Discovering Implicational Knowledge in Wikidata

Tom Hanika[1,2], Maximilian Marx[3], and Gerd Stumme[1,2]
0000-0002-4918-6374 0000-0003-1479-0341 0000-0002-0570-7908

[1] Knowledge & Data Engineering Group, University of Kassel, Germany
[2] ITeG, University of Kassel, Germany
[3] Knowledge-Based Systems Group, TU Dresden, Germany
tom.hanika@cs.uni-kassel.de, maximilian.marx@tu-dresden.de,
stumme@cs.uni-kassel.de

Abstract. Knowledge graphs have recently become the state-of-the-art tool for representing the diverse and complex knowledge of the world. Among the freely available knowledge graphs, Wikidata stands out by being collaboratively edited and curated. Among the vast numbers of facts, complex knowledge is just waiting to be discovered, but the sheer size of Wikidata makes this infeasible for human editors. We apply Formal Concept Analysis to efficiently identify and succinctly represent comprehensible implications that are implicitly present in the data. As a first step, we describe a systematic process to extract conceptual knowledge from Wikidata's complex data model, thus providing a method for obtaining large real-world data sets for FCA. We conduct experiments that show the principal feasibility of the approach, yet also illuminate some of the limitations, and give examples of interesting knowledge discovered.

Keywords: Wikidata, FCA, Property Dependencies, Implications

1 Introduction

The quest for the best digital structure to collect and curate knowledge has been going on since the first appearances of knowledge stores in the form of semantic networks and databases. The most recent, and arguably so far most powerful, incarnation is the *knowledge graph*, as used by corporations like Facebook, Google, Microsoft, IBM, and eBay. Among the freely available knowledge graphs, Wikidata [19, 20] stands out due to its free and collaborative character: like Wikipedia, it is maintained by a community of volunteers, adding *items*, relating them using *properties* and *values*, and backing up claims with *references*. As of 2019-02-01, Wikidata has 52,373,284 items and 676,854,559 statements using a total 5,592 properties. Altogether this constitutes a gargantuan collection of factual data accessible to and freely usable by everyone.

But Wikidata is more than just the collection of all factual knowledge stored within: Wikidata also contains *implicit* knowledge that is not explicitly stated,

Authors are given in alphabetical order. No priority in authorship is implied. We kindly remind the reader that a more detailed version of this paper [9] is available.

but rather holds merely due to certain statements being present (or absent, respectively). While such implicit knowledge can take many shapes, we focus on "rules" (propositional implications) stating that whenever an entity has statements for some given properties, it should also have statements for certain other properties. We believe that such rules can guide editors in enriching Wikidata and discuss potential uses in the full version of the paper. [9]

Previous approaches have studied extracting rules in the form of implications of first-order logic (FOL) as a feasible approach to obtain interesting and relevant rules from Wikidata [5, 10], but the expressive power of FOL comes with a steep price: to understand such rules, one needs to understand not only the syntax, but also advanced concepts such as quantification over variables, and it seems far-fetched to assume that the average Wikidata editor possesses such understanding. We thus propose to use rules that are conceptually and structurally simpler, and focus on extracting Horn implications of *propositional logic* (PL) from Wikidata, trading expressive power for ease of understanding and simplicity of presentation.

While Formal Concept Analysis (FCA) [7] provides techniques to extract a sound and complete basis of PL implications (from which all other implications can be inferred), applying these techniques to Wikidata is not straightforward: A first hurdle is the sheer size of Wikidata, necessitating the selection of subsets from which to extract rules. Secondly, the intricate data model of Wikidata, while providing much flexibility for expressing wildly different kinds of statements, is not particularly amenable to a uniform approach to extracting relevant information.

In this work, we tackle both issues by describing procedures i) for extracting, in a structured fashion, implicational knowledge for arbitrary subsets of properties, and ii) for deriving suitable sets of attributes from Wikidata statements, depending on the type of property. We provide an implementation of these procedures[4], and while incorporating the extracted rules into the editing process is out of scope for this paper, we nevertheless demonstrate that we are able to obtain meaningful and interesting rules using our approach.

