Strategy for Wikidata as a platform

> “Will robots inherit the earth? Yes, but they will be our children.”
> *-- Marvin Minsky*

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

Abstract 2

Background 3

Strategy: Wikidata enables every person to access and share information regardless of language and technology by providing general purpose data about the world 4

Audience 4

Goals 6

Important areas for reaching our goals 6

Opportunities and risks for Wikimedia 8

Why should Wikimedia invest in this now? 8

What are the risks of not acting now? 8

Appendix 9

Existing usage highlights 9

Existing demand 12

Guiding principles and beliefs 13

Capabilities map 14

Abstract

This document lays out the thoughts around the strategy for Wikidata as a platform, the guiding principles and beliefs behind it as well as the needed capabilities. In addition it gives an overview of the progress already made.
The strategy can be summarized as follows: Wikidata is a knowledge base containing general purpose data about the world. This data can be used to build a wide range of applications on top of it, regardless of language and technology. This will help us foster innovation, give underrepresented languages and marginalized groups more leverage and enable machines to give us better information by helping them understand the complexity of the world. In order to make this happen we need to strengthen the core community, provide a high-quality data set, enable contributions by and through re-users, expand the language and culture coverage, provide access to other data by being a hub in the data web, provide data in a reliable environment and make data easily accessible and complex information comprehensible for machines.

Background

Wikidata was started in 2012 with the goal of supporting Wikimedia projects: The idea was to have only one place where people need to update information like the mayor of Cape Town, and that all language Wikipedias would automatically use the up-to-date information. Quickly it became clear that Wikidata not only has the potential of supporting Wikimedia projects, but that the data set of general purpose data about the world is also relevant outside of the Wikimedia projects. By now, there are many projects using Wikidata's data, from applications for learning languages[1] and companies like Mapbox (who use Wikidata to find translations for place names) to most big players like Google or Apple (who use Wikidata for answering users' questions, training AI machines, automated translations and more). For a list of sample users, please see the appendix. Wikidata not only provides a base for any kind of knowledge application, it also allows people to investigate the existing knowledge base much more thoroughly[2]. Since 2018 Wikidata also collects lexicographical data, generating a dictionary between all entered languages of the world, even ones that so far don't have any dictionary (like Estonian to Maltese or Zulu to Tamil).

Wikidata has been successful, because it is curated by a continuously growing, healthy community[3]. It is seeing rapid growth in reuse, and has become an integral linked-data source in the GLAM sector[4]. Wikidata has been awarded by Sir Tim Berners-Lee and Sir Nigel Shadbold with the first ever ODI open data award, won a prize by the German government recognising it as one of 100 innovative places ("Land der Ideen"), and it is an integral part of our effort to provide knowledge as a service and knowledge equity.

With Wikidata we have the next key opportunity for free knowledge in our hand. Now we just have to embrace it.

---

[1] For example Der Die Das to learn articles for German nouns or the Sign Language Browser to learn sign language

[2] For example, with Wikidata we can show the existing gender gap in Wikipedias in a much more nuanced and actionable way (https://www.denelezh.org/gender-gap.php), and equally make statements about geographical imbalances of article contents.

[3] Missing volunteers was one of the reasons why Google gave up on their Freebase project which had pursued a similar goal. If you are interested in more capabilities that make Wikidata as successful as it is, please see the capabilities map in the appendix.
See the editor graph for the number of editors over time that made at least 1, 5 or 100 edits in the previous 30 days.

[4] Just recently in an OCLC survey, Wikidata was named the 5th most-used source of linked data by library users in 2018 (see: "Analysis of 2018 International Linked Data Survey for Implementers" by Karen Smith-Yoshimura. It has also been named as key investment area by the Association of Research Libraries. See also the existing demand section in the appendix.

Strategy: Wikidata enables every person to access and share information regardless of language and technology by providing general purpose data about the world

Audience

Wikidata collects general purpose data about the world, providing

- a data set
- links to other databases
- an ontology around the data set

Wikidata addresses third party application builders. The data is consumed through other services and products. These use:

- data export mechanisms (dumps/ recent changes). These are the easiest ways to consume Wikidata for big players.
- direct access methods (Wikimedia APIs, linked data and other endpoints). These methods work better for smaller players.

