Are Qualifiers Enough? Context-Compatible Information Fusion for Wikimedia Data

Hermann Kroll
Institute for Information Systems,
TU Braunschweig

Wolf-Tilo Balke
Institute for Information Systems,
TU Braunschweig

Abstract

The Wikimedia Foundation provides extensive and well-known collaborative knowledge repositories like Wikipedia and Wikidata. While knowledge in Wikipedia is represented in written narratives on certain items, Wikidata represents this knowledge through triple-shaped statements. However, some statements, and especially, assertions, beliefs, or negations might only become valid when assuming certain context conditions. While those conditions are usually mentioned in written narratives, we state the question whether Wikidata's qualifier data model is enough for a reliable context-compatible information fusion.

Keywords: Context, Context-Compatibility, Information Fusion, Knowledge Graphs, Wikidata

Introduction

Exchanging knowledge through written and oral narratives is a well-established practice in human communication. However, the extensive amount of information published daily asks for novel architecture to handle, manage and provide effective access paths for such information. One way to go is to transform unstructured texts into structured information, e.g., by harvesting structured triple-shaped statements from sentences or asking humans to provide them in a crowd-sourced fashion. Typical examples are a person's birthdate or a politician's citizenship. In addition to the well-known textual knowledge repository Wikipedia, Wikidata (Vrandecic and Krötzsch, 2014) provides a near language independent, and primarily, structured repository of knowledge. First, such a structured representation allows effective overviews of what is known about a specific concept/entity (e.g., a person or a city). And moreover, Wikidata also allows effective retrieval through structured query languages like SPARQL. Those query languages support the fusion of information (also known as joining) to answer a complex question or discover new knowledge. In addition to basic pattern matching, which ensures the compatibility of statements on a structural level, we argue, that such an information fusion should also ensure that statements are context-compatible (Kroll et al., 2022; Kroll et al., 2020).

Context

In our understanding, a context defines the scope in which a piece of information can be fused (joined) with other pieces (Kroll et al., 2022). Briefly speaking, a context comprises every condition that must be considered to *work* with a certain piece, here a statement. Indeed, some pieces of information can be universally applicable, and are thus, easy to connect to other pieces, e.g., the birth date of some person. However, some pieces are only valid within specific semantic settings. For instance, the capital of some countries may change over time, or a person might lose her citizenship. Those pieces should then only be connected to pieces valid in the same time frame (Kroll et al., 2022; Kroll and Balke, 2022).

Implicit and Explicit Contexts

In our previous work, we discussed how the existing literature deals with the context problem (Kroll et al., 2022). Basically, explicit models require knowledge base curators to pre-define context conditions and rules on how to combine statements that were observed under compatible conditions safely. Another option is to enhance the data model by harvesting n-ary relations, for example, (Ernst et al., 2018). However, explicit models usually require extensive domain knowledge and manual curation. In contrast, we introduced the notion of implicit contexts (Kroll et al., 2020). Please consider a publication of scientific findings. Here, scientists usually write about essential conditions (in the form of inclusion and exclusion criteria) and then state their findings. Our argument here is that statements from such a publication may then safely be fused because the context is *quite stable* within a publication (at least in certain sections). In contrast to explicit context models, such an implicit context comes with lower costs but less quality in the end.

Contexts in Wikidata

A close look at Wikidata reveals that Wikidata enriches statements by so-called qualifiers (Hernández et al., 2015). A qualifier is a property-value pair attached to a specific statement. Techniques like reification allow us to represent those qualifiers in the Resource Description Framework (RDF). Examples of those qualifiers are

a *point in time* which may be attached to a birthplace or place of death statement or a time span (through a temporal start and end qualifier), which may be attached to holding a specific political position statement. Provenance information represents a different class of qualifiers, i.e., everything about the origin of a statement, like a reference to its source. However, such a reference usually does not tell us anything about the restricting context of the statement explicitly.

Indeed, we had a close look at the 12-most used (frequency) qualifiers (excluding references) (03/2023)[1]: *series ordinal* (P1545, 167M), *astronomical filter* (P1227, 33M), *catalog* (P972, 23M), *object named as* (P1932, 17M), *point in time* (P585, 11M), *determination method* (P459, 8M), *start time* (P580, 8M), *found in taxon* (P703, 4M), *end time* (P582, 4M), *subject named as* (P1810, 3M), *chromosome* (P1057, 3M) and *language of work or name* (P407, 2M).

