From Freebase to Wikidata: The Great Migration

Thomas Pellissier Tanon*
Google, San Francisco, USA
thomas@pellissier-tanon.fr

Denny Vrandečić
Google, San Francisco, USA
vrandecic@google.com

Sebastian Schaffert
Google, Zürich, Switzerland
schaffert@google.com

Thomas Steiner
Google, Hamburg, Germany
tomac@google.com

Lydia Pintscher
Wikimedia, Berlin, Germany
lydia@pintscher.de

ABSTRACT

Collaborative knowledge bases that make their data freely available in a machine-readable form are central for the data strategy of many projects and organizations. The two major collaborative knowledge bases are Wikimedia's Wikidata and Google's Freebase. Due to the success of Wikidata, Google decided in 2014 to offer the content of Freebase to the Wikidata community. In this paper, we report on the ongoing transfer efforts and data mapping challenges, and provide an analysis of the effort so far. We describe the *Primary Sources Tool*, which aims to facilitate this and future data migrations. Throughout the migration, we have gained deep insights into both Wikidata and Freebase, and share and discuss detailed statistics on both knowledge bases.

General Terms

Human Factors, Documentation

Keywords

Crowdsourcing Systems, Semantic Web, Wikidata, Freebase

1. INTRODUCTION

Large Web-based knowledge bases that make their data available under free licenses in a machine-readable form have become central for the data strategy of many projects and organizations. They find applications in areas as diverse as Web search, natural language annotation, or translation. Exemplary works are those of West *et al.*, who in [33] have proposed a way to leverage Web search-based question answering technology to fill in the gaps in knowledge bases in a targeted way, or research of Gesmundo and Hall [13], who have trained a syntactic parser paradigm that learns from large-scale knowledge bases.

One such collaborative knowledge base is Freebase, publicly launched by Metaweb in 2007 and acquired by Google in 2010. Another example is Wikidata, a collaborative knowledge base developed by Wikimedia Deutschland since 2012 and operated by the Wikimedia Foundation. Due to the success of Wikidata, Google announced in 2014 their intent to shut down Freebase and help the community with the transfer of Freebase content to Wikidata [10].

Moving data between two knowledge bases that do not share a similar design is usually a problematic task and requires the careful mapping between their structures. The migration from Freebase to Wikidata was no exception to this rule: we encountered a number of to-be-expected *structural* challenges. However, even more demanding was the *cultural* difference between the two involved communities. The Freebase and Wikidata communities have a very different background, subtly different goals and understandings of their tasks, and different requirements regarding their data.

In this paper, we describe how we support the Wikidata community with the migration of Freebase content to Wikidata. We programmed the *Primary Sources Tool* and released it as open source software to facilitate the transfer of the Freebase dataset. The *Primary Sources Tool* is also developed with an eye on future data migrations from other sources. We created and published mappings of the Freebase dataset. Throughout the migration, we have gained deep insights into both Wikidata and Freebase, and share and discuss detailed statistics on both knowledge bases.

The remainder of this paper is structured as follows. First, we introduce Freebase and Wikidata in Section 2. Then we describe our methodology and the metrics used to measure the migration in Section 3. In order to support the migration, we have developed a set of open source tools that we present in Section 4. We then show the results of the migration and discuss these statistics in Section 5. Before we close, we report on related work in Section 6, followed by an outlook at proposed next steps and a conclusion in Section 7.

2. BACKGROUND

2.1 Freebase

Freebase[1] is an open and collaborative knowledge base publicly launched in 2007 by Metaweb and acquired in 2010 by Google [3]. It was used as the open core of the Google Knowledge Graph, and has found many use cases outside of Google. Due to the success of Wikidata, Google announced

*The author was an intern at Google when the majority of the work was done.

[1]Freebase: `https://www.freebase.com`

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
`http://dx.doi.org/10.1145/2872427.2874809`.

in 2014 its intent to close Freebase and help with the migration of the content to Wikidata [10, 15].

Freebase is built on the notions of *objects*, *facts*, *types*, and *properties*. Each Freebase object has a stable identifier called a *"mid"* (for Machine ID), one or more types, and uses properties from these types in order to provide facts. For example, the Freebase object for Barack Obama has the mid `/m/02mjmr` and the type `/government/us_president` (among others) that allows the entity to have a fact with the property `/government/us_president/presidency_number` and the literal integer "44" as the value.

Freebase uses Compound Value Types (CVTs) to represent n-ary relations with $n > 2$, *e.g.*, values like geographic coordinates, political positions held with a start and an end date (see Figure 1 for an example), or actors playing a character in a movie. CVT values are just objects, *i.e.*, they have a mid and can have types (although they usually only have the compound value type itself). Most non-CVT objects are called *topics* in order to discern them from CVTs.

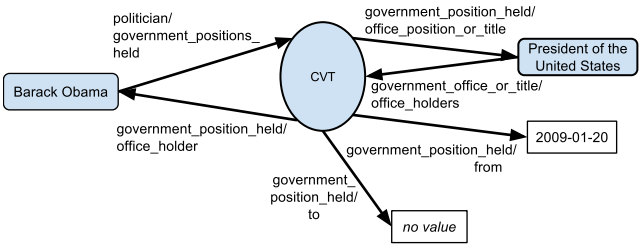


Figure 1: Exemplary Compound Value Type (CVT) in Freebase: Presidency of Barack Obama.

The content of Freebase has been partially imported from various sources such as Wikipedia [1] or the license-compatible part of MusicBrainz [30]. Over the years, the Freebase community and Google have maintained the knowledge base. When Freebase was turned read-only on March 31, 2015, it counted more than 3 billion facts about almost 50 million entities. Freebase data is published as an N-Triples dump in RDF [6] under the Creative Commons CC-BY license.[2]

2.2 Wikidata

Wikidata[3] is a collaborative knowledge base launched in October 2012 and hosted by the Wikimedia Foundation [31]. Its community has been growing quickly, and as of mid-2015, the community comprises about 6,000 active contributors [34].

Wikidata's data model relies on the notions of *item* and *statement*. An item represents an entity, has a stable identifier called "qid", and may have labels, descriptions, and aliases in multiple languages; further statements and links to pages about the entity in other Wikimedia projects—most prominently Wikipedia. Contrary to Freebase, Wikidata statements do not aim to encode true facts, but *claims* from different sources, which can also contradict each other, which, for example, allows for border conflicts to be expressed from different political points of view.

