# What do Others Say About Similar Things – Predicate Comparing for a Linked Data Quality Boost

Benedikt Hitz-Gamper[1]

[1] University of Bern, Institute of Information Systems,
Engehaldenstrasse 8, 3012 Bern, Switzerland,
benedikt.hitz@iwi.unibe.ch

**Abstract.** Crowdsourced knowledge bases like Wikidata allow their contributors to describe information without a schema and therefore in a wide variety of diverse perspectives. This is primarily an advantage but could potentially result in lower data quality concerning uniformity and completeness of the data. Our research shows that despite fewer restrictions, there are structures and patterns becoming apparent concerning the predicates used to describe instances of specific classes like *countries*, which could be used to enhance the data quality of such knowledge bases in favour of the data consumer.

**Keywords:** Semantic Web, Linked Data, Data Quality, Wikidata.

## 1 Problem Statement

Using Linked Data to store semantically enriched information offers flexibility in that there is no need to follow pre-defined schemas. Every Linked Data publisher can decide on its own how to describe different subjects with different predicates and objects. This is especially true for Wikidata as an example of a crowdsourced knowledge base where thousands of individual contributors put information into this knowledge base.

Wikidata acts as a central storage for Linked Data, which then can be reused in other Wikimedia projects. Not only humans but also machines can read and edit the data, which is also accessible via an application programming interface (API) and a SPARQL endpoint.

Wikidata is a bottom-up approach in terms of not having to follow well-defined schemas for describing certain objects or concepts.

On the one hand, the possibility to describe subjects from very different perspectives can result in a broader, more diverse and more ubiquitous representation of the information. On the other hand, this flexibility poses a risk to the data quality (DQ) by potentially resulting in inconsistent and incomplete data.

The challenge is to maintain the flexibility and its associated benefits of Linked Data and at the same time mitigate the risk of inconsistent and incomplete information.

The goal of this work is to support Wikidata contributors in still having maximum flexibility with regard to entering new information without compromising DQ.

## 2 Relevancy

The problem is relevant both for Wikidata consumers, which is the general public, as well as for Wikidata contributors. Users need to be sure, that the data has a high DQ which means that the data is “fit for use” [1]. Contributors face the challenge to enter their information in a way that such fitness for use is ensured for a very broad variety of applications.

Keeping the flexibility of the non-schematic approach of Linked Data and at the same time, mitigating the risk of reduced DQ by inconsistent and incomplete data would be a beneficial contribution to the semantic web.

Also for publishers in a professional context who wish to publish Linked Data, our approach could be promising.

A last party interested in our approach could be semantic web ontologists who would benefit in that they will be able to examine emerging structures and ontologies from a crowdsourced Linked Data knowledge base representing multiple perspectives.

## 3 Related Work

The various fundamental dimension of DQ have been well established before the emerging of linked Open Data [2], [3].

DQ issues have been identified as an impediment for the adoption of Open Data [4] in general.

Measuring DQ in the realm of the Web of Data is in comparison to the Web of Documents on the one hand simpler because many DQ metrics can be calculated automatically. On the other hand, measuring fitness for use for a specified use case is challenging [5].

Low DQ can concern the data itself or the associated meta data [6], [7], [8].

Zaveri et al. [9] presented a systematic review of data quality assessment methodologies applied to linked Open Data (LOD) and classified different aspects of LOD DQ using accessibility dimensions, intrinsic dimensions, trust dimensions, dataset dynamicity dimensions, contextual dimensions and representational dimensions.

Improving DQ with help of a statistical approach which does not depend on domain experts or clean master data is shown in [10].

Our hereby-presented approach focuses on the completeness of data within the contextual DQ dimensions which refers to the degree to which all required information is present in a particular dataset [9]. It is comparable to the work shown in [11] but differs in that we use Wikidata as data source and that we are especially focused on assisting the individual contributors of Wikidata.

## 4 Research Question

Are there any patterns becoming apparent in Wikidata concerning the predicates used to describe instances of a certain specific class and can these structures be used to help the Wikidata contributors to improve DQ?

## 5 Hypotheses

The hypothesis is that in Wikidata, thanks to the flexibility of Linked Data, there will be a diverse and ubiquitous representation of information but still there will be some convergent patterns becoming apparent.

