advance in our quest to bring more knowledge to more people and to foster more equity within our movement. This however requires more investment in some areas:

- Infrastructure: We are hitting capacity and infrastructure limits with the services we provide today. That is due to elements of the current tech stack but also size and speed of our infrastructure. If we want to provide more powerful and reliable APIs and be ready for more and bigger partners, we will need to invest in this area.
- Processes: There are various areas in which we need to improve the way we are currently working. We need to invest in defining and adhering to cross team and organizational processes to ensure knowledge is shared, technical decisions are made in a timely manner and community and contributing groups and institutions are involved at the right point in time, to name a few.
- People: With the current amount of people working to achieve the vision outlined in this paper we will not manage to achieve what we aspire to. The teams currently working on Wikidata and Wikibase are struggling to work on the high priority features for both, the technical teams do not manage to get through the high priority items on the technical backlog while simultaneously doing bug fixes and maintenance let alone experiment or innovate. In order to create thriving new communities and to support the growing groups of contributors in Wikidata we will also need to expand our community support and engagement capabilities. Combined with proper support for partnership management and acquisition we will be able to move towards fulfilling the aspirations of Wikimedia's 2030 strategic direction and a future in which people everywhere in the world are enabled to freely share and contribute in the sum of all knowledge.