





An Observational Study of Equivalence Links in Cultural Heritage Linked Data for *agents*

Nuno Freire¹ , Hugo Manguinhas², and Antoine Isaac^{2,3} 

¹ INESC-ID, Lisbon, Portugal

nuno.freire@tecnico.ulisboa.pt

² Europeana Foundation, The Hague, The Netherlands

{hugo.manguinhas, antoine.isaac}@europeana.eu

³ Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Abstract. This article presents an observational study of the virtual graph formed by equivalence links between *agent* entities across 8 knowledge bases. To evaluate the potential of this linked data graph, we measured the equivalences that it could provide for a real dataset. We crawled the virtual graph by starting from references to *agents* we found in descriptions of objects collected from data of cultural heritage institutions in Europeana. Our study characterizes the current virtual equivalence graph, presenting statistics about the links, their type and origin. Crawling the equivalences for *agent* URIs required several crawling iterations on the virtual equivalence graph. The amount of gathered equivalences grows steeply in the first 3 crawling iterations and stabilizes on the 4th iteration. VIAF was the KB with the highest number of equivalences, reaching 60.7%, and it was followed by Wikidata with 34.5%.

Keywords: Linked data · Agents · Equivalence links · Cultural heritage

1 Introduction

Nowadays, large knowledge bases (KBs) are available as linked data under open licenses, like DBpedia¹ and Wikidata². Exploiting equivalences of entities across these KBs is crucial for data-driven application that require, e.g., to obtain additional data about an entity across several KBs, or to support disambiguation operations.

We conducted an observational study of the virtual graph formed by equivalence relations between entities of 8 open KBs for entities of type *agent* (persons, organizations) in cultural heritage (CH) data. In particular, we measured the quantity of equivalences that this graph could provide for a dataset from Europeana³ containing references to *agents* in descriptions of CH objects.

This study provides insights about the equivalence links across KBs and the potential benefits of crawling this virtual equivalence graph for discovering equivalences of

¹ <https://dbpedia.org/>.

² <https://www.wikidata.org/>.

³ <https://pro.europeana.eu/our-mission>.

agents referred to in datasets. It is informative for future research and for designing innovative applications, such as the case of Europeana who seeks to acquire *agent* name variants/translations or extra biographical information [1].

We follow, in Sect. 2, by describing related work on linked data and equivalence graphs. Section 3 presents how the study was conducted. Section 4 details the results and their analysis. Section 5 highlights our conclusions and presents future work.

2 Related Work

The exploitation of KB equivalence links for specific applications has been addressed earlier. Beek et al. (2018) have gathered the largest dataset of *owl:sameAs* statements from the web of data [2]. Similarly to us, Correndo et al. (2012) have conducted a statistical and qualitative analysis of the graph of instance level equivalences, and explored their use for computing alignments at conceptual level [3].

Research on the quality of linked data equivalence statements is relevant for us. It has especially reported (sometimes incorrect) uses of *owl:sameAs* to represent different degrees of equivalence [4–6]. Work on linked data aggregation and cleaning [7, 8] has also revealed data quality to be a challenge both at the level of semantics and the one of syntax [9, 10]. Especially relevant for us, an empirical study by Asprino et al. (2019) investigated the modelling style and the general structure of linked open data, including issues for the equivalence graphs formed by interlinking [11].

Regarding CH, the creation of KBs has been a long-term practice, and started much earlier than the emergence of the Semantic Web. In this domain however, the stated equivalences between major open KBs have not been studied recently.

3 Design of the Study

We have conducted an observational study gathering the existing equivalence relations between entities across 8 KBs:

- DBpedia - a multilingual KB created by extracting structured data from Wikipedia.
- data.bnf.fr (BnF) - a project by the French National Library that makes available data about bibliographic entities.
- datos.bne.es (BNE) - a KB of bibliographic data by the National Library of Spain.
- Library of Congress Names⁴ (NAF) - a KB that provides authoritative data for names of persons, organizations, events, places, and titles.
- The Union List of Artist Names⁵ (ULAN) - ULAN contains names, relationships, notes, sources, and biographical information for artists.
- Gemeinsame Normdatei⁶ (GND) - an KB for personal names, subject headings and corporate bodies, managed mainly by the German National Library.

