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] Center for Advancing Electronics Dresden (cfaed), TU Dresden
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. Examples include the proprietary knowledge graphs of companies such as Google, Facebook, IBM, or Microsoft, but also freely available ones such as YAGO, DBpedia, and Wikidata. A distinguishing feature of Wikidata is that the knowledge is collaboratively edited and curated. While this greatly enhances the scope of Wikidata, it also makes it impossible for a single individual to grasp complex connections between properties or understand the global impact of edits in the graph. We apply Formal Concept Analysis to efficiently identify comprehensible implications that are implicitly present in the data. Although the complex structure of data modelling in Wikidata is not amenable to a direct approach, we overcome this limitation by extracting contextual representations of parts of Wikidata in a systematic fashion. We demonstrate the practical feasibility of our approach through several experiments and show that the results may lead to the discovery of interesting implicational knowledge. Besides providing a method for obtaining large real-world data sets for FCA, we sketch potential applications in offering semantic assistance for editing and curating Wikidata.

Keywords: Wikidata, Formal Concept Analysis, 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 [28, 29] 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*.

Authors are given in alphabetical order. No priority in authorship is implied.

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.

However, maintaining this large knowledge graph is not an easy task, and retaining active editors poses an important challenge for the community [23]: throughout the six years of Wikidata's existence, Wikidata has amassed over three million registered users, but only 20 thousand editors are still active. An important step towards improving editor retention is to streamline the editing process as much as possible. The largest fraction of edits is made up of *bot edits* (automated editing tools operated by individual users) and *batch edits* (i.e., mass edits done through some tool specifically designed for certain types of edits). These are primarily authored by seasoned editors not usually susceptible to editor attrition [22]. In contrast, casual editors typically do not use tools besides the Wikidata web interface. Towards improving this editing experience, we propose to extract implicational knowledge ("rules") from Wikidata to explicate to the editor the (potentially non-local) consequences of editing a particular item's statements, similarly to the way *property constraints*[4] are already used to highlight potentially conflicting or missing data. Such rules must necessarily be easy to understand for editors that are not already deeply familiar with Wikidata's data model and ontological structure. Previous approaches have studied extracting rules in the form of implications of first-order logic (FOL) is a feasible approach to obtain interesting and relevant rules from Wikidata. [9, 15] The expressive power of FOL comes with a steep price, however: 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) [12] 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[5], 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.

---

[4] `https://www.wikidata.org/wiki/Help:Property_constraints_portal`

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

2 Related Work

In the realm of Wikidata, there have been two prominent applications of FCA so far: The authors in [13] model and predict the dynamic behaviour of knowledge graphs using lattice structures, and [17] attempts to determine obligatory attributes for classes in Wikidata. A more general approach to applying FCA to knowledge graphs was proposed in [8]. A current topic in knowledge bases (and in particular Wikidata) is Rule mining. Several successful approaches to generating lists of FOL rules, e.g., in [15, 9] have been proposed. This task is often connected to ranked lists of rules, sometimes falsely denoted as recommendations, like in [31], or completeness investigations for knowledge graphs, like in [10, 26].

3 Wikidata

Data Model. Wikidata [29] 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:

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

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

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 [20], however, have been proposed specifically to capture Wikidata, and we thus phrase our formalisation in terms of a *multi-attributed relational structure.*

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 $\Delta \coloneqq \mathcal{E} \cup \mathcal{V}$. Now, 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:

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

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

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* ^ *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 Contexts in Wikidata

