On the Analysis of Large Integrated Knowledge Graphs for Economics, Banking, and Finance

Shuai Wang[1]

[1]*Department of Computer Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, the Netherlands*

Abstract

Knowledge graphs are being used for the detection of money laundering, insurance fraud, and other suspicious activities. Some recent work demonstrated how knowledge graphs are being used to study the impact of the COVID-19 outbreak on the economy. The fact that knowledge graphs are being used in more and more interdisciplinary problems calls for a reliable source of interdisciplinary knowledge. In this paper, we study the integration of knowledge graphs in the domains of economics, banking, and finance. Our integrated knowledge graph has over 610K nodes and 1.7 million edges. By performing statistical and graph-theoretical analysis, we demonstrate how the integration results in more entities with richer information. Its quality was examined by analyzing the subgraphs of the identity links and (pseudo-)transitive relations. Finally, we study the sources of error, and their refinement and discuss the benefit of our integrated graph.

Keywords

Integrated knowledge graphs, knowledge graph analysis, knowledge graph refinement

1. Introduction

The 2008 financial crisis urged early detection of systemic risk to national and world economies in derivatives markets. The relative size of these markets is a fundamental risk to geopolitical as well as economic security [1]. One of the trendy tools that can be used for the modelling of relations between companies and their economic behavior is knowledge graph. Knowledge graphs show great potential in use as they can represent companies structured in complex shareholdings, as well as information about investment, acquisition, bankruptcy, etc. Shao et al. used knowledge graphs of real financial data where nodes are customer, merchant, building, etc. The edges can be transactions between customers, residential information about customers, etc. As a benefit of the graphical structure, their knowledge graph captures interrelations and interactions across tremendous types of entities more effectively than traditional methods. They performed extensive experiments and demonstrated the usage of knowledge graphs in the consumer banking sector [2]. Bellomarini et al. address the impact of the COVID-19 outbreak on the network of Italian companies using knowledge graphs of millions of nodes [3]. Such projects require multiple types of domain knowledge, from company ownership to public health policy, from bankruptcy to social resilience. The essence of such knowledge becomes clear for strategy formation and policy making based on the dynamics of complex inter-connected systems. Unfortunately, many sources of knowledge were developed independently of each other. Fusing these independent KGs could lead to a significantly richer source of knowledge which could improve the performance of existing applications. In this paper, we study properties of the integration of knowledge graphs by analyzing the statistical and graph-theoretical properties. More specifically, we study properties of integrated knowledge graphs by combining existing knowledge graphs in the domains of economics, banking, and finance.

Finance The Financial Industry Business Ontology (FIBO) [4] includes formal models that are intended to define unambiguous shared meaning for financial industry concepts. Another popular ontology is the Financial Regulation Ontology (FRO), which has been used as a higher level, core ontology for ontologies such as the Insurance Regulation Ontology[1] (IRO), the Fund Ontology[2], etc.

Economics The STW (Standard Thesaurus Wirtschaft) Thesaurus for Economics was developed by the German National Library of Economics (ZBW) and gained popularity in scientific institutes, libraries and documentation centers, as well as business information providers. The JEL classification system was initially developed for use in the Journal of Economic Literature (JEL) [5] and is now a standard method of classifying scholarly literature in the field of economics.

Banking Knowledge graphs have attracted increasing attention in the banking industry over the past decade. The WBG Taxonomy[3] includes 3,882 concepts. It serves as a small classification schema which represents the concepts used to describe the World Bank Group's topical

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK

shuai.wang@vu.nl (S. Wang)
https://shuai.ai (S. Wang)
0000-0002-1261-9930 (S. Wang)

[1] https://insuranceontology.com/
[2] https://fundontology.com/
[3] https://vocabulary.worldbank.org/PoolParty/wiki/taxonomy

knowledge domains and areas of expertise, providing an enterprise-wide, application-independent framework. In comparison, the Bank Regulation Ontology (BRO) is much bigger and uses two industrial standards, namely FIBO and LKIF [6], as its upper ontology. It was built on top of the FRO ontology, as mentioned above. Unfortunately, many knowledge graphs are developed by banks and are not open source.