[2]Dump: `https://developers.google.com/freebase/data`
[3]Wikidata: `https://www.wikidata.org`

A statement is composed of a claim and zero or more references for this claim. The claim itself is composed of one main property–value pair that encodes the main (claimed) fact like "taxon name is Pantera Leo" and optional qualifiers to add information about it like "taxon author is Carl Linnaeus". Figure 2 illustrates the used Wikidata terminology with an example.

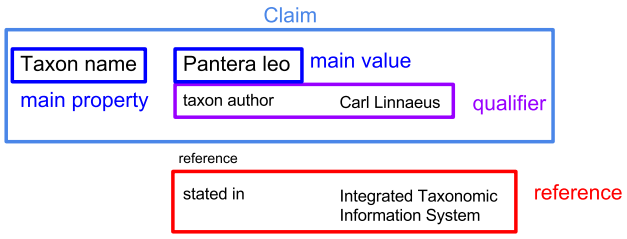


Figure 2: Wikidata statement terminology illustrated with an example.

The content of Wikidata is in the public domain under Creative Commons CC0. As of September 2015, Wikidata counted about 70 million statements on 18 million entities. A more in-depth comparison between Freebase, Wikidata, and other free and open knowledge bases (DBpedia, YAGO, and OpenCyc) is available in [12] by Färber *et al.*

2.3 Motivation for the Migration

When Google publicly launched Freebase back in 2007, Freebase was thought of as a "Wikipedia for structured data". The Knowledge Graph team at Google have been closely watching the Wikimedia Foundation's project Wikidata since its launch, and believe strongly in a robust community-driven effort to collect and curate structured knowledge about the world. The team now think they can serve that goal best by supporting Wikidata, as the project is growing fast, has an active community, and is better-suited to lead an open collaborative knowledge base [10].

In consequence, in mid-2015, the Freebase service as a standalone project was wound down. Freebase always has supported developer access to the data; in order to still cater for this need, a new API for entity search powered by Google's Knowledge Graph was launched.[4] Freebase has as of March 31, 2015, gone read-only, *i.e.*, the website no longer accepts edits and the MQL write API was retired.

3. CHALLENGES OF THE MIGRATION

During our ongoing efforts with the migration of Freebase to Wikidata, we were faced with a number of technical and non-technical challenges that we will outline in the following.

3.1 Licensing

The first challenge concerns the licenses under which the datasets are published. Wikidata is published under the Creative Commons 0 (CC0 1.0) [24] license, which effectively puts the data into the public domain. Freebase is published under a Creative Commons Attribution (CC BY 2.5) license. Google does not own the copyright of some parts of the

[4]Knowledge Graph Search API: `https://developers.google.com/knowledge-graph/?hl=en`

content of the knowledge base, such as images or long entity descriptions extracted from Wikipedia. We filtered the Freebase dump by removing this kind of content before creating a data dump that Google could relicense under CC0. This step reduced the number of Freebase facts that could be republished by about 42 million facts from 3 billion facts in the original corpus, *i.e.*, by about 1.4%.

3.2 References

The second challenge is that the Wikidata community is very eager to have references for their statements, *i.e.*, sources that Freebase usually did not store, except for some specific data like populations and unemployment rates. For these two specific cases, we map the Freebase properties describing sources,[5] expressed as CVTs, into Wikidata references on the appropriate claim. In order to provide the Wikidata community with references for the facts in Freebase, we have reused data from the Google Knowledge Vault [9]. Knowledge Vault aims at extracting facts from the Web.

The result of this step was surprising: the fact extraction was usually correct, in the sense that the fact was indeed being stated on the page. Also the fact itself was usually accurate, in the sense that it corresponded to reality. But the pages the facts were extracted from in many cases did not meet the Wikidata requirements for reliable references: they include pages from online social network sites, reviews on shopping sites, file sharing hosters, *etc.* It became necessary to filter the references before potential inclusion in Wikidata by introducing a domain blacklist.[6]

3.3 Data Quality

The data quality of Freebase was discussed by the Wikidata community, and it was decided that the quality overall was not sufficient for a direct import. For example, a small city in France was assigned an ISO country code,[7] Boston (the city in Massachusetts) had the type `/people/person`, and several other occurrences of data quality issues were revealed. In consequence, a fully automatic upload of all the content from Freebase into Wikidata did not seem advisable as the expectations of the Wikidata community regarding the quality of automatically uploaded data are high (and rightly so). Despite the fact that an automated upload of all the data would have led to accelerated growth in the short term, such a fast increase of available data would certainly have endangered the sense of ownership of the Wikidata community for its data. As an alternative approach, we decided to rely on crowdsourced human curation and created the *Primary Sources Tool*, a widget that displays Freebase statements for curation by the contributor that can be added to the currently shown Wikidata item. The tool will be described in more detail in Section 4.

3.4 Long-Term Maintenance

Wikidata's editing community is in charge of the project's content. This includes collecting the data as well as long-term maintenance of it. Data needs to be updated as the world changes, but also kept free of vandalism. An increase in the amount of data comes with an increase in maintenance cost for the community. Simply ingesting the Freebase data would have meant overwhelming the existing editors and thereby harming the project in the long run. To counter this, a significant increase in the amount of data needs to go along with either an increase in the number of editors or provision of tools to assist the existing editors to allow them to be more effective. With the *Primary Sources Tool*, we aim at helping grow the community in step with the amount of data. It is useful as a means to get more people to edit, as it lowers the barrier to entry by making it easier to find potential edits. At the same time, it helps make some of the existing editors more efficient by allowing them to add statements with a single click.

3.5 Data Topic Mappings

3.5.1 Merger of Existing Mappings

The next challenge concerns the data mappings between Freebase topics and properties, and Wikidata items and properties. Two mappings between Freebase topics and Wikidata items were initially available, and in addition to those, we worked on further mappings. A first one has been created by Google in October 2013[8] and is based on Wikipedia links already present in Freebase: if a Freebase topic and a Wikidata item share at least two Wikipedia links, they are assumed to describe the same subject. The quality of this mapping is considered very good by the Wikidata community, who thus have decided to use it and further curate and maintain it by hand. It maps 1.15 million Freebase topics to their corresponding Wikidata items.