Building upon Definition 1, we now recall basic notions from Formal Concept Analysis and how they relate to the structure of Wikidata. For a thorough introduction, we refer the reader to [12]. A *formal context* is a triple $\mathbb{K} = \langle G, M, I\rangle$ where G is a set of so-called *objects*, M is a set of so-called *attributes*, and $I \subseteq G \times M$ is called the *incidence relation*. This relation gives rise to the definition of two *derivation operations* traditionally sharing the same symbol: $\cdot'\colon \mathfrak{P}(G) \to \mathfrak{P}(M), A \mapsto A' \coloneqq \{m \in M \mid \forall g \in A\colon \langle g, m\rangle \in I\}$ and $\cdot'\colon \mathfrak{P}(M) \to \mathfrak{P}(G), B \mapsto B' \coloneqq \{g \in G \mid \forall m \in B\colon \langle g, m\rangle \in I\}$. Two sets $A \subseteq G$ and $B \subseteq M'$ are called *closed* in $\mathbb{K}$ if $A = A''$ and $B = B''$, respectively. A pair $\langle A, B\rangle$ satisfying $A' = B$ and $B' = A$, where $A \subseteq G$ and $B \subseteq M$, is a *formal concept*, the defining entity for FCA. The set of all concepts is denoted by $\mathfrak{B}(\mathbb{K})$. An *attribute implication* is denoted by $X \to Y$, where $X, Y \subseteq M$. We say $X \to Y$ is valid in $\mathbb{K}$ iff $X' \subseteq Y'$. The set of all valid implications for $\mathbb{K}$ on M is called the *attribute implicational theory*, denoted by $Th_M(\mathbb{K})$. In general, the theory of a formal context can be exponentially large compared to the size of the context. Thus, one employs an implication base, i.e., a sound, complete, and non-redundant set of implications from which the theory can be inferred. Among the various bases used in FCA, the *canonical base* $\mathcal{L}$ stands out due to its minimal size [14].

Reasoning using a canonical base is quite simple: for every attribute set $X \subseteq M$, compute the closure $X^{\mathcal{L}}$, i.e., apply $\mathcal{L}$ to X until the result is stable. This can be done in time linear in the size of the canonical base [3]. Thus, entailment of an implication $X \to Y$ with respect to a base can be decided in linear time: $X \to Y$ is entailed by $\mathcal{L}$ iff $Y \subseteq X^{\mathcal{L}}$ [11, Proposition 16]. In contrast, deciding entailment directly is P-complete with respect to the size of the implicational theory [5]. Computing the canonical base is thus a more efficient way to decide entailment for multiple implications, as the computational effort of computing a base gets amortised over the entailment checks. An implication $X \to Y$ in a formal context $\langle G, M, I\rangle$ has *support*[6] $\operatorname{supp}(X \to Y) = \frac{|X'|}{|G|}$, i.e., the relative number of objects exhibiting the necessary attributes for the rule to be applicable among all objects. A higher support implies that the implication is more relevant to the whole domain of the context. Nevertheless, a valid implication $X \to Y$ may have a support of zero.

---

[6] We use the definition for the support from FCA, which coincides with the definition of the support on *valid* rules in association rules.

5 Property Theory

In the following, we employ these tools and techniques to obtain a more accessible view on the Wikidata knowledge graph and how properties therein depend on each other. Krötzsch [16] 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 introduce four natural problem scenarios for selecting sets of properties from Wikidata, each exploiting different aspects of the rich data model to enhance the understanding of the data.

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 $Th_{\hat{\mathcal{P}}}(\mathcal{E}, \hat{\mathcal{P}}, I^{\mathtt{plain}})$, where

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

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 6.2. 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 $Th_{\hat{\mathcal{P}}}(\mathcal{E}, \hat{\mathcal{P}}, I^{\mathtt{dir}})$, where an entity e coincides with $\hat{p}$ iff it occurs as subject or object (depending on the colour) of some p-statement:

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

Example 4. Let $\hat{\mathcal{P}} = \{\mathtt{\hat{}mother}_{\hat{}P25}, \mathtt{godparent}_{P1290}, \mathtt{mother}_{P25}\}$ be the set of attributes and let $\mathcal{E} = \{\mathtt{Miley\ Cyrus}_{Q4235}, \mathtt{Victoria}_{Q9439}, \mathtt{Naomi\ Watts}_{Q132616}, \mathtt{Angelina\ Jolie}_{Q13909}\}$ be the set of objects. The corresponding formal context $\langle \mathcal{E}, \hat{\mathcal{P}}, I^{\mathtt{dir}} \rangle$ (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.

5.3 Qualified Incidence

While Problem 2 captures Example 4, it is still insufficient to grasp the subtleties of Example 3, since two statements differing only in their annotations are indistinguishable. We thus include annotations into the colours of the properties. For a property p, we define $\mathrm{Annotations}(p) \coloneqq \bigcup_{t \in \mathcal{W}(p)} \operatorname{ann} t$, the set of all individual annotations occurring in statements for p.

