On assessing weaker logical status claims in Wikidata cultural heritage records

Alessio Di Pasquale [a], Valentina Pasqual [b,*], Francesca Tomasi [b] and Fabio Vitali [a]

[a] *Department of Computer Science, University of Bologna, Italy*
E-mails: alessio.dipasquale@studio.unibo.it, fabiovitali@unibo.it

[b] *Department of Italian Studies and Classical Philology, University of Bologna, Italy*
E-mails: valentina.pasqual2@unibo.it, francescatomasi@unibo.it

Abstract. This work presents an analysis of the use of different representation methods in Wikidata to encode information with weaker logical status (WLS, e.g. uncertain information, competing hypothesis, temporally evolving information, etc.). The study examines four main approaches: non-asserted statements, ranked statements, non-existing valued objects, and statements qualified with properties `P5102`:*nature of statement*, `P1480`:*sourcing circumstances* and `P2241`:*reason for deprecated rank*. We analyse their prevalence, success, and clarity in Wikidata. The analysis is performed over cultural heritage artefacts stored in Wikidata divided into three subsets (i.e. visual heritage, textual heritage and audio-visual heritage) and compared with astronomical data (stars and galaxies entities). Our findings indicate that (1) the representation of weaker logical status information is limited, with only a small proportion of items reporting such information, (2) the representation of WLS varies significantly between the two datasets, and (3) precise assessment of WLS statements is made complicated by the ambiguities and overlappings between WLS and non-WLS claims allowed by the chosen representations. Finally, we list a few proposals to simplify and standardize the representation of this type of information in Wikidata, with the hope of increasing its accuracy and richness.

Keywords: Wikidata, ranked statements, weaker logical status, uncertainty, cultural heritage

1. Introduction

Since 2012 Wikidata [1] has been one of the most outstanding platforms for collecting and sharing Linked Open Data through the web.

Through the years Wikidata has developed and provided a variety of representation methods that allow it to encode complex structures much beyond factual descriptive metadata. According to [2], Wikidata encompasses a multitude of facts, including some that may be contrasting since they come from different and disagreeing sources. Additionally, time-sensitive information can also be added through the use of qualifiers and ranks. For instance, structures to represent temporally evolving information (e.g., the number of followers of a YouTube Channel that is updated year after year) or multiple coexisting (and possibly competing) claims over the same subject (e.g., maintaining both the old as well as a new theory over some topic). In many such cases, multiple information items are present, yet newer or better information is not replacing older or less true assertions, but they coexist next to each other, and one or more mechanisms are used to signal their simultaneous presence, and, when appropriate, the currently adopted stance.

[*] Corresponding author. E-mail: valentina.pasqual2@unibo.it.

We understand these statements as enjoying a somehow *weaker logical status* than simply asserted statements: they are neither true nor false, but they are, e.g., true from a specific moment onward but not earlier, or true up to a given moment but not afterwards, or accepted as true by most people but not everybody, etc.

It is a cultural necessity in many (if not all) fields of knowledge to have access to available data about a complex topic completely and objectively, as they evolve, as they are interpreted by different scholars or models, as they represent available hypothesis rather than a positive certainty. For instance, cultural heritage scholars study attributions, the temporal context of events, the temporal evolution of content, and the contradictions of opinions and assertions, so that expressing weak statements, i.e., claims we are not certain about, becomes a necessary tool to increase precise awareness of the currently available data for those who consult or reuse it. Interpretation thus plays a central role in humanities disciplines. Yet, cultural heritage knowledge graphs and domain ontologies frequently limit the formalisation of these phenomena or only partially represent them ([3, 4], cf. section 2). Recently, a rekindled interest has been shown in the formalisation of uncertain statements [5–7], claiming that interpretation constitutes a focal point in humanities data and metadata. Interestingly, these works prove how different motivations for nuanced statements with different degrees of truths end up creating a small and consistent number of approaches to express them, and we conclude that studying the very idea of weak logical status claims *per se*, independently of their different justifications, can help shed some lights over these commonalities and their relative merits and issues. WLS claims are used not only for missing or incomplete information but also for the correct representation of personal opinions or beliefs, for temporally constrained information, for geographically constrained information, etc.

Wikidata supports several patterns to represent situations best expressed with weaker logical status claims. In this paper, we analyse some of these patterns as they are employed in actual collections, both in the humanities and, as a comparison, in hard sciences. A factor that increases complexity is that many of these uses have partially overlapping semantics, i.e., they can be used also for other purposes beyond weaker logical status claims, and this muddles the correct identification and interpretation of the situations we are interested in. We, therefore, want to discuss both the expected application of each approach and its relative success, as well as the impact of their ambiguous applications due to the coexistence of multiple uses for the same techniques.

In particular, we analysed four main families of approaches to the weaker logical status of statements, *asserted vs. non-asserted statements*, *ranked statements*, *unknown objects* and *qualified statements*. In this paper, we try to answer the following research questions:

- (RQ1) How widespread and successful is each of these approaches in the current state of Wikidata?
- (RQ2) How does the cultural domain of the Wikidata topics (and, presumably, of the individuals contributing to the data regarding the entities) affect and reflect on the relative success and richness of some approaches over others?
- (RQ3) How clean and easy to differentiate are the applications of each approach to an actual weaker logical status versus to another of the designed uses of that approach?
- (RQ4) Is there a way to improve the clarity and cleanliness of such differentiation?

To perform such analysis, we accessed and downloaded two large sets of topics from Wikidata, one belonging to the cultural heritage (visual works of art such as paintings and statues, text documents and audio-visual entities), and another from astronomy (celestial bodies such as stars and galaxies). Both make some use of multiple fuzzy assertions and hypotheses and therefore need assertions with weaker status (e.g., attributions uncertainties or physical locations moving over time for paintings, vs. spectral class or radial velocity for stars).

The decision to use a comparative dataset in this study is motivated by the wish to explore the similarities and differences between astronomical and humanities academic practices. Both fields involve studying unique objects, such as stars or books. Yet, the way data is treated differs, with astronomical observations becoming scientific data as soon as they are used as evidence of phenomena [8], while humanities rarely can go beyond learned interpretations.

The sources of data also vary, with humanities researchers using historical documents, literature, art, and oral traditions, each having varying levels of reliability and introducing systemic and insurmountable uncertainty. In astronomy, uncertainty is often related to instrumental limitations and observational conditions. Methodologically, astronomy relies on empirical observation, mathematical modelling, and experimental validation, while humanities research is frequently interpretative, and qualitative and the necessary proof to obtain historical certainty is often unattainable [9]. This difference leads to distinct epistemological foundations, with the humanities acknowledging the subjectivity and cultural bias in interpretations [10], and astronomy seeking to minimize uncertainty through rigorous data collection and adherence to physical principles [11].

Our study's hypotheses and assumptions include the idea that annotators in cultural heritage and astronomy may approach data incompleteness and uncertainty in different ways, with cultural heritage favouring qualitative, context-rich representations of competing hypotheses and astronomy leaning towards more quantitative, data-centric representations. This difference may reflect broader epistemological stances in their respective communities. Additionally, our study assumes that these distinct approaches to handling data incompleteness and uncertainty may impact the ease of integrating data from these fields in interdisciplinary research, with cultural heritage data potentially requiring more effort for reconciliation due to its contextual and subjective nature.

Overall our findings show that the amount of weaker logical status statements in Wikidata seems suspiciously low, as only 0,4% of visual artworks report attribution disputes, a fairly low figure compared to, e.g., a more reasonable 8,5% coming from the RKD images collection[1], a difference that could be attributed to the difficulty and ambiguities in the procedures to report such complex information. We propose here also a way to simplify, streamline and homogenize such complexity, with the hope of increasing the abundance, richness and correctness of the representation of such phenomena in Wikidata.

The paper is structured as follows: in the state of the art (2) relevant data sources KGs and data models are presented when representing weaker logical status claims as well as schema and data assessments proposed over Wikidata. In section 3 we present the approaches provided in Wikidata to encode weaker-logical status claims. In section 4 the research objectives are outlined, the data acquisition process is briefly described and the analysis of our Wikidata sample dataset is presented. In section 5 we present our proposal for improving the quality of annotating weaker-logical status knowledge. Finally, in section, 6 we summarize our findings and outline our conclusion about the work.

2. State of the art

Public Knowledge Graphs such as Wikidata [1], DBpedia [12], Yago [13], and Google Knowledge Graph constitute publicly available collections that can be used for research, either expressing specialist knowledge or general knowledge. In particular, Wikidata is a *collaborative* public platform, built and maintained by a community of contributors.

Weaker logical status statements are a natural occurrence in many contexts covered by these KGs, but the support for their representations varies considerably. Guidelines, data modelling and data harmonization (a particularly relevant need for open platforms) can help in expressing them, i.e. for concurrent opinions or uncertain claims. In the field of cultural heritage studies, the knowledge competition is intriguing. However, some online databases or data models only partially address this issue.

Despite domain ontologies representing the cultural heritage domain hardly managing to integrate support for interpretation (i.e., hermeneutics) into their models [5], there are some exceptions [4, 14].

For instance, CIDOC CRM [4] is a conceptual model, developed and maintained by the International Council of Museums (ICOM), widely adopted by many knowledge graphs in the cultural heritage domain [15, 16]. It offers a formal approach to express weaker logical status claims through the use of instances of classes representing n-ary relations.

Europeana [14] stores approximately 50 million heterogeneous digitized items from museums, libraries, and archives across Europe. Data is collected by content providers (i.e. cultural institutions) using the EDM data model [3] and the use of proxies [17] allows to express conflicting information and track data provenance. However, concurrent statements are not visible on the online pages, and no mechanism is in place to determine which proxy will be made visible when multiple exist.

An interesting instance of an EDM collection is the RKD catalogue, a comprehensive collection of data about Dutch works of art throughout history. By design, RKD allows and gathers contested and discarded attributions of

[1] https://rkd.nl/en/explore/images

paintings and portraits. Although at the moment there is no SPARQL endpoint available for querying the collections, the data can be explored through an online catalogue. Interestingly, about 83.600 artwork descriptions from RKD[2] have been imported into Wikidata, representing ~7,5% of the total of visual artworks in Wikidata.

Despite the support of representational definitions of weaker logical status claims in EDM, CIDOC-CRM and RDK data models, these weaker forms of information are often poorly reported (*reticence*) or are expressed in textual annotations rather than being modelled in the data structure (*dumping*) [18].

