Strategy for the Wikibase Ecosystem

> “The internet explodes when somebody has the creativity to look at a piece of data that's put there for one reason and realise they can connect it with something else.”
> *-- Sir Tim Berners-Lee*

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

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

August 2019

Abstract

This document lays out the thoughts around the strategy for the Wikibase Ecosystem, 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: Wikibase, a software suite for collaborative knowledge bases like Wikidata, enables us to build a thriving ecosystem of interconnected Wikibase instances. This ecosystem will help us to open up data that is currently hidden in silos, find connections in the data that have not been discovered yet, help people collaborate around that data and strengthen the open knowledge ecosystem as a whole. In order to make this happen we need to focus on enabling connections between data and people, partner with the main players in their field and then utilize network effects, leverage mandates to open up data and maximize the competitive advantage gained via Wikidata.

Background

Wikibase is a software suite for running a collaborative knowledge base, built by Wikimedia Deutschland and used by an increasing number of cultural heritage and science institutions around the world. Information entered in Wikibase is machine-readable, making it a base technology that provides knowledge as a service[1] - it is the background that can be used by any application to create diverse ways for people to consume (and share in the sum of) knowledge. When Wikibase's development was started in 2012, its initial purpose was powering Wikidata. However, since the beginning, people have used Wikibase to structure their own repositories of knowledge (see the appendix for some examples of existing Wikibase installations). When it became easier to set up a Wikibase instance, interest grew even more[2]. In 2018 the number of existing Wikibase installations nearly doubled[3].

Example: Wikibase for historical research
University of Erfurt's Gotha Research Centre is researching the historical activities of the Illuminati in the eighteenth century. They set up their own Wikibase instance last year, storing about 16.000 data sets linking to roughly 10.000 documents. Having made great experiences of collaboration through a wiki, they aim to make their data machine-readable and broaden the international cooperation by using multilingual Wikibase. Furthermore, they are eager to connect to other knowledge bases for even better research opportunities and reach. For more information and details, see this blog post.

Wikibase is useful for any kind of heterogeneous data, where connections between the data play an important role. This is relevant for all kinds of fields, starting from (digital) humanities to life sciences. Recently we have seen a lot of activity from the GLAM sector: Not only was Wikibase a trending topic at recent GLAM conferences[4], it also seems to be the future software for storing authority data. Seven national libraries, among them Germany and France, have run substantial pilots[5], and libraries from seven more countries have communicated interest in evaluating or using Wikibase or Wikidata as a platform for creating or participating in linked data

[1] See also Wikimedia's strategic direction

[2] For enthusiastic reactions around the Docker images, see for example a blog post by the author of O'Reilly's Learning SPARQL: *"Many of us have waited years for an open-source framework that makes the development of web-based RDF applications as easy as Ruby on Rails does for web-based SQL applications. This dockerized version of Wikibase looks like a big step in this direction."*

[3] Timeline of Wikibase instances

[4] Wikidata and Wikibase were reportedly hot topics at the SWIB 2018 conference in Bonn, with both the German and French National Library announcing Wikibase pilots focused on authority data.The GNDCon 2018 in Germany also paid much attention to Wikibase, discussing how it can help authority data flourish.

[5] These national libraries are Germany and France, Italy, the Netherlands, Wales, Spain and Sweden.

work for the sector[6]. For an in-depth analysis of the partnership with the German National Library, please see also the appendix.

While Wikibase in itself is already useful, it reveals its full potential when being part of a linked open data network. With Wikibase it will be possible to share information between separate knowledge bases hosted and maintained by others. E.g: One might get all general purpose data from Wikidata, and provide Wikidata and others with information from one's own Wikibase instance, e.g. about important people of the middle ages who were part of the Illuminati. With a network of Wikibase (and other open data) knowledge bases, we have the chance to actually create a new version of the semantic web, increasing the sum of knowledge that is accessible to the world.

With Wikibase we have the next key opportunity for free knowledge in our hands. Now it is on us to embrace it.