The Wikidata editor community consists of many different people[5]. As a broad generalization there are the following editor groups (with overlaps between them):

- Maintainers: The people who are making sure that the project is running, policies are enforced, new properties are created that fit into the larger picture, etc.
- Large-scale editors: The people who run bots and use similar tools to make large scale changes to the knowledge base. These can be data imports or edits to ensure consistency across the knowledge base for example.
- Small-scale editors: The editors who concentrate more on individual Items and make them shine. They often do edits by hand or through special-purpose editing tools like Mix-n-Match for matching external catalogues with Wikidata. Often times, they are not only active on Wikidata, but also Wikipedia or other sister projects[6], and see Wikidata as a way of improving them. They might be part of a wiki project for their special area of interest and expertise.

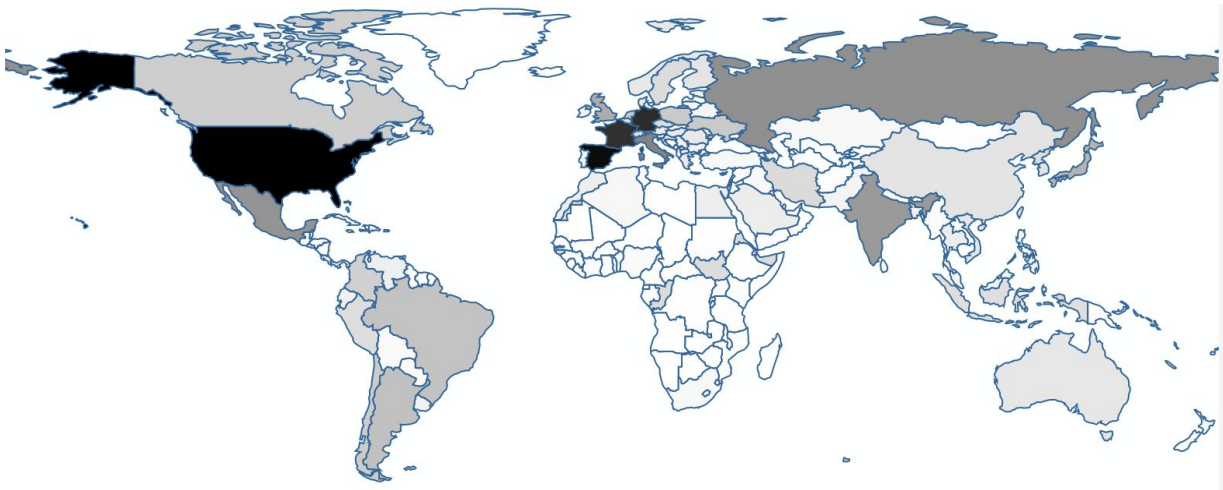
[5] The whole Wikidata community is larger and consists for example also of the data re-users, tool builders and more.

[6] See the co-editor graph for the number of editors who edit on Wikidata and at least one other Wikimedia project.

- Drive-by editors: The people who are making a small change, either on Wikidata directly or through intermediaries, who don't get in contact with the project more. They tend to make a small change like updating some outdated information and then move on.

For all of them we are aiming for a truly international and diverse range of participants. Currently, there is a very strong tendency for contributors to come from Western Europe or the United States of America. Further thought and activities are needed to attract additional contributors from not-yet well represented parts of the world.

Visualisation of where edits on Wikidata are made from currently:



For the data re-use part of our community we currently cluster them into the following groups:

- Blenders: They are using Wikidata's data to augment their own knowledge graph for example by following recent changes or ingesting dumps. Wikidata's data is used as one ingredient for their own data-driven products. They do, for example, build machine translation tools, and question-answering systems. They tend to be less vocal about their use of Wikidata's data among other reasons to not spotlight Wikidata as an entry point into their systems. This type of re-user usually wants to provide feedback on the data, but the necessary processes are not established yet to make it work well.
- Knowledge graph providers: They are using Wikidata's data as an ingredient of a knowledge graph they provide to others as a service.
- Direct re-users: They are using Wikidata's data directly in their products via Wikidata's APIs and the query service. They tend to be smaller and more vocal about their use of Wikidata's data and are valuable advocates. Often this type of re-user also helps by fixing issues they find in the data.