In this paper we study properties of integrated knowledge graphs in the domain of economics, banking and finance. Our results show that even though the integrated knowledge graph has some errors which have been created due to minor mistakes, the overall usefulness has been improved. Our contributions are:

a) We integrate some knowledge graphs in the domain of economics, banking, and finance and present the integrated knowledge graph consisting of over 610K entities and 1.7 million triples[4].

b) We study how the integration can enrich the information of entities with some statistical and graph-theoretical analysis.

c) We discuss the source of error and its refinement of the integrated knowledge graph for future use.

The paper is organised as follows: Section 2 presents the knowledge graphs and their statistics. Section 3 presents details of the integrated knowledge graph with an analysis of the source of error, followed by a discussion. Finally, we draw the conclusion in Section 4.

2. Integrating Knowledge Graphs

A *knowledge graph* $G = \langle V, E, L, l\rangle$ is a directed and labelled graph, where V is the set of nodes, $E \subseteq V \times V$ the set of edges, and L is the set of edge labels. A function $l : E \rightarrow 2^L$ assigns to each edge a set of labels from L. The nodes V can be IRIs, literals, or blank nodes. The edges E are relations between nodes and their types in the form of triples. Ontologies are semantic models of data that define the entities, their properties and types, types and subtyping, as well as relations between entities. An ontology can be represented as a knowledge graph.

An integrated knowledge graph $\mathbf{G} = \langle \mathbf{V}, \mathbf{E}, \mathbf{L}, \mathbf{l}\rangle$ is a combination of a set of N knowledge graphs $\{G_1, \ldots, G_N\}$ where $\mathbf{V} = V_1 \cup \ldots \cup V_N$, $\mathbf{E} = E_1 \cup \ldots \cup E_N$, and $\mathbf{L} = L_1 \cup \ldots \cup L_N$. A function $\mathbf{l} : \mathbf{E} \rightarrow 2^{\mathbf{l}}$ assigns to each edge a set of labels, which is the union of the labels: $\mathbf{l}(e) = l_1(e) \cup \ldots \cup l_N(e)$. For a given set relations $\mathbf{R}$, the subgraph is the graph $\mathbf{G_R}$ with $\mathbf{L} = \mathbf{R}$. When $\mathbf{R} = \{r\}$, $\mathbf{G_R} = \mathbf{G}_r$. Often times, such an integration requires the process of determining correspondences between concepts in ontologies. Such a process is called ontology alignment and the set of correspondences is called a mapping or an alignment.

By integrating knowledge graphs of various domains, we expect more entities and richer information for entities. The following is a list of 11 knowledge graphs we collected from 9 projects in the domains of economics, banking, and finance.

1. the Financial Industry Business Ontology (we collected the FIBO ontology using OWL and FIBO vocabulary using SKOS)[5]
2. the Financial Regulation Ontology (FRO)[6]
3. the Hedge Fund Regulation (HFR) ontology[7]
4. the Legal Knowledge Interchange Format (LKIF) ontology[8]
5. the Bank Regulation Ontology (BRO)[9]
6. the Financial Instrument Global Identifier (FIGI)[10]
7. the STW Thesaurus for Economics (and its mappings)[11]
8. the Journal of Economic Literature (JEL) classification system[12]
9. the Fund Ontology[13]

Not all knowledge graphs are available: some are not open source (e.g., the Italian Ownership Graph [3]), some others are commercial (e.g., the enterprise knowledge graphs by Agnos.ai[14]) and a few are not maintained anymore (e.g., the OntoBacen project [7]).

We used LogMap[15] for the alignment between knowledge graphs [8]. LogMap is a highly scalable ontology matching system with 'built-in' reasoning and inconsistency repair capabilities. It can efficiently match semantically rich ontologies containing tens (and even hundreds) of thousands of classes. Considering the size of our files,

[4] The data and Python scripts are available at https://github.com/shuaiwangvu/EcoFin-integrated.

[5] The product version retrieved from https://edmconnect.edmcouncil.org/fibointerestgroup/fibo-products/fibo-owl (147 files in Turtle format) and https://edmconnect.edmcouncil.org/fibointerestgroup/fibo-products/fibo-voc (1 file in Turtle format) respectively on 14th January, 2022.

[6] 32 Turtle files were retrieved from https://finregont.com/ontology-directory-files-prefixes/ on 14th Janurary, 2022.

[7] 12 Turtle files were retrieved from https://hedgefundontology.com/ontology-files/ on 14th January, 2022

[8] Retrieved from http://www.estrellaproject.org/lkif-core/#download on 30th January, 2022.