A second actively maintained mapping has been created by Samsung.[9] It is based on the same idea, but matches a Freebase topic with a Wikidata item even if there is only a single shared Wikipedia link. The precision of the mapping is naturally lower compared to Google's mapping, as there are more wrong matches (often because there is no clear separation between topics and disambiguation pages in Wikipedia), however, the recall increases to 4.4 million links.

As we have shown in Subsection 3.3, human curation is required before import anyway, which is why we have decided to merge the two mapping dumps. For the 6,000 conflicts that appeared, we decided to prioritize Google's mapping.

3.5.2 Reconciliation Based on External Identifiers

To improve the result, we have also done some reconciliation based on third-party database IDs shared by Freebase and Wikidata, like MusicBrainz [30], VIAF [2], *etc.* With this technique, we were able to match 800,000 external database IDs, creating an additional mapping for 600,000 topics—most of them already in the merged version of Google's and Samsung's mappings. Eventually, this reconciliation effort resulted in an additional 84,000 mapped items.

We also used Google's Knowledge Graph [29] to add about 100,000 further mappings by also matching a topic and an item if they share a Wikipedia link with the Knowledge Graph item. There is some potential to add more data by looking at functional relations like `father` and `mother`, but

[5]The two main properties used in Freebase for sources are `/measurement_unit/dated_integer/source` and `/measurement_unit/dated_percentage/source`

[6]Domain blacklist: `https://www.wikidata.org/wiki/Wikidata:Primary_sources_tool/URL_blacklist`

[7]Freebase topic page of the city of Saint-Martin: `http://www.freebase.com/m/03cc0d0`

[8]Freebase mappings: `http://developers.google.com/freebase/data`

[9]Samsung mappings: `http://github.com/Samsung/KnowledgeSharingPlatform`

these relations only apply to relatively small sets of data (like well known families), and the functional assumptions may run into some edge cases. For example, the `/people/person/parents` property in Freebase also covers stepparents, which may create cases where a person has more than one male or female `/people/person/parents`.

Eventually, with these four sources (Google mapping, Samsung mapping, external IDs, Knowledge Graph), we mapped 4.56 million items in total.

3.5.3 *Topic Mapping Completeness Analysis*

Thanks to these efforts, we have already mapped most of the topics that could be mapped automatically, which is also backed by the comparably small increases gained from the ID reconciliation approach and the Knowledge Graph data described in the previous paragraph. This is no surprise, since the core of both projects was kickstarted by using Wikipedia as a repository of topics.

The *mapped* topics have an average number of 13.9 facts, whereas the topics that were *not* mapped have an average number of only 5.7 facts. We have mapped the majority of all topics with more than 47 facts (191K non mapped topics *vs.* 195K mapped topics). This indicates that we have mapped the most important items of Freebase. Figure 3 provides an overview of mapped and non-mapped items put in relation to the number of facts per item: note that the Figure is log-scaled on both axes, so the seemingly small area between the two lines covers a majority of the topics. It can also be seen that there are a very small number of topics with a high number of facts which have not been covered (in the lower part of the Figure). These have all been manually inspected and found to be specific to the way knowledge in Freebase is modeled. The vast majority of items with more than ten facts—in the center of the Figure—has been mapped. The items that have not been mapped—visible in the upper left corner of the image—are very thin, *i.e.*, mostly have a very small number of facts and would not have added much to the knowledge base.

3.6 Data Property Mapping

For the mapping between Freebase and Wikidata properties, we have chosen to apply a manual approach with the help of the Wikidata community. Note that a mapping of types is not needed, since Wikidata has no real notion of types (or, to be more specific, the property `instance of` (P31) is just another property and connects to items without any special status). With the help of the community, we have been able to quickly map around 360 properties (note that Wikidata had a total of about 1,800 properties in the summer of 2015 when this work was conducted, and Freebase had around 37,700 properties). For the mapping of properties where the domain are other topics or literals, this mapping provides the Wikidata property to use.

For `/people/person/parents`, we implemented a special case in order to map them to the Wikidata properties `father` (P22) and `mother` (P25) based on the person's gender.

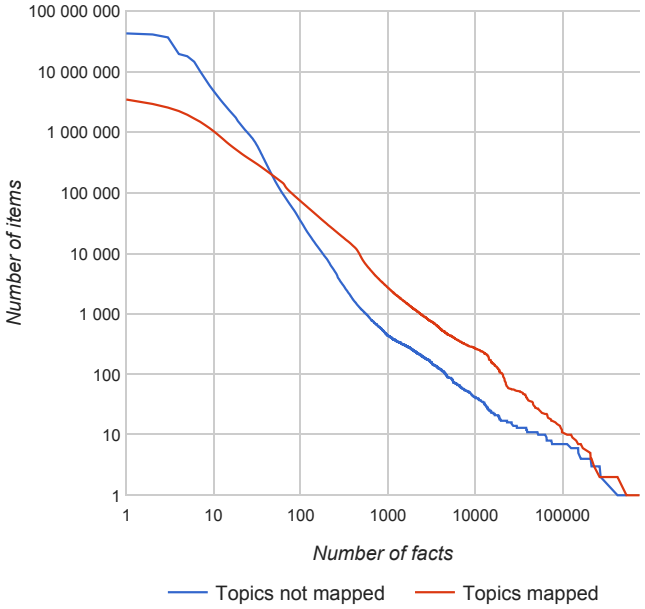


Figure 3: Number of mapped and not mapped topics with more than x facts (log scale on both axes).

The mapping of CVTs (see Figure 1) is more complicated, because it is not possible to have a 1-to-1 relationship between Freebase and Wikidata properties: the CVT is linked to the subject topic by one property and has properties pointing to its component values, whereas the Wikidata statement has a main property value group that is qualified by other such groups. In order to map a CVT to a statement, we have to know which of the CVT's properties should be used as the main value of the statement, with the others being mapped to qualifiers. In order to keep a simple mapping that associates one Wikidata property to one Freebase property, we map both the property that links the topic to the CVT and the main property of the CVT to the same Wikidata property. In a second step, we then merge these two mappings int a single Wikidata claim, on which we add the other properties from the CVT.