[6] Libraries in the United States (See this presentation, and interest of Library of Congress at row 52 in OCLC survey), Canada, Mexico, Finland (see OCLC paper and row 61 in OCLC survey), Portugal, Switzerland and the United Kingdom (see row 15 in OCLC survey)

Strategy: Wikibase powers a thriving linked open data web that is the backbone of free and open knowledge

Audience

- We provide an easy-to-use software bundle that is the base technology for people to make data openly accessible
    - We provide a website where the software can be downloaded. There are a multitude of options how to make Wikibase accessible: Apart from providing users with the means of setting up Wikibase instances by themselves, they range from hosting Wikibase ourselves to making Wikibase a hosting option on platforms of other cloud hosting providers. There are also options in between, like offering specific services such as spam protection for self-hosted Wikibase instances to hook into. At this point, we need to determine which (combination of) options offers the best balance of investment vs. impact. Since the decision where to put our focus and how to make Wikibase sustainably accessible in the long run is about creating long term business relationships, we expect a time frame of about three years until full implementation of a strategy.
- We consult and inform around setting up and using Wikibase
    - We focus on projects that aim to open up closed data, and making already open data even more accessible.
    - Contentwise, we start with GLAM institutions, especially libraries[7], and then move on to other relevant areas, such as science institutions.
    - From the organizational perspective, we focus on institutions with an educational mission, since they are close to our goal of sharing knowledge. We expect foundations and the public sector to be especially interesting for us.

Additional market research will be done over the next months to fully evaluate in which areas Wikibase can have the most impact initially and long-term.

Goals

The Wikibase Ecosystem opens up a lot of new opportunities. With it we can:

- Open up data that is hidden in silos: There is a lot of data to be freed, but often times institutions hesitate. The reasons include licences that allow the unconditional reuse of their content and opening the floodgates by allowing contributions from outsiders. If

[7] There are multiple reasons why GLAMs and libraries specifically should be the starting point: Historically, Wikimedia has had strong relations in that field and so we can build upon existing knowledge and relationships. Furthermore, there is strong overlap between the visions of libraries and ours, and there is a strong need for better software that we can help satisfy.

institutions can set up their own Wikibase instances, they have a good technical base to make their data accessible - on their own terms. While we see institutions (such as libraries) evaluating Wikibase as the new software, replacing an existing solution, we can also imagine Wikibase as filling a not-yet addressed need for data storage in various areas that weren't previously supported by data, and new applications developed for them. With a Wikibase Ecosystem and the standard set of processes, tools and APIs that Wikibase provides we are making it easier for data holders to open up their data and become a part of the linked data web.

- Connect data to surface undiscovered connections: For a long time, people have envisioned a future where data is shared and seamlessly integrated across boundaries. We have the chance to actually enable this. With multiple Wikibase instances, it is much easier to connect data than with different technologies being used. We are connecting data and allowing connections between data and disciplines that have not been thought of before. We allow mashups that have not been imagined before and by doing so we create insights and discoveries that have not been possible before[8]. Connecting the fields of biology and material science brought us self-cleaning surfaces in our bathrooms and kitchens by learning from the Lotus Effect for example. There are many new connections yet to be discovered. Already today Wikidata is connected to over 3600 other databases, websites and projects, making much more information accessible.
- Connect people to enable collaboration: By enabling a vibrant Wikibase Ecosystem we can connect not just data but also people. Today institutions like the German National Library want to collaborate more with others but are hindered by technological barriers. With the Wikibase Ecosystem practitioners inside and across disciplines are brought together to share their data and collaborate on enriching and expanding it. This will lead to new collaborations and innovation that didn't happen before.
- Strengthen the open knowledge movement as a whole in a new, decentralized, linked data ecosystem: This is our chance to significantly strengthen the open knowledge movement as a whole, and the role of Wikimedia as the core player within it. We can be the ones who constantly remind the world to make data open and freely accessible, and who make open and free data an integral part of the net community, and "not just" of the Wikimedia projects. The Wikibase Ecosystem has the power to dramatically strengthen bonds within the open knowledge community, to keep up a strong, common voice for an internet that is not just dominated by commercial interest and private platforms.
- Enable the creation of new products and services: In the 2000s, the tech world saw an explosion in mapping APIs which brought together disparate but powerful data - business information, public transit data, satellite imagery, vector tiles, drawing features, and more. The Map Services innovations of the past 20 years started out simply as ways to do basic navigation, but today they have enabled game-changing new products like Uber/Lyft, Foursquare, Strava, MapMyRun, and countless others. We have the opportunity to be at the forefront of a new but similar phenomenon - linked data services that can power highly useful products we can't even imagine today.