Goals

Wikidata opens up a lot of new opportunities. With it we can:

- Provide more knowledge to foster innovation in a data-driven world: Data is a core part of our lives today but it is concentrated in a few big companies and locked away. This knowledge is not free and open. This makes it hard for new companies, institutions and projects to get started because the technical and legal resources necessary to compete are enormous. This means society is losing out on a lot of knowledge and innovation. We want to open up opportunities for anyone to start projects and innovate in a data-driven world, so that anyone can use data for their services, increasing the shared knowledge and the number of ways to access it.
- Increase the leverage for underrepresented languages and marginalized groups: Focussing on underrepresented languages and marginalized groups is often not considered financially advantageous by companies. Thus too often software is not built towards their needs. With the data provided by Wikidata, we make it easier and less costly to integrate this data into existing applications and underrepresented languages and marginalized groups are empowered to create software specifically for their needs. On top of that Wikidata is also a place for smaller communities to contribute and preserve their knowledge and give it more reach.
- Enable machines to give us better information by helping them understand the complexity of the world: Knowledge graphs as they are run by big companies right now hide a lot of complexity (e.g. the Knowledge Panel for Jerusalem describing it as the capital of Israel without any indication of the complexity around this claim). This leads to applications being built on top of these knowledge graphs that gloss over necessary complexity in our world and dumbed down information being delivered to the end user. By hiding this necessary complexity and not showing people differing views their ability to see and discuss these differences and ultimately resolve them is being taken away. Wikidata allows us to represent different points of view at the same time and to see sources for each of the claims enabling people to interpret the claims in their own way and enables application builders to build on a richer base.

Important areas for reaching our goals

- Strengthen the growing core community: Our community (in the largest sense including data providers and re-users) is our biggest asset. They are contributing the content and are advocates for the project. Wikidata has a lively core community of contributors that hold the project together. It will continue to be a collaborative, forward looking, friendly and inclusive community that is open to challenge and change. This requires conscious and continued effort like providing space for face-to-face meetings, networking and skill exchange as well as efforts to connect editors, data re-users and tool builders closely.
- Provide a high quality data set: Wikidata will be one of the most respected sources of general purpose data. It will help propagate more trustworthy information across the

web. Our data will be used as a an important ingredient for apps, visualisations, websites, artificial intelligence products and more. There will be a large toolbox of data quality tools that help ensure the integrity of our data. They will include parts that find mistakes and inconsistencies in the existing data as well as ones that make sure valid content can't be vandalized as easily. It will be possible to compare Wikidata's data with countless data sources on the web.

- Enable contributions by and through reusers: Wikimedia's content is reused across the internet and more and more of our consumers will not encounter our content in Wikipedia. With Wikidata we are paving the way for re-users of our content to not just use our content but also enable their users to contribute back, thereby increasing our contributor pool.[7] Some of them will make direct changes to the data and others will go through quality assurance steps on either our or the data-reuser's side. On top of that re-users themselves will contribute to Wikidata in order to increase the quality of the content they provide to their users. This can for example happen by making lists of quality issues they found with sophisticated internal algorithms available to our editors or by making edits directly.[8]
- Expand language and culture coverage: With Wikidata we will help increase the coverage of languages and cultures across the web by making it easier to obtain content in them.
- Provide access to other data by being a hub in the data web: Wikidata is a hub that allows machines to jump to many different resources across the web to find information related to a topic. E.g. it allows to easily go from one authority file to another. This instantly makes it easier to access more information about a topic. Some of these resources will be other Wikibase instances in the Wikibase ecosystem. It also makes it possible for a machine to match concepts across them. For example it makes it trivial to find the entry for Douglas Adams in the German National Library's catalog based on his identifier on Goodreads. This matching is a cornerstone of the semantic web.
- Provide data in a reliable environment: Wikidata is part of the Wikimedia movement. This instills a level of trust that helps us attract new re-users because they can count on Wikidata being around for years to come. It takes away their worry that an important piece of their project can be pulled from underneath their feet overnight.
- Make data easily accessible and complex information comprehensible for machines: By consistently modeling and structuring complex data in a globally understood technical standard, Wikidata makes it possible for machines to handle more of the complexity of the world and make use of it downstream. Wikidata's data will be easily accessible for machines through a modern API, SPARQL endpoint, linked data interface and more.

