Communications

Wikidata: From “an” Identifier to “the” Identifier

Theo van Veen

ABSTRACT

Library catalogues may be connected to the linked data cloud through various types of thesauri. For name authority thesauri in particular I would like to suggest a fundamental break with the current distributed linked data paradigm: to make a transition from a multitude of different identifiers to using a single, universal identifier for all relevant named entities, in the form of the Wikidata identifier. Wikidata (https://wikidata.org) seems to be evolving into a major authority hub that is lowering barriers to access the web of data for everyone. Using the Wikidata identifier of notable entities as a common identifier for connecting resources has significant benefits compared to traversing the ever-growing linked data cloud. When the use of Wikidata reaches a critical mass, for some institutions, Wikidata could even serve as an authority control mechanism.

INTRODUCTION

Library catalogs, at national as well as institutional levels, make use of thesauri for authority control of named entities, such as persons, locations, and events. Authority records in thesauri contain information to distinguish between entities with the same name, combine pseudonyms and name variants for a single entity, and offer additional contextual information. Links to a thesaurus from within a catalog often take the form of an authority control number, and serve as identifiers for an entity within the scope of the catalog. Authority records in a catalog can be part of the linked data cloud when including links to thesauri such as VIAF (https://viaf.org/), ISNI (http://www.isni.org/), or ORCID (https://orcid.org/). However, using different identifier systems can lead to having many identifiers for a single entity. A single identifier system, not restricted to the library world and bibliographic metadata, could facilitate globally unique identifiers for each authority and therefore improve discovery of resources within a catalog.

The need for reconciliation of identifiers has been pointed out before.[1] What is now being suggested is to use the Wikidata identifier as “the” identifier. Wikidata is not domain specific, has a large user community, and offers appropriate APIs for linking to its data. It provides access to a wealth of entity properties, it links to more than 2,000 other knowledge bases, it is used by Google, and the number of organisations that link to Wikidata is quantifiably growing with tremendous speed.[2] The idea of using Wikidata as an authority linking hub was recently proposed by Joachim Neubert.[3] But why not go one step further and bring the Wikidata identifier to the surface directly as “the” resource identifier, or official authority record? This has been argued before and the implications of this argument will be considered in more detail in the remainder of this article. [4]

Theo van Veen (theovanveen@gmail.com) is Researcher (retired), Koninklijke Bibliotheek.