The widespread adoption of Wikidata within the cultural heritage community has been well-documented [19][3]. Wikidata is seen not only as a valuable tool for data publishing, alignment and enrichment but also as a means of gaining valuable insights into cultural heritage data and the community itself [20]. Given the significance of comprehensive data in knowledge bases, there has been a focus on improving and evaluating their schema and data quality [21]. In this context, weaker logical status claims may make good use of reification approaches and several studies have been performed to improve their usage e.g., by [22], who compared the efficiency of several reification methods (e.g., singleton properties, n-ary relations, named graphs and standard reification) on Wikidata data.

A widespread adoption of n-ary relationships for WLS claims is provided by CIDOC-CRM itself [4], e.g., via the `crm:E13_AttributeAssignment` class[4]. For example, the painting "Girl reading a letter at an open window"[5], has been attributed over time to Rembrandt, Hooch, and finally to Vermeer (the currently accepted attribution). Listing 1 shows each attribution (`:aa1`, `:aa2` and `:aa3`) as a `crm:E13_AttributeAssignment` which requires the use of three predicates: `crm:P140_assigned_attribute_to` to indicate the item to which an attribute or relation is assigned, and `crm:P141_assigned` to indicate the attribute that was assigned or the item, `crm:P177_assigned_property_of_type` to indicate type of property or relation that this assignment maintains to hold between the item to which it assigns an attribute and the attribute itself.

```
:aa1 a crm:E13_Attribute_Assignment ;
  crm:P177_assigned_property_of_type crm:P14_carried_out_by ;
  crm:P141_assigned ulan:500011051 ; # Rembrandt
  crm:P140_assigned_attribute_to :painting-pr ;
  crm:P4_has_time-span :XVIII_cent.

:aa2 a crm:E13_Attribute_Assignment ;
  crm:P177_assigned_property_of_type crm:P14_carried_out_by ;
  crm:P141_assigned ulan:500020229 ; # Hooch
  crm:P140_assigned_attribute_to :painting-pr ;
  crm:P4_has_time-span :1821.

:aa3 a crm:E13_Attribute_Assignment ;
  crm:P177_assigned_property_of_type crm:P14_carried_out_by ;
  crm:P141_assigned ulan:500032927 ; # Vermeer
  crm:P140_assigned_attribute_to :painting-pr ;
  crm:P4_has_time-span :1860;
  crm:P14_carried_out_by ulan:500326948. # Thore
```

Listing 1: CIDOC CRM use of n-ary relation for encoding concurring attributions

Handling WLS in Semantic Web data can be placed within the larger topic of representing and reasoning over data enriched with metadata, or contextualized data. The topic has been discussed at length from many different angles. A major objective is that of reconciliation or integration of multiple data sources. Indeed, effective representation and reasoning about knowledge with heterogeneous viewpoints is one of the objectives for applications concerned with distributed knowledge sources. Yet, semantic web ontologies force a unique, global view of the represented world, in which the axioms are meant to be interpreted as universally true. Often, though, the same domains are modelled differently depending on the intended use of an ontology. The problem of reconciliation, therefore, is to be able to bring different world views together to create a single, unified model for representation and reasoning.

[2] https://w.wiki/7wfW

[3] The list of cultural institutions involved in Wikidata can be found at http://www.wikidata.org/entity/Wikidata:GLAM

[4] https://cidoc-crm.org/Entity/e13-attribute-assignment/version-6.2.1

[5] https://en.wikipedia.org/wiki/Girl_Reading_a_Letter_at_an_Open_Window

This may be obtained through formal Interoperability Systems [23] extending the expressive reach of Description Logic, or bridge rules mapping separate contexts determining how the local concepts in the two ontologies map onto each other [24], or extended representation models such as RDFS with Annotations [25]. Other approaches, such as colouring [26] or NDFluents [27], or or RDF+ [28] on the other hand, are less interested in obtaining reconciliation and more with representing adequately the semantics of inferences about heterogeneous claims.

The representation of complex data scenarios in knowledge bases often needs to be evaluated according to multiple metrics. For instance, Piscopo and Simperl [29] survey quality metrics from 28 scientific publications on the topic and categorize quality assessments into three dimensions: intrinsic (accuracy, trustworthiness, consistency), context (relevance, completeness and timeliness) and representation (ease of understanding and interoperability). Among quality measures, evaluation of completeness, defined in [30] as the "presence of all required information in a given dataset", has been approached through various methods and assessments as comparing data for similar entities [31], measuring entity relatedness [32], evaluating thoroughness of information by determining the completeness of specific attributes of objects [33], assessing low-quality statements thought the analysis of items' discussion pages, deprecated statements and constraint violations [34], and assessing and comparing data quality across large knowledge bases [30, 35].

Overall, little or no evaluation has been conducted on the representation of weaker logical status claims in Wikidata, nor has a comprehensive analysis been carried out to assess the amount of knowledge related to WLS status in the field of cultural heritage. In the next section, we detail our proposal to address these shortcomings.

3. Representing weaker logical statuses in Wikidata

Wikidata represents weaker logical status statements (e.g. for uncertain or debated assertions) using at least three different representation methods: ranked statements (section 3.1), statements with specific qualifiers (section 3.2) and statements with a non-existing valued object (section 3.3).

3.1. Ranked statements

Ranking of assertions is modelled by the Wikibase data model[6] to express different degrees of the preferability of individual claims.

Claims in Wikidata are expressed through *statements*, a custom reification method[7] to express contextual information (e.g. qualifiers, rankings, references) about it. Statements connect the claim's subject and the claim's predicate to a Statement entity which refers to the claim's object and can be further used as the subject of other triples.

Statements do not assert the corresponding claim. To do so another triple must be added that (using a different prefix) flatly relates the Statement's subject to the Statement's intended object through the Statement's predicate, thus enabling simple query support for asserted facts. The separation between Statements and their assertion is selectively provided, which allows to easily support both claims presented as facts (where both the Statement and the assertion triple exist) as well as claims not meant to be considered facts (the Statement exists, but no assertion triple is added).

The ranking mechanism is enriched with the representation of asserted and non-asserted statements. Rankings [36] communicate the consensus opinion for a statement as reached by the scientific community or Wikidata annotators. Disputes are separately hosted on the corresponding discussion page, in plain text. Many possible combinations of variously ranked competing statements can be found in the Wikidata collection, with various and debatable interpretations. Ranking is assigned to individual Statements using values such as *Preferred*, *Normal* and *Deprecated*).

For sure, whether or not a statement is asserted is determined solely by the statement's rank and the absence of higher-ranked statements using the same predicate – it is automatically provided by the Wikidata engine and is not a conscious choice of the editors.

[6] https://www.mediawiki.org/wiki/Wikibase/DataModel#Statements
[7] http://www.wikidata.org/entity/Help:Statements

3.1.1. Normal statements

The Normal ranking is the default ranking for Statements. A Statement ranked Normal can be either asserted or not depending on the existence and intended meaning of competing Statements that are placed against it. For instance, in listing 2, "The Scream" by Edvard Munch belongs to the Expressionist period[8], and this is expressed as an asserted Normal Statement, to signify that the annotator does not give a WLS status to the statement. In listing 4, on the other hand, the first Statement (lines 1-5) is ranked Normal but not asserted, since the Preferred statement is present and asserted instead.

```
# "The scream" belongs to the Expressionist movement
wd:Q471379 wdt:P135 wd:Q80113 .
wd:Q471379 p:P135 s:Q471379-c3e5c17d-4730-a5dc-85cb-efc9766b7c80 .
s:Q471379-c3e5c17d-4730-a5dc-85cb-efc9766b7c80 a wikibase:Statement,
  wikibase:rank wikibase:NormalRank ;
  ps:P135 wd:Q80113 .
```

Listing 2: Normal rank

3.1.2. Deprecated statements

Deprecated statements are meant for claims with a weak logical status and do not represent a correct value in the view of the editors. Deprecated statements are always automatically non-asserted independently of the ranking of the other concurring statements. For example, listing 3 expresses the concept that "The Lamentation"[9], a print by Albrecht Dürer, was reported to be created in 1504. The Deprecated rank and the lack of an assertion triple indicates that this date is not thought to be valid.

```
# creation date thought to be 1504
wd:Q18338462 p:P571 s:Q18338462-FDDCD91B-3919-450A-B00D-FE3ADA773A11 .
s:Q18338462-FDDCD91B-3919-450A-B00D-FE3ADA773A11 a wikibase:Statement ;
    wikibase:rank wikibase:DeprecatedRank ;
    ps:P571 wdt:P571 "1504-01-01T00:00:00Z"^^xsd:dateTime .# creation date: 1504
```

Listing 3: Deprecated rank

3.1.3. Preferred statements

Preferred statements are meant for claims with a stronger status and representing the currently presumed correct value of a predicate. They are always also asserted. For instance, as shown in listing 4, a retracted attribution of the painting “Madonna with the Blue Diadem” [10] to Raphael is represented only by a Statement ranked as Normal and no assertion triple, while the attribution to Gianfrancesco Penni enjoys both a Preferred rank and the assertion triple.

Even though the first attribution is ranked Normal rather than Deprecated, we must consider it as a superseded claim. This example shows that the nature of Normal statements varies depending on whether they coexist or not with competing Preferred and/or Deprecated claims, and similarly may vary the presence or absence of assertion triples.

```
1      # attribution to Raphael
2      wd:Q738038 p:P170 s:q738038-121B92D0-E6E1-4514-960C-AE34F50054E5 .
3      s:q738038-121B92D0-E6E1-4514-960C-AE34F50054E5 a wikibase:Statement ;
4          wikibase:rank wikibase:NormalRank ;
5          ps:P170 wd:Q5597 .                # creator: Raphael
6
7      # attribution to Gianfrancesco Penni
8      wd:Q738038 wdt:P170 wd:Q2327761 .    # creator: Gianfrancesco Penni (assertion)
```

[8] http://www.wikidata.org/entity/Q471379

[9] http://www.wikidata.org/entity/Q18338462

[10] http://www.wikidata.org/entity/Q738038

```
wd:Q738038 p:P170 s:Q738038-7729b786-4d4f-a0ca-2ded-4ea2c6307e1c .
s:Q738038-7729b786-4d4f-a0ca-2ded-4ea2c6307e1c a wikibase:Statement;
    wikibase:rank wikibase:PreferredRank ;
    ps:P170 wd:Q2327761.            # creator: Gianfrancesco Penni
```

Listing 4: Preferred and Normal ranks

3.2. *Qualifiers*

Statements, independently of rank, can be decorated with additional triples annotating contextual information or specifications about the claim itself[11]. Those annotations may be *additive* when they provide additional information about the fact (e.g., to specify the character played by an actor when listing him or her as a cast member of a movie) or *contextual* when they limit the contexts in which the underlying fact is true (e.g., the claims is a hypothesis) [37].

