Representation of critical discourses in the humanities within Wikidata

Di Pasquale, Alessio
alessio.dipasquale@studio.unibo.it
Department of Computer Science, University of Bologna, Italy

Pasqual, Valentina
valentina.pasqual2@unibo.it
Digital Humanities Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Tomasi, Francesca
francesca.tomasi@unibo.it
Digital Humanities Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Vitali, Fabio
fabio.vitali@unibo.it
Department of Computer Science, University of Bologna, Italy

Introduction

Providing digital representation of critical discourses in addition to and complementing traditional semantic annotations is a central topic in recent knowledge representation discussions (Ontolog, 2018). This includes, e.g., provenance information, evolving knowledge, and metadata versioning. The Cultural Heritage domain (CH) (and LOD datasets thereof (Wilensky 2000; Jia et al. 2014)) is exemplary of this problem, rich in incomplete data, subjective analyses, concurring statements and controversies between annotators. Accurately representing this complexity helps towards computational analyses of critical discourses in the Humanities. Wikidata (Erxleben et al. 2014) supports some complex representation of data, even allowing multiple and possibly competing assertions and versions of data. Within the activities of our research proposal (Daquino et al. 2022), we surveyed Wikidata approaches to represent complex knowledge: (1) ranked statements, (2) "nature of statements" qualifiers, (3) null-valued objects. In this work we examine whether complex knowledge about Cultural Heritage can be satisfactorily represented in Wikidata KB, and whether existing representation methods exhaustively represent CH domain.

Representing complexity in Wikidata

Wikidata represents uncertain or debated statements[1] with at least three different representation methods:

1. *Ranked statements*. Competing statements are represented via a ranking mechanism (e.g., Preferred, Normal and Deprecated). Individual statements are not actually asserted, but an extra triple is added those that are deemed true[2]. For example, the painting "Madonna with the Blue Diadem" (Q738038) has been attributed to Raphael (non asserted statement, ranked as normal) and Gianfrancesco Penni (asserted statement, ranked as preferred and additionally asserted).

Figure 1. Concurring statements for the attribution of "Madonna with the Blue Diadem". Ranking is shown in the icon on the left of the value, as up arrow (*preferred rank*), central dot (*normal*) and down arrow (*deprecated*)

1. *"Nature of statements" qualifiers*. Statements, independently of rank, can be decorated with an additional triple using predicate P5102. 54 terms among 283 available may mark the statement as uncertain or debated (e.g.*debated*, *hypothesis*, *possibly*). For example, the painting "Abstract Speed + Sound" (Q19882431) by Giacomo Balla is deemed to be *possibly* part of a triptych.

Figure 2. Uncertain statement about "Abstract Speed + Sound" using a *nature of statement* (P5102) predicate

1. *Null-valued objects*. A statement can be associated with a blank node[3]. This is meant to imply that the statement is associated with an unknown value, rather than a missing statement[4]. For example, "Missal for the use of the ecclesiastics of Clermont' (Q113302686), an illuminated manuscript from the 14th century, has been recorded with both an unknown creator and author.

Figure 3. Unknown creator and author of 'Missal for the use of the ecclesiastics of Clermont' expressed as *null values*

Analysing complex statements in Wikidata

In this paper we report on an evaluation of how much, how precisely and how satisfactorily have these three methods used to express complex knowledge in Wikidata. Our dataset collects 2 millionsworks of art from Wikidata and their descriptions. Even though critical analyses are a pivotal element in humanities discourses, we discovered that factual descriptive statements are largely the most represented information in the dataset: the vast majority of artworks (>99%) show plainly asserted statements.

Of the methods previously listed, ranked statements are largely the most frequent method for representing competing information (86.9%) if compared with the other surveyed approaches. Null-valued statements amounted to 11.6%, and anecdotal evidence from other domains seem to imply that at least some of them may be the result of poor conversions from empty fields in traditional relational databases. Nature of statement P5102 predicates are only in 1.5% of the dataset, with "possibly" and “presumably ” covering more than 50% of occurrences, and many others (e.g., "unconfirmed”) being present only once in the whole dataset.

A proposal for complex statements in Wikidata

All the methods examined are used to express complex and conflicting statements over the described entities (e.g., debates, conjectures, concurrent attributions). Yet all three methods show evident shortcomings in how they are used and in what they model.

Ranked statements are used to express both competing hypotheses and time-based changes of locations or ownerships. In addition, ranking alone does not justify the reason for the preference/deprecation of one statement over the others, and qualifiers[5] are scantily employed (< 10%).

Null-valued statements lack in terms of expressiveness (providing no justifications for their specifications) and precision (being also used as a fallback mechanism for poor data conversions).

Finally, P5102 statements do provide meaningful justifications, but they allow too many values (283 to our count) with frequent overlapping meanings (e.g., *hypothetically*, *hypothesis*) and unclear status as to the assertedness of the statement.