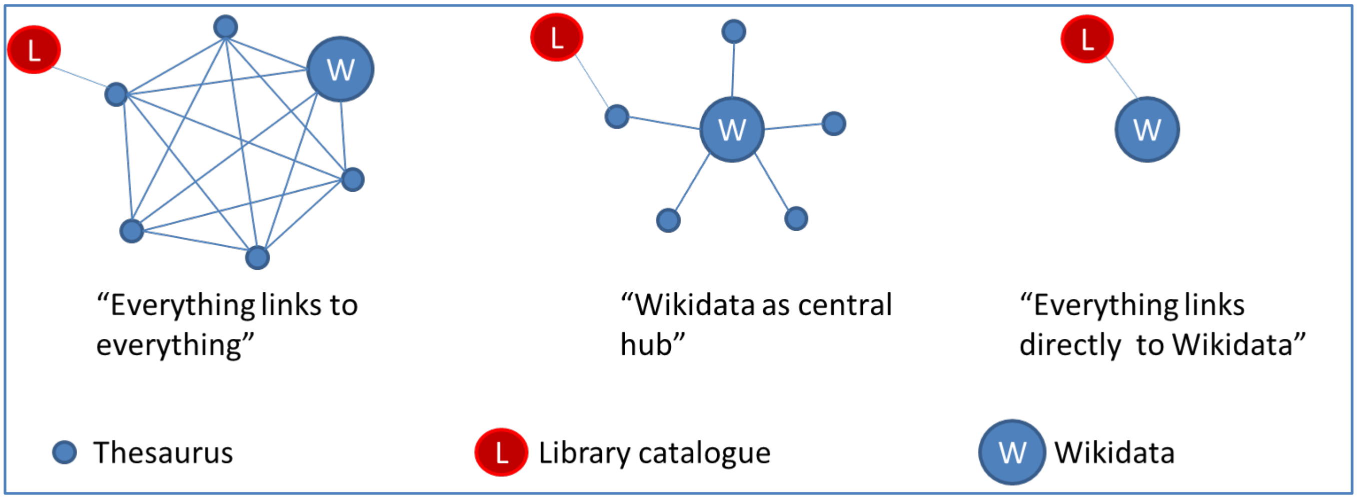

L
W
"Everything links to everything"
L
W
"Wikidata as central hub"
L
W
"Everything links directly to Wikidata"
● Thesaurus
L Library catalogue
W Wikidata

Figure 1. From linking everything to everything to linking directly to Wikidata.

Figure 1 illustrates the differences between a few possible situations that should be distinguished. On the left, the "everything links to everything" situation shows Wikidata as one of the many hubs in the linked data cloud. In the middle, the "Wikidata as authority hub" situation is shown, where name authorities are linked to Wikidata. On the right is the arrangement proposed in this article, where library systems and other systems for which this may apply share Wikidata as a common identifier mechanism.

Of course, there is a need for systems that feed Wikidata with trusted information and provide Wikidata with a backlink to a rich resource description for entities. In practice, however, many backlinks do not provide rich additional information and in such cases a direct link to Wikidata would be sufficient for the identification of entities. Figure 2 shows these two situations and other possible variations by means of dashed lines, i.e. systems that feed Wikidata, but use the Wikidata identifier as resource identifier for the outside world vs. systems that link directly to Wikidata, but keep a local thesaurus for administrative purposes.

It is certainly not the intention to encourage institutions to give up their own resource descriptions or resource identifiers locally, especially not when they are an original or rich source of information about an entity. A distinction can be made between the URL of the description of an entity and the URL of the entity itself. When following the URL of a real-world entity in a browser, it is good practice to redirect to the corresponding description of the entity. This is known as the "HTTPRange-14" issue.[5] This article will not go into any detail about this distinction other than to note that it makes sense to have a single global identifier for an entity while accepting different descriptions of that entity linked from various sources.