Following the example from [38] we examined the 150 most frequently used qualifiers in Wikidata and their most frequently used values. The most used qualifiers to use WLS values are `P1480`:*sourcing circumstances*[12] (47th most used one) and `P5102`:*nature of statement*[13] (134th most used one). Additionally, the Wikidata model provides the properties `P2241`:*reason for deprecated rank*[14] (42th most used qualifier) and `P7451`:*reason for preferred rank*http://www.wikidata.org/entity/Property_talk:P7451 (114th most used qualifier) to annotate contextual information about superseded and preferred claims, respectively.

For example, in listing 5 we see that the painting “Abstract Speed + Sound”[15] by Giacomo Balla is described as *possibly* part of a triptych. The use of a qualifier with a Normal ranking seems to imply that the statement is considered true and therefore it is also asserted.

```
wd:Q19882431 wdt:P361 wd:Q79218 .     # part of: triptych (assertion)
wd:Q19882431 p:P361 s:Q19882431-1ac26ff2-4981-ff79-4fae-9d411ae34296 .
s:Q19882431-1ac26ff2-4981-ff79-4fae-9d411ae34296 a wikibase:Statement;
    wikibase:rank wikibase:NormalRank ;
    ps:P361 wd:Q79218 ;               # part of: triptych
    pq:P5102 wd:Q30230067 .           # circumstance: possibly
```

Listing 5: A qualified statement in Wikidata

Wikidata provides a list of 96 recommended values for *nature of statement* and 83 recommended values for *sourcing circumstances* in their respective *Property Talk* pages, while no list of recommended terms is provided for *reason for deprecated rank* nor *reason for preferred rank*. However, terms that were used with these properties can be retrieved via a simple SPARQL query[16], showing respectively 384 and 83 distinct terms. Even at first glance, it is possible to notice a very wide range of types and specificities (e.g., qualifiers such as *possibly*, *presumably*, and *probably* versus, say, *prosopographical phantom*, *project management estimation* or *archive footage*), and many are not connected to weaker logical status assessments. In addition, semantic overlaps can be noticed on many of these terms, e.g. between *allegation* and *allegedly*, or between *hypothesis*, *hypothetical entity*, *hypothetically* and *scientific hypothesis*. These overlaps support arbitrariness of choice for contributors, increasing the ambiguity of the resulting annotation.

[11] The complete list of available qualifiers in Wikidata is available at https://w.wiki/6TrP

[12] The most frequently used values are: *circa, presumably, allegedly, inference, uncertainty, possibly, near, probably, conventional date, disputed*

[13] The most frequently used values are: *originally, attribution, hypothesis, often, allegedly, expected, possibly, disputed, rarely, mainly*

[14] http://www.wikidata.org/entity/Property_talk:P2241

[15] http://www.wikidata.org/entity/Q19882431

[16] List of terms used in Wikidata with *reason for deprecated rank https://w.wiki/6Tpt* and with *reason for preferred rank* https://w.wiki/7VGf

3.3. Missing values

There are three types of basic information structures used to describe Entities in Wikibase (called SNAKs, or *Some Notation about Knowledge* [17] in Wikidata: actual values (URIs or literals), `someValue` placeholders and `noValue` placeholders. They are used to represent that the statement is associated with an unknown value (mapped as `someValue`) or with a non-existing value (mapped as `noValue`), which is a more precise assessment than simply not recording the statement at all. The use of the same syntactic tool is known to generate precision and correctness issues (e.g., see [39]), since the RDF standard specifically defines blank nodes with an existential semantics while SPARQL does not follow such semantics. As a result, SPARQL queries run over datasets where blank nodes are used as existentials to represent unknown values, the results can be unintuitive or arguably incorrect. Although the RDF representation of Wikidata uses blank nodes for both unknown and non-existing values, a skolemization process separates them conceptually and with a simple filtering query it is possible to distinguish them[18].

3.3.1. Unknown values

Unknown valued statements are claims whose object exists but is not known[19]. For instance, in "The Book of Lismore"[20] there is an unknown value for the `P195`:*collection* property, which is a positive statement that the information existed but it has not been preserved. As mentioned, in the RDF representation unknown values are represented via blank nodes as shown in listing 6.

```
 wd:Q1371647 wdt:P195 _:15518d67963a082b352304a1ab8e016e. # unkown collection
 wd:Q1371647 p:P195 s:Q1371647-B07F6386-A7D0-4C9D-8E77-CC2BD523354E .
 s:Q1371647-B07F6386-A7D0-4C9D-8E77-CC2BD523354E ps:P195 _:0088bc50e53b3902bea74cc2380cbd09 ;
pq:P3831 wd:Q768717 . # the role of this collection is to be a private collection
```

Listing 6: Unknown-valued statement in Wikidata

3.3.2. Non-existing values

Non-existing valued statements[21] are claims whose object is not existent (or not available in Wikidata). For example, the pilot episode of X-files[22] has a non-existing value for the *follows* (`P155`) property, considering that the pilot starts the series. Non-existing values are not conceptually a WLS claim, but we list them in this survey because there exists in practice some overlap between unknown valued and non-existing valued claims. For example, the "Missal for the use of the ecclesiastics of Clermont'[23], an illuminated manuscript from the 14th century, has been recorded with both an unknown creator and unknown author, as shown in listing 7. The example is incorrect as it should use an unknown value, and this leads to confusion about the usage of missing values, further contributing to complications.

```
wd:Q113302686 wdt:P50 _:4c60f23d697d2d89d9fe49824c8f3a01 .   # author: unknown (blank node - assertion)
wd:Q113302686 p:P50 s:Q113302686-032e3cc5-4fd6-1f20-8830-0909945ba683 .
s:Q113302686-032e3cc5-4fd6-1f20-8830-0909945ba683 a wikibase:Statement;
    wikibase:rank wikibase:NormalRank ;
    ps:P50 _:f8c6b698b13ef3dd3738e025df3a2d5d .              # author: unknown (blank node)

wd:Q113302686 wdt:P170 _:759d5c5c7a58a8a286512c257514463a .  # creator: unknown (blank node - asserted)
wd:Q113302686 p:P170 s:Q113302686-8d47e883-4566-bc8b-cd8f-6cffebc5414c .
s:Q113302686-8d47e883-4566-bc8b-cd8f-6cffebc5414c a wikibase:Statement;
    wikibase:rank wikibase:NormalRank ;
    ps:P170 _:28d04a432a3589d30a5c6da79d3fac50 .             # creator: unknown (blank node)
```

Listing 7: non-existing valued statement in Wikidata

[17] https://wikidata.github.io/Wikidata-Toolkit/org/wikidata/wdtk/datamodel/interfaces/Snak.html
[18] https://www.mediawiki.org/wiki/Wikibase/DataModel#PropertysomeValueSnak
[19] https://www.wikidata.org/wiki/Help:Statements#Unknown_or_no_values
[20] https://www.wikidata.org/wiki/Q1371647
[21] https://www.wikidata.org/wiki/Help:Statements
[22] http://www.wikidata.org/entity/Q7194381
[23] http://www.wikidata.org/entity/Q113302686

3.4. Discussion

Even before checking on the actual usage patterns of these methods, we can immediately notice the richness of annotations made possible by them, the subtle nuances they afford, but at the same time the variety of (potential) sources of ambiguities, overlapping connotations and representation vagueness. In particular, we can summarise three specific problems that are worth further discussion:

1. Although the separate uses of Normal, Preferred and Deprecated rankings are clear and practical, there are uncertainties when they coexist on the same predicate, especially for the different representations of Normal statements when Preferred ones are also present, or when all three rankings are present.
2. The sheer number of qualifiers, the differing levels of their respective specificities, and the manifest semantic overlapping of many of them make it quite hard to guarantee homogeneity and precision in their use. The use of contextualizing qualifiers, be they temporal, provenance or otherwise, does not add to the base information, but changes the context within which such information is true. As [37] suggests, contextual qualifiers should not be shown to consumers, but basic tools (visualizers, contextualisers, reasoners) should be written to correctly take such context into account or low-level tools should remove facts that are not valid in the selected contexts.
3. The subtlety in the semantic differences between providing no statement, specifying a `noValue` blank node and providing a `someValue` blank node for a property of a Wikidata item, as well as their other types of applications makes the use of missing values complicated and ambiguous.

In a way, WLS claims can be seen simply as logical disjunctions of competing claims each of which is separately annotated with context, provenance, confidence, temporal boundaries, etc.: "according to α, s p o_1" and "according to β, s p o_2" can be seen as "$[s\ p\ o_1]_\alpha \vee [s\ p\ o_2]_\beta$" with some added annotations connecting the first branch to α and the second to β (e.g., through reification, named graph, or blank nodes). This approach has limitations both from the practical and the conceptual point of view. Practically, RDF has no real way to express disjunctions without some additional baggage to encode predicate calculus employing the systematic use of reification [40]. Conceptually, focusing on the inner statements to the exclusion of the contextualizing information may miss the point that in many scholarly domains, it is not the full list of competing claims to be of interest, but the very existence of the diatribe in the first place. Disjunctions would not help here.

Another way to formally understand WLS claims is to link them to modal statements in modal logic [41], which can be used to understand the coexistence of strong logical status claims, expressed as atomic formulas $p(s,o)$, and weak logical status ones, expressed as modal formulas $K_\alpha p(s,o)$ or $B_\beta p(s,o)$, where K_α and B_β are modal operators guided by specific modal axioms[24]. Various types of modal logics exist and have been used to introduce different operators and represent different semantics, such as possibility and necessity (the *strictu sensu* modal logic), or obligation and permission (*deontic logic*), or temporally bounded predicates (*temporal logic*), or belief (*doxastic logic* or knowledge (*epistemic logic*). Overall, they form a complete formal mechanism to study the characteristics and principles of WLS claims that does not imply the need to proceed to a reconciliation of different world views.

Yet, all these reflections are somehow empty and pointless unless we examine how contributors are using these methods for expressing real WLS claims in their Wikidata contributions. This topic is covered in the next section.

