One Knowledge Graph to Rule them All?

Analyzing the Differences between DBpedia, YAGO, Wikidata & co.

Daniel Ringler and Heiko Paulheim

University of Mannheim, Data and Web Science Group

Abstract. Public Knowledge Graphs (KGs) on the Web are considered a valuable asset for developing intelligent applications. They contain general knowledge which can be used, e.g., for improving data analytics tools, text processing pipelines, or recommender systems. While the large players, e.g., DBpedia, YAGO, or Wikidata, are often considered similar in nature and coverage, there are, in fact, quite a few differences. In this paper, we quantify those differences, and identify the overlapping and the complementary parts of public KGs. From those considerations, we can conclude that the KGs are hardly interchangeable, and that each of them has its strenghts and weaknesses when it comes to applications in different domains.

1 Knowledge Graphs on the Web

The term "Knowledge Graph" was coined by Google when they introduced their knowledge graph as a backbone of a new Web search strategy in 2012, i.e., moving from pure text processing to a more symbolic representation of knowledge, using the slogan "things, not strings"[1].

Various public knowledge graphs are available on the Web, including DBpedia [3] and YAGO [9], both of which are created by extracting information from Wikipedia (the latter exploiting WordNet on top), the community edited Wikidata [10], which imports other datasets, e.g., from national libraries[2], as well as from the discontinued Freebase [7], the expert curated OpenCyc [4], and NELL [1], which exploits pattern-based knowledge extraction from a large Web corpus.

Although all these knowledge graphs contain a lot of valuable information, choosing one KG for building a specific application is not a straight forward task. Depending on the domain and task at hand, some KGs might be better suited than others. However, there are no guidelines or best practices on how to choose a knowledge graph which fits a given problem. Previous works mostly report global numbers, such as the overall size of knowledge graphs, such as [6], and focus on other aspects, such as data quality [2]. The question which KG fits which purpose, however, has not been answered so far.

[1] `https://googleblog.blogspot.de/2012/05/introducing-knowledge-graph-things-not.html`

[2] `https://www.wikidata.org/wiki/Wikidata:Data_donation`

Table 1: Global Properties of the Knowledge Graphs compared in this paper

	DBpedia	YAGO	Wikidata	OpenCyc	NELL
Version	2016-04	YAGO3	2016-08-01	2016-09-05	08m.995
# instances	5,109,890	5,130,031	17,581,152	118,125	1,974,297
# axioms	397,831,457	1,435,808,056	1,633,309,138	2,413,894	3,402,971
avg. indegree	13.52	17.44	9.83	10.03	5.33
avg. outdegree	47.55	101.86	41.25	9.23	1.25
# classes	754	576,331	30,765	116,822	290
# relations	3,555	93,659	11,053	165	1,334

2 Overall Size and Shape of Knowledge Graphs

For the analysis in this paper, we focus on the public knowledge graphs DBpedia, YAGO, Wikidata, OpenCyc, and NELL.[3,4] For those five KGs, we used the most recent available versions at the time of this analysis, as shown in Table 1.

We can observe that DBpedia and YAGO have roughly the same number of instances, which is not surprising, due to their construction process, which creates an instance per Wikipedia page. Wikidata, which uses additional sources plus a community editing process, has about tree times more instances. It is remarkable that YAGO and Wikidata have roughly the same number of axioms, although Wikidata has three times more instances. This hints at a higher level of detail in YAGO, which is also reflected in the degree distributions.

OpenCyc and NELL are much smaller. NELL is particularly smaller w.r.t. axioms, not instances, i.e., the graph is less dense. This is also reflected in the degree of instances, which depicts that on average, each instance has less than seven connections. The other graphs are much denser, e.g., each instance in Wikidata has about 50 connections on average, each instance in DBpedia has about 60, and each instance in YAGO has even about 120 connections on average.

The schema sizes also differ widely. In particular the number of classes are very different. This can be explained by different modeling styles: YAGO automatically generates very fine-grained classes, based on Wikipedia categories. Those are often complex types encoding various facts, such as "American Rock Keyboardists". KGs like DBpedia or NELL, on the other hand, use well-defined, manually curated ontologies with much fewer classes.