[7] This is happening already with Inventaire for example.
[8] This is happening already with Quora for example.

Opportunities and risks for Wikimedia

Why should Wikimedia invest in this now?

- Essential infrastructure of free knowledge: As the Wikimedia movement strives toward the 2030 vision of becoming the essential infrastructure of free knowledge, Wikidata helps free knowledge achieve reach and scale. Wikidata is a platform where the support of diverse languages and cultures is one of the core principles. It is heavily paying into the strategic direction of Knowledge Equity, since Wikidata can be used as the source for creating and displaying information in less common languages. It is also rather easy to do micro contributions on Wikidata, which is especially useful in areas where mobile phones are more prevalent than computers, and thus likely to pay into Knowledge Equity as well. With proper APIs and investment in building a community of people and institutions we will also be able to attract not just more but also more diverse consumers and contributors of our data. Wikidata is also a knowledge base that other services can easily consume, using modern API methods. This is a key aspect towards fulfilling the strategic direction of Knowledge as a Service.
- Trusted Player: Wikimedia is an institution a lot of people trust.[9] Unlike many others we have no interest in monetizing data or changing our content based on political pressure or financial motives. This gives us a unique advantage.

What are the risks of not acting now?

- Values: If we don't provide a reliable, high-quality knowledge base other players will do it that will not be doing this in a way that is compatible with our values and mission.
- Missed window of opportunity: If we don't do it the large AI-driven companies will find other ways to get the data they need. These other ways will very likely not be collaborative and as open and accessible for everyone as Wikidata is. The easier it is to make use of Wikidata the likelier it is that the more proprietary paths will be shunned, allowing everyone to share in the sum of all human knowledge.

[9] Why People Trust Wikipedia More Than the News and Brand awareness, attitudes, and usage

Appendix

Existing usage highlights

Project	**Why it matters**
Histropedia: a website that enables you to build timelines	The timelines can for example be used in a classroom to teach about a certain era in history or by a GLAM institution to show highlights in the life of one of the artists in their exhibition. They are great advocates of Wikidata and prominently show the source of their data.
Wolfram Alpha (blog post): a computational question answering engine	Wolfram Alpha offers Wikidata as one source of data for computational question answering and integrates it into their existing tools and workflows making it readily available to researchers and programmers.
Inventaire (example): a website for cataloging one's books and making them available to others	It is a great showcase of how our data can be used to enrich the content of another free culture project, making it more usable and valuable. At the same time it is one of the first examples of a re-user of our data encouraging their users to contribute back to Wikidata directly and making it technically possible in their website.
Chronos (blog post): a visual arts browser	Chronos is a great example for a mash-up of our existing content that becomes possible through machine-readable descriptions of Wikimedia content. It is building on an effort by the Sum of All Paintings project that collects data about visual art in Wikidata. It provides a great opportunity for learning about art.
OpenAI (blog post): a research organisation using Wikidata for entity disambiguation	OpenAI uses Wikidata's concepts to build intelligent machine learning systems that can determine which concept is meant in a text if there are several options (e.g. Jaguar the car vs. the animal). This is a crucial steps to text understanding.
Chronas: a history map application	It is an educational website. It is one of the most elaborate uses of Wikidata's data outside large tech companies.
Google: a search engine augmenting their internal knowledge graph with data from Wikidata	Google uses Wikidata's data as one of the sources for its knowledge graph that powers the search engine and gives instant answers in the form of knowledge panels.