Problem 3. For $\mathcal{W}$ and $\hat{\mathcal{P}} \subseteq \bigcup_{p \in \mathcal{P}}(\{p\} \times \{\mathrm{subj}, \mathrm{obj}\} \times \mathrm{Annotations}(p))$, compute the canonical base for $Th_{\hat{\mathcal{P}}}(\mathcal{E}, \hat{\mathcal{P}}, I^{\texttt{qual}})$, where

$$\begin{aligned} \langle q, \hat{p} \rangle \in I^{\texttt{qual}} :\Longleftrightarrow\ & \big(\hat{p} = \langle p, \mathrm{obj}, a \rangle \wedge \exists t \in \mathcal{W}(p).\ (\operatorname{obj} t = q) \wedge (a \in \operatorname{ann} t)\big) \\ & \vee \big(\hat{p} = \langle p, \mathrm{subj}, a \rangle \wedge \exists t \in \mathcal{W}(p).\ (\operatorname{subj} t = q) \wedge (a \in \operatorname{ann} t)\big),\ \text{ i.e.,} \end{aligned} \tag{7}$$

an entity e coincides with $\hat{p} = \langle p, d, a \rangle$ iff it occurs as subject or object (depending on d) of some p-statement t, and the annotation $\operatorname{ann} t$ of t includes a.

5.4 Classified Incidence

Another natural approach to distinguishing properties is to consider 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”). We thus define for a property $p \in \mathcal{P}$ the set of all classes that objects of p-statements are instances of: $\mathrm{Classes}(p) \coloneqq \{\operatorname{obj} t \mid t \in \mathcal{W}(\texttt{instance of}_{P31}), s \in \mathcal{W}(p), \operatorname{obj} s = \operatorname{subj} t\}$.

Table 1. Property selection in data sets

data set	properties in class ("Wikidata property for . . .")
`awards`	Q56150830 ("... awards, prizes and honours")
`family`	Q22964231 ("... human relationships")
`math`	Q22988631 ("... mathematics")
`space`	Q28104992 ("... spacecraft")
`time`	Q51077473 ("... time and duration")

Problem 4. Given $\mathcal{W}$ and some set $\hat{\mathcal{P}} \subseteq \bigcup_{p \in P}(\{p\} \times \{\text{subj}, \text{obj}\} \times \text{Classes}(p))$, compute the canonical base for the implicational theory $Th_{\hat{\mathcal{P}}}(\mathcal{E}, \hat{\mathcal{P}}, I^{\texttt{class}})$, where

$$\begin{aligned}\langle q, \hat{p}\rangle \in I^{\texttt{class}} :\iff\ & \big(\hat{p} = \langle p, \text{subj}, c\rangle \wedge \exists s \in \mathcal{W}(p).\\ & \exists t \in \mathcal{W}(\texttt{instance of}_{P31}).\ (\text{subj}\, s = q)\ \wedge (\text{obj}\, s = \text{subj}\, t) \wedge (\text{obj}\, t = c)\big)\\ & \vee \big(\hat{p} = \langle p, \text{obj}, c\rangle \wedge \exists t \in \mathcal{W}(\texttt{instance of}_{P31}).(q \in \text{obj}\, \mathcal{W}(p))\\ & \quad \wedge (\text{subj}\, t = q) \wedge (\text{obj}\, t = c)\big),\ \text{i.e.,}\end{aligned} \tag{8}$$

an entity e coincides with $\hat{p} = \langle p, d, c\rangle$ if there is a p-statement t with subject (or object, respectively) e, and the object of t is an instance of class c.

Generalised Incidence. Any of the incidences so far may be combined into one generalised incidence, since the point-wise union of two formal contexts is again a formal context. In the same spirit, one could investigate further incidences emphasising other aspects of the Wikidata data model.

6 Computations

Solving Problems 1 to 4 should, in theory, be straightforward: merely apply one of the two known algorithms for computing the canonical base, or first compute some other base like a *direct base* from which one can deduce the canonical base. However, for $|\hat{\mathcal{P}}| \geq 200$, computing the canonical base in reasonable time on affordable hardware might already be impossible. Nonetheless, to demonstrate that it is indeed possible to derive meaningful formal contexts from Wikidata with the approaches of Problems 1 and 2, we conducted experiments with classical and more recent methods for computing canonical bases. We discuss a range of different selections of properties to illustrate these techniques.

