The Workshops of the Tenth International AAAI Conference on Web and Social Media
Wiki: Technical Report WS-16-17

State of the Union: A Data Consumer's Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities

Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz and Johanna Geiß
Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 205
69120 Heidelberg, Germany
{spitz, gertz, geiss}@informatik.uni-heidelberg.de
{dixit, ludwig.richter}@stud.uni-heidelberg.de

Abstract

Wikipedia is one of the most popular sources of free data on the Internet and subject to extensive use in numerous areas of research. Wikidata on the other hand, the knowledge base behind Wikipedia, is less popular as a source of data, despite having the "data" already in its name, and despite the fact that many applications in Natural Language Processing in general and Information Extraction in particular benefit immensely from the integration of knowledge bases. In part, this imbalance is owed to the younger age of Wikidata, which launched over a decade after Wikipedia. However, this is also owed to challenges posed by the still evolving properties of Wikidata that make its content more difficult to consume for third parties than is desirable. In this article, we analzye the causes of these challenges from the viewpoint of a data consumer and discuss possible avenues of research and advancement that both the scientific and the Wikidata community can collaborate on to turn the knowledge base into the invaluable asset that it is uniquely positioned to become.

Introduction

In the pursuit of information, journalists are taught to follow the so called *Five Ws*, which are five simple questions that serve as a structured approach to uncovering the developments of a newsworthy event after the fact. The name derives from the questions themselves, which are: *Who was involved?*, *When did it happen?*, *Where did it happen?*, *What happened?* and *Why did it happen?*. Naturally, answering the last question is impossible without first figuring out the answers to the previous four questions. The first three questions in particular serve to put the event into a recognizable context that is easy for the human mind to process. In Natural Language Processing, the task of Information Extraction (IE) from unstructured text is similar to journalism in this regard, as it also entails the uncovering of information after the fact, where the only evidence is a textual source from which structured information must be derived. It is therefore obvious, how the task of Named Entity Recognition (NER), which answers the first three of the Five Ws for a given text, is of central importance to the process of Information Extraction. Consequently, the extraction and classification of named entities is a well established field in which numerous approaches such as learning methods or linguistic analysis can be utilized (for an overview see (Nadeau and Sekine 2007)). All of these approaches, however, stand to benefit directly from the structured information in a knowledge base that enables the classification of entities and the linking of entity mentions in a text to entities in the knowledge base. Here, the combination of Wikipedia and Wikidata in particular provides a unique repository of text in which entity mentions are already linked to a knowledge base. Especially with respect to the open question of language-independent named entity recognition (Tjong Kim Sang and De Meulder 2003), the multilingual nature of Wikipedia and the links provided by Wikidata constitute a potentially invaluable resource.

In contrast to this potential, currently only a very limited number of research articles actually make use of Wikidata, and almost none in the field of Information Extraction. The Wikimedia research newsletter[1], which collects and summarizes scientific and scholarly research articles related to Wikimedia projects, only features a total of seven articles in 2015 that either use or analyze Wikidata in some way. Of these few articles, only three use Wikidata as a data source, such as the automatic generation of career profiles (Firas, Simon, and Nugues 2015) or the analysis of networks that are implicitly given by the co-occurrence of named entities in texts, where Wikidata is used for entity resolution (Geiß, Spitz, and Gertz 2015) and (Geiß et al. 2015). The focus of research into Wikidata appears to be on the process of inputting data, such as in the example of a pilot study for automatically updating drug information (Pfundner et al. 2015), or research into the usage behaviour of both automated and human users. Despite the young age of Wikidata, this activity is substantial and while over 88% of activity is automated, the participation of human users is significant (Steiner 2014). This activity, however, is input-related and most of the attention that has been given to Wikidata concerns the addition of new information to the knowledge base or the transfer of structured information from Wikipedia. Much of this is owed to the central goal of Wikidata, which is to serve as a source of information for and database behind Wikipedia. Therefore, both the focus on the development of user interfaces for Wikidata and much

[1] https://meta.wikimedia.org/wiki/Research:Newsletter

of the existing research on Wikidata is concerned with establishing it as a platform for cooperative building of structured knowledge. A recent study analyzed Wikidata in regard to the collaborative building effort and describes it from two perspectives of inputting data (Müller-Birn et al. 2015), namely the perspective of the peer-production community and the perspective of collaborative ontology engineering. One of the essential observations is that the creation of structured data is not equivalent to the creation of data structures themselves. The latter, however, is especially important from the perspective of possible end-users of the knowledge base, as design decisions in the input phase directly relate to the way in which it can later be used. This perspective of potential data consumers is therefore an equally important piece of the collaborative effort that is so far largely missing.