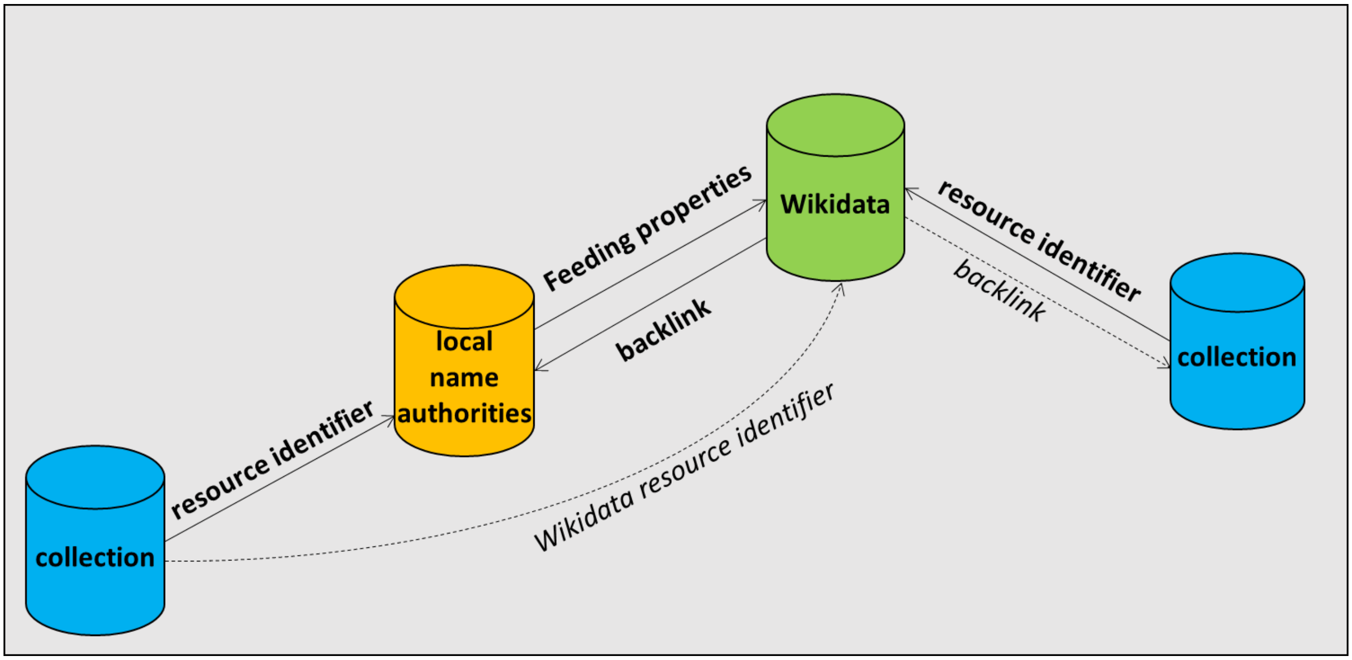


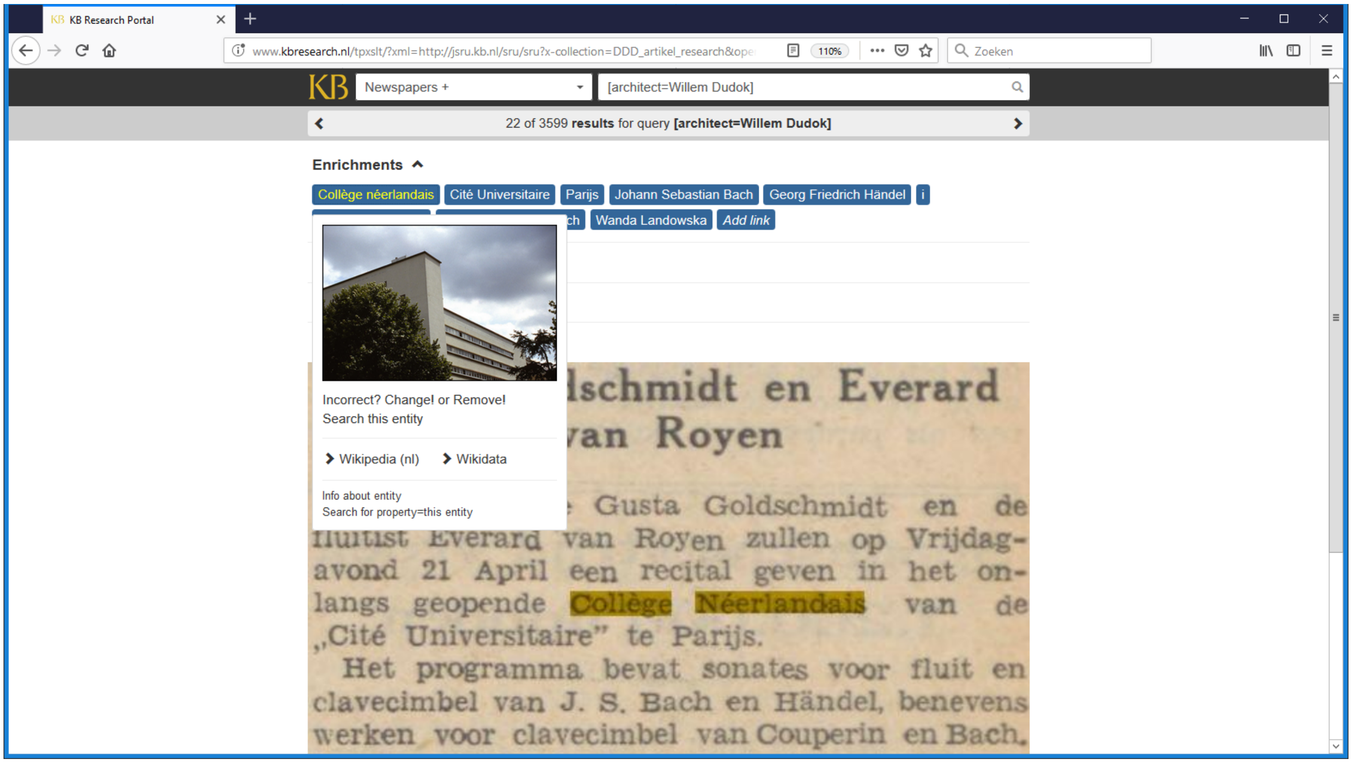
Figure 2. Feeding properties connecting collections to Wikidata (left) and direct linking to Wikidata using resource identifier (right). The dashed lines show additional connecting possibilities.