⁴ <http://id.loc.gov/authorities/names.html>.

⁵ <https://www.getty.edu/research/tools/vocabularies/ulan/>.

⁶ https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html.

- Virtual International Authority File⁷ (VIAF) - a cooperation of OCLC with mainly national libraries, combining multiple KBs from libraries, archives and museums.
- Wikidata - a collaborative KB hosted by the Wikimedia Foundation.

By considering the transitive closure of the resulting compound set of equivalence statements, one obtains a virtual equivalence graph with entities from all the KBs as nodes. Our study was divided into two parts.

First, we measured the amount of stated equivalence relations between KBs by considering all the equivalences asserted by at least one KB, not using any additional external sources. The statements were collected preferably via SPARQL, or via a file-based RDF distribution of the KB. We collected all statements where the property was one of⁸: *owl:sameAs*; *skos:exactMatch*; *skos:closeMatch*; or *schema:sameAs*. This selection was based on a preliminary profiling of the KBs, where we found these standard properties to be the most often used for representing equivalence.

In the second part of the study, we focused on the entity type *agent*, and measured the quantity of equivalences that the joint equivalence graph could provide for a dataset containing references to *agents* in descriptions of CH objects.

The first task was to create a set of URIs referring to *agents*. For this purpose, we used the APIs⁹ for accessing and querying the dataset aggregated by Europeana. We located 1,164,323 unique RDF resources about *agents* used by the Europeana data providers¹⁰. From these we excluded all anonymous (blank) nodes and all the URIs that contain a URI fragment appended to the URI of the CH object. These resources without a “real” identifier are likely to correspond to cases where the *agent* does not come from a pre-existing controlled, “authoritative” KB, but are just created ad-hoc for the description of the cultural object. The resulting set contains 286,090 unique *agent* URIs, and the majority of them belong to a KB in our study, as Table 1 shows.

The set of *agent* URIs was then used to initiate a series of crawling iterations of the equivalence graph. The crawler was instructed to crawl the statements with any of the properties mentioned in Sect. 3. It assumes that all properties are transitive, including *skos:closeMatch*, and that transitivity applies across all types of properties¹¹. In the first iteration, we crawled directly the *agent* URIs and gathered all the equivalence relations their KB contained for them. From the second iteration and onwards, the crawler obtained equivalent *agent* URIs by searching in the KBs for any URI that was collected in previous crawling iterations and adding the URIs that these KBs declared to be equivalent to the original ones. At the end of each iteration, the crawler generated a report about the newly

⁷ <https://viaf.org>.

⁸ For readability purposes, in this article we abbreviate namespaces as follows: owl for <http://www.w3.org/2002/07/owl#>; skos for <http://www.w3.org/2004/02/skos/core#>; schema for <http://schema.org/>; wdt for <http://www.wikidata.org/prop/direct/>.

⁹ <https://pro.europeana.eu/resources/apis>.

¹⁰ We aim to provide insights that could be beneficial to providers and users of the original metadata, therefore, we have excluded the URIs used in automatic enrichment by Europeana (cf. <https://pro.europeana.eu/page/europeana-semantic-enrichment>).

¹¹ This goes beyond the actual formal semantics of these properties, but we wanted to experiment with it nonetheless, to get an upper bound of the level of benefit obtainable from the equivalences - and experience shows that the biggest data quality issues actually lie elsewhere.

found equivalent URIs. We repeated the crawling process for newly found equivalences several times until the increase of URIs resulting from one iteration was negligible.

Table 1. Amounts of unique URIs in the set from Europeana that belong to a KB in the study

	DBpedia	BnF	BNE	NAF	ULAN	GND	VIAF	Wikidata	Other KBs (or none)
URI uses in Europeana	2	2,010	30,449	0	7,451	242,297	2,174	0	1707

4 Results

The study provided informative results on four aspects of the KBs and their virtual equivalence graph. Each aspect is presented in the following subsections.