For example, if we map the CVT shown in Figure 1, we first use the Wikidata property `position held` (P39) to map `/politician/government_positions_held` and `/government_position_held/office_position_or_title`. Second, for the qualifiers, we use `start time` (P580) for `/government_position_held/from` and `end time` (P582) for `/government_position_held/to`. With the then resulting property mapping, the created Wikidata statement looks like the one presented in Figure 4.

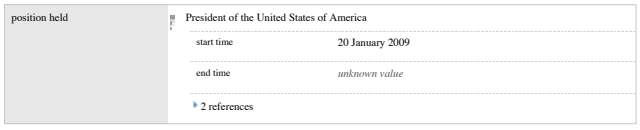


Figure 4: Wikidata statement for a Freebase Compound Value Type (CVT). See Figure 1 for comparison.

For CVTs that include sources, we map the property that links the CVT to the source as a Wikidata reference instead of a qualifier. The hardest part then is to actually

create the full topic mapping by considering CVTs and the complex datatypes of Wikidata. For CVTs, when we have a triple whose value is a CVT, we just retrieve it, map all its component triples (while ignoring ones that cannot be mapped), and apply the above mechanism to get its main value. For values, we use the datatype information stored with Wikidata properties in order to cast values to the right type. For quantities, we assume that the amounts are precise, as there is no precision information in Freebase. For globe coordinates, we have hardcoded the mapping of the `/location/geocode` CVT into the globe-coordinate value type of Wikidata and use the same precision guessing algorithm as the one used by the Wikidata user interface. For dates, we just map the date and time values (encoded using XSD types [4]) to the Wikidata `time` datatype. However, we remove dates before 1920 that have a precision higher than year because we do not know if they are relative to the Julian or the Gregorian calendar. In such cases, we are unable to fill the `calendar` value of the Wikidata `time` datatype.

4. PRIMARY SOURCES TOOL

As outlined earlier, the *Primary Sources Tool* is a crowdsourced human curation software solution that displays Freebase statements for verification to the Wikidata contributor so that these statements can be added to the currently shown Wikidata item. With just one click, the contributor can reject or approve a statement, and, in case of approval, add the statement to Wikidata. The code of the *Primary Sources Tool* is openly available (`https://github.com/google/primarysources`) under the terms of the Apache 2.0 license. The tool is deployed as a gadget, so that it can be easily enabled as an add-on feature in Wikidata. It is independent from Freebase and can be—and already has been—reused to import other datasets into Wikidata.

We have implemented an upload API in the back-end and a dataset filter in the front-end that allows users to display only statements from a selected dataset. This flexibility is indeed much-wanted, and during the development period of the project another researcher approached us in order to upload a dataset extracted from natural language text. We expect more datasets to be added to the *Primary Sources Tool* in the future.

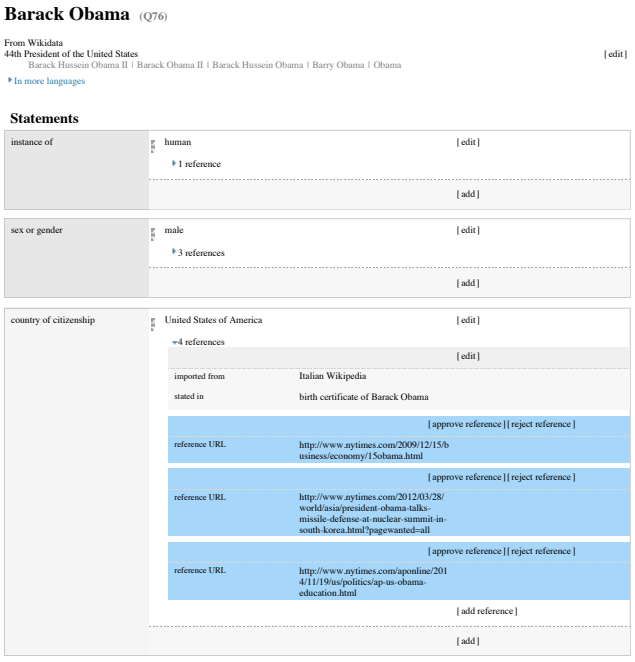
At the time of writing (January, 2016), the tool has been used by more than a hundred users who performed about 90,000 approval or rejection actions. More than 14 million statements have been uploaded in total. Figure 5 shows two screenshots of the tool and illustrates how it integrates nicely with the Wikidata user interface. To visualize the migration progress, we have created a realtime dashboard.[10]

Figure 6 shows the overall architecture of the *Primary Sources Tool*. In the following, we describe the back-end, the front-end, and the chosen data rollout plan.

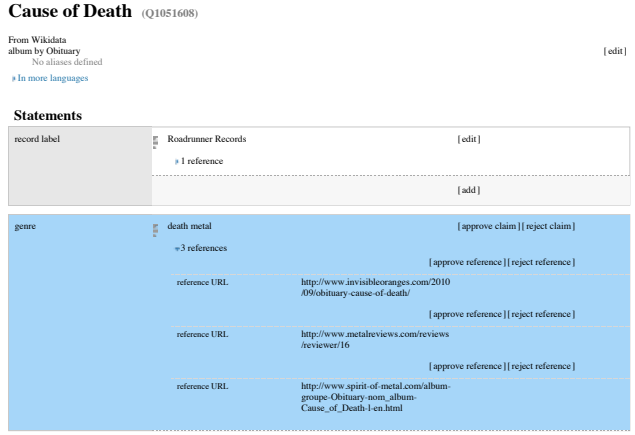
4.1 Back-End

The objective of the back-end was to have a REST API allowing us to serve data efficiently to the front-end of the tool and providing statistical data about the approvals and rejections done over time. As the resources of the Wikimedia Foundation are limited, the back-end was implemented in highly optimized C++ and runs as a FastCGI application in the lighttpd Web server.

Entities and statements are stored in a relational database (first SQLite, later migrated to MariaDB) and follow the Wikidata data structure with an additional status field to keep track of statements that either already have been approved, rejected, or not yet visited. Statements can be grouped in "datasets" (set of statements from the same origin) and "uploads" (set of statements uploaded at the same time) in order to distinguish data from different sources.[11]



(a) Incoming references.