[8] Story about how Wikidata uncovered a forgotten sister city connection

Important areas for reaching our goals

- Focus on enabling connections between data and people: Wikibase's main selling points for institutions are its collaborative nature and the possibility of connecting with other data. While the collaborative aspect is well represent through the existing software, it is important to focus on making the most out of the connections in the data, e.g. through better technical possibilities to connect knowledge bases and fostering the social aspects around it. Only then will people be able to collaborate better, and to uncover yet unknown connections.
- Partner with the main players in their field, utilize network effects and branch out: Main players set examples for many other institutions in their field. For example, it will be much more attractive for libraries to be part of the Wikibase Ecosystem once the German National Library and other major libraries adopt our platform. The more hubs exist in the Wikibase Ecosystem, the more useful and compelling it is to join. By focusing on specific areas, we will reach the point of usefulness for institutions much quicker. We will see even bigger breakthrough benefits when we reach the next step. Once there are multiple areas covered, we can create interdisciplinary connections - something that is very hard to do right now.

Example: Wikibase for authority files
WMDE is currently supporting the German National Library to bring their integrated authority file to Wikibase. Sparked by their interest, multiple other European national libraries have followed suit piloting and evaluating the software. We are expecting to see an even bigger demand from national libraries once the first authority files have successfully been set up on Wikibase instances. For more information and details, see the flagship partnership with GND and national libraries in the appendix

- Leverage mandates to open up data: If grant makers make releasing their data as open data a requirement for grant recipients, there will be more open data in the world that can be connected to the Wikibase Ecosystem.[9]
- Maximize the competitive advantage gained via Wikidata: Wikibase comes paired with a powerful, well-maintained, built-in advantage: Wikidata. This advantage helps institutions get their new Wikibase instance off the ground quickly with thousands of ready-to-use entities for both common and esoteric concepts.

[9] This is already happening for example with grants from the Bill and Melinda Gates Foundation.

Opportunities and risks for Wikimedia

Why should Wikimedia invest in this now?

- More usable data for Wikimedia projects: A strong Wikibase Ecosystem will greatly increase the amount of freely accessible, maintained data the Wikimedia projects can use. And since that data will be primarily in the hands of its stewards at the host organizations, this benefit comes without burdening existing communities with the maintenance of vast amounts of new open data.
- Take pressure off Wikidata: Wikidata is under a lot of pressure from various sides. For example it is asked to handle more data than it can handle and wants because importing to Wikidata is the only viable option data donors see. In addition there is pressure to change its license in order to accommodate existing data that is not published under CC-0. With a strong Wikibase Ecosystem, the pressure on Wikidata will be drastically relieved.
- Deeper connections to institutions and social actors like GLAMs: So far, we have mainly been asking institutions to donate content to us. By fostering a Wikibase Ecosystem, we also have something significant and useful to give back to institutions. This kind of mutually beneficial relationship would strengthen our existing bonds with institutions, and create new partnerships with institutions we haven't reached yet.
- Support linked open data, a very sought after part of the open knowledge movement: Linked open data is currently a very important topic in the open knowledge movement. Supporting it would not only strengthen the movement, but also open up opportunities for grants from other foundations.