6.1 Data Sets

Our data sets were generated from the full Wikidata dump (in JSON format) from 2018-10-22[7]. From this dump, we have extracted different subsets by selecting fixed sets of properties related to certain domains of interest (arbitrarily chosen

---

[7] `https://dumps.wikimedia.org/wikidatawiki/entities/20181022/`

by the authors due to personal preference), which are represented as classes of properties in Wikidata. Since Wikidata comprises knowledge from a vast number of distinct and unconnected domains, this is a natural simplification, as the properties used for, e.g., spacecraft are disjoint from those properties used for mathematics. Table 1 describes which properties were selected for which data set; Table 2 shows the sizes of the various data sets. For each data set, we have extracted formal contexts corresponding to the incidences of Problems 1 to 4, respectively. In generating these contexts, we ignore statements that are i) deprecated, since these are no longer considered valid, ii) have an unknown value, or iii) have no value. All other statements contribute to populating the context according to the corresponding incidence relation. Finally, we remove empty rows and columns, since these do not influence implications.

For comparison, we have also included the data set `wiki44k` as provided by [15], a small subset of simple statements extracted from a Wikidata dump from December 2014. Meanwhile, though, the usage of some properties on Wikidata has changed, and, in particular, eight properties used in this data set have since been deleted on Wikidata. Hence, we have also generated `wiki44k-tr`, where these properties have been replaced by their modern equivalents: **P7 ("brother"), P9 ("sister"):** replaced by P3373 ("sibling"), **P45 ("grandparent"):** replaced by P1038 ("relative"), with a P1039 ("type of kinship") qualifier with value Q167918 ("grandparent"), **P70 ("order"):** replaced by P171 ("parent taxon"), where the object gets an additional P105 ("taxon rank") statement with value Q36602 ("order"), **P71 ("family"):** replaced by P171 ("parent taxon"); where the object gets an additional P105 ("taxon rank") statement with value Q35409 ("family"), **P107 ("main type GND"), P132 ("administrative entity"):** replaced by P31 ("instance of"), and **P133 ("language family"):** replaced by P279 ("subclass of"). Both data sets were converted to JSON format and then processed analogously to the other data sets. Data sets `wiki44k-2018` and `wiki44k-2018-tr` are subsets of the 2018 dump obtained by dropping all items and properties (and statements connecting those) not appearing in `wiki44k` and `wiki44k-tr`, respectively.

Table 2. Size of data sets

data set	items	properties	statements
`awards`	429,207	27	892,723
`family`	307,330	10	728,669
`math`	36,913	45	84,255
`space`	7,693	20	30,212
`time`	216,865	9	219,803
`wiki44k`	45,021	101	295,352
`wiki44k-2018`	44,915	92	382,725
`wiki44k-tr`	45,021	95	300,687
`wiki44k-2018-tr`	44,919	94	384,700

Table 3. Canonical bases for contexts in Problem 1

data set	density	$\lvert CanBase(\cdot)\rvert$	# supported
`awards`	0.039	280	17
`family`	0.163	46	46
`math`	0.040	752	71
`space`	0.195	157	125
`time`	0.112	27	0
`wiki44k`	0.045	7,040	3,556
`wiki44k-2018`	0.053	8,179	5,550
`wiki44k-tr`	0.043	6,408	3,261
`wiki44k-2018-tr`	0.053	9,422	6,641

6.2 Empirical Evaluation

Table 3 depicts the results of our computations for Problem 1. The computation time for those results varied from seconds in the case of `family` to approximately five hours for the `wiki44-*` data sets. The program code for our experiments was written in *Clojure* and builds on the existing *conexp-clj*.[8]. A *GitHub* repository[9] holds data sets and computed results. We used a state-of-the-art server system with two Intel® Xeon® Gold 5122 CPUs and 768 GB RAM.

A first observation is that the canonical bases remain small in relation to the data sets, even though the data is inherently noisy and a considerable number of special items provokes peculiar rules. This enhances the applicability of these bases. Secondly, we observe that the number of supported rules varies strongly with different data sets, e.g., all rules are supported in the `family` data set, whereas no rule in `time` is. Thus, `time` does not admit a non-trivial propositional Horn logic theory, but still, the canonical base comprises a plenitude of valid rules. This could be due to incorrectly or insufficiently contributed data, an unusually high number of exceptions in the domain, or even due to a conscious design decision in the data modelling with respect to the properties in `time`. Nonetheless, this canonical base may still be used to validate that particular implications hold in the domain.

We now list some interesting, exemplary implications that we discovered for Problems 1 and 2. In the data set names, `<data set>-0` corresponds to a context for Problem 1, whereas `<data set>-1` denotes a context for Problem 2.

`awards-0`: $\{\texttt{Nobel prize ID}_{P3188}\} \to \{\texttt{award received}_{P166}\}$ is an implication supported by 0.2% of the data set. While this is a reasonable implication, it is hardly surprising. Altogether, we obtained a set of 280 valid implications from which 17 are supported in the data set.