4.1 Existing Equivalences Between Knowledge Bases

We did two measurements on the equivalence statements between the KBs. The first measurement considered all types of equivalences, and the second measurement was made considering solely *skos:closeMatch* equivalences. Our motivation for measuring separately the *skos:closeMatch* equivalences was because this property expresses equivalence with a degree of uncertainty, while the three others seek to capture exact equivalence, which may be an important aspect for many applications.

Table 2 presents the results considering the 4 properties for equivalence, showing the amounts of statements when a KB publishes an equivalence to another KB and when other KBs publish an equivalence to the KB being considered. The table also shows the number of KBs linked by equivalences to each KB. A total amount of 60,307,328 equivalences are stated in the 8 KBs.

The results show high interconnection between KBs. All KBs express equivalences to at least one other KB, and all KBs are the target of equivalences stated in at least one KB. An interesting observation is that 3 out of the 8 KBs are focused only on *agents* (VIAF, NAF and ULAN), and 2 of them, VIAF and NAF, are among the 3 most linked KBs. GND is the second most linked KB, and the most linked of the KBs that cover more than one entity type, followed by Wikidata and DBpedia.

skos:closeMatch equivalences are much less frequent than the exact equivalences and only two KBs use them: BnF and ULAN. They represent only 1.5% of the total amount of equivalences stated by BnF. ULAN applies *skos:closeMatch* more frequently, reaching nearly 50% of the equivalences published. Overall, 192,300 statements use the *skos:closeMatch* predicate, which represents only 0.3% of all the equivalences stated by the studied KBs (Table 3).

Table 2. The amounts of equivalence statements involving each knowledge base.

KB	As subject of equivalences		As object of equivalences		Total statements
	Statements	to KBs	Statements	from KBs	
VIAF	25,118,745	7	21,666,779	6	46,785,524
GND	11,313,935	4	9,454,213	5	20,768,148
NAF	6,101,051	1	14,216,491	6	20,317,542
Wikidata	4,624,309	6	9,785,342	4	14,409,651
DBpedia	7,396,520	3	977,907	5	8,374,427
BnF	4,505,773	5	3,124,674	3	7,630,447
BNE	997,183	5	698,329	2	1695,512
ULAN	249,812	2	383,593	2	633,405

Table 3. The amounts of *skos:closeMatch* statements involving each knowledge base.

KB	As subject of equivalences		As object of equivalences	
	Statements	to KBs	Statements	from KBs
GND			25,952	1
NAF			150,224	2
Wikidata			6	1
BnF	67,746	3		
BNE			16,118	1
ULAN	124,554	2		

4.2 Crawling of the Equivalences for *agent* URIs

The results of the crawling iterations on the URIs of Europeana are shown in Table 4. After the 1st iteration (i.e., crawling beginning from the URIs in the Europeana set alone) we found 50,112 equivalent URIs. The amount of gathered equivalences has increased steeply in the first 3 crawling iterations. From the 1st crawl to the 2nd, the number of equivalences increased by 588%, and it increased by 42% on the 3rd iteration. The number of newly acquired equivalences was 0.76% in the 4th iteration, and under 0.1% in the 5th, so we opted to analyse and report on the results up to the 4th iteration (included). Only 3 iterations were needed to collect 99% of the equivalences. Although not all KBs are directly connected by equivalences, this shows that equivalent *agent* instances are closely connected in the equivalence graph.

VIAF was the KB with the highest number of equivalent URIs found. After the 4th crawling iteration, 60.7% of the set had equivalent VIAF URIs. Wikidata had the 2nd highest number of equivalences, reaching 34.5%.

For 3 KBs, less than 10% of the set had equivalences: ULAN, BNE and GND. The lower result for ULAN was expected since it is focused on artists. GND was the KB with the most URIs in the Europeana set, therefore, this result can be explained by the fact that for all GND URIs in the set, only equivalences to other KBs could be found. The results of BNE may be also explained by its high presence in the Europeana set.

For researchers and practitioners designing innovative systems based on *agent* linked data, the choice for using one or more KBs will always be highly influenced by the specific domain of application. Nevertheless, the results of the study indicate VIAF as the most linked KB, and therefore, in future work we would like to further exploit its data and equivalences.

Table 4. The results of the 4 crawling iterations of the Europeana set of *agent* URIs.