Since Wikidata provides live updates, it is the most timely source (together with DBpedia Live, which is a variant of DBpedia fed from an update stream of Wikipedia). From the non-live sources, NELL has the fastest release cycle, providing a new release every few days. However, NELL uses a fixed corpus of Web pages, which is not updated as regularly. Thus, the short release cycles do not necessarily lead to more timely information. DBpedia has biyearly releases, and YAGO and OpenCyc have update cycles longer than a year.

[3] Freebase was discarded as it is discontinued, and non-public KGs were not considered, as it is impossible to run the analysis on non-public data.

[4] Scripts are available at `https://github.com/dringler/KnowledgeGraphAnalysis`.

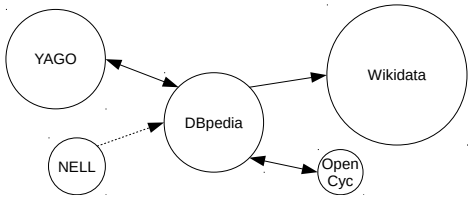


Fig. 1: Knowledge Graphs inspected in this paper, and their interlinks. Like for the Linked Open Data Cloud diagrams [8], the size of the circles reflects the number of instances in the graph (except for OpenCyc, which would have to be depicted an order of magnitude smaller).

3 Category-Specific Analysis

When building an intelligent, knowledge graph backed application for a specific use case, it is important to know how fit a given knowledge graph is for the domain and task at hand. To answer this question, we have picked 25 popular classes in the five knowledge graphs and performed an in-depth comparison. For those, we computed the total number of instances in the different graphs, as well as the average in and out degree. The results are depicted in figure 2.

While DBpedia and YAGO, both derived from Wikipedia, are rather comparable, there are notable differences in coverage, in particular for events, where the number of events in YAGO is more than five times larger than the number in DBpedia. On the other hand, DBpedia has information about four times as many settlements (i.e., cities, towns, and villages) as YAGO. Furthermore, the level of detail provided in YAGO is usually a bit larger than DBpedia.

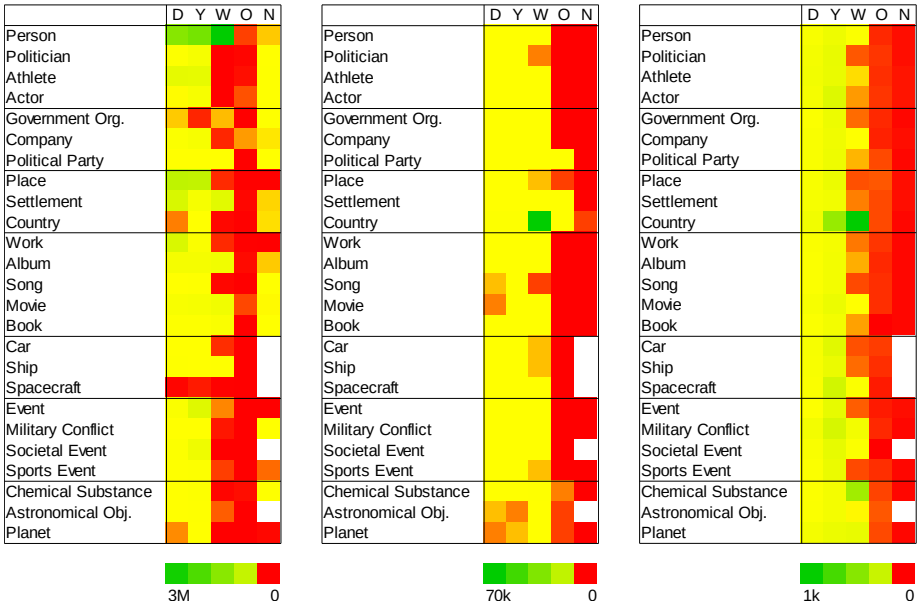
The other three graphs differ a lot more. Wikidata contains twice as many persons as DBpedia and YAGO, and also outnumbers them in music albums and books. Furthermore, it provides a higher level of detail for chemical substances and particularly countries. On the other hand, there are also classes which are hardly represented in Wikidata, such as songs.[5] As far as Wikidata is concerned, the differences can be partially explained by the external datasets imported into the knowledge graph.