`awards-1`: We found the following implication, stating that everything that someone has been nominated for and that has an associated category for the award is also an award received by some entity, supported by 0.03% of the data set:

---

[8] `https://github.com/exot/conexp-clj`

[9] `https://github.com/wikiexploration/wikiexploration`

$\{\texttt{\textasciicircum nominated for}_{\text{\textasciicircum}P1411}, \texttt{category for recipients of this award}_{P2517}\}$ $\rightarrow \{\texttt{\textasciicircum award received}_{\text{\textasciicircum}P166}\}$. Beyond this, implications from `awards-1` do not seem to shed more light on the set of investigated properties.

`family-0:` Here we computed a canonical base in which all implications are supported. For example, the implication $\{\texttt{godparent}_{P1290}, \texttt{partner}_{P451}\} \rightarrow \{\texttt{sibling}_{P3373}\}$ is supported by 7 out of 306,908 entities. It states that an entity that has a godparent and has a partner also has a sibling. However, this implication is not necessarily true for family relations.

`family-1:` With 0.03% support, the implication $\{\texttt{\textasciicircum father}_{\text{\textasciicircum}P22}, \texttt{\textasciicircum relative}_{\text{\textasciicircum}P1038}, \texttt{spouse}_{P26}\} \rightarrow \{\texttt{child}_{P40}\}$ is unsurprising, but witnesses that the more general $\{\texttt{\textasciicircum father}_{\text{\textasciicircum}P22}\} \rightarrow \{\texttt{child}_{P40}\}$ has counterexamples in the data set: in fact, there are 1,634 non-fictional humans serving as counterexamples.

`math-0:` Among 752 implications, we discovered, with 0.01% support, the implication $\{\texttt{has vertex figure}_{P1678}, \texttt{base}_{P3263}\} \rightarrow \{\texttt{has facet polytope}_{P1678}\}$. It is clearly helpful to obtain such rules when unfamiliar with polytope theory.

`math-1:` We observe a large number of rules relating mathematical identifiers, or at least using them. One such example, with support 0.02%, is the implication $\{\texttt{has vertex figure}_{P1678}, \texttt{MathWorld identifier}_{P2812}, \texttt{\textasciicircum dual to}_{\text{\textasciicircum}P1322}\}$ $\rightarrow \{\texttt{has facet polytope}_{P1678}\}$. Polytope theory is well represented in Wikidata. Other fields of mathematics are lacking data, however, and our hopes for numerous implications from diverse fields remained unfulfilled.

`space-0:` 4% of the data set supports the implication $\{\texttt{type of orbit}_{P522}, \texttt{periapsis}_{P2244}\} \rightarrow \{\texttt{apoapsis}_{P2243}\}$. From our point of view, there are several more elements of the computed canonical base contributing to a better understanding of the properties in this domain.

`space-1:` The rule (support 0.01%) $\{\texttt{apoapsis}_{P2243}, \texttt{\textasciicircum type of orbit}_{\text{\textasciicircum}P522}\}$ $\rightarrow \{\texttt{orbital period}_{P2146}, \texttt{type of orbit}_{P522}, \texttt{periapsis}_{P2244}\}$ states that all other relevant orbital facts are present for types of orbit with an apoapsis.

`wiki44k-0:` We find the rule $\{\texttt{producer}_{P162}, \texttt{country}_{P17}\} \rightarrow \{\texttt{genre}_{P136}\}$ with support 0.04%, stating that knowing the producer and the country of an item we can infer the genre associated to this item. Since this data collection was constructed with the explicit goal of having a very dense data set [9], the probability for sparse counterexamples is low.

We claim that the implications discovered by us are readable and comprehensible by humans, at least in cases where a user is familiar with the domain of knowledge. Our experiments shed light on two particular kinds of errors entailed in Wikidata. First, we found that implications that should be valid, yet cannot be inferred from the data set due to the presence of counterexamples. By incorporating a background ontology, one may apply our method to identify missing implications and therefore possible errors in Wikidata. These can be fixed by editing statements or, in more serious cases, by introducing new properties and deprecating (particular uses of) others. Secondly, we observed valid implications that can easily be refuted by inquiring a domain expert for a counterexample. Attribute exploration on parts of Wikidata could be harnessed to systematically obtain such counterexamples, although the sheer size of Wikidata requires a collaborative exploration method. At this point, we refrain from a thorough statistical evaluation, as, e.g., done in [15], since, from a logical standpoint, the computed bases are not only sound and complete, but also unique, methods such as cross-validation via, e.g., embedding models, are inappropriate to obtain meaningful measurements. Rather, we focus on obtaining bases suitable for verifying implicational assumptions on the data, as well as enhancing the comprehension of properties by users of Wikidata.