What are the risks of not acting now?

- Missing the opportunity to lead the next stage of growth in this space: In recent years, the interest in and demand for Wikibase has greatly increased. Between 2012 and 2017, 12 Wikibase instances were created. But in 2018 alone 11 new instances were set up. We believe this increase comes from having created a mature product as well as from investing in the Wikibase community and supporting potential users. There is already a community, that even organized funding for a three part-workshop series. If we don't keep up and extend our game, this is likely to be taken over by other players and we will lose the ability to steer the development.
- Losing the chance of creating an open linked data network: Already at this stage we can see players collaborate with top universities to further develop Wikibase for specific use cases. These are not as good news as it sounds - so far these are not planning to contribute back to the larger ecosystem in any way. If we don't invest in Wikibase now, there is a chance commercial players will take over, and the data network will not be free and open.

Appendix

Existing usage highlights

Project	**Why it matters**
Lingualibre (blog post): a website that allows people to generate lists of words, record them, and upload them automatically on Commons	LinguaLibre is a very good example of how Wikibase can be used for a 3rd-party project that is connected to Wikidata and helps us improve and expand our content. It is based on Mediawiki and Wikibase, and makes edits on Wikidata and other wikis with a bot.
Factgrid (blog post): a website for researchers in the humanities to collect their research findings initially with a focus in the Illuminati	Factgrid is a trailblazer project that helps iron out a lot of the issues of running Wikibase to run a website other than Wikidata. It brought Wikibase to the next level where new partners like the German National Library are ready to try it. It will also be a valuable source for high-quality data for Wikidata and a place to reference this data to.
Rhizome (blog post): a catalog of digital-born art	Based at the New Museum in New York City, Rhizome is dedicated to the preservation and promotion of digital art. They have been among the earliest adopters of Wikibase, using it since 2015 to describe their own catalog of internet artworks with specialized preservation metadata.
DroidWiki: a website about Android phones and apps	DroidWiki uses Wikibase as a data backend for their project in a similar way to Wikipedia.
EAGLE project: a multi-lingual online collection of millions of digitised Items from European museums, libraries, archives and multimedia collections, which deal with inscriptions from the Greek and Roman World	The EAGLE project was the first production user of Wikibase outside Wikimedia, years before anyone else and their users were very satisfied with the way they can collect data even at that point in time already.

Wikidata: the first and largest Wikibase instance	Wikidata is the first and largest Wikibase instance. It is the most prominent node in the Wikibase Ecosystem.
Linked Jazz: a project that develops an open knowledge graph of jazz music (see also Wikibase for Research Infrastructure - Part 1)	They are currently experimenting with Wikibase and say they plan to migrate the project over to it as a new backend.
Enslaved: People of the Historic Slave Trade: a research project at Michigan State University	Enslaved is using Wikibase as a linking and reconciliation hub to unify seven separate scholarly databases with information documenting the international slave trade through history.

Flagship partnership: GND, national libraries and authority files

Why libraries: Libraries have been natural allies for at least a decade. But only recently, with the rise of Wikidata, have we glimpsed the possibility of a larger strategic collaboration around shared open infrastructure. The seeds of this can be seen most overtly in the Wikidata community's mapping of authority files from leading libraries. These authority files are like the 'topic maps' of libraries and research fields, and mapping them to Wikidata helps expose library resources to the wider web, while connecting Wikimedia contributors and readers to authoritative information on a vast array of topics. Libraries have taken notice of this beneficial relationship and have started to look seriously at Wikibase as a tool they might use to turn their authority data into an open infrastructure that they themselves can maintain —while continuing to connect to Wikidata.

For a number of reasons, libraries make obvious sense as a first vertical investment in sector-specific Wikibase adoption. Like Wikimedia, libraries are global curators of knowledge and a universal fixture in the global knowledge landscape. In many countries around the world, there is a national library with a responsibility for organizing and making discoverable the totality of recorded knowledge in its cultural and linguistic sphere, alongside countless local and specialized libraries organizing the knowledge in deeper contexts. In short: libraries, as a community, match Wikimedia's scope of aspiration toward the sum of all knowledge.