4. Usage patterns of WLS in Wikidata datasets

To generate some analysis about the actual usage of WLS claims, and to provide an initial answer to our research questions, we collected three datasets of Wikidata items, one about Cultural Heritage items (visual arts, text documents and audio-visual entities), another about Astronomical objects (galaxies and stars) and one with a selection of random entities reflecting the actual distribution of entities in classes in the whole Wikidata as discussed in section 1.

The datasets were selected to be approximately comparable in size and the number of individual statements, and under evidence that many types of entities rely on weaker logical status claims when entities undergo re-evaluations due to new pieces of evidence or the recording of different opinions.

[24] e.g., **T** ($K_\alpha\phi \rightarrow \phi$) for epistemic logic or **N** ($\vdash \phi \implies \vdash B_\alpha\phi$) for doxastic logic.

4.1. Data Acquisition

```
SELECT DISTINCT ?artwork ?type
WHERE {
    ?artwork wdt:P31 ?type.
    ?type (wdt:P279*) wd:Q838948.
    hint:Prior hint:rangeSafe true
    }
```

Listing 8: SPARQL query retrieving Wikidata entities to subclasses of work of art (`Q838948`)

The first dataset contains Cultural Heritage items (CH), a complete snapshot of the entirety of Wikidata records of these cultural assets, under the assumption that discipline-oriented datasets give access to domain-specific annotation habits of scholars better than a sample of random entities. All Wikidata entities belonging to the class *work of art* [25]) or any of its sub-classes were collected using a SPARQL query (listing 8). The statements for all selected entities were downloaded in JSON format[26]. Data is stored in numerous files in JSON format and each file contains a complete representation of at most 50 Wikidata entities with their labels, descriptions and statements. This Cultural Heritage dataset has been semi-automatically divided into three sub-datasets, due to the wide diversity of cultural properties and their associated claims:

- *Audio-Visual heritage* (CHav): This collection holds information about audio-visual materials that have cultural, historical, or artistic value. They include movies, videos, recordings of music or spoken words, and other audio-visual materials that provide a record of a particular event in a specific time or place. The dataset contains 1.251.626 entities and 17.141.394 statements organized in 25.033 JSON files.
- *Visual heritage* (CHv): This collection holds information about visual artefacts that have cultural, historical, or artistic value. They include paintings, drawings, sculptures, photographs, decorative arts, etc. The dataset contains 1.078.855 entities and 12.850.825 statements organized in 21.579 JSON files.
- *Textual heritage* (CHt): This collection holds information about written and printed materials that have historical or cultural significance. They include books, manuscripts, letters, and other written documents. The dataset contains 625.110 entities and 4.584.444 statements organized in 12.503 JSON files

We also downloaded Wikidata entities of architecture-related classes; they were later discarded due to their fairly lower number as well as for the presence of many statistical ambiguities that could make their evaluation useless (e.g., many entities belonging to these classes should not be considered relevant to cultural heritage collections).

The second dataset, chosen to verify our assumptions using a different collection with a similar size, is a collection of astronomical entities, organized into two datasets:

- *Stars* (ANs): This collection holds a random selection of 1.199.950 Wikidata entities (of the ~3.3 million existing) belonging to the class *Star*[27], The dataset contains 27.470.140 statements in 23.999 JSON files[28].
- *Galaxies* (ANg): This collection holds a random selection of 1.200.000 Wikidata entities (of the ~2 million existing) belonging to the class *Galaxy*[29], The dataset contains 14.439.421 statements in 24.000 JSON files.

We decided to limit the number of astronomical entities to 1.200.000 to approximately balance them to each other (although the CHt is about half in size with 625.110 entities), as well as the average number of statements for each entity (CHav: 13,7, CHv: 11,9, CHt: 7,3, ANs: 22,9, ANg: 12).

[25] http://www.wikidata.org/entity/Q838948

[26] via http://www.wikidata.org/entity/Wikidata:Data_access

[27] http://www.wikidata.org/entity/Q523

[28] the ANs dataset was meant to be composed of 24.000 files with 50 entities each, but after running our tests we noticed that a file was corrupt and we chose to simply discard that contribution.

[29] http://www.wikidata.org/entity/Q318

	Cultural Heritage			Astronomy		
	Audio-visual (*CHav*)	Visual (*CHv*)	Textual (*CHt*)	Stars (*ANs*)	Galaxies (*ANg*)	Random (R)
Entities	1.251.626	1.078.855	625.110	1.199.950	1.200.000	1.159.800
Statements	17.141.394	12.850.825	4.584.444	27.470.140	14.439.421	61.798.072
Weaker Logical Status (*WLS*)	**50.193**	**227.218**	**17.216**	**7.532.169**	**721.504**	**1.101.014**
% WLS / Statements	**0,29%**	**1,77%**	**0,37%**	**27,42%**	**5,00%**	**1,78%**
Non-asserted statements	43.211	9.056	14.055	7.532.107	721.503	1.089.469
Ranked as Deprecated	7.622	3.057	1.568	2.768.829	189.691	721.870
Deprecated with a reason	4.949	769	715	2	0	8.993
Non-existing values	50.611	1.969	1.356	4	0	3.857
Unknown value	4.896	106.521	1.843	0	0	5.139
Qualified statements	2.406	114.674	1.556	532	1	7.716
WLS qualified statements	2.086	111.641	1.318	62	1	6.406
WLS qualifiers w/o *circa*	719	3.988	330	35	0	1.724

Table 1

Entities, statements and types of WLS statements

The third dataset is a selection of randomly chosen entities from Wikidata. This dataset was acquired to compare WLS claims in the other datasets with a randomized subset designed to mimic the overall distribution of WLS claims in Wikidata.

- *Random* (R): This dataset comprises 1.159.800 Wikidata entities (starting from a selection of 1,2 million entities from which we removed duplicates) chosen randomly from the most numerous 100 classes, to reflect the proportional distribution of entities found in Wikidata[30]. This dataset encompasses 61.798.072 statements distributed across 23.196 JSON files.

In table 1 we show a summary of basic information about these collections. All these datasets can be accessed and downloaded from Zenodo[31] [42] and all Python scripts are accessible in GitHub[32].

4.2. Analysis

In the following, we will describe as WLS statements all Wikidata statements showing the use of each method described in section 3, regardless of whether they have actually been used to make weaker logical status claims. A tabular presentation of our analysis is shown in table 1.

Even though critical analysis is a pivotal part of humanities discourses, plainly stated statements with no competing claims are largely the most represented information in the CH dataset: the vast majority of statements here (>99%, in particular 99,74% in CHav, 99,92% in CHv and 99,69% in CHt) are plainly asserted statements with no WLS additions. In contrast, the Astronomical datasets show a reasonably different situation, 83% overall of plainly asserted statements, and specifically ANs at 72,58% and ANg at 95%. The overall distribution of the Random (R) dataset showcases a low percentage of WLS claims (1,78%), closer to the CH and the AN datasets. Yet, interestingly, almost the whole percentage is made of non-asserted statements (98,95%) matching a similar distribution in the AN dataset.

When analysing the Random (R) dataset, we can notice that the simplicity of the ranking system leads to a clear predominance of deprecated items and, consequently, of non-asserted claims. The other WLS methods appear to be underutilized in a proportion closer to the AN dataset. Possibly this is again a reflection that, in the CH community, the historical uncertainty and the representation of interpretation are more frequent and typical than in other disciplines.

[30] https://w.wiki/7iCR

[31] https://doi.org/10.5281/zenodo.7624783

[32] https://github.com/alessiodipasquale/Wikidata_WLS

To further explore these data, we can notice that:

Non-asserted statements: Of the methods previously listed (cf. section 3), non-asserted statements (i.e. variously ranked statements with no corresponding asserted triples) are largely the most frequent method for representing competing information in both AN and R. The situation is fairly different in the CH collections, non-asserted statements being the most frequently used method in CHt (81,64%) and CHav (only 86,09%) and almost unused in CHv (3,99%).

Deprecated statements: Deprecated claims are visibly a small portion of the overall non-asserted statements, occurring only in 20% of the non-asserted statements of the Cultural Heritage entities, in 30% of the non-asserted statements of Astronomical entities and in the 66% of the non-asserted statements of Random entities. At the same time, about half of the deprecated statements were annotated with the corresponding *reason for deprecated rank* qualifier (in particular, 45,59% CHt, 25,15% CHv, 64,93% CHav - compare this with basically 0% in both AN datasets and 1,24% in R dataset), proving that scholars in the humanities have a solid interest in annotating provenance of WLS claims on CH data. Yet, only less than 1% of preferred statements have been annotated with the corresponding qualifier *reason for preferred rank*.

Unknown values: Unknown valued statements are not used at all in Astronomical data (absolute 0 in both ANg and ANs out), poorly adopted in the R dataset (0,47%), and sparsely used in the Humanities as well (9,75% in CHav and 10,71% in CHt). Higher is the result for the CHv dataset, with 46,88% of the overall WLS claims using this method.

Non-existing values: Even if we do not consider them as part of WLS claims as a method of representation, we examined them in our datasets for contiguity to unknown values. Non-existing values are almost unused in Astronomical data (exactly 4 occurrences in ANs and an absolute 0 in ANg out of more than 7 million WLS claims), and very sparsely used in the Humanities and Random datasets as well: 1.969 statements in CHv, 1.356 statements in CHt and 3.857 statements in R dataset. Fairly higher is the result for the CHav dataset, with 50.611 statements using this method. This outlier value is probably justified and will be commented on later in this section.

Qualifiers: Statements qualified with *nature of statement* and *sourcing circumstances* predicates are the least employed representation method out of the surveyed ones, being used in 7,66% of the WLS statements in CHt, in 0,58% of the WLS statements in R and in 4,16% of the CHav statements, present in 0,0008% of the ANs statements and only in one ANg statement. Yet this approach is used in 49,13% of the WLS statements of the CHv dataset. This value will be commented later on in this section.

We further surveyed the terms actually used as values for the qualifiers.

We witnessed the use of respectively 200 different values for qualifier *nature of statement*, 419 for *sourcing circumstances* and 588 for *reason for deprecated rank*. These values largely exceed the proposed values specified in the corresponding Wikidata property talk pages (respectively, 194 values for *nature of statement* and 175 for *sourcing circumstances*) or property constraints as for the 384 values for *reason for deprecated rank*). Furthermore, the three sets of actual terms show considerable overlap of values between them (in our datasets, but also over all of Wikidata), as shown in figure 1. This seems to imply that the semantics associated to these values, and indeed to the properties themselves, may have been unclear to contributors, who then in some cases selected the qualifier in non-predictable ways. Therefore, we took the decision to group all three sets into a single category (shown as *WLS qualified statements* in table 1).