The availability of data and the possibility of using it in novel ways was one of the core conceptual ideas behind Wikidata (Krötzsch et al. 2007). A large part of the popularity of Wikipedia is due to the easy availability of data that it offers to its users. At least for the time being, however, this is not the case for Wikidata, as the targeted users are different. Wikipedia provides data for human consumption, while Wikidata provides data for more data-driven, structured and possibly automated applications. For the data in a knowledge base to be useful and reliable in such a setting, a clean and structured schema is required. While the correctness of a collaboratively managed knowledge base will never be perfect due to the risk of vandalism (Heindorf et al. 2015), the principles behind the data structure itself should be considered to be more important than the existence of a few wrong pieces of information in the knowledge base. If the structure of the data is not well documented and intuitive enough to allow researchers to use it without first researching the structure itself, then the initial energy that is required to use Wikidata is likely too high. This constitutes the major problem that we currently see with Wikidata, as the hierarchical relations between entities in the knowledge base are still evolving.

For the classification of entities to support the task of Information Extraction, the structure of a knowledge base has to reflect the simple categories that correspond to *who*, *where* and *when*. While the specific information that *Barack Obama is president of the United States* can be useful, it would not do us any good if we are unable to identify Obama as a person, the US as a location, or find the time frame of validity for this claim, because we would never arrive at a step where this information can be used. This aspect, which also reflects the way in which we as humans think about the involvement of named entities, can serve as a guideline to the creation of simpler structures from which the classes of entities are immediately clear to the user and easier to transfer to third-party applications. In the following, we present the current state of conceptual hierarchies in Wikidata along with the problems in its structure that we encountered during our extensive use of Wikidata for IE over the past year. We discuss possible avenues of improving the hierarchies of entity classes towards a more unified state in a knowledge base that is designed to contain the knowledge of all supported Wikipedias.

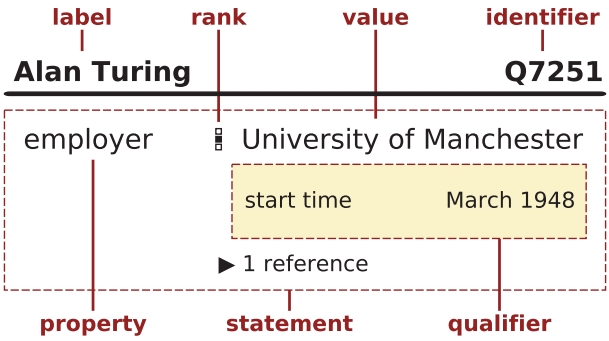

Figure 1: Example of an entry in Wikidata for the item that corresponds to Alan Turing.

Wikidata Data Model

Wikidata is an open, collaboratively edited knowledge base that is directly associated with Wikipedia and intended to serve as a central data base for information that is used in all language specific versions of Wikipedia. New data is entered into Wikidata either manually by users or through extraction from Wikipedia and other sources. In the following, we give a brief overview of the data model that Wikidata employs and discuss its relation to other well known knowledge bases. More detailed information can be found in the original paper (Vrandečić and Krötzsch 2014).

Wikidata Data Model

The data model of Wikidata consists of two major components, namely *items* and *properties*. In contrast to other semantic technologies, items represent both entities as well as classes and are denoted by IDs that start with the letter *Q*. In fact, each item is associated explicitly with its ID, as labels are not unique due to the multilingual nature of Wikidata and serve only as additional information (in this article, we show English labels for readability). Properties on the other hand connect items and have IDs that start with the letter *P*. Aside from having the same multilingual notion as items, properties correspond closely to RDF properties, i.e., they link items to other items or values. To create hierarchical structures of classes from items, class membership is predominantly modelled by the properties *instance of* (P31), in which the second item corresponds to what can be seen as a class analogue, while class items are connected through the property *subclass of* (P279).

In Figure 1, we show a typical data entry in Wikidata. In general, properties are assigned to items as so called *statements*, which are supported by *references* that confirm the claim. Wikidata uses statements instead of facts since it may contains several identical statements with different values due to its collaborative nature. As a result, conflicting statements can exist for an item and are ranked by the community to select the preferred option. Statements contain the property itself, the associated target value and a set of qualifiers that further specify the property. In the case of a city, for example, the property *population* (P1082) would be assigned a

value that corresponds to the population size, while a qualifier denotes the year of the census. This structure enables a flexible organization of statements but requires careful preparation of the retrieval of information from the knowledge base, due to possible overlapping statements.