(b) Incoming claims and references.

Figure 5: *Primary Sources Tool* user interface. Statements coming from Freebase that are not yet available in Wikidata are shown in light blue. Users can approve or reject these statements using the "approve" or "reject" buttons.

[10]Primary Sources Dashboard: `https://tools.wmflabs.org/wikidata-primary-sources/status.html`

[11]At the time of writing, the back-end contained five different datasets.

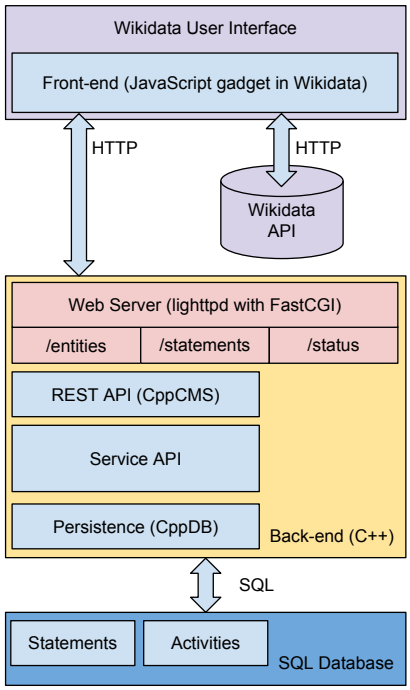


Figure 6: *Primary Sources Tool* architecture. The front-end (JavaScript) sends HTTP requests to the back-end (C++) to retrieve and update entities or statements, stored in a relational database. Approved statements are in addition written by the front-end to the Wikidata REST API.

The REST API of the back-end supports the following main types of requests:

- Get an entity (*i.e.*, its statements) by Wikidata QID (*e.g.*, `GET /entities/Q2529789`).
- Get all entities with unapproved (or approved or rejected) statements (`/entities/all`).
- Get a random entity with unapproved (or approved or rejected) statements (`/entities/any`).
- Get a statement by database ID (`GET /statements/1`), get all statements (`GET /statements/all`), or get a list of random statements (`GET /statements/any`).
- Set a new state for a statement (`POST /statements/1?state=approved&user=Alice`).

All GET requests support an additional `&state={state}` query parameter to filter by state. This allows Wikidata developers to examine statements that have already been updated (*e.g.*, "return 10 random rejected statements") for further analysis or reinvestigattion. The default value of this parameter is `unapproved`.

In addition to these main request types, the REST API also provides endpoints for importing and deleting datasets, as well as for retrieving the system's status.[12]

All update activity around statements is logged in an activity log for statistical purposes using the Wikidata user name of the user who performed the update. Activity information can be displayed in an aggregated leader board form (top users, top rejecters, trends, *etc.*) to add aspects of gamification to the migration process.

4.2 Front-End

For the front-end, our objective was to model the migration process as close to the natural Wikidata editing flow as possible. Ideally, the user should barely notice the different underlying datasources when navigating Wikidata.

To achieve this, we have in a first step created a Wikidata *user script*. Wikidata user scripts are part of the Wikidata tool chain[13] and are created by users, but unlike gadgets—to be explained in the following—do not appear in a user's preferences. Instead, users have to add a line of code into their `common.js` file in order to activate and use the script. User scripts can be created by anyone and—unlike gadgets—might not always be stable.

Once a user script has matured, it can be converted into a *gadget*. Gadgets are scripts that are likewise created by users, but which can be simply enabled in user preferences under the section "Gadgets". They can only be edited by administrators and are assumed to be stable.

With our front-end, we went through this maturation steps and started with a user script called `freebase2wikidata.js`[14] that we later converted into a gadget.[15] Figure 5 illustrates the user interface in action. Existing Wikidata statements are shown as usual, Freebase facts that could potentially be imported into Wikidata are shown in light blue. Instead of an "edit" button they feature two new buttons, "approve reference" and "reject reference" (Figure 5a) or "approve claim" and "reject claim" (Figure 5b) respectively.

4.3 Data Rollout

In order to prepare the data for integration into the *Primary Sources Tool*, we have created a set of scripts which map the content of the last Freebase dump to Wikidata statements.[16] The *Primary Sources Tool* was developed in parallel with the mapping scripts. We first created a small dataset of around 1 million statements based on few select properties (*e.g.*, `/people/person/birth_place`) and deployed both the first dataset and a basic version of the tool in order to gather initial user feedback.

In the following months, we progressively added more statements to the tool's database back-end. Thereby, we were able to slowly upscale the tool's database back-end without risking back-end performance bottlenecks, and so far, scale or throughput have not been a problem. The process of adding data in small batches of 1 million to 2 million statements per batch further allowed us to detect some potential issues in the front-end and fix them before they became actual problems. Another positive side-effect of this approach was that it allowed us to gather quick feedback from the community about the already mapped data and

[12]The system's status is also presented in more convenient form at `http://tools.wmflabs.org/wikidata-primary-sources/status.html`.

[13]Wikidata tool chain: `https://www.wikidata.org/wiki/Wikidata:Tools`

[14]Initial user script: `https://www.wikidata.org/wiki/User:Tomayac/freebase2wikidata.js`

[15]Final gadget: `https://www.wikidata.org/wiki/MediaWiki:Gadget-PrimarySources.js`

[16]The source code of theses scripts is also available under the terms of the Apache 2.0 license: `https://github.com/google/freebase-wikidata-converter`

thus to adapt the mapping scripts early and correct minor issues with the data in the back-end.

Upon approval or rejection of either a claim or a reference, the Wikidata item page in question reloads in order to reflect the potentially changed state of the migration, as, in case of incoming claims with references, freshly accepted claims can require incoming references to be attached elsewhere in the item, which is hard to keep track of without a refresh.

4.4 Usage with Other Datasets

As outlined earlier, the *Primary Sources Tool* from the start was designed to be used with other datasets apart from the Freebase dataset. A concrete first example is the "FBK StrepHit Soccer" dataset,[17] which, among other datasets, can be activated in the tool by clicking on its gears icon. The dataset selection dialog is depicted in Figure 7, as well as facts coming from this dataset. It is also possible to activate all datasets at once in parallel.