2 Related Work

FCA has been applied to Wikidata in [8] to model and predict the dynamic behaviour of knowledge graphs using lattice structures, and in [12] to determine obligatory attributes for classes. Another related topic is rule mining, and several successful approaches to generating lists of FOL rules, e.g., in [10, 5] have been proposed. This task is often connected to ranked lists of rules, like in [21], or to completeness investigations for knowledge graphs, as in [6, 18]. Rule mining using FCA has been proposed for RDF graphs [1], but the extensive use of reification in the Wikidata RDF exports prohibits such an approach. Rudolph [16] describes a general method for deriving attributes from properties of relational structures, where the property can be expressed by a concept description in the description logic $\mathcal{FLE}$. We note that there is no such concept description capturing Problem 2.

[4] https://github.com/mmarx/wikidata-fca

3 Wikidata

Data Model. Wikidata [20] is the free and open Knowledge Graph of the Wikimedia foundation. In Wikidata, *statements* representing knowledge are made using *properties* that connect *entities* (either *items* or other properties) to *values*, which, depending on the property, can be either items, properties, *data values* of one of a few data types, e.g., URIs, time points, globe coordinates, or textual data, or either of the two special values *unknown value* (i.e., *some* value exists, but it is not known) and *no value* (i.e., it is known that there is no such value).

Example 1. Liz Taylor was married to Richard Burton. This fact is represented by a connection from item Q34851 ("Elizabeth Taylor") to item Q151973 ("Richard Burton") using property P26 ("spouse"). But Taylor and Burton were married twice: once from 1964 to 1974, and then from 1983 to 1984.

To represent these facts, Wikidata enriches statements by adding *qualifiers*, pairs of properties and values, opting for two "spouse" statements from Taylor to Burton with different P580 ("start time") and P582 ("end time") qualifiers.

Metadata and Implicit Structure. Each statement carries metadata: *references* track provenance of statements, and the statement *rank* can be used to deal with conflicting or changing information. Besides *normal* rank, there are also *preferred* and *deprecated* statements. When information changes, the most relevant statement is marked preferred, e.g., there are numerous statements for P1082 ("population") of Q1794 ("Frankfurt"), giving the population count at different times using the P585 ("point in time") qualifier, with the most recent estimate being preferred. Deprecated statements are used for information that is no longer valid (as opposed to simply being outdated), e.g., when the formal definition of a planet was changed by the International Astronomical Union on 2006-09-13, the statement that Q339 ("Pluto") is a Q634 ("Planet") was marked deprecated, and an P582 ("end time") qualifier with that date was added.

Example 2. We may write down these two statements in *fact notation* as follows, where qualifiers and metadata such as the statement rank are written as an *annotation* on the statement:

$$\begin{split}\texttt{population}_{P1082}(\texttt{Frankfurt}_{Q1794}, 736414)@[\texttt{determination method}_{P459}\!:\\ \texttt{estimation}_{Q965330}, \texttt{point in time}_{P585}\!: 2016\text{-}12\text{-}31, \texttt{rank}\!: \texttt{preferred}]\end{split} \tag{1}$$

$$\begin{split}\texttt{instance of}_{P31}(\texttt{Pluto}_{Q339}, \texttt{Planet}_{Q634})@[\texttt{end time}_{P582}\!: 2006\text{-}09\text{-}13,\\ \texttt{rank}\!: \texttt{deprecated}]\end{split} \tag{2}$$

Further structure is given to the knowledge in Wikidata using statements themselves: Wikidata contains a class hierarchy comprising over 100,000 *classes*, realised by the properties P31 ("instance of") (stating that an item is an *instance* of a certain class) and P279 ("subclass of"), which states some item q is a *subclass* of some other class q', i.e., that all instances of q are also instances of q'.

Formalisation. Most models of graph-like structures do not fully capture the peculiarities of Wikidata's data model. The generalised Property Graphs [15], however, have been proposed specifically to capture Wikidata, and we thus phrase our formalisation in terms of a *multi-attributed relational structure.*[5]