Further Knowledge Bases

Aside from Wikidata, there are a couple of other knowledge bases that are also populated from Wikipedia information and are used more frequently as data sources in research. DBpedia is a prominent node in the Linked Open Data cloud (Bizer et al. 2009) and contains information from over 100 different language editions of Wikipedia. It is intended to take the human-readable content of Wikipedia and transform it into structured information in RDF format, where central entries correspond to Wikipedia pages. YAGO (Suchanek, Kasneci, and Weikum 2008) extracts structured information from Wikipedia and combines it with relations from WordNet (Miller et al. 1990), with classes derived from Wikipedia categories. We also list Freebase (Bollacker et al. 2008) for the sake of completeness, although this knowledge base is in the process of being shut down, with its contents being moved to Wikidata (Tanon et al. 2016). While these knowledge bases contain information that is extracted from Wikipedia, they are not collaboratively edited by the Wiki community. This difference provides both unique challenges as well as opportunities for Wikidata.

The State of Entity Classification in Wikidata

For a semantic analysis of natural language texts, one of the key steps is the extraction and disambiguation of entity mentions in the text, such as persons or locations. Ideally, this is followed by linking the discovered mentions to entities in a knowledge base to establish further connections between them, a step which is known as *resolution*. This extraction, disambiguation and resolution of entities relies on the properties on the entities in question. While some properties are important for all types of entities, others are entity specific. For example, person name resolution is likely to include a number of different properties in comparison to toponym resolution for place mentions, simply due to the inherently distinct naming conventions for persons and places. Thus, a limitation of the data in the knowledge base to the desired type is required in order to use knowledge bases as support for these IE processes. In the following, we discuss the challenges that we encountered in the classification of Wikidata entities into such commonly used classes. We collected these issues during our work on data from Wikidata in 2015 and used the Wikidata version of February 2016 to confirm their existence where necessary.

Persons

Issues that arise in regard to persons, their extraction, and everything that serves as an attribute for them are mostly related to their (non-) existence. A nice example of this is given in a study about the automatic extraction of career profiles (Firas, Simon, and Nugues 2015), which uses Wikidata for the extraction of professions and finds that "The search performed to find all the occupations collects anything remotely related to the *Occupation* node." This includes, among others, fictional professions of superheroes which are hard to distinguish from real occupations as a set of Wikidata entities. In Wikidata, the distinction between real and fictional beings is made through a separate item *fictional human* (Q15632617), which is then used with the property *instance of* (P31) to set the fictional status of persons that do not exist in the real world. The item *fictional human* is then a *fictional analog* (P1074) of *human* (Q5), which is a subclass of *person* (Q215627). As a result, there is a difference in the knowledge base between human and fictional human, but no such distinction exists for persons (the closest fictional analog would be *fictional character)*. Constructing a set of real persons for a disambiguation task is thus a time consuming matter of selecting numerous valid and invalid subclasses by hand with a high probability of erroneously including a fictional subclass by accident. It is also something that requires a good level of prior knowledge about the structure in Wikidata that is not necessarily given or easy to obtain. This goes well beyond persons, as similar classes exist for *fictional animal character* (Q3542731) and *fictional city* (Q1964689). It is especially critical for the topic of religion, where the fictionality of entities is not agreed upon (and of course the claim of fictionality may be considered offensive for active religions while it is generally not an issue for historic religions).

The need for a distinction between fictional and non-fictional entities is well reflected on the application side depending on the context. The current approach, however, when taken to its ultimate conclusion, would require having a fictional and a non-fictional category for literally everything. Furthermore, the distinction would have to be made in a way that is clear to the user and allows a separation into the two classes without forcing the user to manually separate them. As such, this approach seems unnecessarily complicated due to its many moving parts. Here, Wikidata stands to profit from the underlying schema that does not consider classes as their own type of entity. As a result, we see the addition of a property *is fictional*, which would be equally applicable to all fictional entities regardless of their primary class, as a simple solution to avoid needlessly complicated splits between classes in the hierarchy.

Organizations and Groups

Organizations and groups are useful concepts in the analysis of affiliations, i.e., the association of persons to their social and professional groups. To enable such analyses, a knowledge base has to support the extraction of groups of persons in a precise manner. In the case of Wikidata, groups of people can be found mostly as (direct or indirect) subclasses of *organization* (Q43229). Here, an extraction is complicated primarily by two issues, namely the overlap of organizations with other entity types and a rather convoluted scheme of subclasses that makes the extraction of groups of people difficult.

The most significant overlap occurs between organizations and locations due to the direct annotation of organizations. For an example consider Table 1 (top), which shows

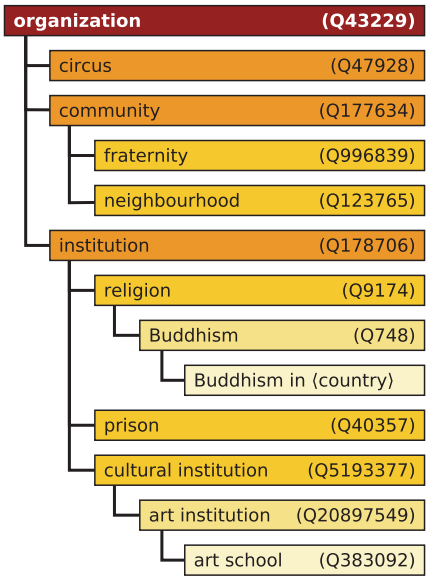

Figure 2: Selection of classes in the subtree of *organization*. Note the definition of *community* as "a group of interacting living organisms sharing a populated environment; a social unit of human organisms who share common values."