They are also a professional community that has long embraced data standardization and exchange. Though their data formats have become antiquated, libraries are endeavoring to transition to a modern linked open data model that would better serve the complexity of their collections and do a better job of surfacing their resources on the popular web. If Wikidata already is enabling that surfacing through the mappings described earlier, Wikibase offers

libraries a practical and empowering solution for developing their own open data infrastructure. Libraries have tried building this infrastructure themselves and it is cumbersome, expensive, and too technical for broad numbers of librarians to maintain.

No doubt many sectors could potentially benefit from Wikibase, but no other professional community is already stepping up and proactively experimenting with Wikibase in the way that we are seeing with libraries.

Though research should undoubtedly be undertaken to identify other professional opportunity spaces for Wikibase, libraries offer the most direct (and urgent) path now.

What is the GND: The Integrated Authority File, also called "Gemeinsame Normdatei" (GND) is the authority file of the German National Library (DNB) in cooperation with German-speaking libraries in Germany, Austria, and Switzerland and the magazine database ZDB. The GND organizes data about people, subject headings and corporate bodies from catalogues. It is used mainly for documentation in libraries and increasingly also by archives and museums. The GND holds more than 14 million records and is licenced under CC-0.

What the GND is currently struggling with:

- The software the GND currently uses is not well suited for their needs anymore: e.g. it is not built for easy linking of records, it does not have a record history (each save completely overwrites the previously existing information), and it only allows the entering of one point of view, with only one reference per record.[10]
- Although, it is published under CC-0, the GND is currently very difficult to access, because there is fear of unqualified people tampering with a very well curated data set. However, it is at the same time increasingly difficult for DNB staff to curate the data - it is simply too much.
- Currently the authority of the GND comes mainly from the reputation of the German National Library and its staff. However, this means that often times it is not really possible to fact check - data that was imported from existing library repositories often does not even have one reference per record[11].

Why they are interested:

- The Wikibase software would suit their needs better, and allow them to open up more - shifting from authority by institution's reputation to authority by proof.
- They have already made great experiences with the Wikimedia movement:

[10]FactGrid's Olaf Simons wrote a great analysis of the current state of the software in his blog post about the GNDCon (in German).
[11] See Olaf Simon's blog post about the GNDCon (in German).

    - Since 2005 the Wikipedia community has been connecting articles with the corresponding GND entries, substantially supporting the German National Library in uncovering errors[12].
    - The GND is already now linking to Wikipedia articles from each of its people entries.
    - Wikidata has become a repository for authority file matching, since it stores not only links from their Items to the GND, but also to the French equivalent and others.
- They would like to make use of Wikidata's Items more, e.g. through federation.
- Being part of a Wikibase Ecosystem would bring them closer to their goal of making use of authority files as a backbone of a semantic web of culture and science:
    - As part of its development program 2017- 2021 the German National Library started an initiative for the connecting of authority files[13]. They highlight the potential of authority files to become the backbone of a machine-readable, semantic web of culture and science. Opening up the GND for all GLAM areas and a seamless collaboration with open community projects like Wikidata and Wikipedia is one of the goals of the program.

Where are we now: In 2018 the first steps for a partnership between WMDE and the German National Library were taken[14], culminating in a very Wikidata/Wikibase-heavy GND Con, whose main talk was about the authority files of the national libraries of Sweden, France and Germany, and Wikidata with a following panel discussion. Since then WMDE has had workshops and regular syncs with the German National Library, and will continue doing so, to support them in migrating to Wikibase. The current work is an assessment phase to evaluate Wikibase for the GND. It will lead to a second phase in 2019/2020 where the requirements and needs will be implemented depending on funding.
Apart from the German National Library, six other national libraries have run substantial pilots [15], and libraries from seven more countries have communicated interest in evaluating or using Wikibase or Wikidata as a platform for creating or participating in linked data work for the sector[16]. We are already in closer contact with the French National Library and with a Canadian working group that plans to pilot a national authorities system for Canada.