Definition 1. *Let $\mathcal{Q}$ be the set of Wikidata items, $\mathcal{P}$ be the set of Wikidata properties, and let $\mathcal{V}$ be the set of all possible data values. We denote by $\mathcal{E} \coloneqq \mathcal{Q} \cup \mathcal{P}$ the set of all entities, and define $\Delta \coloneqq \mathcal{E} \cup \mathcal{V}$. The Wikidata knowledge graph is a map $\mathcal{W} \colon \mathcal{P} \to \mathfrak{P}(\mathcal{E} \times \Delta \times \mathfrak{P}(\mathcal{P} \times \Delta))$ assigning to each property p a ternary relation $\mathcal{W}(p)$, where a tuple $\langle s, v, a \rangle \in \mathcal{W}(p)$ corresponds to a p-statement on s with value v and annotation a.*

Thus, $\langle \Delta, (\mathcal{W}(p))_{p \in \mathcal{P}} \rangle$ is a *multi-attributed relational structure*, i.e., a relational structure in which every tuple is annotated with a set of pairs of attributes and annotation values. While technically stored separately on Wikidata, we will simply treat references and statement ranks as annotations on the statements. In the following, we refer to the Wikidata knowledge graph simply by $\mathcal{W}$. Furthermore, we assume that deprecated statements and the special values *unknown value* and *no value* do not occur in $\mathcal{W}$. This is done merely to avoid cluttering formulas by excluding these cases, and comes without loss of generality.

Example 3. Property P26 ("spouse") is used to model marriages in Wikidata. Among others, $\mathcal{W}(\texttt{spouse}_{P26})$ contains the two statements corresponding to the two marriages between Liz Taylor and Richard Burton from Example 1:

$$\begin{aligned} \Big\langle &\texttt{Elizabeth Taylor}_{Q34851}, \texttt{Richard Burton}_{Q151973}, \\ &\quad \big\{ \langle \texttt{start time}_{P580}, 1964 \rangle, \langle \texttt{end time}_{P582}, 1974 \rangle \big\} \Big\rangle \end{aligned} \tag{3}$$

$$\begin{aligned} \Big\langle &\texttt{Elizabeth Taylor}_{Q34851}, \texttt{Richard Burton}_{Q151973}, \\ &\quad \big\{ \langle \texttt{start time}_{P580}, 1983 \rangle, \langle \texttt{end time}_{P582}, 1984 \rangle \big\} \Big\rangle \end{aligned} \tag{4}$$

Next, we introduce some abbreviations for when we are not interested in the whole structure of the knowledge graph.

Definition 2. *Let $R \subseteq S^3$ be a ternary relation over S. For $t = \langle s, o, a \rangle \in S^3$, we denote by $\operatorname{subj} t \coloneqq s$ the* subject *of t, by $\operatorname{obj} t \coloneqq o$ the* object *of t, and by $\operatorname{ann} t \coloneqq a$ the* annotation *of t, respectively. These extend to R in the natural fashion: $\operatorname{subj} R \coloneqq \{\operatorname{subj} t \mid t \in R\}$, $\operatorname{obj} R \coloneqq \{\operatorname{obj} t \mid t \in R\}$, and $\operatorname{ann} R \coloneqq \{\operatorname{ann} t \mid t \in R\}$, respectively. We indicate with $\hat{}$ that a property is incident with an item as object: $\mathcal{W}(\hat{}\,\texttt{spouse}_{P26})$ contains $\langle \texttt{Richard Burton}_{Q151973}, \texttt{Elizabeth Taylor}_{Q34851}, \{\langle \texttt{start time}_{P580}, 1964 \rangle, \langle \texttt{end time}_{P582}, 1974 \rangle\} \rangle$.*

4 Formal Concept Analysis

We assume familiarity with the basic notions from FCA and kindly refer the reader to the full version of the paper [9], and to [7] for an introduction to FCA.

[5] This is merely a formalisation of Wikidata's actual data model (cf. https://mediawiki.org/wiki/Wikibase/DataModel), not a new model for conceptual data.

5 Property Theory

We now describe how to harness FCA to obtain a more accessible view of the Wikidata knowledge graph and how the properties therein depend on each other. Krötzsch [11] argues that knowledge graphs are primarily characterised by three properties: i) normalised storage of information in small units, ii) representation of knowledge through the connections between these units, and iii) enrichment of the data with contextual knowledge. In Wikidata, properties serve both as a mechanism to relate entities to one another, as well as to provide contextual information on statements through their use as qualifiers. Taking the structure and usage of properties into account is thus crucial to any attempt of extracting structured information from Wikidata. We now introduce the two most basic problem scenarios for selecting sets of properties from Wikidata. We outline several other approaches to exploiting different aspects of Wikidata's rich data model, but defer a detailed discussion to the full version of the paper. [9]

5.1 Plain Incidence

We start by constructing the formal context that has a chosen set $\hat{\mathcal{P}} \subseteq \mathcal{P}$ as its attribute set and the entity set $\mathcal{E}$ as the object set.