WikiGenomes: freely open, editable, and centralized model organism database for the biological research community	WikiGenomes is a website built on top of data in Wikidata. The GeneWiki community are at the forefront of providing high-quality structured data about genes and setting up continuous data import and maintenance processes. WikiGenomes is a good example of a tool helping researchers explore new connections that were previously unknown.
Scholia: a website providing insights into scholarly literature	It makes the networks around scholarly literature easily understandable through simple queries to Wikidata. It makes it possible to find related literature for a topic or understand who are the prolific publishers in a field.
Quora (blog post): a question answering website using Wikidata's ontology to clean up and augment their topic tree	Wikidata's ontology helps Quora provide a better user experience by having a cleaner topic tree. At the same time they help clean up the ontology upstream in Wikidata, leading to better data for everyone.
MySociety: a civic engagement project using Wikidata for data about politicians	MySociety is heavily engaged with Wikidata to bring data about politicians and legislatures up-to-date and complete it. They are a moving away from keeping their own data in order to make the data more widely available and encourage contributions by more people in order to get a more complete picture of politics.
Siri: a personal digital assistant	Siri makes use of Wikidata's data to provide answers to general knowledge questions. It has significant reach through Apple's iPhone and other devices.
Libraries	
Library.Link network: an initiative by library linked data services provider Zepheira (see Library.Link staffer presentation from WikiCite 2018)	Library.Link is increasingly leveraging Wikidata for disambiguation and enrichment. They recently rolled out an "author widget" tool that has been incorporated in the online catalogs of several major US public library systems.
Library of Congress + JSTOR Labs	The Library of Congress and JSTOR Labs used Wikidata for entity reconciliation in a prototype discovery tool focused on traversing different repositories to find resources related to the history of baseball. See example page for Jackie Robinson.
A network of software conservators in research libraries (Yale University,	They have been using Wikidata as a comprehensive technical registry of metadata relating to software and computing environments, demonstrating how this

Bibliothèque nationale de France) and other organizations (Open Preservation Foundation) for digital preservation	infrastructure can be used to run "emulation as a service" of historical software via library catalogs and other access points. See Modeling the Domain of Digital Preservation in Wikidata and Introducing Wikidata for Digital Preservation.
Museums	
Smithsonian's American Women's History initiative: an initiative of the Smithsonian, a group of museums and research centers administered by the US government	Forthcoming: the Smithsonian's American Women's History initiative will be leveraging Wikidata to identify underrepresented women in their collections.
Science Stories: a dynamic discovery prototype with the goal of making unrepresented women of science more visible	Science Stories is aimed at broad audiences which leverages Wikidata to aggregate dispersed resources and archival material about underrepresented women of science. It explicitly mentions using Wikidata on their website.
8 Flemish museums and art collections, coordinated by Belgian digital heritage center PACKED vzw	They have comprehensively documented their holdings in Wikidata. See: https://www.wikidata.org/wiki/Wikidata:Flemish_art_collections,_Wikidata_and_Linked_Open_Data
Metropolitan Museum of Art	As part of its open access program, the Met is steadily developing work-level data about its collection on Wikidata, and has talked about this as a key part of its digital strategy. See: Wikimedia and The Met: A Shared Digital Vision
Cleveland Museum of Art	Modeling on the Met Museum's work, the CMA just did a high-profile open access release in January 2019 and plans to build out representation of its collection in Wikidata as part of its ongoing workflows.
National Library of Wales	They used Wikidata to translate labels to concepts and items in their collection from English labels into Welsh and other languages and used data on Wikidata to geolocate its collections.[10]

[10] Jason Evans and Simon Cobb, "How the World's First Wikidata Visiting Scholar Created Linked Open Data for Five Thousand Works of Art"