[9] 16 Turtle files were retrieved from https://bankontology.com/ontology-directory-files-prefixes/ on 30th January, 2022.

[10] 4 RDF files were retrieved from https://www.omg.org/spec/FIGI/ on 22nd December, 2021.

[11] The paper used STW v9.12 based on the SKOS ontology. The ontology and its 9 mappings files were retrieved from https://zbw.eu/stw/version/latest/download/about.en.html on 30th Janurary, 2022.

[12] The Turtle file was retrieved from https://zbw.eu/beta/external_identifiers/jel/about on 30th January, 2021.

[13] The paper used 8 Turtle files retrieved from https://fundontology.com/ontology-files/ on 28th December, 2021.

[14] https://agnos.ai/services

[15] http://krrwebtools.cs.ox.ac.uk/logmap/

Table 1
Alignment of knowledge graphs

	FIBO-vD	FIBO-OWL	LKIF	FIGI	STW	JEL	Fund
FIBO-vD	-	599	1	147	12	204	11
FIBO-OWL	-	-	24	516	5	57	70
LKIF	-	-	-	1	0	0	23
FIGI	-	-	-	-	0	34	2
STW	-	-	-	-	-	2	0
JEL	-	-	-	-	-	-	1
Fund	-	-	-	-	-	-	-

Table 2
General statistics of knowledge graphs

Name	|V|	|E|	Size
FIBO-vD	17,547	28,128	3.1MB
FIBO-OWL	103,288	250,002	16MB
FRO	94,215	283,976	16MB
HFR	14,235	34,771	2.6MB
LKIF	1,005	2,363	141KB
BRO	259,074	838,007	43MB
FIGI	12,180	16,434	822KB
STW	51,128	113,276	3.4MB
JEL	12,109	177,57	1.1MB
Fund	10,119	35,005	3.2MB
STW-mappings	78,398	177,603	11MB
alignment	2,327	1,698	255KB
integrated	610,866	1,778,755	93MB

we used the version with mapping repair but not the aid of any reasoner. Unfortunately, FRO, BRO, and HFR failed to load due to parsing errors in some files they import. Table 1 summarizes the number of pairs of entities generated by LogMap. Overall, 1,698 unique identity links of `skos:exactMatch` were added to the integrated graph.

All the knowledge graphs were first converted to Turtle format and then used the RDFpro[16] [9] for the integration process with duplicated triples removed. RDFpro is an open source stream-oriented toolkit for the processing of RDF triples. We used RDFpro (version 0.6) without smushing. The integration took 23 seconds on a 2.2 GHz Quad-Core i7 laptop with a 16GB memory running Mac OS. All the files were then converted to their HDT format for further experiments. The integrated knowledge graph consists of 1,778,755 unique triples (edges) and 610,866 nodes. It has 93MB and 22MB in its Turtle and HDT format respectively. Table 2 summarize the statistics of the number of nodes, edges and the size of their Turtle files. For the sake of speed, when studying properties of these knowledge graphs, we use files in their HDT format.

3. Analysis of the Integrated knowledge graph

In this section, we first study how the information of entities can be enriched with some statistical analysis of graph structure (Section 3.1). We then examine identity links (e.g. `skos:exactMatch`) in the integrated graph **G** and their corresponding subgraphs (Section 3.2). Finally, we study transitive and pseudo-transitive relations such as concept generalisation (Section 3.3) followed by a discussion (Section 3.5).

3.1. Statistical analysis

We study how the information of entities can be enriched when combining different resources. When an entity is described in different domains, its in- and out-degree are expected to increase. Figure 1 illustrates the in-/out-degree of the knowledge graphs and the integrated knowledge graph. Both the in- and out-degrees of the integrated graph show a power-law distribution. Moreover, the figures show that the integration increases both the number of degrees in general and the number of nodes with high degrees, which demonstrates how this integration can enrich the information of entities. For example, `lkif-core-norm:allowed_by` has an out-degree of 7 in the integrated graph but the three graphs that contain information about it has out-degrees of 2, 5, and 1 respectively[17].