Problem 1. Given the Wikidata knowledge graph $\mathcal{W}$ and some subset $\hat{\mathcal{P}} \subseteq \mathcal{P}$, compute the canonical base for the implicational theory $Th_{\hat{\mathcal{P}}}(\mathcal{E}, \hat{\mathcal{P}}, I^{\texttt{plain}})$, where

$$\langle e, \hat{p} \rangle \in I^{\texttt{plain}} :\Longleftrightarrow\; e \in \operatorname{subj} \mathcal{W}(\hat{p}),\ \text{i.e.,} \tag{5}$$

an entity e coincides with property $\hat{p}$ iff it occurs as a subject in some $\hat{p}$-statement.

Although this is the most basic problem we present, with growing $\hat{\mathcal{P}}$ it may quickly become computationally infeasible, cf. Section 6. More importantly, however, entities occurring as objects are not taken into account: almost half of the data in the knowledge graph is ignored, motivating the next definition.

5.2 Directed Incidence

We endue the set of properties $\mathcal{P}$ with two colours $\{\text{subj}, \text{obj}\}$ signifying whether an entity coincides with the property as subject or as object in some statement.

Problem 2. Given $\mathcal{W}$ and some set $\hat{\mathcal{P}} \subseteq \mathcal{P} \times \{\text{subj}, \text{obj}\}$ of directed properties, compute the canonical base for $Th_{\hat{\mathcal{P}}}(\mathcal{E}, \hat{\mathcal{P}}, I^{\texttt{dir}})$, where an entity e coincides with $\hat{p}$ iff it occurs as subject or object (depending on the colour) of some p-statement:

$$\begin{aligned} \langle e, \hat{p} \rangle \in I^{\texttt{dir}} :\Longleftrightarrow &\; \big(\hat{p} = \langle p, \text{subj} \rangle \wedge e \in \operatorname{subj} \mathcal{W}(p)\big) \\ &\vee \big(\hat{p} = \langle p, \text{obj} \rangle \wedge e \in \operatorname{obj} \mathcal{W}(p)\big). \end{aligned} \tag{6}$$

Example 4. Let $\hat{\mathcal{P}} = \{\texttt{\^{}mother}_{\hat{}P25}, \texttt{godparent}_{P1290}, \texttt{mother}_{P25}\}$ be the set of attributes and let $\mathcal{E} = \{\texttt{Miley Cyrus}_{Q4235}, \texttt{Victoria}_{Q9439}, \texttt{Naomi Watts}_{Q132616}, \texttt{Angelina Jolie}_{Q13909}\}$ be the set of objects. The corresponding formal context $\langle \mathcal{E}, \hat{\mathcal{P}}, I^{\texttt{dir}} \rangle$ (as extracted from Wikidata) is given by the following cross table:

Example	^P25 (“^mother”)	P1290 (“godparent”)	P25 (“mother”)
Q13909 (“Angelina Jolie”)	×	×	×
Q4235 (“Miley Cyrus”)		×	×
Q132616 (“Naomi Watts”)	×		×
Q9439 (“Victoria”)	×	×	×

Observe that the only valid (non-trivial) implication (and hence sole constituent of the canonical base) is $\{\} \to \{\texttt{mother}_{P25}\}$: every entity has a mother.

Generalised Incidence. Even though Problem 2 can cope with Example 4, it still does not capture the subtleties of Example 3, as, e.g., two statements differing only in their annotations are indistinguishable, even though the meaning of, e.g., statements for P1038 (“relative”) can vary wildly among statements with different values for the P1039 (“type of kinship”) qualifier. Similarly, we might distinguish properties by the classes that objects of the property are instances of: having a P25 (“mother”) that is a Q22989102 (“Greek deity”) is significantly different from one that is merely a Q5 (“human”). In practice, we likely want to use different incidences for different properties, and we thus introduce generalised incidences. Again, we refer to the full paper [9] for a discussion of these approaches.