What needs to happen:
Apart from supporting the German National Library in opening their integrated authority file, we want to create a pilot cohort of adopters of the same use case, France and Canada

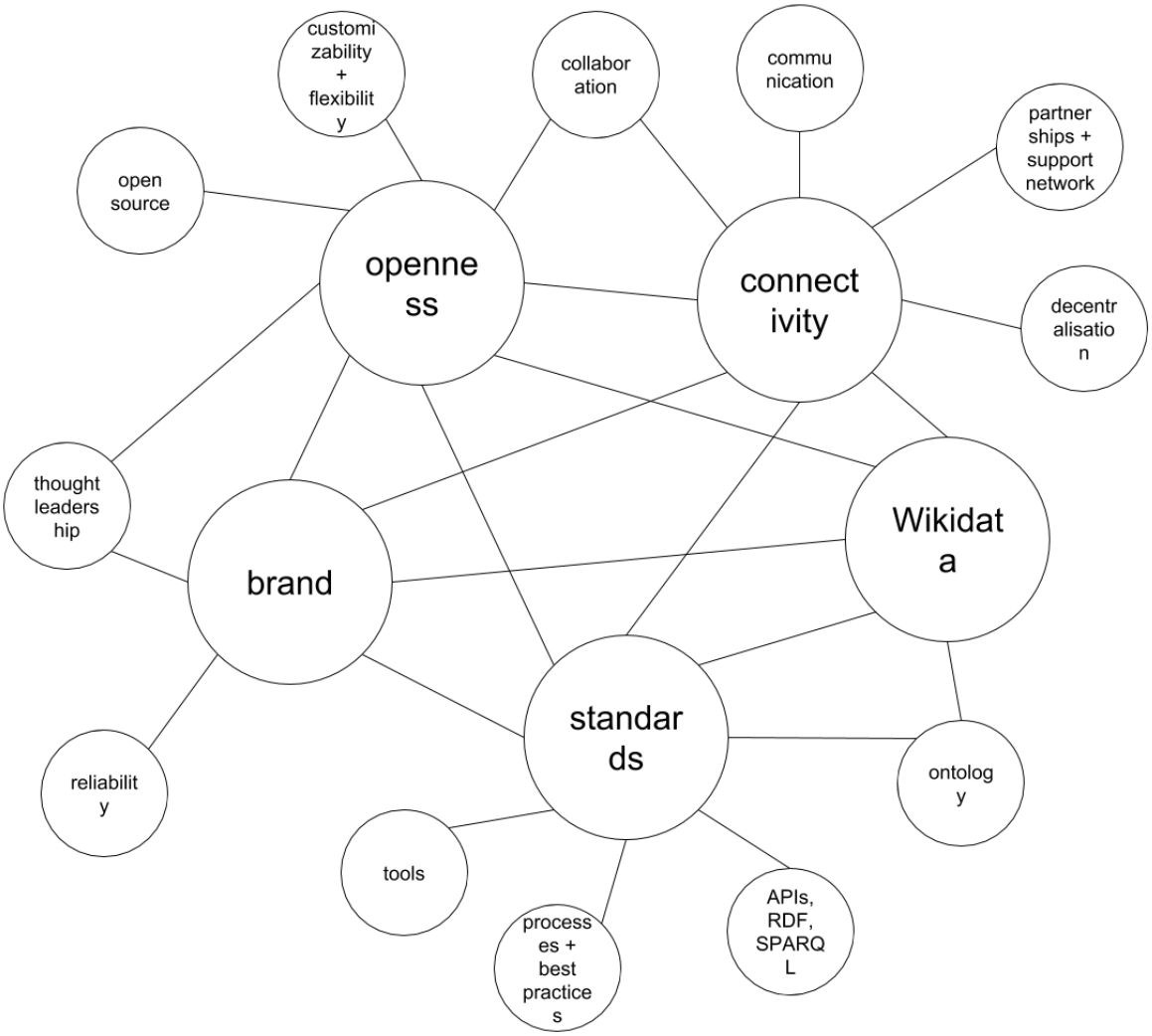
[12] See the report (in German) here and the way how errors are being reported (in German also) here.
[13] You can find the initiative's plans (in German) here.
[14] See also the blog post about a workshop between WMDE and the German National Library to clarify how to "wikify" the integrated authority file.
[15] These national libraries are France, Italy, the Netherlands, Wales, Spain and Sweden
[16] Libraries in the United States (See this presentation, and interest of Library of Congress at row 52 in OCLC survey), Canada, Mexico, Finland (see OCLC paper and row 61 in OCLC survey), Portugal, Switzerland and the United Kingdom (see row 15 in OCLC survey)