OpenCyc and NELL are generally smaller and less detailed. However, NELL has some particularly large classes, e.g., actor, song, and chemical substance, and for government organizations, it even outnumbers the other graphs. On the other hand, there are classes which are not covered by NELL at all.

4 Overlap of Knowledge Graphs

Knowledge graphs on the Web are equipped with links connecting identical entities between those graphs. However, due to the *open world assumption*, those

[5] The reason why so few politicians, actors, and athletes are listed for Wikidata is that they are usually not modeled using explicit classes.



(a) Number of instances (b) Average indegree (c) Average outdegree

Fig. 2: Number of instances (a), avg. indegree (b) and avg. outdegree (c) of selected classes. D=DBpedia, Y=YAGO, W=Wikidata, O=OpenCyc, N=NELL.

links are notoriously incomplete. For example, from the fact that 2,000 cities in knowledge graph A are linked to cities in knowledge graph B, we cannot conclude that this is the number of cities contained in the intersection of A and B.

Links between knowledge graphs can be determined using entity linkage approaches [5], e.g., interlinking all entities with the same name.

Given that there is already a certain number of (correct) interlinks between two knowledge graphs, we can also compute the quality of a linking approach in terms of recall and precision. Given that the actual number of links is C, the number of links found by a linkage rule is F, and that the number of correct links in F is F^+, recall and precision are defined as

$$R := \frac{|F^+|}{|C|} \tag{1}$$

$$P := \frac{|F^+|}{|F|} \tag{2}$$

By resolving both to $|F^+|$ and combining the equations, we can estimate $|C|$ as

$$|C| = |F| \cdot P \cdot \frac{1}{R} \tag{3}$$

For our analysis, we use 16 combinations of string metrics and thresholds on the instances' labels: string equality, scaled Levenshtein (thresholds 0.8, 0.9,

and 1.0), Jaccard (0.6, 0.8, and 1.0), Jaro (0.9, 0.95, and 1.0), JaroWinkler (0.9, 0.95, and 1.0), and MongeElkan (0.9, 0.95, and 1.0). Furthermore, to speed up the computation, we exploit token-based blocking in a preprocessing step (where each instance is only assigned to the block of the least frequent token), and discarding blocks larger than 1M pairs.

As incomplete link sets for estimating recall and precision, we use the links between the knowledge graphs, if present. If there are no links, we exploit transitivity and symmetry, and follow the link path through DBpedia (see Fig. 1). NELL has no direct links to the other graphs, but links to Wikipedia pages corresponding to DBpedia instances, which we use to create links to DBpedia (indicated by the dashed line in the figure).

Fig. 3 depicts the pairwise overlap of the knowledge graphs, using the 25 classes also inspected above, according to two measures: potential gain by joining the two knowledge graphs (i.e., the relation of the union to the larger of the two graphs), and the overlap relative to the existing KG interlinks.

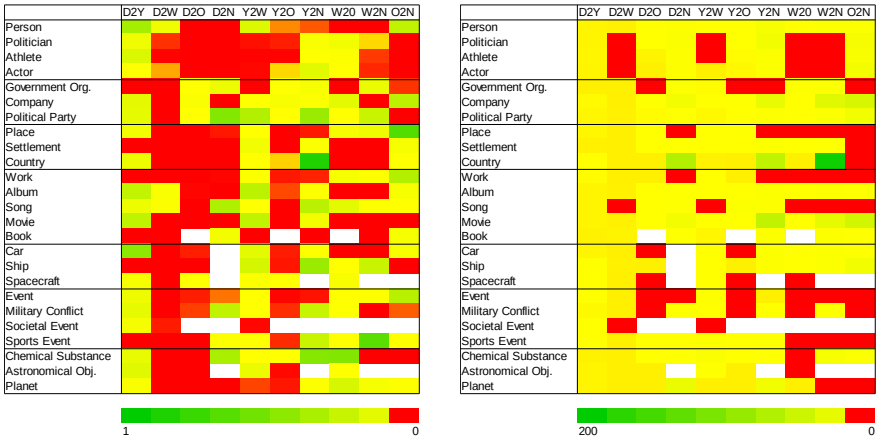
Overall, we can observe that merging two graphs would usually lead to a 5% increase of coverage of instances, compared to using one KG alone. The largest potential gain most often comes from merging the larger knowledge graphs with NELL. We can therefore conclude that NELL is rather complementary to most of the other KGs under consideration. The most complementary classes, with an average gain of more than 10% across all pairs of knowledge graphs, are political parties and chemical substances. When looking at the overlap relative to the number of existing links, NELL has the weakest interlinking: e.g., for YAGO and NELL, the estimated overlap is more than eight times larger than the number of interlinks. The classes with the weakest degree of interlinking are countries (32 times larger overlap than explicit interlinks), movies (13 times larger), and companies (10 times larger).[6]