Wikidata items that have also been tagged with geolocations. It is evident that many of these cases describe organizations, for which the coordinates correspond to the location of the headquarters or head office. However, we also find examples of bars or restaurants that are tagged as organizations. While the overlap between a (small) company and the building it occupies is reasonable, this assumption is invalid for larger companies. In the case of Amazon.com, a company with hundreds of locations worldwide, there really is no reason to directly apply a single coordinate location directly to the company. Instead, a strict distinction between an organization and the locations that its offices occupy seems to be more reasonable. After all, an organization is a conceptual entity, rather than a physical one, as reflected in the Wikidata description which reads "social entity with a collective goal." Similarly, we observe an overlap between persons and organizations, which most notably happens when persons incorporate a company under their own name. One example of this is *Charles Brigham* (Q5075781), which is an instance of both *human* and *architectural firm*. Here, we argue that a company should always contain a person, even if it consists of only the owner or founder itself. On the other hand, a person should never be identified with a company or organization, but rather constitute one of its parts.

We show an example of the second and more pronounced issue in Figure 2, which contains a selection of items that can be found as subclasses of *organization*. Note that the number of subclasses is much larger than shown here (the entire tree contains over 7,500 entries) as the structure of subclasses is quite complex and rather intricate. Here, we find first-level subclasses of *organization* such as *circus*, which are directly usable groups of people that conform with an intuitive interpretation of social group. Other valid groups can be found at deeper levels in the hierarchy, such as *art schools*. Between them, however, we find classes that do not correspond to cooperative social groups. Even broad concepts of geographical entities such as *neighbourhood* can be found, which as a class contains local neighbourhoods within cities in which individuals share no connection besides spatial proximity. However, even though the definition of *community* is vague enough to allow it to contain groups of people based on spatial proximity or even groups of animals, it cannot be excluded entirely since it also contains social groups of people like *fraternities*. The listed *cultural institution* provide a good example of inconsistencies in the hierarchy. A circus is certainly more specialized than a cultural institution in general, yet it appears at a higher level in the hierarchy. Buddhism as an example of a religion is especially noteworthy, as it contains a distinct organization for each individual country, which is not reflected in existing organizations. This mixture and blending of useful and questionable classes of organizations of differing importance levels turns the extraction of groups into a daunting task. In practice, a researcher has no alternative to considering all entries in the class tree by hand and selecting appropriate subclasses.

item label	ID	instance of
Opera Software	Q215639	software house
Bank of Japan	Q333101	central bank
Fellini's Pizza	Q16993200	pizza chain
Amazon.com	Q3884	public company
UN Office at Geneva	Q680212	organization
Ich bin ein Berliner	Q443	speech
I Have a Dream	Q192341	speech
September 11 attacks	Q10806	terrorist attack
'05 Bali bombings	Q86584	suicide attack
'02 Sumatra earthquake	Q4600516	earthquake

Table 1: Selection of items with geocoordinates. Top: items in the subtree of *organization* (Q43229). Bottom: items in the subtree of *event* (Q1190554).

Locations

Based on the notion that a location is a point in space and thus possesses geo-coordinates, the extraction of locations as a class from Wikidata constitutes a task that should be simple. In practice, however, it turns out to be rather complex, even without considering that many locations have spatial extent and cannot cannot be represented by points. The primary reason for this is the lack of geo-coordinates on a number of locations on the one hand, and a very complicated hierarchy of places on the other. Especially the structure of the hierarchical relationship between locations is difficult to extract and requires the user to have previous knowledge. For example, the property *country* (P17) is described as "sovereign state of this item", which is sensible but leads to the problem of requiring a proper definition of *sovereign state*. Here, the problem arises due to the different connotations of *country*, *state* or similar descriptions in different contexts and languages. For example, a *state* in the United