Since the R dataset is not a disciplinary dataset we deemed that the variety of situations occurring across disciplinary boundaries would inevitably pollute any analysis deeper than mere counting, and therefore in the following sections, we will focus only on the disciplinary datasets.

We further surveyed the terms actually used as values for the qualifiers.

Overall, the three sets contain a variety of terms such as generic contextual information items, e.g., provenance details, as well as domain-specific terms not relevant to our purposes (e.g. *show election*, *declared deserted*, or *text exceeds character limit*), as well as qualifiers we can truly consider suggesting weaker logical statuses (e.g., *possibly*, *disputed*, *expected*, etc.).

Therefore, we decided to ignore the list of suggested values provided by the *Property Talk* pages and focused on the actual values found in our datasets. We surveyed the list of terms and selected a subset of 101 terms that seem to concretely refer to WLS claims. This subset of WLS terms seems to be widespread in CH and Random datasets

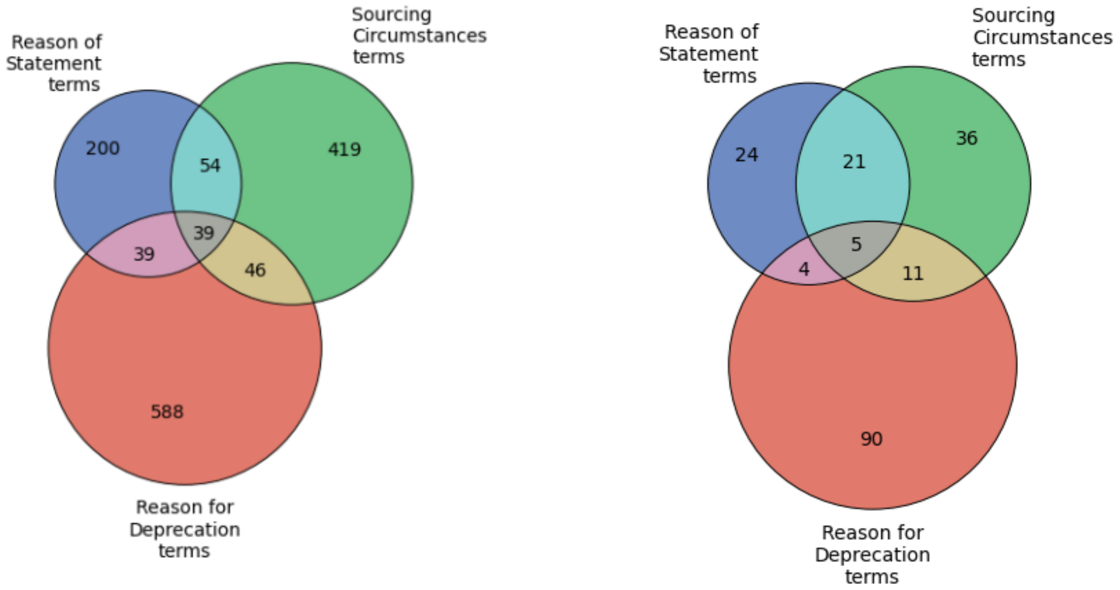

Fig. 1. Terms used in qualifiers *nature of statement*, *sourcing circumstances* and *reason for deprecated rank* throughout Wikidata (*left*) and in the CH datasets (*right*)

(2.086 occurrences in CHav, 111.641 occurrences in CHv, 1.318 occurrences in CHt and 6406 occurrences in R), while almost not employed in Astronomical datasets (62 occurrences in ANs and only 1 in ANg).

The distribution of methods for WLS claims in the CH dataset is not homogeneous, as unknown values and WLS-qualified statements are both highly used in the CHv dataset while non-asserted statements for CHav and CHt. An obvious outlier is the use of one specific qualifier. Indeed, the value *circa*[33] is by far the most employed value in CHv, appearing 107.653 times in *sourcing circumstances*. This brings the overall count of this value completely out of scale concerning other values (e.g., the second most frequent WLS term in CHv is *probably*, occurring only 1.676 times). By removing specifically the value "circa" from the others in the last line of table 1, we see a much homogeneous distribution of values across the three CH datasets. On the contrary, many other terms in the list are present only once in the whole dataset and contribute very little to the overall impact of the qualified statements method.

Another outlier seems to be the number of non-existing valued statements, which are present in the CHav dataset with a much higher proportion than elsewhere. In this dataset, non-existing valued statements seem to be heavily employed correctly in specific properties that appear frequently here and not elsewhere, such as `P364`:*original language of film or TV show*, `P155`:*follows*, and `P156`:*followed by*. This is the correct use of the non-existing valued predicates, while in the other datasets, these properties do not appear with the same frequency and we observe a more heterogeneous distribution of methods (cf. figure 2).

In theory, the methods for WLS are **not** meant as alternatives to each other and to be used exclusively. It would be perfectly acceptable and reasonable to use them on the same statement for the same entity, e.g., to describe a deprecated non-existing valued statement that then results as non-asserted. Yet, method co-occurrence in the surveyed datasets is very poorly represented and datasets demonstrate very few cases of use of multiple WLS methods for the same statements. In particular, no co-occurrence can be found in the AN dataset because almost all WLS claims are expressed via ranked statements except for a little co-occurrence of deprecated statements marked with a WLS qualifier in the ANs dataset (0,01%). Co-occurrences between WLS representation methods seem to be poorly implied in CH datasets. Almost no co-occurrence could be found between unknown and deprecated statements (0,1% in CHav, 0,04% in CHv and none (0%) in CHt), as well as the co-occurrence of deprecated and

[33] http://www.wikidata.org/entity/Q5727902

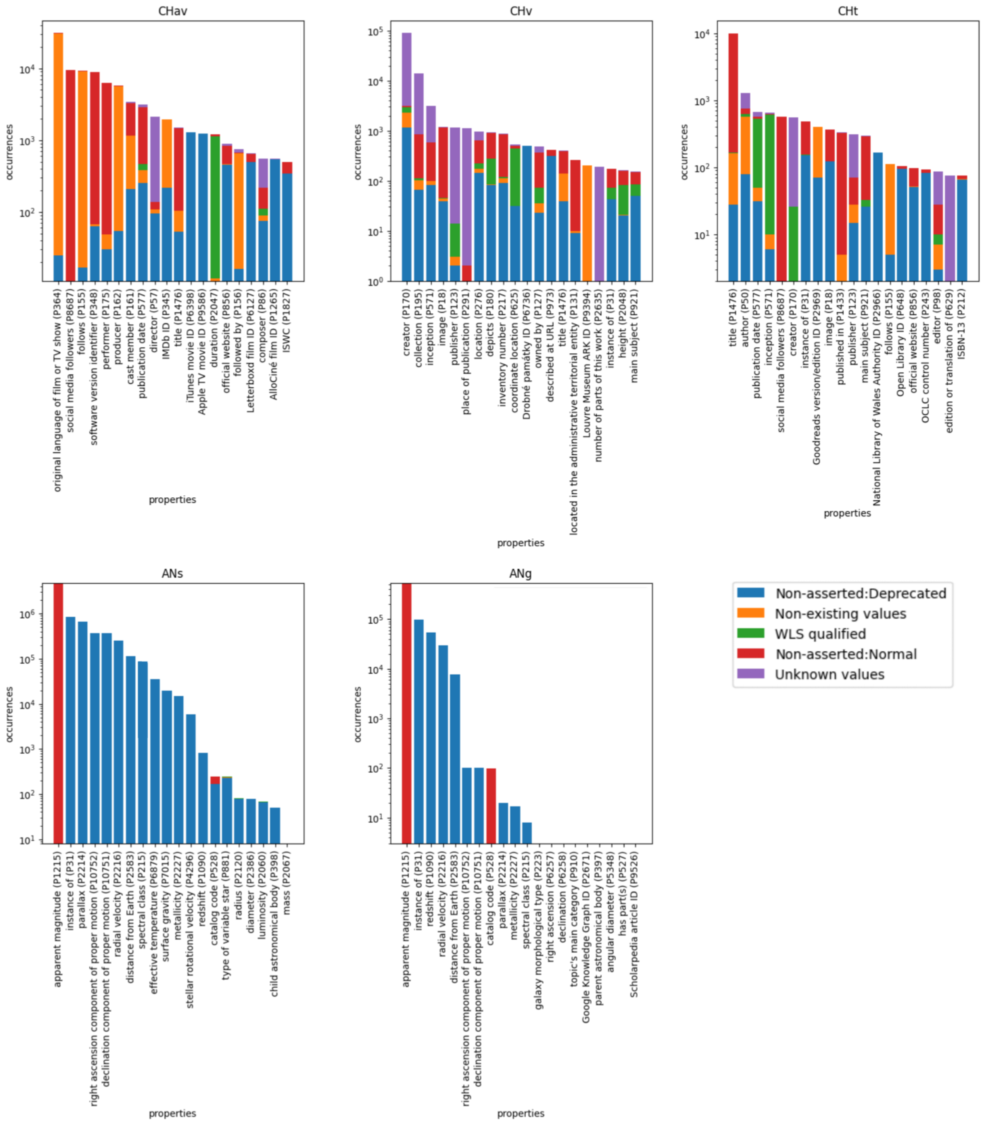

Fig. 2. Top 10 most recurrent properties implied in WLS claims in each disciplinary dataset

WLS qualified statements (0,04% in CHav, 0,01% in CHv, 0,07% in CHt), as well as the co-occurrence of unknown and WLS qualified statements (in 0% in CHav, 1,14% in CHv and 0,13% in CHt).

To summarise, it becomes manifest that the WLS representation methods employed are quite diverse, even between the datasets of the same domain. Specifically, in CHav the most commonly used WLS representation method is non-asserted (86,09%), in CHv it is the WLS Qualified statement (49,13%) followed by unknown value (46,88%), and in CHt it is non-asserted (81,64%). In the Astronomy datasets, non-asserted statements overwhelmingly represent WLS claims, but deprecated statements have a much larger impact on them than in the Cultural Heritage domain.