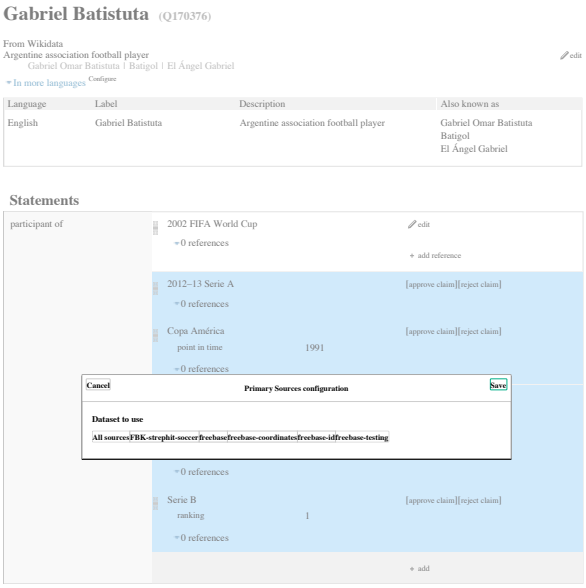


Figure 7: *Primary Sources Tool* used with the "FBK StrepHit Soccer" dataset, the screenshot shows the dataset selection dialog and incoming facts from this dataset.

5. STATISTICS ON THE MIGRATION

5.1 Quantitative Comparison

In order to put the statistics in this section in context, we first provide an overview of the size of the last dump of Freebase from March 2015:

- 48 million topics
- 2,997 million triples
- 442 million "useful" facts
- 68 million labels

"Useful" facts are the triples that remain after filtering out all triples about labels (`/type/object/name`, identifiers (`/type/object/id`), types (`/type/object/type`), descriptions (`/common/topic/description`), *etc.*), and triples removed for licensing reasons (the latter were about 42 million, see Subsection 3.1). The triples removed in this step are not considered to be potentially useful for Wikidata, since IDs, labels, and descriptions are not handled through statements.

As a direct comparison, in the following we provide the raw statistics for Wikidata as of August 2015.

- 14.5 million items
- 66 million statements
- 82 million labels

Why are these numbers so different? Regarding the number of topics, Freebase has a lot of topics about subjects that do not match Wikidata's notability criteria.[18] For example, Freebase holds data about 11.3 million musical recordings, 8.3 million music canonical versions, 2.6 million ISBNs, *etc.* that are not contained in Wikidata.

We also cannot compare the number of facts and the number of statements directly, even considering the lower number of topics. Freebase encodes its knowledge far more redundantly than Wikidata. To illustrate this, properties in Freebase often have a *reverse* property that is used in order to be able to traverse the Freebase graph easily in both directions. For example, the property `/people/person/place_of_birth` has the corresponding reverse property `/location/location/people_born_here` that encodes exactly the same semantic information. Such reverse properties sometimes exist in Wikidata (like `children` for `father` and `mother`), but are far less common and also not automatically handled.

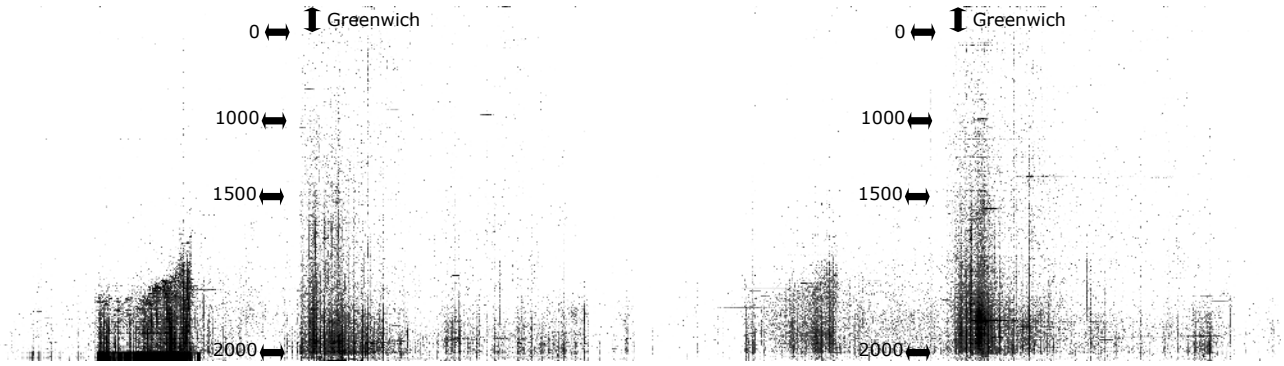
Another point is that CVTs use a lot of facts to convey the exact same information that can be represented with a single Wikidata statement. For example, to encode that Barack Obama is the president of the United States since January 20, 2009, Freebase requires more than six facts as shown in Figure 1. If we attempted to encode Wikidata statements as if they were Freebase facts, *i.e.*, by removing sources, representing statements with qualifiers using CVTs, and adding reverse properties, this would lead to a number of 110 million facts, *i.e.* an increase of 167% over the raw number of statements. The representation of Wikidata in RDF and its impact on the number of triples is discussed in more detail in [17]. In a considerable amount of cases, Freebase also contains duplicate data. For example, many cities in the United States have duplicate CVTs for the population, where the only difference is the encoding of the source.

5.2 Spatio-Temporal Comparison

In order to get an idea of the coverage of the two knowledge bases, we have further created a spatio-temporal visualization of Freebase and Wikidata that is shown in Figure 8. We extracted the earliest date and the longitude linked to each of the entities of the two knowledge bases. We then propagated them to the connected entities that do not have such data. To implement this propagation, we set as value (date or longitude) the average of the values of the entities

[17]"FBK StrepHit Soccer" dataset: `https://lists.wikimedia.org/pipermail/wikidata/2015-September/007014.html`

[18]Wikidata notability criteria: `https://www.wikidata.org/wiki/Wikidata:Notability`



(a) Spatio-temporal distribution of Freebase.

(b) Spatio-temporal distribution of Wikidata.

Figure 8: Spatio-temporal distribution of Freebase and Wikidata items, the x axis shows longitude degrees from 180° E to 180° W, the y axis shows time from the year 0 to 2015, following a power law ($y' = y^8$).

whose distance is < 10 with a weight of 2^{-d} (with d being the distance between the two entities).