6 Experimental Results

We have computed canonical bases for Problems 1 and 2 on selected subsets of Wikidata, obtained by fixing a set of properties $\hat{\mathcal{P}}$ and restricting to statements involving them. We have selected the sets of properties by picking thematically related properties (as specified in Wikidata), see the full paper [9] for descriptions of methodology, data sets, and discovered rules. Already this preliminary evaluation shows several limitations: For the `family` datasets, we find the valid rule $\{\texttt{godparent}_{P1290}, \texttt{partner}_{P451}\} \to \{\texttt{sibling}_{P3373}\}$, stating that an entity with a godparent and a partner must also have a sibling, even though this need not be true in the real world; Wikidata is lacking a counterexample. Contrarily, the rule $\{\texttt{\^{}father}_{\hat{}P22}, \texttt{\^{}relative}_{\hat{}P1038}, \texttt{spouse}_{P26}\} \to \{\texttt{child}_{P40}\}$ witnesses that the more general $\{\texttt{\^{}father}_{\hat{}P22}\} \to \{\texttt{child}_{P40}\}$ has counterexamples: indeed there are 1,634 non-fictional humans contradicting it. Besides such (unavoidable) noise and lack of completeness in the underlying data, another problem is that of computational infeasibility: computing canonical bases on state-of-the-art hardware takes several hours for subsets of one hundred properties; for larger sets (much less all of Wikidata), computation time is prohibitively long. Still, we obtain a plenitude of interesting and meaningful rules about the subject domains. [9]

7 Conclusion and Outlook

We have shown how to extract, in a structured fashion, subsets of Wikidata and represent them as formal contexts. This provides a practically limitless source of real-world data to the FCA community, to which the full range of tools and techniques from FCA may be applied. Most importantly, we can obtain relevant, meaningful, and perspicuous implications valid for Wikidata, which may be useful in different ways: valid rules can highlight other entities that must be edited to avoid introducing counterexamples, and, conversely, absence of an expected rule may indicate the presence of unwanted counterexamples in the data.

Our approach is feasible for subsets of Wikidata that capture interesting domains of knowledge. For larger subsets, we propose to compute Luxenburger bases of association rules [13, 17] or PAC bases [2], both of which are feasible for Wikidata as a whole. Another approach would be to employ conceptual exploration to compute canonical bases for generalised incidences over Wikidata. The missing ingredient here is a method to query Wikidata for possible counterexamples to a proposed implication, e.g., via the SPARQL endpoint, enabling Wikidata to be used as an *expert* for the exploration. Ultimately, Wikidata could be employed alongside human experts in a collaborative exploration setting to stretch the boundaries of human knowledge. Possible future work includes: i) implementing the exploration approach, ii) devising further incidence relations that capture aspects of the data model (e.g., aggregation for quantitative values), iii) integrating completeness [3] tools such as COOL-WD [4] to ensure that incomplete data does not induce counterexamples, and incorporating background knowledge in the form of the MARPL rules [15] proposed for ontological reasoning on Wikidata [14].

Acknowledgements. This work is partly supported by the German Research Foundation (DFG) in CRC 248 (Perspicuous Systems), CRC 912 (HAEC), and Emmy Noether grant KR 4381/1-1 (DIAMOND).

References

[1] M. Alam et al. “Mining Definitions from RDF Annotations Using Formal Concept Analysis”. In: *Proc. 24th Int. Joint Conf. on Artificial Intelligence (IJCAI’15)*. Ed. by Q. Yang and M. Wooldridge. AAAI Press, 2015.

[2] D. Borchmann, T. Hanika, and S. Obiedkov. “On the Usability of Probably Approximately Correct Implication Bases.” In: *ICFCA*. Ed. by K. Bertet et al. Vol. 10308. LNCS. Springer, 2017, pp. 72–88.

[3] F. Darari et al. “Completeness Management for RDF Data Sources”. In: *TWEB* 12.3 (2018), 18:1–18:53.

[4] F. Darari et al. “COOL-WD: A Completeness Tool for Wikidata”. In: *Proc. 16th Int. Semantic Web Conf. (ISWC’17): Posters & Demonstrations and Industry Tracks*. Ed. by N. Nikitina et al. Vol. 1963. CEUR WS Proceedings. CEUR-WS.org, 2017.