The property analysis provides valuable insights, too, as shown in Figure 2. We divided the actual usage of WLS methods by the property in which said method appears. The x-axis contains, for each dataset, the ten most frequent properties in which WLS statements appear, and the y-axis shows in logarithmic scale the number of occurrences of such statements, organized by colour: non-asserted statements (with rank normal), non-asserted statements (with rank Deprecated), statements with qualifiers (only WLS-related qualifiers), and non-existing valued statements.

The analysis of the datasets was performed through a systematic evaluation of the properties associated with the WLS representation methods. Each dataset was analysed to identify (1) the most prominent properties of each dataset, and (2) the most prominent properties of each dataset with each method.

Normally ranked, yet non-asserted statements appear in large numbers in CHav for `P8687`:*Social media followers*, `P348`:*software version identifier*, `P175`:*performer* and `P1476`:*title*. They represent peculiar uses of the non-asserted Normal ranks for statements that represent multiple, independent values for the same property, none of which is "more important" than the others. Similar reflections can be made for `P18`:*image* on dataset CHv, and properties `P1433`:*published in* and `P921`:*main subject* in dataset CHt. The property `P1215`:*apparent magnitude* dominates this category for astronomical data. Most of the remaining properties employ a Deprecated rank for evolving or uncertain information.

Qualified statements are largely present in CHv and CHt on properties `P571`:*inception*, `P577`:*publication date*, and `P625`:*coordinate location*, where, as mentioned, the *circa*[34] value for qualifier dominates the occurrences.

Unknown valued statements are mostly used in CHv and CHt datasets and only sparsely in CHav dataset. Their usage seems to be mostly implied in the description of agents in roles in all CH datasets (e.g. `P170`:*creator*, `P98`:*editor*, `P123`:*publisher*, `P50`:*author*, `P86`:*composer*, `P57`:*director*). In CHv and CHt datasets their usage includes also locations (e.g. `P195`:*collection*, `P291`:*place of publication*), time (e.g. `P571`:*inception*) and the artworks' description (e.g. `P2635`:*number of parts of this work*, `P629`:*edition or translation of*). The significant prevalence of unknown values when annotating agents in roles related to artworks is evident in the CHv dataset, reflecting the paramount relevance of authorship attributions given by scholars in the field of art history.

We can also notice the predominance of non-existing valued statements in CHav (`P364`:*original language of film or TV show*, `P155`:*follows*, `P156`:*followed by*, `P162`:*producer* and `P345`:*IMDb ID*), which goes to prove the peculiarity of the use of non-existing valued statements in the CHav dataset previously described. The dataset CHt has a considerable number of non-existing valued statements, too, but only on properties `P1476`:*title* and `P50`:*author*, for untitled and/or anonymous documents.

Besides this, we registered some co-occurrences of the use of unknown and non-existing valued statements with the same properties (e.g. `P57`:*director* in CHav, `P170`:*creator*, `P291`:*publisher*, `P180`:*depicts*, `P571`:*inception*, `P127`:*owned by* in CHv and `P50`:*author*, `P98`:*editor*, `P123`:*publisher*, `P577`:*publication date*, as shown in figure 2). Since these two methods should be employed alternatively, this co-occurrence on the same properties might indicate that annotators are using these two types of blank nodes imprecisely.

To summarize, we show here a list of some of the he complexities and ambiguities we identified in both the CH and the AN datasets.

- **Ranked statements**
  - *Evolving situation*: The claim is not true at the moment but was correct at some point in the past, and keeping this information is deemed interesting to maintain. For example, the number of `P8687`:*social media followers* of artists and politicians, the change of `P276`:*location* of a movable cultural object such as a painting or a statue, or the change of its `P6216`:*copyright status*, may change over time and this change is recorded via differently ranked statements. For instance, the print "Races: Anteriel"[35] *star* recently shifted from copyrighted to the public domain. In this case, the deprecated statement was correct up to a given moment in time, but it has not been correct anymore since then.
  - *Evolving knowledge*: Because of a new observation or theory, a previous value is considered superseded. This situation is mainly connected to new observations, theories, measurements, guesses and interpretations. For example, the introduction of a new accepted attribution of a work of art means that the previous one is now deemed as false or at least deprecated, or, in astronomy, the object "15 Orionis"[36] was previously

[34] https://www.wikidata.org/entity/Q5727902
[35] http://www.wikidata.org/entity/Q79471408
[36] http://www.wikidata.org/entity/Q6675

considered an P31:*instance of* an *infrared source*[37], but it is now fully considered as a *star*[38]; in this case, the deprecated statement has always been incorrect, but it has been decreed as such only after a specific moment in time.

  ∗ *Less favoured versions*: Similar claims are ranked not because they are either false or true, but because one of them is preferred over the others so that they are marked as Preferred and asserted while the others are non-asserted. For example, the P1476:*titles* of textual works are often provided in different languages, and the title in the original language is marked as the preferred version, while the translated titles in other languages are not asserted. In this case, the Deprecated statement is not incorrect, but it has been demoted to prioritize another one. This is not strictly speaking a WLS situation, but uses the same ranking approach of truly WLS ones.

- **Qualified statements**

  ∗ *Uncertainty*: For instance, the painting "Madame Antoine Arnault"[39] has P170:*creator* set to Jean-Baptiste Regnault[40] with a P5102:*nature of statement* qualifier *disputed*; Here, the statement is not certain, and competing (and incompatible) statements may be present or at least expected.

  ∗ *Caution*: For instance, the "Frontispiece to Christopher Saxton's Atlas of the Counties of England and Wales State I"[41] has the P170:*creator* property set to Remigius Hogenberg[42], with the contributor cautioning through a P5102:*nature of statement* qualifier that this is only an *attribution*[43]. Here, the statement is not certain, but the statement is implying that the proposed value may be wrong rather than positively asserting that there are disagreements on it.

  ∗ *Imprecision*: For instance, the hypothetical entity "IRAS 17163-3907"[44] has an observed P2060:*luminosity* property set to "500.000 solar luminosity" with a P1480:*sourcing circumstances* qualifier *circa*; similarly, the painting "Girl Reading a Letter at an Open Window"[45] by Johannes Vermeer is dated (P571:*inception*) 14th century with a *sourcing circumstances* qualifier *circa*. For instance, the star "Altair" (Q12975) has a P1102:*flattening* property set to 0,2 with a *nature of statement* qualifier *greater than*; Here, the statement is certain but the value is inherently loose. One may wonder if this is truly a WLS statement or a positive statement of an imprecise value.

- **Missing value statements**.

  ∗ *Data entry errors*: Data include errors probably introduced during the annotation. For instance, the novel "Invisible Monsters"[46] is both attributed to Chuck Palahniuk (the actual author) and an unknown and probably erroneous entity. Here, there is a clear error in the dataset. Whether a someValue or a noValue blank node is used is not important as they would both be errors.

  ∗ *Dumping from pre-existing databases*: Some non-existing values may be the result of an error in the conversion or an empty field of a record after importing an existing database into Wikidata. For example, the painting "Marshy Landscape"[47] has a non-existing valued statement for the P528:*catalog code* property. Again, this represents an error in the dataset and the corresponding statement should be omitted.

  ∗ *The value does not exist*: For instance, the first and last entities of a sequence use properties P155:*follows* and P156:*followed by* with a non-existing value. For example, the first episode of a TV series, or the last song of a recording, should have non-existing values for the corresponding properties. This is a correct use of a noValue blank node and it is not a WLS claim.

[37] http://www.wikidata.org/entity/Q67206691
[38] http://www.wikidata.org/entity/Q523
[39] https://www.wikidata.org/entity/Q109252498
[40] https://www.wikidata.org/entity/Q453485
[41] https://www.wikidata.org/entity/Q105949375
[42] https://www.wikidata.org/entity/Q18576859
[43] https://www.wikidata.org/entity/Q230768
[44] https://www.wikidata.org/entity/Q540167
[45] https://www.wikidata.org/entity/Q700251
[46] https://www.wikidata.org/entity/Q2600527
[47] https://www.wikidata.org/entity/Q6773948

- ∗ *Model fitting*: When the model does not fully support the situation to be described, some arrangements were taken, such as the use of a non-existing value for the property *original language of film or TV shows* `P364` when the entity is a silent movie. For example see "Silent Tests"[48], whose `P364`:*original language of film or TV show* predicate is non-existing valued and additionally qualified with `P518`:*applied to part dialogue* [49]). Here, a non-existing value is correctly used for a value that was not felt necessary in the model (e.g., a specific property *language of dialogue* to be used in sound fields and omitted for silent movies). This is again a correct use of the method, yet not a WLS claim.
- ∗ *The value exists but is not known*: For example, the painting "The Welcome Home"[50] is marked to have an unknown `P170`:*creator* as a `someValue` blank node. This is probably the only true WLS use of missing value statements.

The previous list shows a series of situations where the same methods are used for different purposes. All such purposes (with the exception of data entry errors) are perfectly legitimate, yet we fear at the same time that users of data may have trouble differentiating the purpose of each use because the method chosen is not sufficiently precise to distinguishing the specific situation in a clear and unambiguous way. Rather than suggest forcing all different situations into a single over-encompassing method, in section 5 we list some increasingly impactful approaches to solving this ambiguities without overly revolutionizing the data model.

4.3. Discussion

The datasets presented in the previous section and the analysis we performed on their content allow us to reach some conclusions on the research questions specified in the introduction.

RQ1 - *How widespread and successful in each of these approaches in the current state of Wikidata?* - The current state of WLS claims in Wikidata is poor. Even though Wikidata focus on established knowledge (community consensus), rather than conjectural or controversial information[51], in many cases it is objective and scientifically precise to represent the complexity of uncertainty and evolving knowledge, rather than omitting information because they are not completely established. In these cases, Wikidata seems to be doing poorly, as <1% of the claims we analysed in CH datasets show weaker logical status characteristics, a much lower figure than the 5% in the ANg dataset or the 27,41% figure of ANs data. Does this show an intrinsic difference in the two cultural domains or is there something else underneath? To provide an answer to this further question, we turned to the RKD database.