6.3 Limitations

These experiments are subject to two limitations of our method: i) Already for a formalism as simple as horn PL rules, computing canonical bases is only feasible for small subdomains of Wikidata. This is hardly surprising, as recognising elements of the canonical base is coNP-complete [2] (but becomes even harder for more expressive formalisms, if bases even exist). We investigate two known approaches for coping with this limitation in the next section. ii) The collaborative editing process is prone to introducing noise into the data, but the canonical base is sensitive to small changes and thus not well-suited for noisy data. This naturally leads to a weaker notion of base. While technically also a limitation of the approach, we consciously limit ourselves to implications in propositional Horn logic which cannot express some of the rules obtained by more expressive frameworks such as [15, 9]: computing bases of implications for more expressive logics is computationally infeasible at best, and impossible at worst. Moreover, propositional Horn rules are arguably easier to grasp for untrained editors of Wikidata than, e.g., first-order rules.

7 Association Rules and PAC Bases

Various approaches to overcome the limitations stated in the last section are known. A well-investigated and mature procedure is the computation of *association rules* [1]. While closely related to implications in the FCA sense [21, 24, 30], they were developed independently. For $X, Y \subseteq M$, the *confidence* of an association rule $X \to Y$ is $\operatorname{conf}(X \to Y) \coloneqq \operatorname{supp}(X \cup Y)/\operatorname{supp}(X)$. The classical problem then is to compute all association rules with minimum support `minsup` and confidence at least `minconf` in a domain. This usually leads to exponentially many rules. A plenitude of extensions of association rules have been proposed, e.g., different kinds of support, head-confidence, etc. For our goal of obtaining a base of implications, the most interesting extension was done in [25], relaying on the *Luxenburger base of association rules* [18]. From this base one can infer all valid association rules in the domain with respect to `minsup` and `minconf`.

As a short empirical evaluation, we computed the Luxenburger base for the data sets `awards`, `family`, `math`, `space`, and `time` for the setting of Problem 2 with `minsupp` $= 0.0001$ and `minconf` $= 0.6$. We list some interesting elements of these bases: **awards (429,126 items, 38 properties):** The base has 28 rules, the rule $\{\texttt{National Medal of Arts winner ID}_{P5719}, \texttt{nominated for}_{1411}\} \to$

$\{\texttt{award received}_{P166}\}$ has confidence 99.8% and support 0.017%, with the sole counterexample Q15782045 ("Albert Maysles").[10] **family (306,908 items, 19 properties):** The rule $\{\texttt{mother}_{P25}, \texttt{father}_{P22}\} \rightarrow \{\texttt{\textasciicircum child}_{\text{\textasciicircum}P40}\}$ has a support 11.7% and confidence 99.8%. The size of the base is 11,359. **math (36,904 items, 63 properties):** Among the 482 rules is $\{\texttt{input set}_{P1851}, \texttt{domain}_{P1568}\} \rightarrow \{\texttt{codomain}_{P1571}\}$ which has support 0.024% and confidence 88.8%, with the only counter example being Q3075242 ("inverse function"), due to a missing statement on this item. **space (7,693 items, 30 properties):** The base has 666 elements. The rule $\{\texttt{orbital period}_{P2146}, \texttt{apoapsis}_{P2243}\} \rightarrow \{\texttt{periapsis}_{P2244}\}$ is supported by 18% of the data set and has a confidence of 99.9% (note that this rule is not an element of the canonical basis). **time (216,856 items, 15 properties):** The base has 4 elements. An interesting rule with support 0.03% and confidence 97.4% is $\{\texttt{temporal range end}_{P524}\} \rightarrow \{\texttt{temporal range start}_{P523}\}$. Except for `space`, the Luxenburger bases are of reasonable size, i.e., the size of the base is about 1% of the number of entities. We also computed the Luxenburger base for `wiki44k` and `wiki44k-2018`, resulting in bases of 359,745 and 293,038 rules, respectively. While this is in sharp contrast to the positive results above, there is a simple explanation for this effect: the properties in our data sets were chosen from a common domain, leading to a data set that includes items predominantly from this domain. The `wiki44k` data set, however, was constructed by searching for frequently-used (but not necessarily related) properties [9]. This more arbitrary set of properties results in a large number of rules crossing semantically independent subsets of the property set.