We see that the two knowledge bases display the same basic patterns, with their coverage mostly concentrated on the longitudes of Europe and North America. There are no strong differences between them, yet coverage of *post*-year-2000 entities seems stronger in Freebase.

Our theory is that this is *(i)* due to the absence of a lot of musical data in Wikidata that is present in Freebase, and *(ii)* due to the US population CVTs in Freebase. The big horizontal lines in Figure 8a and even more in Figure 8b can most probably be traced back to places that do not have any linked date, and in consequence inherit the foundation date of their country. For Wikidata (Figure 8b), two lines that are definitely caused by this effect are the big line over the USA around 1780 and the one covering Russia around 850.

One possible conclusion from this visualization could be that Wikidata, despite its smaller absolute size, appears to have a nearly as complete coverage of the world as Freebase. Nonetheless, it is very challenging to find objective ways to compare the two databases and good metrics of success for the migration. We certainly can note that for data sets to be added to a knowledge bases, the raw number of triples is *not* an adequate measure of quality.

5.3 Raw Statistics

From the Freebase dump, we have been able to create more than 17 million Wikidata claims (*i.e.*, statements without references), including 1.5 million IDs. For that, we have used 31 million Freebase facts. Including the references, we obtain 19.6 million statements, and, after removing duplicates and facts already contained in Wikidata, we obtain 14 million new statements. If all these statements were added to Wikidata, we would see a 21% increase of the number of statements in Wikidata.

This raises the question why these numbers are so low compared to the size of Freebase of 3 billion facts. The main reason is that we have only mapped 4.56 million items to Wikidata, *i.e.*, only 9.5% of Freebase's topics in total. The mapped topics are the subject of only 64 million facts. So even under otherwise ideal circumstances, we cannot map more than 64 million statements. This assumes that we could map all reverse properties, which, in fact, we could not. So leaving aside the not mapped topics, we have created a statement for more than 24% of the facts.

If we restrict ourselves to reviewed facts—a set of 1.6 million human curated facts—we have far better results. Of the human curated facts, 0.58 million facts have their subject mapped to Wikidata. Based on our property mappings, 0.52 million of these facts (*i.e.*, 92%) are converted to Wikidata statements. Finally, 58% of the statements created from reviewed fact are already in Wikidata, allowing us to add 0.25 million new reviewed statements to Wikidata.

6. RELATED WORK

For the challenges listed in Section 3, there are a number of works dealing with it. In many cases, the problems we have encountered in this migration have turned out to be much simpler than the problems solved in existing research.

Data licensing. Although many books and tutorials on publishing data on the Web of Data (e.g. [16, 19]) mention the importance of licenses, they are often added as an afterthought. There have been remarkably few dedicated publications on the topic [21, 26], none of them written by lawyers. Due to the lack of understanding surrounding these topics, we have been cautious and chose to simply remove any facts from the to-be-migrated data if there were doubts.

Schema mappings. There has been extensive research on the topic of mapping the structures of two databases or knowledge bases (some surveys are [5, 8, 25]). The current work is, in many technical respects, rather simple: we are looking for a one-directional mapping (since the case of data flowing from Wikidata back to Freebase is not relevant), a one-off data flow (since Freebase is now read-only), mostly individually evaluated by human contributors (therefore not requiring a perfect mapping). The effort most similar to ours is the DBpedia Mapping Wiki [20]. The DBpedia community is manually writing and maintaining the mapping from the DBpedia extraction process (working on Wikipedia) to the DBpedia ontology. The main difference is that the DBpedia mapping is running continuously and does not allow for point-fixes, thus necessitating a higher quality of the mapping, whereas for the Freebase to Wikidata mapping every single result is expected to be curated. The Schema.org structured data project is also exploring mappings to and from Wikidata, to facilitate the use of Wikidata's vocabulary as Schema.org markup across the

wider Web. This involves similar issues mapping Schema.org's CVT-like statement qualification mechanism and roles to Wikidata conventions.[19]

Entity mappings. Entity mapping (often called entity linking) deals with finding the objects in several sources that refer to the same entity in the domain of discourse. There is extensive work in this area (e.g. [7, 28]) but we avoided most of the challenges by relying on the simple fact that both Freebase and Wikidata have historically been bootstrapped by using Wikipedia. Wikipedia is often used as a benchmark for entity mapping systems [14]. As described in Subsection 3.5 this leads to sufficiently good results, and we argue it would be surprising if further effort in this area would lead to reasonable gains.

Data maintenance. The Wikidata requirements [32] state that the data uploaded to Wikidata must not overwhelm the community. Research in understanding how systems for collaborative knowledge creation are impacted by events like this data migration are still in its early stages [23], in particular for structured knowledge [18]. Most of the research is focused on Wikipedia [11], which is understandable considering the availability of its data sets, in particular the whole edit history [27] and the availability of tools for working with Wikipedia [22].

7. FUTURE WORK AND CONCLUSIONS

The largest gains for the migration can be achieved by extending the mapping to more Freebase topics. A possible way of realizing this is to create a user interface—possibly leveraging elements of gamification—that would allow users to create new Wikidata items for a suggested topic or to add a mapping for them to an existing Wikidata item. In order to suggest interesting topics to add to Wikidata, we could rank the topics that are not mapped yet by the number of incoming links from already mapped topics and filter less interesting types like ISBNs. Another area for improvement is to upload high quality datasets using a bot, like the reviewed facts or some sets for external IDs, in order to speed up the integration of Freebase content into Wikidata. We have already started to upload simple reviewed facts about humans like birth date, death place or gender using a bot. We have started to import labels that are in Freebase but not in Wikidata. In the 17.2 million Freebase labels for mapped topics, only 0.9 million, *i.e.*, 5%, lack from Wikidata.

Concluding, in a fairly short amount of time, we have been able to provide the Wikidata community with more than 14 million new Wikidata statements using a customizable and generalizable approach, consisting of data preparation scripts and the *Primary Sources Tool*, which is well integrated into the Wikidata user interface. The effort needed to map two fairly different knowledge bases has also been a good occasion to highlight the difficulty of having adequate metrics to measure the size of knowledge bases in a meaningful way, and, in consequence, the "value" of such collaborative knowledge bases and the datasets added to them. We hope that with the help of the *Primary Sources Tool* and the Freebase dataset—and in future even more datasets—we will increase the completeness and accuracy of Wikidata.