RKD[52] holds detailed data about Dutch and Flemish paintings, drawings and prints throughout the ages, from XVI Centuries artworks to modern ones. Overall, more than 260.000 items belonging to the image collections are described in the database, and through the use of an EDM-inspired data model particular attention is given to multiple competing assertions, e.g., incompatible authorship attributions, containing more than 317.000 recorded attributions, i.e, an average of 1,2 attributions per artwork. For instance, deprecated authorship attributions are present in about 8,5% of the works in the RKD image collection (e.g., about 290.000 current attributions vs. 27.000 discarded ones in the RKD images collection), a conspicuously higher figure than the meagre 1,77% WLS statements of the CHv dataset. Although Dutch and Flemish collections may not be representative of the full scale of worldwide types of artworks represented in the CHv dataset, they can provide an interesting comparison. To improve our understanding of these phenomena, we created an age-appropriate sub-dataset. Since RKD stores mostly Dutch paintings from the 17th to the 20th century, we retrieved in Wikidata all paintings created in the same temporal period and excluded RKD artworks[53]. Wikidata stores 501.049 paintings in the interval 17-20th century not uploaded from RKD, for a total of 340.661 attributions[54]. The results of such comparison are shown in 2. Out of the total number of

[48] https://www.wikidata.org/entity/Q390207
[49] https://www.wikidata.org/entity/Q131395
[50] https://www.wikidata.org/entity/Q110041706
[51] http://www.wikidata.org/entity/Help:Ranking#What_ranks_are_not
[52] see https://rkd.nl/en/
[53] https://w.wiki/7VRg
[54] The count of attributions is calculated over the number of claims having the predicate `P170`:*creator*

	RKD images 17-20th c. Dutch paintings	Wikidata 17-20th c. paintings
Paintings	267.238	501.049
Attributions	317.165	340.661
Current attributions	289.918	340.213
Discarded attributions	27.247	448
% Discarded	8,5%	0,13%

Table 2

Comparison between attributions in RKD images collection, CHv dataset and CHv selection of paintings from 17th to 20th century

Wikidata statements, only the 0,13% of the items are discarded attributions (448)[55]. This fact may indicate a radical under-representation of complex attributions within Wikidata entities. We may conclude that WLS statements are not particularly widespread nor successful in Wikidata collections within the Cultural Heritage domain, and they are possibly misrepresenting the complexity and variety of situations that exist in this domain.

RQ2 - *How does the cultural domain of the Wikidata topics (and, presumably, of the individuals contributing to the data regarding the Wikidata topics) affect and reflect on the relative success and richness of some approaches over others?* - Our analysis of data highlighted several peculiarities between the Cultural Heritage datasets and the Astronomical ones. The two families of datasets present many different representational artefacts: while the CH datasets seem to employ, with variable proportions, all the listed methods, the astronomical datasets employ almost exclusively ranked statements. Additionally, while WLS statements in AN datasets affect a fairly small number of properties, they cover a much wider range of properties in CH, as shown in figure 2. These aspects seem to highlight key differences in what the two communities consider weaker logical status: without committing too much to interpretations outside our competency, we may hypothesize that deprecations in astronomical data mostly reflect the result of newer and better data, while the humanities community uses WLS statements for a much larger set of uncertainties of due to ignorance, scholarly interpretations and disagreements as hypothesised in section 1. Thus, it may occur that the specification of the `P5102`:*nature of statement* and the `P2241`:*reason for deprecated rank* qualifiers may seem overkill in astronomical data, and a real necessity for some annotations in the humanities.

RQ3 - *How clean and easy to differentiate are the applications of each approach to an actual weaker logical status versus to another of the designed uses of that approach?* - Unfortunately, there is much noise and ambiguity in how Wikidata contributors have used WLS methods in the datasets we studied. This makes it very difficult to differentiate and search WLS data. The variety of cases listed at the end of section 4.2 summarizes a probably lacking yet vast panorama of WLS and non-WLS knowledge situations modelled through WLS representation methods. It is therefore fairly difficult not only to search for specific data patterns over the full dataset but even to interpret correctly individual entities properly.

Furthermore, the use of the same methods for WLS and non-WLS-related characterizations makes complex patterns very hard to express and identify. For instance, if an artwork AW was *supposedly* moved from location X to location Y, but we are not certain, then both location X and location Y must be represented as WLS, the first because of an evolving situation (AW is not at location X anymore) and the second because of uncertainty since the new location Y is only guessed. Therefore, none of these assertions can be asserted, and none can be ranked as preferred. We need a complete and thorough contextual annotation (e.g. why each claim is discarded) without which disambiguation and full understanding of the state and truth of the relevant predicate are impossible. In section 5, we suggest a possible pattern to represent such situation (cf. point 7, in particular, *Normal rank + non-asserted*).

RQ4 - *Is there a way to improve the clarity and cleanliness of such differentiation?* - Getting down to detailing workable solutions to improve the situation for WLS statements in a project as large and as complex as Wikidata is always running the risk of becoming an exercise in futility. In the next section, we will try to propose a list of

[55] The number of discarded attributions is calculated over the number of claims having `P170`:*creator* as predicate and are not asserted

possible actions for WLS statements, starting from very conservative proposals with limited impact, up to bolder and more impacting changes.

5. Towards a leaner and harmonic support for WLS in Wikidata

In this section, we enumerate a short list of possible remediation activities to be performed over the Wikidata data model and the collection itself to simplify and disambiguate WLS assertions from the rest. We approach such a complex endeavour with humility and caution, as we are well aware that it may be hard to assess from our vantage point both the impact and the difficulty of the implementation of each suggested step.

For this reason, we express our suggestions as an ordered list whose first items are meant as simple cleaning-up activities of little impact and then progress to bolder and more impacting actions that sometimes require not just a modification in the data model, but possibly also the systematic update of small, but still numerically relevant, selections of the current datasets.

1. Require a `P7452`:*reason for preferred statement* qualifier in all preferred statements and a `P2241`:*reason for deprecated statement* qualifier in all deprecated statements. Provide simple-to-use interface widgets for their specification. Make sure that no such statements can be saved without a qualifying proposition.
2. Require the specification of `P5102`:*nature of statement* and `P1480`:*sourcing circumstances* qualifiers for all WLS-related rankings: only asserted statements with Normal rank are allowed to remain without qualifiers.
3. Create a new and separate *Certainty Degree* qualifier specifically for WLS statements, separating the reason for the chosen qualification from the certainty or confidence degree of the qualification. Such certainty degree should be scalar and use a limited number of values, avoiding any complexity in distinguishing between, e.g.: possibly, presumably, hypothetical, dubious, etc.). A 5- or 7-item scale would suffice, e.g.: *non accepted, highly unlikely, unlikely, possible, probable, almost surely*, and *accepted*. Different labels, and even the use of numerical values instead of labels, would be perfectly acceptable.
4. Reorganize the values of `P5102`:*nature of statement* and `P1480`:*sourcing circumstances* to remove values merely representing an uncertainty (replaced by the new *Certainty Degree* qualifier). To this end, an initial list of values is being created. The current list has been generated by following a Grounded Theory approach [43]: first, labels, definitions and usage data of suggested and used qualifiers have been collected and categorized to represent different macro-themes or concepts. These concepts allowed theories to emerge and be developed from the coded data with an iterative process that continued until the theory was "grounded" in the data. The resulting list in its current state, collecting the surveyed terms from the Wikidata *Property Talk* pages and the terms actually used in the CH datasets, contains 150 values referring to WLS claims and organized in 18 theories, and can be accessed in the GitHub folder of the project[56].
5. Restrict ranking for competing statements to just three (possibly four) different patterns, and prevent any other variant:

   - *Preferred + Deprecated*: To be used whenever there is several competing statements and some of them are chosen to be the best ones. Accepted statements are set to Preferred (and asserted) while the rest is set to Deprecated (and not asserted); there are no Normal ranks. Both Preferred and Deprecated statements are fully qualified with `P5102`:*sourcing circumstances*, `P2241`:*reason for deprecated statement* and `P7452`:*reason for preferred statement* respectively, and the new Certainty qualifier. Preferred statements would be assigned a *accepted* or *almost surely* degree, while Deprecated ones would be assigned a *not accepted* or *highly unlikely* certainty degree. Intermediate degrees would not be used.
   - *Normal rank + asserted*: This would be the default situation, to be used when no dispute or disagreement exists and the statement(s) are all equally accepted. All statements are also asserted. Since this is the default no qualifier is necessary, but it is still possible to specify a `P5102`:*nature of statement* or a `P1480`:*sourcing circumstance* value. No certainty degree is necessary.

[56] https://github.com/alessiodipasquale/Wikidata_WLS

– *Normal rank + non-asserted*: To be used when there is several competing statements but none of them stands above the rest as being the most likely. This would be the case, for instance, of a work of art not definitely attributed to anyone, but for which several competing hypotheses exist, but none seem more convincing than the others. No statement is asserted, and `P5102`:*nature of statement* and/or a `P1480`:*sourcing circumstance* values are required. All statements would be assigned a value from the central ones, from *highly unlikely* to *probable*, to the exclusion of the extremes.

A fourth pattern could be in theory allowable, that of a claim for which the only reported value is wrong, but no acceptable alternatives exist. In this case, we could use a deprecated statement for the reported wrong value, and a non-existing valued statement with a Normal rank to represent the unknown correct value.

6. Conclusions and future works

Our work is the first systematic study about the representation of weaker logical status claims (WLS) over cultural heritage data in Wikidata. Through WLS claims it is possible to express uncertain information, competing hypotheses, temporally evolving information, etc. for which a plain and direct assertion is inappropriate. We analysed four patterns used in Wikidata for WLS claims, asserted vs. non-asserted statements, ranked statements, missing values and qualifiers.

In our analysis, we found out several interesting facts. First of all, very few statements are expressed using weaker logical status than could have been expected by comparing other similar sources. Second, the Wikidata data model, far from being too poor for expressing WLS claims, has been shown to provide, in fact, an overabundance of methods, but there seems to be a large overlapping in uses between themselves and also towards non-WLS applications. Finally, there are important differences in how different datasets coming from different domains employ these methods for weaker logical status claims. It seems that domain-specific non-WLS situations can be considered as a justification for much of this variety, and this contributed to the idea that WLS-specific features should be introduced in the Wikidata model to address specifically weaker logical status claims. We proposed a set of increasingly impacting modifications to the data model aiming towards a leaner and more accurate representation of these phenomena with the expectation that they can manage to improve data quality and information retrieval specifically over uncertain, evolving and competing statements.

We are still working toward a full taxonomy of values for qualifying ranked predicates, as this seems to be to our eyes the most rapid and solid way to fully represent both the weaker logical status of a claim and its underlying nature and justification. We plan to publish such taxonomy with a proposal for mapping existing data points into such taxonomy to lose no information in the conversion.