A more recent approach is to employ the idea of PAC learning, as introduced with the seminal paper by Valiant [27]. The authors in [4] present a procedure for retrieving *probably approximately correct implication bases* from a formal context. For fixed probability δ and accuracy ε, a PAC basis can be computed in time polynomial in the size of the input context and the size of the resulting basis. In contrast to association rules, this approach still yields a canonical base (of some approximation of the theory) consisting of (possibly) unsupported, yet correct, implications. A thorough investigation of the applicability of PAC learning for Wikidata based formal contexts requires its own research work and is out of scope for this paper. Nevertheless, we provide first computational results in the data repository. Both of the methods discussed in this section are more suitable for coping with the scale of Wikidata and the noise inherently present in the data.

8 Conclusion and Outlook

We have demonstrated in this work how to extract, in a structured fashion, subsets of Wikidata and represent them as formal contexts, thus opening a practically limitless source of real-world data to the FCA community. This allows us (and indeed the whole community) to apply the full range of tools and techniques from

---

[10] This error in the data has already been automatically flagged as a property constraint violation: "*item requires statement constraint*: An entity with *National Medal of Arts Winner ID* should also have a statement *award received National Medal of Arts*."

FCA to (arbitrary) parts of Wikidata. Most importantly, we are now able to obtain relevant and meaningful implications in a form that is readily understood by untrained editors of Wikidata. Such rules are useful in many different ways: i) they can further the understanding of knowledge implicit in Wikidata, ii) they can make explicit how editing a statement interacts with the implicational theory on properties, iii) they may highlight the need to edit further items to avoid introducing new counterexamples to valid rules, and iv) absence of expected rules serves as an indicator for errors present in the knowledge graph. These qualities are highly desirable for streamlining the editing experience on Wikidata, not only for casual editors, but also for curators of Wikidata.

While we have shown that the direct computation of a canonical base is feasible for small subsets of the data, this becomes infeasible as the number of properties under consideration increases. We discussed two different approaches to overcome this limitation: computing Luxenburger bases of association rules, and computing a PAC basis, both of which remain feasible on the scale of Wikidata.

A complementary approach would be to employ the well-known attribute exploration algorithm to compute canonical bases for generalised incidences over Wikidata. The key ingredient required for this is a method to query Wikidata for possible counterexamples to proposed implications, e.g., via the SPARQL endpoint. This enables Wikidata to be used as an *expert* for the exploration. Further expanding on this, a collaborative exploration algorithm may employ both Wikidata and human experts to stretch the boundaries of human knowledge.

Possible further directions for future work include i) a practical study on the usefulness of integrating implicational knowledge into the editing process, ii) integrating with completeness [6] tools such as COOL-WD [7] to ensure that only counterexamples above a certain completeness threshold are considered, iii) extending the structured approach to include further incidence relations adapted to other aspects of the Wikidata data model, such as grouping properties for quantities by intervals of their values, and iv) extend the approach to incorporate background knowledge given, e.g., in the form of the MARPL rules [20] that have been proposed for ontological reasoning on Wikidata [19].

References

[1] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". In: *Proc. Int. Conf. on Management of data (SIGMOD'93)*. New York, NY, USA: ACM Press, 1993, pp. 207–216.

[2] M. A. Babin and S. O. Kuznetsov. "Recognizing Pseudo-intents is coNP-complete." In: *CLA*. Ed. by M. Kryszkiewicz and S. A. Obiedkov. Vol. 672. CEUR Workshop Proceedings. CEUR-WS.org, 2010, pp. 294–301.

[3] C. Beeri and P. A. Bernstein. "Computational Problems Related to the Design of Normal Form Relational Schemas". In: *ACM Trans. Database Syst.* 4.1 (Mar. 1979), pp. 30–59. ISSN: 0362-5915. DOI: `10.1145/320064.320066`.

[4] 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. ISBN: 978-3-319-59271-8.

[5] E. Dantsin et al. “Complexity and expressive power of logic programming”. In: *ACM Comput. Surv.* 33.3 (2001), pp. 374–425.