[19]On Roles: `http://schema.org/Role`. Mappings from Schema.org to Wikidata: `https://github.com/schemaorg/schemaorg/issues/280`. On the interaction of Schema.org and Wikidata: `https://meta.wikimedia.org/wiki/Wikidata/Notes/Schema.org_and_Wikidata`

8. REFERENCES

[1] P. Ayers, C. Matthews, and B. Yates. *How Wikipedia Works: And How You Can Be a Part of It*. No Starch Press, Sept. 2008.

[2] R. Bennett, C. Hengel-Dittrich, E. T. O'Neill, and B. B. Tillett. VIAF (Virtual International Authority File): Linking die Deutsche Bibliothek and Library of Congress Name Authority Files. In *World Library and Information Congress: 72nd IFLA General Conference and Council*, 2006.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[4] C. E. Campbell, A. Eisenberg, and J. Melton. XML Schema. *SIGMOD Rec.*, 32(2):96–101, June 2003.

[5] N. Choi, I.-Y. Song, and H. Han. A Survey on Ontology Mapping. *ACM Sigmod Record*, 35(3):34–41, 2006.

[6] R. Cyganiak, D. Wood, and M. Lanthaler. *RDF 1.1 Concepts and Abstract Syntax*. World Wide Web Consortium, Feb. 2014. `https://www.w3.org/TR/rdf11-concepts/`.

[7] H.-J. Dai, C.-Y. Wu, R. Tsai, W. Hsu, et al. From Entity Recognition to Entity Linking: A Survey of Advanced Entity Linking Techniques. In *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 1–10, 2012.

[8] A. Doan and A. Y. Halevy. Semantic Integration Research in the Database Community: A Brief Survey. *AI Magazine*, 26(1):83, 2005.

[9] X. Dong, E. Gabrilovich, et al. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610. ACM, 2014.

[10] J. Douglas. Announcement: From Freebase to Wikidata, Dec 2014. `https://groups.google.com/d/msg/freebase-discuss/s_BPoL92edc/Y585r7_2E1YJ`.

[11] F. Flöck, D. Laniado, F. Stadthaus, and M. Acosta. Towards Better Visual Tools for Exploring Wikipedia Article Development—The Use Case of "Gamergate Controversy". In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[12] M. Färber, B. Ell, C. Menne, and A. Rettinger. A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web Journal*, July 2015. `http://www.semantic-web-journal.net/content/comparative-survey-dbpedia-freebase-opencyc-wikidata-and-yago` (submitted, pending major revision).

[13] A. Gesmundo and K. Hall. Projecting the Knowledge Graph to Syntactic Parsing. In G. Bouma and Y. Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 28–32. The Association for Computer Linguistics, 2014.

[14] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. Evaluating Entity Linking with Wikipedia. *Artificial intelligence*, 194:130–150, 2013.

[15] D. Harris. Google is Shutting Down its Freebase Knowledge Base. *GigaOM*, Dec. 2014. https://gigaom.com/2014/12/16/google-is-shutting-down-its-freebase-knowledge-base/.

[16] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan and Claypool, 2011.

[17] D. Hernández, A. Hogan, and M. Krötzsch. Reifying RDF: What Works Well With Wikidata? In T. Liebig and A. Fokoue, editors, *Proceedings of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems*, volume 1457 of *CEUR*, pages 32–47. CEUR-WS.org, 2015.

[18] M. Horridge, T. Tudorache, C. Nuylas, J. Vendetti, N. F. Noy, and M. A. Musen. WebProtege: A Collaborative Web Based Platform for Editing Biomedical Ontologies. *Bioinformatics*, pages 1–2, May 2014.

[19] B. Hyland, G. Atemezing, and B. Villazon-Terrazas. *Best Practices for Publishing Linked Data*. W3C Working Group Note. World Wide Web Consortium, Jan. 2014. http://www.w3.org/TR/ld-bp/.

[20] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al. DBpedia—A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 5:1–29, 2014.

[21] P. Miller, R. Styles, and T. Heath. Open Data Commons, a License for Open Data. In *Proceedings of the Linked Data on the Web workshop*, Beijing, China, Apr. 2008.

[22] D. Milne and I. H. Witten. An Open-Source Toolkit for Mining Wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

[23] J. Moskaliuk, J. Kimmerle, and U. Cress. Collaborative Knowledge Building with Wikis: The Impact of Redundancy and Polarity. *Computers & Education*, 58(4):1049–1057, 2012.

[24] D. Peters. *Expanding the Public Domain: Part Zero*. Creative Commons, Mar. 2009. http://creativecommons.org/weblog/entry/13304.

[25] R. Press. Ontology and Database Mapping: A Survey of Current Implementations and Future Directions. *Journal of Web Engineering*, 7(1):001–024, 2008.

[26] V. Rodríguez-Doncel, M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Poveda. License Linked Data Resources Pattern. In *Proceedings of the 4th International Workshop on Ontology Patterns*, Sydney, Australia, 2013.

[27] M. Schindler and D. Vrandečić. Introducing New Features to Wikipedia: Case Studies for Web Science. *IEEE Intelligent Systems*, (1):56–61, 2011.

[28] W. Shen, J. Wang, and J. Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, 2015.

[29] A. Singhal. Introducing the Knowledge Graph: Things, not Strings. Official Google Blog, May 2012. http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

[30] A. Swartz. MusicBrainz: A Semantic Web Service. *IEEE Intelligent Systems*, 17:76–77, Jan. 2002.

[31] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[32] D. Vrandečić. *Wikidata Requirements*. Wikimedia Foundation, Apr. 2012. https://meta.wikimedia.org/w/index.php?title=Wikidata/Notes/Requirements&oldid=3646045.

[33] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin. Knowledge Base Completion via Search-based Question Answering. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 515–526, New York, NY, USA, 2014. ACM.

[34] E. Zachte. *Statistics Wikidata*. Wikimedia Foundation, Sept. 2015. http://stats.wikimedia.org/wikispecial/EN/TablesWikipediaWIKIDATA.htm.