A strongly connected component (SCC) of a directed graph is a maximal subgraph where there is a path between all pairs of vertices. A weakly connected component (WCC) is a subgraph of the original graph where all vertices are connected to each other by some path, ignoring the direction of edges. Table 3 summarizes the graph-theoretical statistics. Let maxSCC and maxWCC represent the number of nodes in the largest strongly connected component and weakly connected component respectively. In addition, we compute the fraction of nodes in the biggest SCC and WCC, denoted p_S and p_W respectively. The high values of p_W in the table show that the graphs are mostly connected. More specifically, $p_W = 99.98\%$ for the integrated graph, which is due to the overlapping domains of the knowledge graphs and the mappings. The low values of p_S indicate that the underlying structure of these graphs is mostly hierarchical, especially that of JEL, BRO, and FIBO-vD.

3.2. Analysis of identity links

Identity links are relations between entities that are considered identical and intended to refer to the same

[16] http://rdfpro.fbk.eu/

[17] The prefix lkif-core-norm corresponds to the namespace http://www.estrellaproject.org/lkif-core/norm.owl#.

Table 3
Graph-theoretical statistics of knowledge graphs

Name	maxSCC	p_S(%)	maxWCC	p_W(%)
FIBO-vD	1	0.01	17,535	99.93
FIBO-OWL	297	0.29	103,208	100
FRO	17	0.02	94,015	99.79
HFR	849	5.96	14,230	99.96
LKIF	88	8.76	963	95.82
BRO	13	0.01	258,982	99.96
FIGI	13	0.11	12,180	100
STW	6777	13.25	51,128	100
JEL	1	0.01	12,099	99.92
Fund	109	1.08	10,111	99.92
STW-mappings	617	0.79	78,398	100
alignment	3	0.13	119	5.11
integrated	36,853	6.03	610,792	99.98

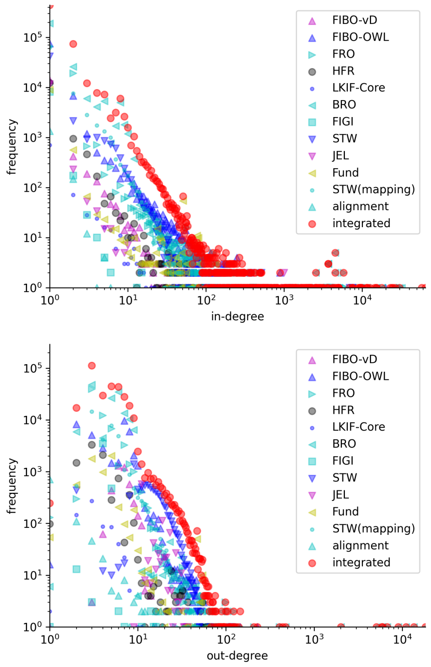

Figure 1: Distribution of in-/out-degree of nodes in knowledge graphs

real-world entities. Typical identity links use relations such as `owl:sameAs` and `skos:exactMatch`. We first study identity links in **G** and their corresponding subgraphs. In contrast to the statistics reported by Raad et al., where `owl:sameAs` is much more popular than `skos:exactMatch` [10], our analysis shows that only 5,253 triples about `owl:sameAs` are in **G** against 31,254 triples about `skos:exactMatch`. In addition, there are 8,172 triples about `skos:relatedMatch`, and 6,418 triples about `skos:closeMatch`. Figure 2 shows the frequency distribution of the weakly connected components in their corresponding subgraphs.

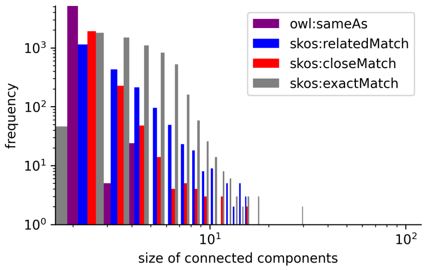

Figure 2: Frequency distribution of connected components in the integrated graph

The largest two connected components of the subgraph of `owl:sameAs` are with 8 and 6 entities each. In contrast, the largest two connected components of `skos:exactMatch` are much bigger, with 119 and 45 entities respectively. For `skos:relatedMatch`, the largest weakly connected component consists of 21 entities. That of `skos:closeMatch` consists of 52 entities. A manual examination below shows that there are errors in these large connected components. The mis-use of these SKOS mapping properties can have less implications than the `owl:sameAs` since `skos:exactMatch` indicates only “a high degree of confidence that the concepts can be used interchangeably across a wide range of applications”[10]. Moreover, `lkif-core:mereology.owl#strictly_equivalent` is a equivalence relation but corresponds to no triple[18]. More discussion is included in Section 3.4.