References

[1] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandečić, Introducing Wikidata to the Linked Data Web, in: *The Semantic Web - ISWC 2014*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz and C. Goble, eds, Springer International Publishing, Cham, 2014, pp. 50–65. ISBN 978-3-319-11964-9. doi:doi.org/10.1007/978-3-319-11964-9_4.

[2] C. Möller, J. Lehmann and R. Usbeck, Survey on English Entity Linking on Wikidata: Datasets and Approaches, *Semantic Web* **13** (2022). doi:10.3233/SW-212865.

[3] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini and H. Van de Sompel, The europeana data model (EDM), in: *World Library and Information Congress: 76th IFLA general conference and assembly*, Vol. 10, 2010, p. 15.

[4] M. Doerr, C.-E. Ore and S. Stead, The CIDOC conceptual reference model-A new standard for knowledge sharing, in: *26th international conference on conceptual modeling (ER 2007)*, 2007.

[5] M. Piotrowski and M. Neuwirth, Prospects for computational hermeneutics, in: *Proceedings of the 9th AIUCD Annual Conference*, 2020. http://amsacta.unibo.it/6316/.

[6] M. Fafinski and M. Piotrowski, Modelling Medieval Vagueness, in: *INFORMATIK 2020*, R.H. Reussner, A. Koziolek and R. Heinrich, eds, Gesellschaft für Informatik, Bonn, 2021, pp. 1317–1326. doi:10.18420/inf2020_123.

[7] M. Daquino, V. Pasqual and F. Tomasi, Knowledge Representation of digital Hermeneutics of archival and literary Sources, *JLIS.it* (2020), 59–76. doi:https://doi.org/10.4403/jlis.it-12642.

[8] C.L. Borgman and M.F. Wofford, From Data Processes to Data Products: Knowledge Infrastructures in Astronomy, *Harvard Data Science Review* **3**(3) (2021), https://hdsr.mitpress.mit.edu/pub/xfgywa6x.

[9] A. BLAU, UNCERTAINTY AND THE HISTORY OF IDEAS, *History and Theory* **50**(3) (2011), 358–372. doi:https://doi.org/10.1111/j.1468-2303.2011.00590.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2303.2011.00590.x.

[10] H.-G. Gadamer, *Truth and method*, A&C Black, 2013.

[11] P.T. Darch and A.E. Sands, Uncertainty about the Long-Term: Digital Libraries, Astronomy Data, and Open Source Software, in: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2017, pp. 1–4. doi:10.1109/JCDL.2017.7991584.

[12] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, Dbpedia: A nucleus for a web of open data, in: *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, Springer, 2007, pp. 722–735. doi:doi.org/10.1007/978-3-540-76298-0_52.

[13] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey and G. Weikum, YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames, in: *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15*, Springer, 2016, pp. 177–185. doi:doi.org/10.1007/978-3-319-46547-0_19.

[14] V. Petras, T. Hill, J. Stiller and M. Gäde, Europeana–a Search Engine for Digitised Cultural Heritage Material, *Datenbank-Spektrum* **17** (2017), 41–46. doi:10.1007/s13222-016-0238-1.

[15] E. Delmas-Glass and R. Sanderson, Fostering a community of PHAROS scholars through the adoption of open standards, *Art Libraries Journal* **45**(1) (2020), 19–23–. doi:10.1017/alj.2019.32.

[16] E.E. Fink, American Art Collaborative (AAC) Linked Open Data (LOD) Initiative, Overview and Recommendations for Good Practices (2018). https://repository.si.edu/bitstream/handle/10088/106410/OverviewandRecommendationsAccessible.pdf.

[17] A. Isaac et al., Europeana data model primer (2013). https://pro.europeana.eu/page/edm-documentation.

[18] G. Barabucci, F. Tomasi and F. Vitali, Supporting complexity and conjectures in cultural heritage descriptions, in: *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, CEUR Workshop, 2021, pp. 104–115. http://ceur-ws.org/Vol-2810/paper9.pdf.

[19] A. Stinson, S. Fauconnier and L. Wyatt, Stepping Beyond Libraries: The Changing Orientation in Global GLAM-Wiki, *JLIS.it* **9**(3) (2018), 16–34–. doi:10.4403/jlis.it-12480. https://www.jlis.it/index.php/jlis/article/view/95.

[20] M. Zhitomirsky-Geffet and S. Minster, Cultural information bubbles: A new approach for automatic ethical evaluation of digital artwork collections based on Wikidata, *Digital Scholarship in the Humanities* (2022). doi:10.1093/llc/fqac076.

[21] M. Mora-Cantallops, S. Sánchez-Alonso and E. García-Barriocanal, A systematic literature review on Wikidata, *Data Technologies and Applications* **53**(3) (2019), 250–268. doi:doi.org/10.1108/DTA-12-2018-0110.

[22] D. Hernández, A. Hogan and M. Krötzsch, Reifying RDF: What works well with wikidata?, *SSWS@ ISWC* **1457** (2015), 32–47.

[23] S. Klarman and V. Gutiérrez-Basulto, Two-Dimensional Description Logics for Context-Based Semantic Interoperability, *Proceedings of the AAAI Conference on Artificial Intelligence* **25**(1) (2011), 215–220. doi:10.1609/aaai.v25i1.7854. https://ojs.aaai.org/index.php/AAAI/article/view/7854.

[24] P. Bouquet, F. Giunchiglia, F. van Harmelen, L. Serafini and H. Stuckenschmidt, C-OWL: Contextualizing Ontologies, in: *The Semantic Web - ISWC 2003*, D. Fensel, K. Sycara and J. Mylopoulos, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 164–179. ISBN 978-3-540-39718-2.

[25] A. Zimmermann, N. Lopes, A. Polleres and U. Straccia, A general framework for representing, reasoning and querying with annotated semantic web data, *Journal of Web Semantics* **11** (2012), 72–95.

[26] G. Flouris, I. Fundulaki, P. Pediaditis, Y. Theoharis and V. Christophides, Coloring RDF Triples to Capture Provenance, in: *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, A. Bernstein, D.R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta and K. Thirunarayan, eds, Lecture Notes in Computer Science, Vol. 5823, Springer, 2009, pp. 196–212. doi:10.1007/978-3-642-04930-9_13.

[27] J.M. Giménez-García, A. Zimmermann and P. Maret, NdFluents: An Ontology for Annotated Statements with Inference Preservation, in: *The Semantic Web*, Springer International Publishing, 2017, pp. 638–654. doi:10.1007/978-3-319-58068-5$_3$9.

[28] R. Dividino, S. Sizov, S. Staab and B. Schueler, Querying for provenance, trust, uncertainty and other meta knowledge in RDF, *Journal of Web Semantics* **7**(3) (2009), 204–219.

[29] A. Piscopo and E. Simperl, What We Talk about When We Talk about Wikidata Quality: A Literature Survey, in: *Proceedings of the 15th International Symposium on Open Collaboration*, OpenSym '19, Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450363198. doi:10.1145/3306446.3340822.

[30] M. Färber, F. Bartscherer, C. Menne, A. Rettinger, A. Zaveri, D. Kontokostas, S. Hellmann and J. Umbrich, Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO, *Semantic Web* **9**(1) (2018), 77–129–. doi:10.3233/SW-170275.

[31] V. Balaraman, S. Razniewski and W. Nutt, Recoin: Relative Completeness in Wikidata, in: *Companion Proceedings of the The Web Conference 2018*, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, pp. 1787–1792–. ISBN 9781450356404. doi:10.1145/3184558.3191641.

[32] M. Ponza, P. Ferragina and S. Chakrabarti, A Two-Stage Framework for Computing Entity Relatedness in Wikipedia, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1867–1876–. ISBN 9781450349185. doi:10.1145/3132847.3132890.

[33] L. Galárraga, S. Razniewski, A. Amarilli and F.M. Suchanek, Predicting Completeness in Knowledge Bases, in: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 375–383–. ISBN 9781450346757. doi:10.1145/3018661.3018739.

[34] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe and P. Szekely, A study of the quality of Wikidata, *Journal of Web Semantics* **72** (2022), 100679. doi:https://doi.org/10.1016/j.websem.2021.100679. https://www.sciencedirect.com/science/article/pii/S1570826821000536.

[35] D. Abián, F. Guerra, J. Martínez-Romanos and R. Trillo-Lado, Wikidata and DBpedia: A Comparative Study, in: *Semantic Keyword-Based Search on Structured Data Sources*, J. Szymański and Y. Velegrakis, eds, Springer International Publishing, Cham, 2018, pp. 142–154. ISBN 978-3-319-74497-1.

[36] Help:Ranking - Wikidata. https://www.wikidata.org/wiki/Help:Ranking.

[37] P.F. Patel-Schneider, Contextualization via qualifiers., in: *CKGSemStats@ ISWC*, 2018.

[38] S. Aljalbout, G. Falquet and D. Buchs, Handling Wikidata Qualifiers in Reasoning, *arXiv preprint arXiv:2304.03375* (2023).

[39] D. Hernández, C. Gutierrez and A. Hogan, Certain Answers for SPARQL with Blank Nodes, in: *The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I*, Springer-Verlag, Berlin, Heidelberg, 2018, pp. 337–353–. ISBN 978-3-030-00670-9. doi:10.1007/978-3-030-00671-6$_{2}$0.

[40] D.V. McDermott and D. Dou, Representing Disjunction and Quantifiers in RDF, in: *Proceedings of the First International Semantic Web Conference on The Semantic Web*, ISWC ’02, Springer-Verlag, Berlin, Heidelberg, 2002, pp. 250–263–. ISBN 3540437606.

[41] J. Garson, Modal Logic, in: *The Stanford Encyclopedia of Philosophy*, Spring 2023 edn, E.N. Zalta and U. Nodelman, eds, Metaphysics Research Lab, Stanford University, 2023.

[42] A.D. Pasquale, F. Vitali and V. Pasqual, Wikidata selection of Cultural Heritage, Stars, Galaxies and Random entities and claims, Zenodo, 2023. doi:10.5281/zenodo.7624783.

[43] B.G. Glaser and A.L. Strauss, *The discovery of grounded theory: Strategies for qualitative research*, Routledge, 2017. ISBN 1351522167.

[44] A. Blau, Uncertainty and the history of ideas, *History and Theory* **50**(3) (2011), 358–372. http://www.jstor.org/stable/41300100.