THE MOTIVATING USE CASE

The idea of using the Wikidata identifier as a universal identifier was born at the research department of the National Library of the Netherlands (KB) while working on a project aimed at automatically enriching newspaper articles with links to knowledge bases for named entities occurring in the text.[6] These links include the Wikidata identifier and, where available, the Dutch and English DBpedia (http://dbpedia.org) identifiers, the VIAF number, the Geonames number (http://geonames.org), the KB thesaurus record number, and the identifier used by the Parliamentary Documentation Centre (https://www.parlementairdocumentatiecentrum.nl/). The identifying parts of these links are indexed along with the article text in order to enable semantic search, including search based on Wikidata properties.

For demonstration purposes the enriched "newspapers+" collection was made available through the KB Research Portal, which gives access to most of the regular KB collections (figure 3).[7] In the newspaper project, linked named entities in search results are clickable to obtain more information. As most users are not expected to know SPARQL, the query language for the semantic web, the system offers a user-friendly method for semantic search: a query string entered between square brackets, for example "[roman emperor]", is expanded by a "best guess" SPARQL query in Wikidata, in this case resulting in entities having the property "position held=roman emperor.". These in turn are used to do a search for articles containing one or more mentions of a Roman emperor, even if the text "roman emperor" is not present in the article. In another example, when a user searches for the term "[beatles]" the "best guess" search yields articles mentioning entities with the property "member of=The Beatles". For ambiguous items, as in the case of "Guernica,", which can be the place in Spain or Picasso's painting, the one with the highest number of occurrences in the newspapers is selected by default, but the user may select another one. For the default or selected item, the user can select a specific property from a list of Wikidata properties available for that specific item.

The possibilities of this semantic search functionality may inspire others to use the Wikidata identifier for globally known entities in other systems as well.



[architect=Willem Dudok]

22 of 3599 **results** for query **[architect=Willem Dudok]**

lschmidt en Everard
van Royen

Gusta Goldschmidt en de fluitist Everard van Royen zullen op Vrijdagavond 21 April een recital geven in het onlangs geopende Collège Néerlandais van de „Cité Universitaire" te Parijs.

Het programma bevat sonates voor fluit en clavecimbel van J. S. Bach en Händel, benevens werken voor clavecimbel van Couperin en Bach.

Figure 3. Screenshot of the KB Research Portal with a newspaper article as result of searching "[architect=Willem Dudok]". The results are articles about buildings of which Willem Dudok is the architect. The name of the building meeting the query [architect=Willem Dudok] is highlighted.

USAGE SCENARIOS

Two usage scenarios can be considered in more detail: (1) manually following links between Wikidata descriptions and other resource descriptions, and (2) a federated SPARQL query can be performed by the system to automatically bring up linked entities.

In the first scenario, in which resource identifiers link to Wikidata, the user can follow the link to all resource descriptions having a backlink in Wikidata. But why would a user follow such a link? Reasons may include wanting more or context-specific information about the entity, or a desire to search in another system for objects mentioning a specific entity. In the latter case, the information behind the backlink should provide a URL to search for the entity, or the backlink should be the search URL itself. Wikidata provides the possibility to specify various URI templates. These can be used to specify a link for searching objects mentioning the entity, rather than just showing a thesaurus entry. When the backlink does not provide extra information or a way to search the entity, the backlink is almost useless. Thus, when systems provide resource links to Wikidata they give users access to a wealth of information about an entity in the web of data and, potentially, to objects mentioning a specific entity. Some systems only provide backlinks from

Wikidata to their resource descriptions but not the other way around. Users from such systems cannot easily benefit from these links.