If predicates used to describe instances of the same class (e.g. “countries”) are compared, there will be predicates, which are more often used than others are. If a predicate used for almost every instance is missing for a certain instance, this is an indication for potentially missing data or an inconsistency in having used another predicate to describe the same information.

## 6 Evaluation Plan

We have to show which predicates are used how many times in describing all the subjects of a certain class. We call this the predicate distribution of a certain class. Our goal, which is to assist the contributors of Wikidata, will work, when such a predicate distribution shows, that certain predicates are used for a large majority of subjects of the examined class.

## 7 Approach

For testing our hypothesis and thereby trying to answer our research question, we examine Linked Data from Wikidata. Wikidata is an open knowledge base with people and machines being able to read and edit data [12]. The approach of Wikidata is a collaborative effort by a community of contributors [13]. It profits from the flexibility of Linked Data in that every contributor can describe information in different ways and from different perspectives. There are no schemas given to enter certain information, which increases the danger of incomplete and inconsistent information. This puts Wikidata in an ideal position to test our approach.

Our approach consists of the following steps: Decision on the class (in the sense of an object-oriented approach, eg. “countries”) from which the instances belong. Collecting all the triples where the instances appear on the subject position. Listing all the predicates used in these triples and counting for which instances they occur how many times.

The technical realisation of these steps is done with help of the SPARQL endpoint of Wikidata (http://query.wikidata.org) which allows to query the knowledge base data in an easily manner. The aggregation of the data is done with some simple scripts within the statistical program ‘R’.

## 8 Preliminary Results

To explore the viability of our ideas, we did a manual examination of a few examples, which we present here. As a first step, we chose to examine instances of a class with only a few members. We decided to take a closer look at all the seven members of the Swiss Federal Council, which is the government of Switzerland.

The following tables 1-3 show our preliminary results using our approach for all the seven members of the Swiss Federal Council. Table 1 shows all the used predicates. It turns out that predicates exist which are used for every but one subject. This could be a hint that there is a missing value for the subject in question.

Indeed, it turns out, that the three predicates, which are missing only once, are really missing values: Alain Berset surely has a work location and a native language. Guy Parmelin on the other hand has no GND ID (yet). But this could (and probably should) also be given in a triple stating that there is no such value for mister Parmelin.

Table 2 shows summarized data concerning the predicates used to describe all the members of the Swiss Federal Council.

**Table 1.** Members of the Swiss Federal Council (columns) acting as subjects in Linked Data triples. Used predicates are shown in rows. The numbers denote the count of objects for a certain subject/predicate combination. NA stands for “not available” and implies that the predicate in question was not used in this case. The yellow filled boxes are NA-values in cases where all other subjects have used this predicate.

| | **Didier Burkhalter** | **Doris Leuthard** | **Guy Parmelin** | **Simonetta Sommaruga** | **Alain Berset** | **Ueli Maurer** | **Johann Schneider-Ammann** |
|---:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|
| position held | 4 | 6 | 1 | 4 | 3 | 4 | 3 |
| work location | 1 | 1 | 1 | 1 | NA | 1 | 1 |
| member of political party | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| name in native language | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| instance of | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| native language | 1 | 1 | 1 | 1 | NA | 1 | 1 |
| place of origin (Switzerland) | 2 | 2 | 1 | 2 | 1 | 2 | 1 |

| | | | | | | | |
|---:|---|---|---|---|---|---|---|
| occupation | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| languages spoken, written or signed | 2 | 2 | 2 | 2 | 2 | 2 | 3 |
| GND ID | 1 | 1 | NA | 1 | 1 | 1 | 1 |
| SUDOC authorities | 1 | NA | NA | NA | 1 | NA | NA |
| NUKAT (WarsawU) authorities | 1 | NA | NA | NA | NA | NA | NA |
| Munzinger IBA | 1 | NA | NA | 1 | 1 | NA | NA |
| date of birth | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| image | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| sex or gender | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| LCAuth ID | 1 | 1 | NA | 1 | NA | NA | NA |
| country of citizenship | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| given name | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| Commons category | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ISNI | 1 | 1 | NA | 1 | NA | NA | NA |
| VIAF ID | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Swiss parliament ID | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| place of birth | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Freebase ID | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| father | NA | 1 | NA | NA | NA | NA | NA |
| HDS ID | NA | 1 | NA | 1 | NA | 1 | 1 |
| educated at | NA | 1 | 1 | 1 | 1 | NA | 2 |
| military rank | NA | NA | 1 | NA | NA | NA | NA |
| spouse | NA | NA | 1 | 1 | NA | NA | 1 |
| official website | NA | NA | 1 | NA | 1 | NA | NA |
| family name | NA | NA | 1 | NA | NA | NA | NA |
| residence | NA | NA | 1 | NA | NA | NA | NA |
| birth name | NA | NA | NA | 1 | NA | NA | NA |
| relative | NA | NA | NA | 1 | NA | NA | NA |
| number of children | NA | NA | NA | NA | NA | 1 | 1 |
| Twitter username | NA | NA | NA | NA | NA | NA | 1 |
| award received | NA | NA | NA | NA | NA | NA | 1 |