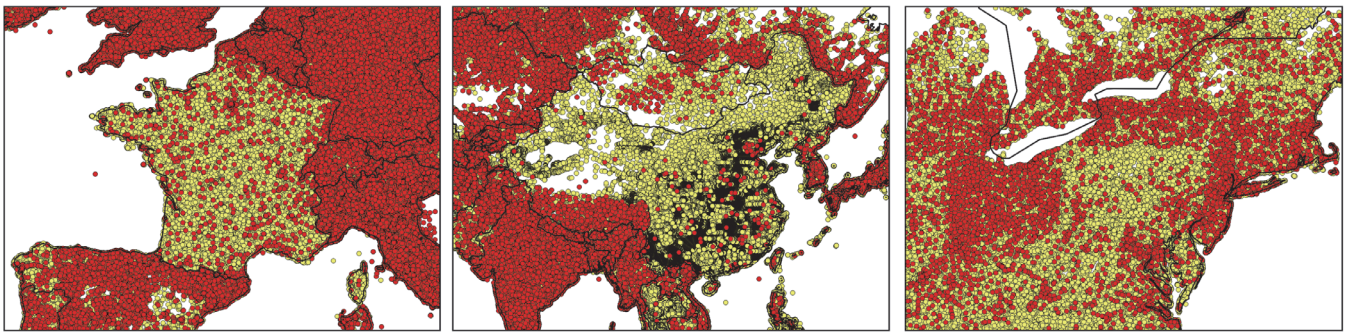

Figure 3: Vizualization of instances of *human settlements* in Wikidata (red). Shown for comparison are instances of *populated places [PPL]* from GeoNames (yellow). Examples of areas in which settlements are misclassified in Wikidata because they are only contained in the municipal hierarchy include France (left), China (center) and the Northeastern United States (right).

States or Germany usually refers to a part of the federal union as a whole. In other contexts, however, it may refer to a country. Therefore, a distinction has to be made between *sovereign state*, *constituent state* and *state*, which can be either. In Wikidata, all three exist such that *constituent state* (Q5164076) is a subclass of *state* (Q7275), which is *said to be the same as* (P460) a *country* (Q6256). However, a *state* is a subclass of *organization* (Q43229), while a *country* is not. While these relations are factually correct, they could be considered incomplete and rather complicated to parse into a hierarchy of places, especially if we consider that there are over 3,200 other items in the tree below the *administrative territorial entity* (Q56061) that have to be included. While we agree that these administrative relations between countries and their components are important and should be reflected, it is not the same as a hierarchy of places. Currently, the only way of extracting geographic relationsships is through complete analysis of a majority of worldwide administrative hierarchies.

An unfortunate result of the inclusion of these administrative roles in the Wikidata hierarchy is the problem that one faces in the extraction of something as simple as cities. In Figure 3, we show vizualizations of places that we extracted from Wikidata in relation to places that are contained in GeoNames[2]. It is evident that cities are missing in large regions of the World. The reason behind this is not a chunk of missing data in Wikidata, but that they are not instances of subclasses of *human settlement* (Q486972) and instead belong to subclasses of local municipal hierarchies such as *commune of France* (Q484170), *town of the United States* (Q15127012) or *town in China* (Q735428). As a result of our analysis, we find that it would be preferable to not include administrative specifics in the class hierarchy but add it as additional information to a statement. For example, a single property such as *administrative parent* could be used to describe any political hierarchy at any level of granularity. Such a relation would be entirely sufficient to reconstruct a complete municipal hierarchy as well as a geographic hierarchy. The type of relation can then be used as additional information. In essence, such a hierarchical skeleton would enable a much simpler reconstruction of any desired hierarchy and avoid many of the problems we encountered.

A second problem is the construction of classes through the discretization of location attributes that could be represented by scalar values instead. An example of this is the class *big city* (Q1549591) as subclass of *city* (Q515), which is labelled as a "city with a population of more than 100,000 inhabitants." Since this issue is related to the problem of discretizing time intervals, we discuss it in the next section.

Temporal Aspects

Time as a concept lends itself well to discretization, which is something that we as humans use frequently when we split it into arbitrarily small or large chunks, depending on the task at hand. For Wikidata, this poses the challenge of having to represent time in such a way that allows the reconstruction of points and intervals in time at arbitrary levels of granularity as required by an application. Therefore, the data model of Wikidata includes the option of assigning a *start time* and an *end time* as qualifiers to a statement, which is especially useful in determining the validity period that a statement covers. However, this is not the only place where temporal information can be found in Wikidata, which also includes instances of temporal information that has already been discretized.

One such example is the item *former entity* (Q15893266) and its subclasses, which reduce temporal information to a binary classification and include it in the hierarchical structure given by the properties. In these cases, such discretized information of *things in the past* poses a challenge for both curators and users of Wikidata, due to the sliding window that is caused by the perpetual movement of time. On the one hand, it necessitates updates and readjustments of entities in the database on a constant basis, since today's novel development is tomorrow's former event. In NLP on the other hand, such information is virtually unusable for tasks relating to corpora that cover a time interval in the past and are frequently used as a standard for the evaluation of new methods for information retrieval such as the well known New York Times annotated corpus (Sandhaus 2008). In additon to the entire corpus covering a timespan in the past from 1987 to 2007, each individual news article in the cor-

[2]http://www.geonames.org/

pus has its own publication date. In practice, this results in a sliding window for the relative *now*, which is completely different from the information that is given by former entities in Wikidata. From our perspective, it is unclear why such a discretization would even be desirable for a knowledge base. The information whether an entity is currently valid or already a former entity can always be reconstructed relative to any desired point in time as long as the start and end dates are provided. In cases of former entities for which the exact end time is not known, an *unspecified time in the past* as end time can serve as indication that the entity is no longer valid in the present, without enforcing this choice relative to other points in time.