The second scenario of a federated SPARQL query applies when searching objects in one system based on properties coming from other systems. Formulating such a SPARQL query is not easy because doing so requires a lot of knowledge about the linked data cloud. The alternative is to put the complete linked data cloud in a unified (triple store) database. The technology of linked data fragments might solve the performance and scaling issues but not the complexity.[8] Using a central knowledge base like Wikidata could reduce complexity for the most common situation of searching objects in other systems using properties from Wikidata. This use case requires these systems to take the users query and automatically formulate a SPARQL search. There are many systems that are linked to Wikidata that do not support SPARQL at all or only support it in a way that is not intended for the average user. Those systems can still let users benefit from Wikidata by offering a simple add-on to search in Wikidata for entities that meet some criteria and use the identifiers for a conventional search in the local system as shown for the case of the historical newspapers.

These two use cases illustrate how the use of a Wikidata identifier can lower the barrier to access information about an entity and to finding objects related to an entity by minimizing the number of hubs, minimizing the required knowledge and minimizing the required technology. This is achieved by linking resources to Wikidata and, even more so, by making objects searchable by means of the Wikidata identifier.

ADVANTAGES OF USING THE WIKIDATA IDENTIFIER AS UNIVERSAL IDENTIFIER

Summarizing the above, a number of significant advantages of using the Wikidata identifier as universal identifier can be seen. These include:

- Using the Wikidata identifier as resource identifier makes Wikidata the first hub. Applications therefore have in the first instance to deal with only one description model. From there, it is easy to navigate further: most information is only “one hub away,” so less prior knowledge is required to link from one source to another.
- Wikidata identifiers can be used for federated search based on properties in Wikidata, so there is less need to know how to access properties in other resource descriptions.
- Wikidata identifiers facilitate generating “just in case” links to systems having the Wikidata identifier indexed.
- Complicated SPARQL queries using Wikidata as primary source for properties can be shared and reused more easily compared to a situation with many diverse sources for properties.
- Wikidata offers many tools and APIs for accessing and processing data.
- Some libraries and similar institutions may even decide to use Wikidata directly for authority control when it reaches a critical mass, relieving them from maintaining a local thesaurus.

IMPLEMENTATION

Institutions can gradually adopt the use of Wikidata identifiers without needing to make radical changes in their local infrastructure. A simple first step is automatically generating links to

Wikidata in the presentation of an object or to the object description to provide contextual information and navigation options.

As a next step, the Wikidata Q-number of an entity could be indexed along with the descriptions containing it, so these objects become findable via a Wikidata identifier search, e.g. of the form:

https://whatever.local/wdsearch?id=***Q937***

The Wikidata identifier could then be used in conventional as well as federated searches for a resource, regardless of the exact spelling of a resource name. A search may be refined using Wikidata properties without further requirements with respect to local infrastructures. Institutions having a SPARQL endpoint can allow for a federated SPARQL query for combining local data with data from Wikidata. As SPARQL is not easy for the end user this requires a user interface that can formulate a SPARQL query to protect the user from knowing SPARQL.

Those institutions willing to start using the Wikidata identifier as resource identifier can unify references in their bibliographic records. Currently, for example, a reference to Albert Einstein, in a simplified, RDF-like (https://www.w3.org/RDF/) XML fragment in a bibliographic record, could look quite different for different institutions, e.g.:

```
<creator rdf:Resource=”http://data.kb.nl/thesaurus/068350767”>Albert Einstein</creator>
<creator rdf:Resource=” http://bnb.data.bl.uk/id/concept/person/lcsh/EinsteinAlbert1879-
   1955”>Albert Einstein</creator>
<creator rdf:Resource=” http://data.bnf.fr/11901607/albert_einstein/”>Albert
   Einstein</creator>
<creator rdf:Resource=” http://id.loc.gov/authorities/names/n79022889”>Albert
   Einstein</creator>
```