**Table 2.** Summarized data for instances of class “members of the Swiss Federal Council”

| | |
|---:|---|
| number of instances of class “members of the Swiss Federal Council” | 7 |

| | |
|---|---|
| number of different predicates used | 38 |
| number of predicates used for all members | 17 |
| number of predicates used for all but one member | 3 |
| number of predicates used for only one member | 7 |

As a second class, we selected all the 202 countries, which are listed in Wikidata. Table 3 shows summarized data for the used predicates.

One interesting observation is, that if the number of potential missing values for a certain country is very high, it turns out, that this probably is not because of low DQ but is a special case of either a historical (non-existing anymore) or a disputed country. A high number of potential missing values for a certain instance could therefore mean, that this instance could belong to another class ("historical country" instead of "country").

**Table 3.** Summarized data for instances of class "countries"

| | | |
|---|---|---|
| number of instances of class "countries" | 202 | |
| number of different predicates used | 266 | |
| number of predicates used for more than 95% but less than 100% of the countries | 45 | |
| number of predicates used for less than 1% of the countries | 72 | |
| number of countries with more than 3 potential missing values | 8 | |
| countries with more than 3 potential missing values | country | # values |
| | Eastern Kingdom of Women | 45 |
| | Turkish Republic of Northern Cyprus | 22 |
| | Somaliland | 22 |
| | Sahrawi Arab Democratic Republic | 22 |
| | Palestine | 22 |
| | Transnistria | 15 |
| | Kingdom of the Netherlands | 11 |
| | Kosovo | 9 |

Fig. 1 shows the distribution of predicates for six selected classes. For all the examined classes, there is only a small fraction of predicates, which do occur in conjunction with a high percentage of instances.

The yellow curve concerning the predicates used in *countries*, which is s-shaped with a long tail, is the one to be expected for a class with a high number of predicates and instances. It means that most predicates are only scarcely used. These predicates could be very interesting in depicting instances with special properties.

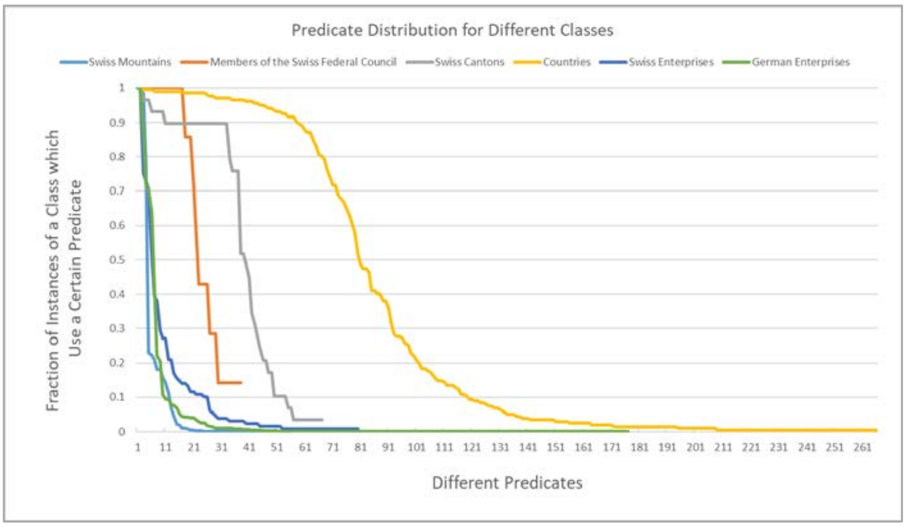

**Fig. 1.** Predicate distribution for different classes. Reading example: There are a little bit more than 60 predicates, which are used in conjunction with more than 90% of the countries.