3.3. Analysis of transitive and pseudo-transitive relations

Transitive relations are widely used in knowledge graphs on the definition of class subsumption, concept generalisation, organisation composition, etc. Due to transitivity, entities in cycles imply some equivalence relation, which could be erroneous. Take `lkif-core:component_of` for example. A triple specifies that “some thing is a (functional) component of some other thing”. Entities in a cycle of `lkif-core:component_of` indicate that all they are components of each other, which could be erroneous. Some past work showed how strongly connected components can be used to locate errors when refining knowledge graphs [11, 12].

[18] The prefix `lkif-core` corresponds to the namespace http://www.estrellaproject.org/lkif-core/.

There are in total 20 relations typed owl:TransitiveProperty in **G**. We also study the pseudo-transitive relations: those relations that are not typed owl:TransitiveProperty but shows transitivity in their intended semantics [11]. In this study, we focus on two pairs of such relations: skos:broader and its inverse skos:narrower, skos:broaderMatch as well as its inverse relation skos:narrowerMatch. This section excludes relations of identity links such as skos:exactMatch, which was discussed in Section 3.2.

Take skos:broadMatch for example. A manual analysis of the largest three SCC shows the edges could be erroneous. These SCCs are: a component with four entities about plebiscite, referendum, and popular initiative; a component with three entities about insurance and private insurance; a component with three distinct entities about the CARICOM countries, Caribbean countries, and the Caribbean Community.

Let $\mathbf{G_B}$ be the subgraph of the integrated graph **G** with **B** = {skos:broader, skos:broaderMatch} and $\mathbf{G_N}$ for **N** = {skos:narrower, skos:narrowerMatch}. Next, we combine the $\mathbf{G_B}$ with the graph $\mathbf{G'_N}$, where $\mathbf{G'_N}$ is a graph with each edge of **G** reversed in direction. After performing the same analysis, we discover a new strongly connected component with four entities about adjustable peg, fixed exchange rate, exchange rate regime and internationales Währungssystem, respectively. Moreover, the resulting graph has 44 connected components of two entities, which are more than that of the subgraphs corresponding to each individual relation. This indicates that such integration can result in more complex errors which do not exhibit in stand-alone graphs.

Our analysis shows that rdfs:subClassOf is a popular relation with 47,597 triples. However, there is no SCC with more than one component, which implies that the underlying class hierarchy is a directed acyclic graph. In addition, lkif-core:component, fro:divides[19], and its inverse fro:divided_by are also popular transitive relations. Finally, none of them has strongly connected components of size greater or equal to two.

3.4. Source of Error and Refinement

When tracing back to the sources of each edge, we found that skos:broader and skos:narrower are mostly from three sources: STW, JEL, and FIBO-vD. When combined with the subgraph of skos:broadMatch and skos:narrowMatch, there are in total 44 SCCs of two entities, two SCCs of three entities, and two SCCs of four entities. It is feasible that some domain experts manually examine all these small SCCs without employing any refinement algorithm.

Our analysis also shows that the identity links come solely from two sources: the owl:sameAs triples are from the FIBO-OWL knowledge graph, the triples about skos:exactMatch, skos:closeMatch, and skos:relatedMatch are from STW-mappings and our alignment. Mapping files about the STW subject categories were created by the alignment tool Amalgame[20]. Our manual examination shows that these identity links are closely related concepts and requires knowledge from experts for refinement.

3.5. Discussion

As shown above, this integration results in new statistical and graph-theoretical properties. Next, we compare how these problems exhibit in our graph and the LOD-a-lot[21] [13]. LOD-a-lot is a dataset that integrates over 28 billion triples from 650K files of the LOD Cloud into a single ready-to-consume file. While our integrated knowledge graph has 1.7 million unique triples, LOD-a-lot is much larger with 28.3 billion triples. For LOD-a-lot, 356.9K edges out of 11.8 million edges of skos:broader are involved in SCCs [11]. In contrast, we have no SCC with two or more entities among 17,868 edges of skos:broader. For LOD-a-lot, 1.4K edges out of 4.4 million edges of rdfs:subClassOf are involved in SCCs [11, 12]. In contrast, there is no cycle for our corresponding subgraph. This confirms the quality of the knowledge graphs we used. The identity graph of the LOD-a-lot graph regarding owl:sameAs consists of 558.9 million triples with the largest connected component consisting of 177.8K entities [10]. In contrast, our identity graphs of both owl:sameAs and skos:exactMatch are small and can be manually refined.