First, compared to the size of Wikidata[2] which has around 102M items with 1.4B statements, it becomes evident that not all statements have qualifiers attached. Second, the previous qualifiers fell into different semantic categories: While some of some indeed describe a temporal context or a location (e.g., on a chromosome), other qualifiers like a very broad *determination method*, *astronomical filter* or *catalog* might not represent a context which we are looking for in this paper. Even if every single statement in Wikidata would be attached with all suitable qualifiers which apply to the statement's property, *how can we reliably use those qualifiers in the end?*

Discussion

While humans may recognize those qualifiers when browsing Wikidata, SPARQL queries may not ask for them. And, precisely for those SPARQL queries, we see an open problem. How *can* and *should* those qualifiers be used automatically to ensure a context-compatible information fusion in Wikidata? For instance, one static rule could define a *temporal-based* fusion, i.e., to only fuse statements that are valid in the same time span (qualified by start and end times). However, the obvious question arises whether such rules scale with the size of different qualifiers and different combinations between those (e.g., temporal/geospatial contexts).

And even worse, contexts like temporal/geospatial ones might only be a small class of contexts. Think, for instance, about event-centric knowledge and its representation. While some properties of events might be static, like the point in time or the location, subjective attributions might be stated differently depending on someone's stance and viewpoint (Plötzky and Balke, 2022). For instance, while one party may agree on a certain war aggressor, the war party might not agree. While viewpoints and stances are specific problems on their own, Suchanek argued that there is a need to move beyond triples (Suchanek, 2020). From his viewpoint, a machine must also consider assertions, negations, and beliefs. And even worse, some statements may only be valid in their temporal or causal chain. Even though Wikidata introduced the properties *nature of statement* (P5102) and *sourcing circumstances* (P1480), in our eyes, the question still remains how those properties can automatically be used for a context-compatible information fusion.

That is why we conclude with the following research question: *Are qualifiers in Wikidata enough for a reliable context-compatible information fusion?*

[1] https://sqid.toolforge.org/

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

References

[Ernst et al.2018] Patrick Ernst, Amy Siu, and Gerhard Weikum. 2018. Highlife: Higher-arity fact harvesting. In *Proc. of the World Wide Web Conf. on World Wide Web, WWW 2018*, pages 1013–1022. ACM.

[Hernández et al.2015] Daniel Hernández, Aidan Hogan, and Markus Krötzsch. 2015. Reifying RDF: what works well with wikidata? In *Proc. of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems@ISWC2015*, volume 1457 of *CEUR Workshop Proc.*, pages 32–47. CEUR-WS.org.

[Kroll and Balke2022] Hermann Kroll and Wolf-Tilo Balke. 2022. On Design Principles for Narrative Information Systems. In *Proc. of the Workshop on Semantic Techniques for Narrative-Based Understanding@IJCAI-ECAI 2022)*, volume 3322 of *CEUR Workshop Proc.*, pages 11–18. CEUR-WS.org.

[Kroll et al.2020] Hermann Kroll, Jan-Christoph Kalo, Denis Nagel, Stephan Mennicke, and Wolf-Tilo Balke. 2020. Context-Compatible Information Fusion for Scientific Knowledge Graphs. In *Proc. of the 24th International Conf. on Theory and Practice of Digital Libraries, TPDL 2020*, LNCS, pages 33–47. Springer.

[Kroll et al.2022] Hermann Kroll, Florian Plötzky, Jan Pirklbauer, and Wolf-Tilo Balke. 2022. What a Publication Tells You—Benefits of Narrative Information Access in Digital Libraries. In *Proc. of the 22nd ACM/IEEE Joint Conf. on Digital Libraries*. ACM.

[Plötzky and Balke2022] Florian Plötzky and Wolf-Tilo Balke. 2022. It's the same old story! enriching event-centric knowledge graphs by narrative aspects. In *14th ACM Web Science Conf. 2022*, pages 34–43. ACM.

[Suchanek2020] Fabian M. Suchanek. 2020. The need to move beyond triples. In *Proc. of Text2Story@ECIR 2020*, volume 2593 of *CEUR Workshop Proc.*, pages 95–104. CEUR-WS.org.

[Vrandecic and Krötzsch2014] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.