[6] F. Darari et al. “Completeness Management for RDF Data Sources”. In: *TWEB* 12.3 (2018), 18:1–18:53. DOI: `10.1145/3196248`.

[7] 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 Workshop Proceedings. CEUR-WS.org, 2017.

[8] S. Ferré. “A Proposal for Extending Formal Concept Analysis to Knowledge Graphs”. In: *Formal Concept Analysis*. Ed. by J. Baixeries, C. Sacarea, and M. Ojeda-Aciego. Cham: Springer International Publishing, 2015, pp. 271–286. ISBN: 978-3-319-19545-2.

[9] L. Galárraga et al. “Fast Rule Mining in Ontological Knowledge Bases with AMIE++”. In: *The VLDB Journal* 24.6 (Dec. 2015), pp. 707–730. ISSN: 1066-8888. DOI: `10.1007/s00778-015-0394-1`.

[10] L. Galárraga et al. “Predicting Completeness in Knowledge Bases”. In: *Proc. 10th Int. Conf. on Web Search and Data Mining (WSDM’17)*. New York, NY, USA: ACM, 2017, pp. 375–383. ISBN: 978-1-4503-4675-7.

[11] B. Ganter and S. Obiedkov. *Conceptual Exploration*. Springer, 2016. ISBN: 978-3-662-49290-1. DOI: `10.1007/978-3-662-49291-8`.

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

[13] L. González and A. Hogan. “Modelling Dynamics in Semantic Web Knowledge Graphs with Formal Concept Analysis”. In: *Proc. of the 2018 World Wide Web Conf. (WWW’18)*. Ed. by P. Champin et al. ACM, 2018, pp. 1175–1184. DOI: `10.1145/3178876.3186016`.

[14] J.-L. Guigues and V. Duquenne. “Famille minimale d’implications informatives résultant d’un tableau de données binaires”. In: *Mathématiques et Sciences Humaines* 24.95 (1986), pp. 5–18.

[15] 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. Vrandecic et al. Vol. 11136. LNCS. Springer, 2018, pp. 72–90. ISBN: 978-3-030-00670-9.

[16] 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 Workshop Proceedings. CEUR-WS.org, 2017.

[17] J. Lajus and F. M. Suchanek. “Are All People Married?: Determining Obligatory Attributes in Knowledge Bases”. In: *Proc. of the 2018 World Wide Web Conf. (WWW’18)*. Ed. by P. Champin et al. ACM, 2018, pp. 1115–1124. ISBN: 978-1-4503-5639-8.

[18] M. Luxenburger. "Implications partielles dans un contexte". In: *Math. Inform. Sci. Humaines* 113 (1991), pp. 35–55. ISSN: 0995-2314.

[19] 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 Workshop Proceedings. CEUR-WS.org, 2017.

[20] 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. ISBN: 978-0-9992411-0-3.

[21] N. Pasquier et al. "Closed Set Based Discovery of Small Covers for Association Rules." In: *Proc. 15èmes Journées Bases de Données Avancées, BDA*. Jan. 3, 2002, pp. 361–381.

[22] A. Piscopo and E. Simperl. "Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (Nov. 2018), 141:1–141:18. ISSN: 2573-0142. DOI: `10.1145/3274410`.

[23] C. Sarasua et al. "The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits". In: *Computer Supported Cooperative Work (CSCW)* (Dec. 2018). ISSN: 1573-7551. DOI: `10.1007/s10606-018-9344-y`.

[24] G. Stumme. *Conceptual Knowledge Discovery with Frequent Concept Lattices*. FB4-Preprint 2043. TU Darmstadt, 1999.

[25] 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, pp. 335–350. ISBN: 978-3-540-45422-9.

[26] 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. ISBN: 978-3-319-68288-4.

[27] L. G. Valiant. "A Theory of the Learnable". In: *STOC*. Ed. by R. A. DeMillo. ACM, 1984, pp. 436–445.

[28] D. Vrandecic. "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. ISBN: 978-1-4503-1230-1.

[29] D. Vrandečić and M. Krötzsch. "Wikidata: a free collaborative knowledgebase". In: *Commun. ACM* 57.10 (2014), pp. 78–85. DOI: `10.1145/2629489`.

[30] M. J. Zaki and C.-J. Hsiao. *ChARM: An efficient algorithm for closed association rule mining. Technical Report 99–10*. Tech. rep. Computer Science Dept., Rensselaer Polytechnic, Oct. 1999.

[31] 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. ISBN: 978-1-4503-4451-7.