4. Conclusion

In this paper, we presented an integrated knowledge graph in the domain of Economics, Finance, and Banking. We demonstrated how the integrated graph has more entities with richer information. We discussed subgraphs of (pseudo-)transitive and identity relations as well as their refinement. The overall usefulness has been improved despite minor errors introduced due to integration.

Our integrated knowledge graph can be used to evaluate data interoperability. Also, it can enrich the features of entities, which may increase the accuracy of pattern recognition using Machine Learning for the detection of takeovers, money laundering, insurance fraud, counterfeiting, etc. Furthermore, it can also be used to improve the quality of suspicious activity reports, recommendation systems, conversational agents, etc.

[19] The prefix fro corresponds to the namespace http://finregont.com/fro/ref/LegalReference.ttl#.

[20] https://github.com/jrvosse/amalgame

[21] http://lod-a-lot.lod.labs.vu.nl/

References

[1] S. Malik, The ontology of finance: Price, power, and the arkhederivative, in: Collapse Vol. VIII: casino real, Falmouth: Urbanomic, 2014, pp. 629–811.

[2] D. Shao, R. Annam, Translation embeddings for knowledge graph completion in consumer banking sector, in: A. El Fallah Seghrouchni, D. Sarne (Eds.), Artificial Intelligence. IJCAI 2019 International Workshops, Springer International Publishing, Cham, 2020, pp. 5–17.

[3] L. Bellomarini, M. Benedetti, A. Gentili, R. Laurendi, D. Magnanimi, A. Muci, E. Sallinger, COVID-19 and company knowledge graphs: Assessing golden powers and economic impact of selective lockdown via AI reasoning, CoRR abs/2004.10119 (2020). URL: https://arxiv.org/abs/2004.10119. arXiv:2004.10119.

[4] M. Bennett, The financial industry business ontology: Best practice for big data, Journal of Banking Regulation 14 (2013) 255–268.

[5] B. Cherrier, Classifying economics: A history of the jel codes, Journal of economic literature 55 (2017) 545–79.

[6] R. Hoekstra, J. Breuker, M. Di Bello, A. Boer, The lkif core ontology of basic legal concepts, 2007, pp. 43–63.

[7] F. F. Polizel, S. J. Casare, J. S. Sichman, Ontobacen: A modular ontology for risk management in the brazilian financial system, in: Proceedings of the Joint Ontology Workshops, 2015.

[8] E. Jiménez-Ruiz, B. Cuenca Grau, Logmap: Logic-based and scalable ontology matching, in: L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, E. Blomqvist (Eds.), The Semantic Web – ISWC 2011, Springer Berlin Heidelberg, 2011, pp. 273–288.

[9] F. Corcoglioniti, M. Rospocher, M. Mostarda, M. Amadori, Processing billions of RDF triples on a single machine using streaming and sorting, in: R. L. Wainwright, J. M. Corchado, A. Bechini, J. Hong (Eds.), Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, April 13-17, 2015, ACM, 2015, pp. 368–375. URL: https://doi.org/10.1145/2695664.2695720. doi:10.1145/2695664.2695720.

[10] J. Raad, N. Pernelle, F. Saïs, W. Beek, F. van Harmelen, The sameas problem: A survey on identity management in the web of data, CoRR abs/1907.10528 (2019). URL: http://arxiv.org/abs/1907.10528. arXiv:1907.10528.

[11] S. Wang, J. Raad, P. Bloem, F. Van Harmelen, Refining transitive and pseudo-transitive relations at web scale, in: European Semantic Web Conference, Springer, 2021, pp. 249–264.

[12] S. Wang, J. Raad, P. Bloem, F. van Harmelen, Submassive: Resolving subclass cycles in very large knowledge graphs, in: Workshop on Large Scale RDF Analytics, 2020.

[13] W. Beek, J. D. Fernández, R. Verborgh, Lod-a-lot: A single-file enabler for data science, Association for Computing Machinery, New York, NY, USA, 2017.