Both the W3C (Laskey at al. 2008) and ISIF[6] (Blasch et al. 2019) propose uncertainty categorizations for semantic datasets. Based on this, we categorised typical CH complex statements in four categories, “evolving knowledge”, “actual uncertainty/debate”, “other than preferred version”, and “other or not applicable” (table 1). In our dataset uncertainty is mainly used for agents involved in the artwork (e.g. creator, manufacturer), time (e.g. inception), category (e.g. type), and interpretation (e.g. subject, depicts), while evolving knowledge occurs mainly with locations.

Property	Label	Metadata category	Complex knowledge type	Method of representation	Number of occurrences
P50	author	Agents in roles	Actual uncertainty or debate	Nature of statement OR ranked statements OR null valued	Vast majority
P276	location	Locations	Evolving Knowledge	Ranked Statements	Mostly
P276	location	Locations	Actual debate or uncertainty	Nature of statement	Rarely
P1476	title	Titles	Other than preferred version	Ranked statements	Vast majority
P1476	title	Titles	Other or not applicable	Null valued	Rarely

Table 1. Selection of properties with their categorisation with respect to metadata areas and complex knowledge types

All in all, despite the limited presence of the "nature of statement" triples in the surveyed data, qualitative analysis shows this to be the most precise approach to mark uncertain or debated statements. Yet a fundamental aspect of ranked statements, besides being easy to understand, is that best ranks are actually asserted, and the others are not. Thus, it becomes easy to query and find asserted statements rather than those provided for completeness.

Conclusions

Wikidata offers several representations to complex statements (especially uncertainty and debates), but the approaches are too many and data use them in a fragmented and unreliable ways: fairly complex queries are necessary to retrieve complex statements. What is worrying, looking at data, is that Wikidata annotators are often reticent in providing complex information (such as evolving knowledge, actual uncertainty, or debate, and other than preferred version), possibly because of the many similar approaches, and no clear guideline exists to this end. Indeed, the categorisation exemplified in table 1 can be considered a representation of frequent phenomena in Cultural Heritage: providing clear guidelines on the best ways to represent them would have a beneficial effect on the richness of data provided and help reduce reticence. In particular, the analysis demonstrates that contextual information (motivation, provenance, certainty degree) are required in recording and retrieving disputed or subjective information. A richer dataset with more nuances would enhance computational analysis in the Humanities discourse.

Notes

1. Wikidata represents many types of complex assertions as instances of the Statement class to which related statements are associated. See https://www.wikidata.org/wiki/Help:Statements
2. See Wikidata example query at https://w.wiki/5pE$, and Rankings documentation at https://www.wikidata.org/wiki/Help:Ranking
3. See Wikidata example query at https://w.wiki/5pEz and Unknown or No values documentation at https://www.wikidata.org/wiki/Help:Statements#Unknown_or_no_values
4. The Wikdata interface represents these occurrences with the string "unknown value".
5. See for example “reason of deprecation” (P2241) and “reason of preferred rank” (P7452)
6. The Evaluation of Techniques for Uncertainty Representation Working Group (ETURWG) is an official activity of the International Society of Information Fusion (ISIF), https://eturwg.c4i.gmu.edu/

Bibliography

Blasch, Erik P. / **Insaurralde, Carlos C.** / **Costa, Paulo C. G.** / **de Waal, Alta** / **de Villiers, Johan Pieter** (2019): “Uncertainty Ontology for Veracity and Relevance”, in: 2019 22th International Conference on Information Fusion (FUSION). pp. 1-8.

Daquino, Marilena / **Pasqual, Valentina** / **Tomasi, Francesca** / **Vitali, Fabio** (2022): “Expressing Without Asserting in the Arts”, in: IRCDL.

Erxleben, Fredo / **Günther, Michael** / **Krötzsch, Markus** / **Mendez, Julian** / **Vrande#i#, Denny** (2014): “Introducing Wikidata to the Linked Data Web”, in: The Semantic Web – ISWC 2014. Cham: Springer International Publishing, pp. 50-65. DOI: 10.1007/978-3-319-11964-9_4x.

Jia, Yuan / **Niu, Xi** / **Bharali, Reecha** / **Bolchini, Davide** / **De Tienne, Andre** (2014): “Collaborative Online Research Platform for Scholars in Humanities”, in: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. New York, NY,

USA: Association for Computing Machinery, pp. 181-184. DOI: 10.1145/2556420.2556507.

Laskey, Kenneth J. / **Laskey, Kathryn B.** (2008): “Uncertainty Reasoning for the World Wide Web: Report on the URW3-XG Incubator Group”, in: URSW.

Ontolog (2018): “Ontology Summit 2018 Communiqué. Contexts in Context”. http://ontologforum.org/index.php/OntologySummit2018

Wilensky, Robert (2000): “Digital Library Resources as a Basis for Collaborative Work”, in: Journal of the American Society for Information Science 51(3), pp. 228-245. DOI: 10.1002/(SICI)1097-4571(2000)51:3<228::AID-ASI3>3.0.CO;2-5.