If the Wikidata identifier is used as resource identifier, this could for all institutions become the same:

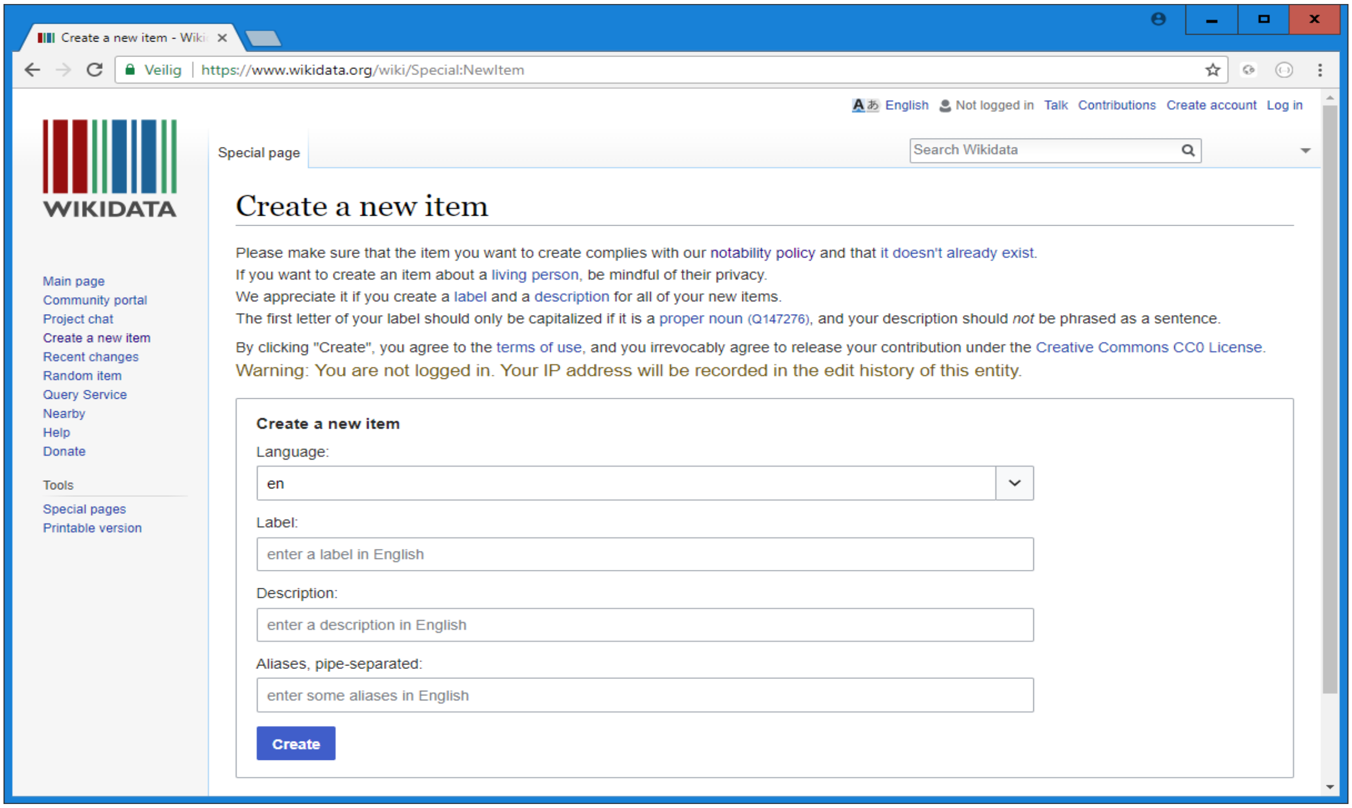
```
<creator rdf:Resource=”https://www.wikidata.org/wiki/Q937”>Albert Einstein</creator>
```

In this case it becomes easy to navigate the web, to create common bookmarklets, and provide additional functionality using the Wikidata identifier.