KB	Initial Europeana set (a)	New equivalences found after each iteration (b)				% of the initial Europeana set with equivalences (c)
		1 st crawl	2 nd crawl	3 rd crawl	4 th crawl	
DBpedia	2	4,407	34,968	47,031	47,410	16.57%
BnF	2,010	6,282	9,803	53,280	54,554	19.07%
BNE	30,449	3,321	9,952	12,471	12,934	4.52%
NAF	0	11,935	15,554	77,702	78,207	27.34%
ULAN	7,451	1,737	3,439	12,137	12,701	4.44%
GND	242,297	7,684	8,596	14,939	15,100	5.28%
VIAF	2,174	13,095	170,057	173,608	173,613	60.68%
Wikidata	0	1,651	92,588	98,450	98,813	34.54%
Total	284,383	50,112	344,957	489,618	493,332	–
Δ from previous crawl	–	–	588%	42%	0.76%	–

a - number of URIs of each KB in the Europeana set

b - number of equivalences found after each iteration

c - percentage of the Europeana URIs that after the 4th iteration have an equivalence to the KB considered.

4.3 Compliance with Semantic Web Standards

One of our initial observations during the study was that Wikidata is the only KB which does not use the standard equivalence properties. In fact, in an earlier study on Wikidata's data about CH resources [12], we have observed that it uses a very limited number of the standard Semantic Web “meta-modeling” properties. During the current study, we observed that *owl:sameAs* is in use only for internal equivalences between Wikidata's entities. None of *skos:exactMatch*, *skos:closeMatch* nor *schema:sameAs* are used.

Instead, Wikidata uses its own *wdt:P2888* (exact match), and a set of properties categorized as *External identifiers*¹². Each of these External identifier properties represents the local identifier for a Wikidata resource within the external information space of a particular institution or dataset. The values of statements with these properties are usually not URIs, and when a local identifier can be transformed into a URI, the definition of the property contains the formatting string for deriving the URI from the local identifier¹³. We have identified 159 properties of type *External Identifier* from which a URI could be derived.

We collected Wikidata's equivalences via its SPARQL endpoint, therefore we adapted our SPARQL queries to use the corresponding Wikidata properties. Another adaptation was done in the tools for analysis of the equivalence graph, so that the Wikidata properties would be considered as exact equivalences.

4.4 Data Quality of the Equivalence Statements

Our study did not have the objective to address the quality of equivalence statements, but we did come across a problem that blocked our crawling experiment, forcing us to find a solution. This problem was caused by four URIs used in 77,379 equivalence statements by VIAF, which seem plainly wrong¹⁴. Besides establishing wrong equivalences, this problem posed difficulties for crawling the equivalence graph. It would take several (probably many) additional iterations for the number of equivalent URIs to stabilize, and very large groups of equivalent URIs would be formed. To bypass the problem, we tried to filter out such incorrect URIs by detecting major outliers in terms of the mean of equivalences/URI. The mean of equivalences/URI in VIAF was of 1.006 and each of these four URIs were present in thousands of equivalence statements. The outlier URIs were discarded when we repeated the crawling process, therefore they were excluded from our study.

5 Conclusion and Future Work

The results obtained in our study confirm that the *agents* in KBs are highly interlinked. This high level of interlinking is in accordance with earlier studies of *owl:sameAs* general usage [3, 11] and the reports from the publishers of the CH KBs on the work they have carried out¹⁵. The study highlights also that the majority of equivalences are expressed with exact equivalence predicates (like *owl:sameAs*), while matches with uncertainty (*skos:closeMatch*) are a minority of 0.3%.

¹² The list of Wikidata properties for external identifiers is available at <https://www.wikidata.org/wiki/Special:ListProperties/external-id>.

¹³ The properties will contain an attribute *wdt:P1921* (formatter URI for RDF resource).

¹⁴ These 4 URIs are: <http://data.bnf.fr/#foaf:Person>; <http://data.bnf.fr/#foaf:Organization>; <http://data.bnf.fr/#owl:Thing>; and <http://data.bnf.fr/#spatialThing>. None correspond to an actual *agent* at BnF. We have mailed VIAF maintainers about it.