With regard to spatial information that is encoded in items and properties, this blurs the distinction between space and time. While the concept of spacetime is valid from a physicist's perspective, it does not correspond to the intuitive human understanding of space and time as separate entities as it is reflected in language. Therefore, entities such as *former country* (Q3024240) or *former building or structure* (Q19860854) as subclasses of *former entity* add a level of complexity to the extraction of spatial hierarchies that is unnecessary due to the naturally ordering aspect of time. A user has to be aware of their existence to avoid them, independently of their realization. If such former entities are included in the hierarchy, they have to be excluded by the user if the data is to be used for tasks in the present. If they are excluded and form a separate class, they have to be included manually for tasks in the past. Thus, it is never a simple matter of not using a temporal property or item relation, as its existence always forces the user to consider its implications.

Events

From the perspective of NLP, events are frequently considered as "something that happens at a given place and time between a group of actors" (Cieri et al. 2002), which already highlights the temporal and spatial component as well as the potential involvement of persons. Due to this combination of different named entities, it is one of the most complicated and involved tasks in NLP, which stands to benefit from annotated data in Wikipedia and Wikidata as a connected knowledge base. In Wikidata, events as item are currently quite rare: we find about 4,000 meaningful instances of a subclass of *event* (Q1190554), the majority of which are instances of *natural disaster* (Q8065) and *concert* (Q182832). This list already suggests that events can be either instantaneous occurrences or have a duration. Events are labelled as "occurrences in space and time" in Wikidata and as such constitute a subclass of *point in time* (Q186408). This is quite contradictory, as it rules out events with a duration, such as earthquakes or concerts. However, earthquakes are a subclass of *natural disasters* which are a subclass of events, so strictly speaking they would be points in time.

With respect to location, a similar problem to what we already observed for organizations arises, namely the direct attribution of geo-coordinates to events. In Table 1 (bottom), we show a selection of different events that possess geo-coordinates in Wikidata. While the meaning of the coordinates can obviously be interpreted as the point in space where the event took place, this also makes it rather similar to a location. Instead of being treated in the same way as a location, an event should rather be connected to a location. Only such an abstraction then allows the proper modelling of events that are spread across multiple or larger locations, such as the Olympic Games, wars, or even earthquakes.

To simplify matters, we thus argue that a strict separation of events into components would resolve these problems. Especially due to the currently low number of events that is sure to rise sharply in the future, this would be advisable. Instead of directly tagging events as points in time or places, events should be associated with a time frame and a location where they took place. The concepts that would be required to do this already exist in Wikidata. Adding further data such as involved persons or organizations would be a trivial matter as well.

Discussion

Based on the issues presented in the previous section, we now discuss their implications as well as possible avenues of action and research that we consider to be of interest to both the Wikidata and the research community with the aim of improving the usability of Wikidata.

Towards a Skeleton Class Hierarchy

The probably most significant problem that we encountered in our use of Wikidata pertains to the complicated hierarchy, in which classes appear at somewhat arbitrary levels. This makes it difficult to obtain subsets of the data that correspond to basic types of named entities. While this issue may partially be rooted in the missing distinction between classes and items in the Wikidata scheme, all class hierarchies can, in principle, be modelled in the given scheme. However, the currently observable hierarchy is very complex. An immutable and well documented set of basic properties that a Wikidata user can rely on for simple classification tasks would therefore greatly increase the usability of the data. To include classes of entities from Wikidata in a research project, one currently has to invest a considerable amount of effort into the extraction of data from an already structured source, which is incredibly time consuming. Here, a well defined and fixed skeleton of basic classes would be beneficial. Given the current focus of Wikidata, we conjecture that the majority of existing structural relations currently arise from the necessity to support Wikipedia's primary task, i.e., the management of data for Wikipedia. This is reflected by the inclusion of Wikipedia categories in the hierarchy of Wikidata. While this is a natural evolution given the circumstances, there is an argument to be made for striving towards a separation of items at the top of the hierarchy that reflect the phenomena in natural language more closely, i.e., a separation into named entities, which directly corresponds to the understanding of entity classes that we employ in everyday life. Most importantly, we find that such a hierarchy is not mutually exclusive with the existing structure but rather an optional addition, so existing structures would be preserved. From the perspective of NLP, this would allow for an extraction and classification of entities at a scale and level of precision that is simply not possible at the current time.