[5] L. Galárraga et al. “Fast Rule Mining in Ontological Knowledge Bases with AMIE++”. In: *The VLDB Journal* 24.6 (Dec. 2015), pp. 707–730.

[6] L. Galárraga et al. “Predicting Completeness in Knowledge Bases”. In: *Proc. 10th Int. Conf. on Web Search and Data Mining (WSDM’17)*. ACM, 2017.

[7] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.

[8] L. González and A. Hogan. “Modelling Dynamics in Semantic Web Knowledge Graphs with Formal Concept Analysis”. In: *Proc. 2018 World Wide Web Conf. (WWW’18)*. Ed. by P. Champin et al. ACM, 2018.

[9] T. Hanika, M. Marx, and G. Stumme. “Discovering Implicational Knowledge in Wikidata”. In: *CoRR* abs/1902.00916 (2019).

[10] V. T. Ho et al. “Rule Learning from Knowledge Graphs Guided by Embedding Models”. In: *Proc. 17th Int. Semantic Web Conf. (ISWC’18)*. Ed. by D. Vrandečić et al. Vol. 11136. LNCS. Springer, 2018, pp. 72–90.

[11] M. Krötzsch. “Ontologies for Knowledge Graphs?” In: *Proc. 30th Int. Workshop on Description Logics (DL’17)*. Ed. by A. Artale, B. Glimm, and R. Kontchakov. Vol. 1879. CEUR WS Proceedings. CEUR-WS.org, 2017.

[12] J. Lajus and F. M. Suchanek. “Are All People Married?: Determining Obligatory Attributes in Knowledge Bases”. In: *Proc. 2018 World Wide Web Conf. (WWW’18)*. Ed. by P. Champin et al. ACM, 2018.

[13] M. Luxenburger. “Implications partielles dans un contexte”. In: *Math. Inform. Sci. Humaines* 113 (1991), pp. 35–55.

[14] M. Marx and M. Krötzsch. “SQID: Towards Ontological Reasoning for Wikidata”. In: *Proc. 16th Int. Semantic Web Conf. (ISWC’17): Posters & Demonstrations and Industry Tracks*. Ed. by N. Nikitina et al. Vol. 1963. CEUR WS Proceedings. CEUR-WS.org, 2017.

[15] M. Marx, M. Krötzsch, and V. Thost. “Logic on MARS: Ontologies for Generalised Property Graphs”. In: *Proc. 26th Int. Joint Conf. on Artificial Intelligence (IJCAI’17)*. Ed. by C. Sierra. ijcai.org, 2017, pp. 1188–1194.

[16] S. Rudolph. “Exploring Relational Structures Via FLE”. In: *Proc. 12th Int. Conf. on Conceptual Structures (ICCS’04)*. Ed. by K. E. Wolff, H. D. Pfeiffer, and H. S. Delugach. Vol. 3127. LNCS. Springer, 2004, pp. 196–212.

[17] G. Stumme et al. “Intelligent Structuring and Reducing of Association Rules with Formal Concept Analysis”. In: *KI 2001: Advances in Artificial Intelligence*. Ed. by F. Baader, G. Brewka, and T. Eiter. Springer, 2001.

[18] T. P. Tanon et al. “Completeness-Aware Rule Learning from Knowledge Graphs”. In: *Proc. 16th Int. Semantic Web Conf. (ISWC’17)*. Ed. by C. d’Amato et al. Vol. 10587. LNCS. Springer, 2017, pp. 507–525.

[19] D. Vrandečić. “Wikidata: a new platform for collaborative data collection”. In: *Companion of the 21st World Wide Web Conf. (WWW’12)*. Ed. by A. Mille et al. ACM, 2012, pp. 1063–1064.

[20] D. Vrandečić and M. Krötzsch. “Wikidata: a free collaborative knowledgebase”. In: *Commun. ACM* 57.10 (2014), pp. 78–85.

[21] E. Zangerle et al. “An Empirical Evaluation of Property Recommender Systems for Wikidata and Collaborative Knowledge Bases”. In: *Proc. 12th Int. Symp. on Open Collaboration (OpenSym’16)*. Ed. by A. I. Wasserman. ACM, 2016, 18:1–18:8.