CATALOGUING PROCESS AND CRITERIA FOR NEW WIKIDATA ENTRIES

For institutions that decide to link their entities directly to Wikidata, their catalog software would have to be configured to support Wikidata lookups. Catalogers would not have to know about linked data or RDF to create links to Wikidata; they would simply have to query Wikidata and select the appropriate entry to link. The cataloging software would then add the selected identifier to the record being edited.

If a query in Wikidata does not yield any results the item would first then have to be created by the cataloger. Creating a new item using the Wikidata user interface (figure 4) is straightforward: create an account, add a new item, and add statements (fields) and values.



Create a new item

Please make sure that the item you want to create complies with our notability policy and that it doesn't already exist.
If you want to create an item about a living person, be mindful of their privacy.
We appreciate it if you create a label and a description for all of your new items.
The first letter of your label should only be capitalized if it is a proper noun (Q147276), and your description should *not* be phrased as a sentence.

By clicking "Create", you agree to the terms of use, and you irrevocably agree to release your contribution under the Creative Commons CC0 License.

Warning: You are not logged in. Your IP address will be recorded in the edit history of this entity.

Create a new item

Language: en

Label: enter a label in English

Description: enter a description in English

Aliases, pipe-separated: enter some aliases in English

Figure 4. Data entry screen for entering a new item in Wikidata.

Catalogers must be aware of some rules when creating items. Wikidata editors may delete items that fall under one of Wikidata's exclusion criteria, such as vandalism, empty descriptions, broken links, etc. In addition, the item must refer to an instance of a clearly identifiable conceptual or material "notable" entity. Notable means that the item must be mentioned by at least one reliable, third-party published source. Here, common sense is required: being mentioned in a telephone book or a newspaper is in itself not considered as notability. Entities that are not notable enough to be entered into Wikidata would then remain identified by a link to a local or other thesaurus.

POSSIBLE OBJECTIONS TO WIKIDATA AS AUTHORITY CONTROL MECHANISM

Although it is, at least at the present moment, not the intention of this article to propose the use of Wikidata as the primary local authority control mechanism, some institutions may nonetheless consider the opportunity to do so. There are numerous objections to this idea to note, including:

1) Institutions may consider themselves authoritative sources of information, and may therefore want to keep control over "their" thesaurus. The idea that the greater community can make changes to "their" thesaurus may not be tenable to them.

   Quality control and error detection certainly are important issues, but experts from outside the library can sometimes provide more and better information about a resource than cataloguing professionals. For misuse and erroneous input, the community can be relied on and trusted to correct and add to Wikidata entries. Information that is critical for local usage, such as access control, may still be managed locally. Despite possible objections to using Wikidata for universal authority control, national libraries and other institutions can work together with Wikidata to share responsibility of maintaining the resource, to optimize and harmonize the shared use of Wikidata, and maintain validity and authority. This might imply a more rigorous quality control.

2) Existing systems like VIAF and ISNI already, at present, still contain more persons than Wikidata, so why use Wikidata? VIAF and ISNI are domain specific and are more restrictive with respect to updates of their content and the availability of tools and APIs. In Wikidata both VIAF and ISNI are just one hub away and for internal use the VIAF and ISNI identifiers remain available. The question here is whether there will be a moment that Wikidata reaches a critical mass and supersedes VIAF and ISNI.

3) There may be disagreement about a certain entity, especially when it concerns political events or persons whose role is perceived differently by different political parties.

   Wikidata contains neutral properties. The properties that may contain subjective qualifications or might suffer bias are mostly behind the backlinks, like the abstract in Wikipedia. A fundamental difference between Wikipedia and Wikidata is that Wikipedia doesn't have to be consistent across languages. Wikidata is much more structured and therefore more useful for semantic applications. It doesn't allow for the different nuances in descriptions like Wikipedia articles do and therefore Wikidata doesn't reflect different opinions in descriptions and is less subject to bias.[9] Furthermore, the cataloguing practices in libraries are subject to bias and subjectivity too. Perception and political view may, for example, be reflected in some subject headings and may also change over time.[10] It is debatable whether a cataloger is more neutral and less biased than a larger user community.