In addition to a manual extension of the base hierarchy, the automatic generation of such skeleton hierarchies is possible and has previously been demonstrated for the automated refinement of infobox ontologies in Wikipedia (Wu and Weld 2008). Similar approaches have also been applied to bootstrapping simple ontologies from Wikipedia categories (Mirylenka, Passerini, and Serafini 2015). The adaptation of such methods to Wikidata would help to ensure a simple hierarchy that is easy to maintain.

Integration in the Semantic Web

As an alternative to the reduction of the Wikidata hierarchies to a more condensed level, one can also consider the mapping of the entire hierarchy to existing structures, such as those of existing knowledge bases. A primary candidate in this respect is the Semantic Web with its RDF standards, which is in part already under way. In 2014, an RDF export function was introduced to Wikidata (Erxleben et al. 2014), which allows a mapping of Wikidata properties to a different schema. This approach has the advantage of relying on well-established schemata and years of previous research, as is the case in the transfer of data from Wikidata to DBpedia (Ismayilov et al. 2015). However, the task is highly complex and the mapping of properties serves as an additional source of errors that have to be accounted for. While such a mapping directly enables the formulation of entity queries in known query languages (Hernández, Hogan, and Krötzsch 2015), the issue remains that a manual mapping of individual properties from Wikidata to the relations of the knowledge base is required. This is further complicated by the fact that Wikidata properties do not correspond directly to RDF properties (Erxleben et al. 2014) and an inversion of the mapping to extract sets of entities is therefore nontrivial in itself. Thus, we do not see the integration of Wikidata into the Semantic Web as a short-term solution, since it just moves the manual work that is necessary for the extraction of entities from one domain to another. Here, further research into standardized and dynamic mappings is required to ensure that they remain robust in the face of changes to the collaboratively edited Wikidata.

Frequency of Updates

One of Wikidata's unique features, which we encountered repeatedly over the course of our year-long use of Wikidata for entity extraction and resolution, is the perpetually changing content that is subject to constant updates. This in itself is problematic for any third party user of Wikidata, especially in light of the enormous work that data extraction from Wikidata currently requires. Here, we see the need for research into a set of legacy properties that stay constant while only the involved items are modified or deleted. Ideally, these should correspond to the minimal skeleton hierarchy. As an example, the proposed geographic *parent* relationship is likely a good candidate that can be enriched with further information but is unlikely to require changes itself.

Avoiding discrete reductions

In the case of temporal information as well as population numbers for cities, we encountered a number of properties and items that represent discrete reductions of the scalar values of statements. We found that they primarily dilute the hierarchy of entities and were unable to determine a practical function. Furthermore, all such discretized items that we found could be reproduced from the scalar attribute values if necessary. Due to the subjective or cultural interpretation behind such discretized relations, a reconstruction is less error prone. As a side effect, the required maintenance to keep such statements like *instance of big city* up to date for continually changing population numbers appears to be immense. We therefore suggest the removal of any such relations. Further extensions of Wikidata such as query support can easily serve to replace and surpass their current functionality.

Property Constraints

Constraints in knowledge bases can serve to limit the number of possible different relations or the set of attributes that can be assigned to an item. In the Wikidata community, constraints for properties are used for this purpose, but only on an informal basis by the users since they are not enforced in the underlying data model (Erxleben et al. 2014). Given some of the arguments made above, such constraints could be used to ensure that some properties are infused with additional information. Especially in the case of geo-coordinates in general and events in particular, the application of further constraints would serve to simplify the existing hierarchy. A similar argument can be made for the relation between persons and organizations. In this regard, tools that directly support constraints (e.g., as suggestions for users during data input) would make this approach more viable, even without integrating constraints in the model itself.

User Interfaces for Data Output

Finally, we note what we missed most in our use of Wikidata: a comprehensive tool for browsing both hierarchies as well as contained items within a hierarchy. We are aware of the existence of a selection of tools that extend the Wikidata Query Service[3], such as WMFlabs' tree for Wikidata[4], which greatly increase the user's capability of researching class hierarchies. Given the current complexity of the hierarchy, however, such tools fall short. Based on the primary goal of Wikidata and the current status of development, it is understandable that the tools that Wikidata provides to the user are geared towards entering data instead of retrieving it. As Wikidata grows, such tools are increasingly required. A good, integrated tool for browsing and selecting classes and subclasses of entities would go a long way towards making data selection and entity classification more viable. To our knowledge, there currently is no tool that allows the selection of entire classes of entities for extraction. Where the support of SPARQL queries is geared towards the extraction of specific pieces of knowledge from Wikidata, an interactive browsing of the hierarchies for the purpose of data selection would enable the extraction of entire sets of data for use in research projects or IE applications.