## 9 Reflections

Our preliminary results show that there are examples of classes, which show a potentially useful distribution of the predicates used to describe the various instances of the class. Of course, this finding has to be systematically expanded to other classes. This will show, for which kind of classes, the distribution of the predicates used is especially helpful in supporting the Wikidata contributors. This research has to take into account the fact, that information in Wikidata is not solely sourced by human contributors but also by automated importing from other sources. This will have an effect on the predicate distribution for sure.

Further research has to include practical aspects like how to pass information about other subjects of the same class to the Wikidata contributors during editing of information in the knowledge base in a helpful but unrestricting way.

Obviously, with our approach, low DQ concerning missing values can only be detected if the missing predicates are used for other instances of the same class. However, the community aspect of Wikidata increases the probability that the important predicates are present for other instances of the same class.

Our approach can contribute to high DQ in Wikidata and other crowdsourced Linked Data knowledge bases without limiting the freedom of different and diverse perspectives which is one of the strength of such crowdsourced approaches.

## Acknowledgements

I would like to give thanks to my PhD advisor Prof. Dr. Thomas Myrach from the Institute of Information Systems of the University of Bern and to my mentor Dr. Yolanda Gil from the Information Sciences Institute of the University of Southern California. Furthermore, I would like to give credit to Mikhail Popov from the Wikimedia Foundation for his "WikidataQueryServiceR" library for accessing the Wikidata Query Service (WDQS) through 'R'.

## References

[1] C. Bizer and R. Cyganiak, 'Quality-driven information filtering using the WIQA policy framework', *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 1, pp. 1–10, Jan. 2009.

[2] R. Y. Wang and D. M. Strong, 'Beyond Accuracy: What Data Quality Means to Data Consumers', *Journal of Management Information Systems*, vol. 12, pp. 5–33, 1996.

[3] D. M. Strong, Y. W. Lee, and R. Y. Wang, 'Data Quality in Context', *Commun. ACM*, vol. 40, no. 5, pp. 103–110, May 1997.

[4] A. Zuiderwijk, M. Janssen, S. Choenni, R. Meijer, R. S. Alibaks, and R. Sheikh_Alibaks, 'Socio-technical impediments of open data', *Electronic Journal of eGovernment*, vol. 10, no. 2, pp. 156–172, 2012.

[5] J. Debattista, Sö. Auer, and C. Lange, 'Luzzu—A Methodology and Framework for Linked Data Quality Assessment', *Journal of Data and Information Quality*, vol. 8, no. 1, pp. 1–32, Oct. 2016.

[6] S. Neumaier, J. Umbrich, and A. Polleres, 'Automated Quality Assessment of Metadata across Open Data Portals', *Journal of Data and Information Quality*, vol. 8, no. 1, pp. 1–29, Oct. 2016.

[7] J. Umbrich, S. Neumaier, and A. Polleres, 'Quality Assessment and Evolution of Open Data Portals', in *2015 3rd International Conference on Future Internet of Things and Cloud*, 2015, pp. 404–411.

[8] J. Kučera, D. Chlapek, and M. Nečaský, 'Open Government Data Catalogs: Current Approaches and Quality Perspective', in *Technology-Enabled Innovation for Democracy, Government and Governance*, 2013, pp. 152–166.

[9] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, 'Quality assessment for Linked Data: A Survey', *Semantic Web*, vol. 7, no. 1, pp. 63–93, Jan. 2016.

[10] S. De, Y. Hu, V. V. Meduri, Y. Chen, and S. Kambhampati, 'BayesWipe: A Scalable Probabilistic Framework for Improving Data Quality', *Journal of Data and Information Quality*, vol. 8, no. 1, pp. 1–30, Oct. 2016.

[11] C. Böhm *et al.*, 'Profiling linked open data with ProLOD', in *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, 2010, pp. 175–178.

[12] 'Wikidata'. [Online]. Available: https://www.wikidata.org/wiki/Wikidata:Main_Page. [Accessed: 26-Apr-2017].

[13] D. Vrandečić and M. Krötzsch, 'Wikidata: A Free Collaborative Knowledgebase', *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014.