   Although the use and acceptance of Wikipedia as a true source of information may be arguable, in the light of the current "fake news" discussion it is extremely important to guard the correctness of information in Wikipedia. In this context it is interesting to note that "according to a study in Nature, the correctness of Wikipedia articles is comparable to the *Encyclopaedia Britannica*, and a study by IBM researchers found that vandalism is repaired extremely quickly."[11]

4) Some objections have to do with the discussion of "centralization versus decentralization." Some institutions may not want a central system perceptively having control over their local data.

   The idea of using Wikidata as a common authority control mechanism is not that different from the use of any other thesaurus or identifier framework like ISBN, ISSN, etc., except for its use of a central resource description.

5) What if Wikidata disappears?

   There are solutions in terms of mirrors and a local copy of Wikidata. Moreover, national libraries and other, similar institutions that are already responsible for long-term preservation of digital content can take responsibility for keeping Wikidata alive to maximize its viability

CONCLUSION

Reconciliation of linked data identifiers in general, and using the Wikidata identifier as universal identifier in particular, has been shown to have many advantages. Libraries and similar institutions can gradually start using the Wikidata identifier without needing to make radical changes in their local database infrastructure. When Wikidata reaches a critical mass, libraries and similar institutions may want to switch to using Wikidata identifiers as the default resource identifiers or authority records. However, given the enormous growth of the number of collections that link entities to Wikidata that is already taking place, we might end up in a situation where the perception is that "if an item is not in Wikidata, it doesn't exist" stimulating putting more items in Wikidata and making local descriptions less relevant. From a strategic point of view for adopting Wikidata decision makers may pose the question: "Why do we have a local thesaurus when we already have Wikidata?" The next question, then, will probably not be "Should we go this way?" but rather "When should we go this way and start using the Wikidata identifier as The Identifier?"

REFERENCES

[1] Robert Sanderson, "The Linked Data Snowball and Why We Need Reconciliation," SlideShare, Apr. 4, 2016, https://www.slideshare.net/azaroth42/linked-data-snowball-or-why-we-need-reconciliation.

[2] Karen Smith-Yoshimura, "The rise of Wikidata as a linked data source," Hanging Together, Aug. 6, 2018, http://hangingtogether.org/?p=6775.

[3] Joachim Neubert, "Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings," in *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop*, 2017, 14-25, http://ceur-ws.org/Vol-1937/paper2.pdf.

[4] Theo van Veen, "Wikidata as universal library thesaurus," presented Oct. 2017 at WikidataCon 2017, Berlin, https://www.youtube.com/watch?v=1_NxKBnCOHM.

[5] "HTTPRange-14," Wikipedia, accessed Mar. 15, 2019, https://en.wikipedia.org/wiki/HTTPRange-14.

[6] Theo van Veen et. al., "Linking Named Entities in Dutch Historical Newspapers," in *Metadata and Semantics Research, MTSR 2016*, ed. Emmanouel Garoufallou (Cham: Springer, 2016), 205–10, https://doi.org/10.1007/978-3-319-49157-8_18.

[7] Video demonstration of "KB Research Portal," KB | National Library of the Netherlands, http://www.kbresearch.nl/xportal, accessed Apr. 26, 2019, https://www.youtube.com/watch?v=J5mCem-hEMg.

[8] Ruben Verborgh, "Linked Data Fragments: Query the Web of data on Web-scale by moving intelligence from servers to clients," accessed Mar. 15, 2019, http://linkeddatafragments.org/.

[9] Mark Graham, "The Problem with Wikidata," Apr. 6, 2012, https://www.theatlantic.com/technology/archive/2012/04/the-problem-with-wikidata/255564/.

[10] Candise Branum, “The Myth of Library Neutrality,” May 15, 2014, https://candisebranum.wordpress.com/2014/05/15/the-myth-of-library-neutrality/.

[11] “The Reliability of Wikipedia,” Wikipedia, accessed Mar. 15, 2019, https://en.wikipedia.org/wiki/Reliability_of_Wikipedia.