[3] https://query.wikidata.org/

[4] http://tools.wmflabs.org/wikidata-todo/tree.html

Conclusion

In this paper, we reported on our experiences with using Wikidata hierarchies to extract and classify sets of entities for the support of Information Extraction tasks. Specifically, we found that Wikidata as a collaboratively edited and perpetually growing knowledge base follows unique dynamics in its structural growth, many of which are geared towards the input of data, while data retrieval is still in its infancy. The current, complex hierarchy does not reflect the simple classification of entities that underlies the identification of named entities in Information Extraction tasks. Based on our findings of these structural problems, we discussed possible avenues of improvement and future research for increasing the effectiveness and ease with which Wikidata can be used in research projects. We see great potential in the inclusion of simple skeleton hierarchies in support of such classification tasks, which require further academic research into the principles behind the automatic generation and maintenance of such structures.

While we are aware that many of our suggested changes to conventions are of a kind that is typically established by the userbase of Wikidata themselves, we argue that the set of Wikidata users and third party data consumers is likely to become more disparate as Wikidata grows than is the case for Wikipedia. Thus, we hope that the outside perspective which we provide here serves as insight into possible paths of improvement that would otherwise not be evident and that it ultimately helps to improve a unique knowledge base that stands to inherit the union of knowledge of Wikipedia.

References

Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; and Hellmann, S. 2009. DBpedia - A Crystallization Point for the Web of Data. *Web Semantics* 7(3):154–165.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *SIGMOD International Conference on Management of Data*.

Cieri, C.; Strassel, S.; Graff, D.; Martey, N.; Rennert, K.; and Liberman, M. 2002. Corpora for Topic Detection and Tracking. In *Topic Detection and Tracking*. Springer.

Erxleben, F.; Günther, M.; Krötzsch, M.; Mendez, J.; and Vrandečić, D. 2014. Introducing Wikidata to the Linked Data Web. In *The Semantic Web–ISWC 2014*. 50–65.

Firas, D.; Simon, L.; and Nugues, P. 2015. Extraction of Career Profiles from Wikipedia. In *First Conference on Biographical Data in a Digital World*.

Geiß, J.; Spitz, A.; Strötgen, J.; and Gertz, M. 2015. The Wikipedia Location Network - Overcoming Borders and Oceans. In *Proceedings of the 9th Workshop on Geographic Information Retrieval (GIR' 15)*.

Geiß, J.; Spitz, A.; and Gertz, M. 2015. Beyond Friendships and Followers: The Wikipedia Social Network. In *Advances in Social Networks Analysis and Mining (ASONAM)*.

Heindorf, S.; Potthast, M.; Stein, B.; and Engels, G. 2015. Towards Vandalism Detection in Knowledge Bases: Corpus Construction and Analysis. In *SIGIR Conference on Research and Development in Information Retrieval*.

Hernández, D.; Hogan, A.; and Krötzsch, M. 2015. Reifying RDF: What Works Well With Wikidata? In *International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS)*.

Ismayilov, A.; Kontokostas, D.; Auer, S.; Lehmann, J.; and Hellmann, S. 2015. Wikidata through the Eyes of DBpedia. *arXiv preprint arXiv:1507.04180*.

Krötzsch, M.; Vrandečić, D.; Völkel, M.; Haller, H.; and Studer, R. 2007. Semantic Wikipedia. *Web Semantics* 5(4):251–261.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. J. 1990. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography* 3(4):235–244.

Mirylenka, D.; Passerini, A.; and Serafini, L. 2015. Bootstrapping Domain Ontologies from Wikipedia: A Uniform Approach. In *International Conference on Artificial Intelligence (AAAI)*.

Müller-Birn, C.; Karran, B.; Lehmann, J.; and Luczak-Rösch, M. 2015. Peer-production System or Collaborative Ontology Engineering Effort: What is Wikidata? In *International Symposium on Open Collaboration (OpenSym)*.

Nadeau, D., and Sekine, S. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes* 30(1):3–26.

Pfundner, A.; Schönberg, T.; Horn, J.; Boyce, R. D.; and Samwald, M. 2015. Utilizing the Wikidata System to Improve the Quality of Medical Content in Wikipedia in Diverse Languages: A Pilot Study. *Journal of medical Internet research* 17(5).

Sandhaus, E. 2008. The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia*.

Steiner, T. 2014. Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A global Study of Edit Activity on Wikipedia and Wikidata. In *International Symposium on Open Collaboration (OpenSym)*.

Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2008. YAGO: A Large Ontology from Wikipedia and Wordnet. *Web Semantics* 6(3):203–217.

Tanon, T. P.; Vrandecic, D.; Schaffert, S.; Steiner, T.; and Pintscher, L. 2016. From Freebase to Wikidata: The Great Migration. In *International Conference on World Wide Web*.

Tjong Kim Sang, E. F., and De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Conference on Natural Language Learning at HLT-NAACL*.

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

Wu, F., and Weld, D. S. 2008. Automatically Refining the Wikipedia Infobox Ontology. In *International Conference on World Wide Web*.