Existing demand

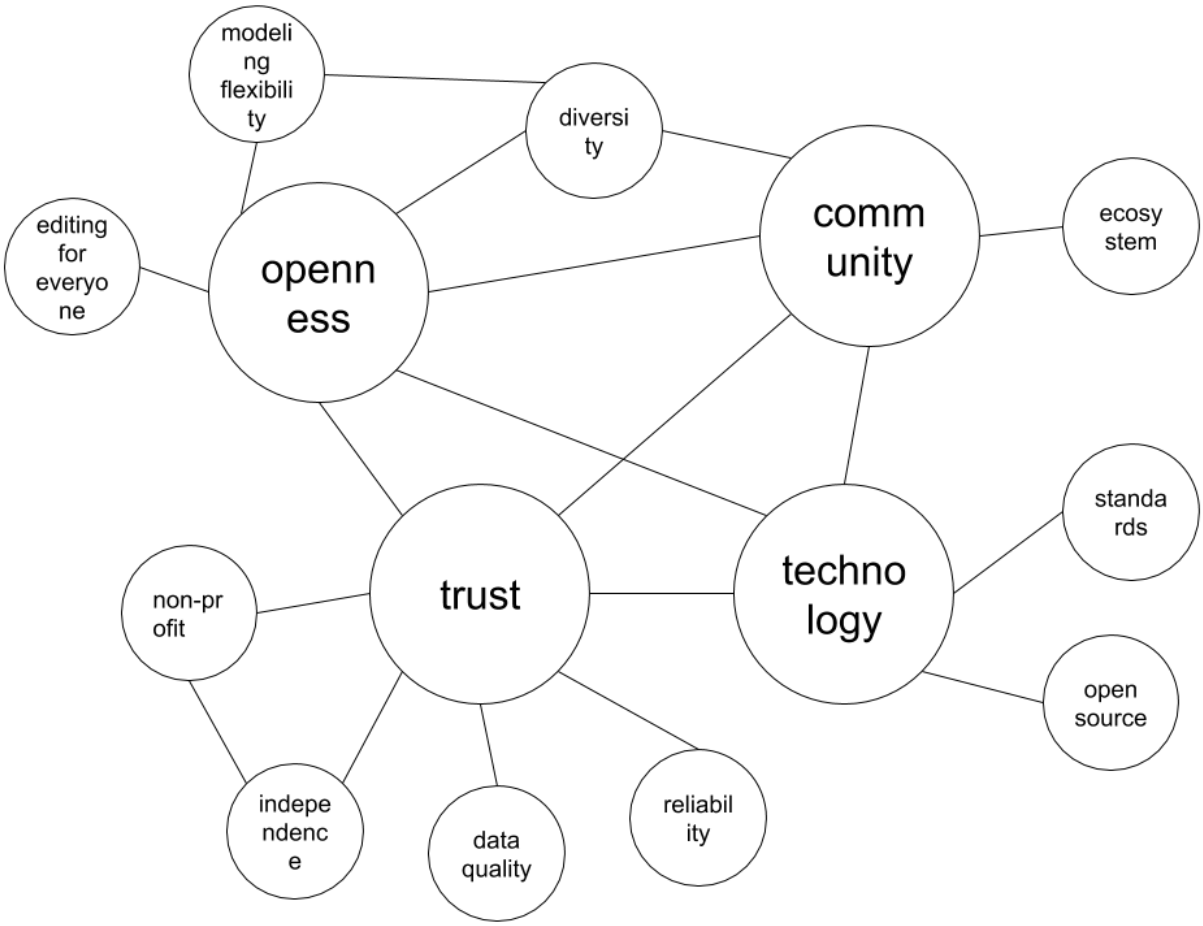
- OCLC survey reveals rapid growth in Wikidata usage by top adopters of Linked Data: When OCLC surveyed 90 top library users of Linked Data, Wikidata went from the 15th most-used source of linked data in 2015 to the 5th most-used source in 2018.
- Forthcoming ARL white paper will recommend Wikidata as key investment area: The Association of Research Libraries (ARL) is a nonprofit organization of 124 of the most prestigious and well resourced research libraries in the US and Canada. A draft of a forthcoming white paper was already widely circulated in the Library Tech space during an open commenting period in late 2018. It contains a set of concrete recommendations for additional investment in Wikidata and/or Wikibase. ARL white paper recommendations are closely heeded by research library directors, so this has the potential for significant influence of the North American research library space (and beyond).
- IFLA White Paper recommends Wikidata for research libraries: Before the latest wave of library experimentation and adoption witnessed in the past two years, the International Federation of Library Associations (IFLA), the most prominent international network of libraries, published a 2016 white paper on opportunities for research libraries with Wikipedia. It made substantial recommendations about the incorporation of Wikidata for authorities reconciliation and catalog enrichment. It pointed to various early adopter examples, and also discussed the GeneWiki and WikiCite initiatives as examples of high-scale potential.
- Wikidata article is most-viewed of 2018 in Code4Lib journal: The Code4Lib article "Wikidata: a platform for your library's linked open data" by Stacey Allison Cassin and Dan Scott, was the most viewed article in 2018 for the journal (even though it was only published in May), and after only 8 months was the 39th most viewed page ever for the open access journal (which has been on the web since 2007 and is widely recognized as the leading library/tech journal).
- Wikidata project incorporation requested for funding by Mellon Foundation: LD4P is a $4 million Mellon Foundation-funded project by Stanford University and several prominent U.S. academic libraries (including Harvard and Cornell) to develop shared linked data tools and workflows. When they approached the Mellon Foundation for a third round of funding last year, Mellon specifically requested that they incorporate Wikidata into their project. WMF GLAM team helped them to scope a set of investigations and to resource a Wikimedian in Residence position, which will be onboarded soon.
- Europeana exploring Wikidata as enrichment source for its core entity collection: Europeana is a virtual library to preserve the cultural and scientific heritage of Europe. Europeana's publishing guide for contributing institutions encourages the use of Wikidata to enrich metadata records (e.g., to translate titles of works), and is exploring Wikidata as an enrichment source for its core entity collection.
- Wikidata trending at recent GLAM conferences: Wikidata has been central to several recent GLAM conferences, among them:

    - EuropeanaTech 2018: Wikidata figured prominently at last year's EuropeanaTech conference, with popular sessions on Structured Data on Commons and a keynote by WMF Community Programs lead Ben Vershbow. On a fun "zeitgeist" note, during a closing plenary session reflecting on the conference, attendees used a web app to assemble a word cloud of major themes and takeaways. Wikidata was at the center.
    - SWIB (Semantic Web in Libraries) 2018: Wikidata and Wikibase were reportedly hot topics at the recent SWIB conference in Bonn. Both the German and French national libraries announced Wikibase pilots focused on authorities data. A Wikibase workshop delivered by Stacy Allison-Cassin and Dan Scott was among the most popular sessions.
- Professional development around Wikidata: Interest in the GLAM sector is growing and Wiki Edu is currently developing Wikidata professional education aimed at librarians.

Guiding principles and beliefs

- People before data! The community is the most valuable thing we have. We do everything in collaboration with the community.
- We provide people with the tools they need to help themselves. We can't and shouldn't solve all the problems for them but need to put them in a position where they can.
- We build a healthy ecosystem around Wikidata. We don't do it all ourselves.
- We generally aim for sustainability over speedy growth and data quality over quantity.
- The world is complex and we need to deal with that complexity instead of hiding necessary complexity. We are not looking for truth but verifiability.
- We want to make structured data available for everyone to level the playing field.

Capabilities map



- Openness: Wikidata is built on openness. This openness comes with the desire to empower everyone to contribute and allows us to bring in diversity in our data as well as our community. Both is supported by the flexibility in modelling that the Wikibase software allows for.
- Community: The community is what makes Wikidata possible. Comprehensiveness comes from diverse areas of knowledge and points of view, and those come from a diverse community. The Wikidata community also supports the Wikibase ecosystem by promoting and championing it and benefits from it through the collaborations it enables.
- Trust: Wikidata benefits from the trust people have in the Wikimedia projects. They trust that as a non-profit it is independent. Wikimedia has a track-record of reliability, assuring them that Wikidata will not go away anytime soon and endanger the project they built on top of Wikidata. This trust needs to be sustained by providing high-quality data.
- Technology: Wikidata needs to rest on a solid technological base that is open source. By building on or tying into existing standards like SPARQL, Wikidata's data becomes more accessible to developers.