¹⁵ For space reasons we cannot refer to all presentations and articles here. Some of them are accessible on the online documentation for the KB considered, given as earlier references.

Although each KB is not directly linked to all other KBs, all KBs are a source and a target of equivalence links. Crawling of the *agent* URIs used in Europeana shows that only a few crawling iterations of the equivalence graph are needed to acquire a nearly complete set of equivalences from all KBs. Three iterations were enough to collect 99% of the equivalences gathered after five iterations.

VIAF is the KB with the highest number of *agent* equivalences, followed by Wikidata. An equivalent VIAF URI was found for 60.7% of Europeana's *agent* URIs, and for Wikidata, equivalences were found in 34.5% of Europeana's *agent* URIs.

Future work includes the detection of possibly incorrect equivalences, since this study, like earlier research [4], has detected some quality issues in the (*owl:sameAs*) links. Conversely, it would be interesting to estimate recall issues, i.e. whether many new links could (and should) be created across KBs via automatic or manual alignment.

Acknowledgments. This work was partly supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020 and by the European Commission under contract number 30-CE-0885387/00-80.

References

1. Charles, V., Manguinhas, H., Isaac, A., Freire, N., Gordea, S.: Designing a multilingual knowledge graph as a service for cultural heritage – some challenges and solutions. In: International Conference on Dublin Core and Metadata Applications, 2018 (2018)
2. Beek, W., Raad, J., Wielemaker, J., van Harmelen, F.: *sameAs.cc*: the closure of 500 M *owl:sameAs* statements. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 65–80. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_5
3. Correndo, G., Penta, A., Gibbins, N., Shadbolt, N.: Statistical analysis of the network for aligning concepts in the linking open data cloud. In: Liddle, S.W., Schewe, K.-D., Tjoa, A.M., Zhou, X. (eds.) DEXA 2012. LNCS, vol. 7447, pp. 215–230. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32597-7_20
4. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When *owl:sameAs* isn't the same: an analysis of identity in linked data. In: Patel-Schneider, P.F., et al. (eds.) ISWC 2010. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17746-0_20
5. Ding, L., Shinavier, J., Shangguan, Z., McGuinness, Deborah L.: *SameAs* networks and beyond: analyzing deployment status and implications of *owl:sameAs* in linked data. In: Patel-Schneider, P.F., et al. (eds.) ISWC 2010. LNCS, vol. 6496, pp. 145–160. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17746-0_10
6. Papaleo, L., Pernelle, N., Saïs, F., Dumont, C.: Logical detection of invalid *SameAs* statements in RDF data. In: Janowicz, K., Schlobach, S., Lambrix, P., Hyvönen, E. (eds.) EKAW 2014. LNCS (LNAI), vol. 8876, pp. 373–384. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13704-9_29
7. Beek, W., Rietveld, L., Schlobach, S., van Harmelen, F.: LOD Laundromat: why the semantic web needs centralization (even if we don't like it). *IEEE Internet Comput.* **20**(2), 78–81 (2016)
8. Fernández, J.D., Beek, W., Martínez-Prieto, M.A., Arias, M.: LOD-a-lot. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10588, pp. 75–83. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68204-4_7

9. Rietveld, L.: Publishing and consuming linked data: optimizing for the unknown. In: *Studies on the Semantic Web*, vol. 21. IOS Press (2016)
10. Radulovic, F., Mihindikulasooriya, N., García-Castro, R., Gomez-Pérez, A.: A comprehensive quality model for linked data. *Seman. Web* **9**(1), 3–24 (2018)
11. Asprino, L., Beek, W., Ciancarini, P., van Harmelen, F., Presutti, V.: Observing LOD using equivalent set graphs: it is mostly flat and sparsely linked. In: Ghidini, C., et al. (eds.) *ISWC 2019*. LNCS, vol. 11778, pp. 57–74. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30793-6_4
12. Freire, N., Isaac, A.: Technical usability of Wikidata’s linked data: evaluation of machine interoperability and data interpretability. In: Abramowicz, W., Paschke, A. (eds.) *Lecture Notes in Business Information Processing*. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-36691-9_47