standing out as potential candidates. We also see strong potential to leverage the WMF's strong relationship with the International Federation of Library Associations (IFLA) to identify and engage a more globally diverse set of national libraries. With a multi-national library authorities project there are not only funding opportunities, we could also make use of WikiCite or the SWIB conference for strengthening bonds and growing the early adopter community. With the German National Library, and other pilots making use of Wikibase, not only the ecosystem of open knowledge will have gained more extremely valuable players. It will also be much more attractive for other national libraries to follow suit. For example, once multiple authority files are connected, the national librarians will be able to discuss and plan the future of their authority files together in much more detail. This effect increases as more libraries join in sharing their authority files. We are convinced that the German National Library's authority file with its more than 14 million files is a great first partner that will speed up the avalanche of Wikibase interest in the field of libraries - especially when we manage to quickly create an international pilot cohort.

Guiding principles and beliefs

- Data does not need to be stored in any single place like Wikidata but we do want as much data as possible to be freely available, machine-readable, and connected among others to Wikidata.
- No single project, company or institution is an expert for all kinds of niche data. It is better for everyone if this data is maintained by the people close to it and then shared with the world as appropriate.
- If projects, companies and institutions receive support in building out their own knowledge base from an open movement like Wikimedia there is a higher chance of opening up more data in a collaborative manner.
- If more data is available in a structured and machine-readable format (ideally even using the same software and ontology) then comparing and exchanging data becomes significantly easier.
- Institutions, companies and other projects have different requirements for permissions to access and edit their data, licensing, organisational structures and more. Giving them the ability to open up on their own terms is good.

Capabilities map



- Openness: The Wikibase Ecosystem is built on open data, open software, and open processes which enable easy connections, collaborations, and exchanges. Our openness also allows for extensive customizability and flexibility so that Wikibase can be used in a wide range of projects.
- Connectivity: The Wikibase instances are connected to each other. These connections bring value by linking data that was previously disparate, leading to the discovery of new knowledge. This connectivity also allows for more decentralization of knowledge, making the whole ecosystem more robust. The connections in the world of Wikibase don't stop with the data though - there's also a strong partnership and support network of people, and this simplifies finding solutions and the "getting started" process.

- Wikidata: Wikidata is an asset for the Wikibase Ecosystem because it provides a basic ontology and is a well-known entry point into the ecosystem.
- Standards: The Wikibase Ecosystem is strengthened by a common set of tools, open processes, and best practices that provide flexible but consistent ways to access data. The unified and standardized data management methodologies allow considerable amounts of work to be shared between different actors in the ecosystem.
- Brand: The Wikibase Ecosystem relies on the strong reputations of Wikimedia, Wikipedia, and Wikidata. These brands are seen as reliable and trustworthy, and this assures Wikibase users that the platform will not vanish anytime soon. But the growing Wikibase brand also provides a potential branding boon to Wikimedia as an opportunity to position Wikimedia as a leader and advocate for more open data and collaboration.