5 Conclusions and Recommendations

We have compared the coverage and level of detail for 25 popular classes. Some key findings from this comparison include:

- For person data, Wikidata is the most suitable source, containing twice as many instances as DBpedia or YAGO, at a similar level of detail.
- Organizations, such as companies, are best described in YAGO.
- DBpedia contains more places than the other KGs, including almost four times more cities, villages etc. than YAGO.
- While DBpedia and YAGO contain much more countries than Wikidata (due to the inclusion of historic countries, such as the Roman Empire), Wikidata holds the most detailed information about countries.

[6] Note that it is not necessary that the linking approach is particularly good, as long as we can estimate its quality reasonably well. In our experiments, the agreement about the estimated overlap is rather high, showing an intra-class correlation coefficient (ICC) of 0.969. In contrast, the size of the actual alignments found by the different approaches differs a lot more, showing an ICC of only 0.646.



(a) Overlap as potential gain (b) Overlap relative to existing links

Fig. 3: Number as potential gain (a) and relative to existing interlinks (b) of selected classes. D=DBpedia, Y=YAGO, W=Wikidata, O=OpenCyc, N=NELL.

- Overall, DBpedia contains the largest number of artistic works, although details differ for subclasses: Wikidata contains more music albums and movies, while YAGO contains more songs. The most detailed information about artistic works is provided by YAGO.
- Cars and spacecraft are best covered in YAGO, while DBpedia is the better resource for ships.
- For events, YAGO is the most suitable source, both in terms of coverage and level of detail.
- NELL contains the largest number of chemical substances. The highest level of degree for chemicals, however, is provided in Wikidata.
- YAGO contains the largest number of astronomical objects.

Note that those numbers are not exhaustive, they merely demonstrate the need for a careful analysis of KGs before exploiting them for a project at hand.

In addition to the question which knowledge graph serves a certain task best, another question is whether it makes sense to use *more than one* combined. Here, we have observed that there is often a considerable complementarity. Especially NELL is very complementary to the other KGs, although a lot less rich in details. Thus, the coverage can often be extended significantly by combining different KGs. This, however, requires refinement of the interlinking, since the interlinks are usually incomplete.

Summarizing: Although DBpedia, YAGO, Wikidata & co. are often perceived at somewhat similar to one another, our analysis has revealed that there are considerable differences. Hence, when deploying a public KG in a project, it makes sense to look at the details first before selecting one KG.

References

1. Carlson, A., Betteridge, J., Wang, R.C., Hruschka Jr, E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 101–110 (2010)
2. Färber, M., Ell, B., Menne, C., Rettinger, A., Bartscherer, F.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web (to appear) (2016)
3. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal 6(2) (2013)
4. Lenat, D.B.: CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM 38(11), 33–38 (1995)
5. Nentwig, M., Hartung, M., Ngonga Ngomo, A.C., Rahm, E.: A survey of current link discovery frameworks. Semantic Web 8(3), 419–436 (2017)
6. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 8(3), 489–508 (2017)
7. Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., Pintscher, L.: From freebase to wikidata: The great migration. In: Proceedings of the 25th International Conference on World Wide Web. pp. 1419–1428 (2016)
8. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data Best Practices in Different Topical Domains. In: International Semantic Web Conference. LNCS, vol. 8796 (2014)
9. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: 16th international conference on World Wide Web. pp. 697–706 (2007)
10. Vrandečić, D., Krötzsch, M.: Wikidata: a Free Collaborative Knowledge Base. Communications of the ACM 57(10), 78–85 (2014)