Using logical constraints to validate statistical information about disease outbreaks in collaborative knowledge graphs: the case of COVID-19 epidemiology in Wikidata

Houcemeddine Turki [1] , **Dariusz Jemielniak** [2] , **Mohamed A Hadj Taieb** [1] , **Jose E Labra Gayo** [3] , **Mohamed Ben Aouicha** [1] , **Mus'ab Banat** [4] , **Thomas Shafee** [5, 6] , **Eric Prud'hommeaux** [7] , **Tiago Lubiana** [8] , **Diptanshu Das** [9, 10] , **Daniel Mietchen**
Corresp. [11, 12, 13, 14]

[1] Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Sfax, Tunisia

[2] Department of Management in Networked and Digital Societies, Kozminski University, Warsaw, Masovia, Poland

[3] Web Semantics Oviedo (WESO) Research Group, University of Oviedo, Oviedo, Asturias, Spain

[4] Faculty of Medicine, Hashemite University, Zarqa, Jordan

[5] La Trobe University, Melbourne, Victoria, Australia

[6] Swinburne University of Technology, Melbourne, Victoria, Australia

[7] World Wide Web Consortium, Cambridge, Massachusetts, United States of America

[8] Computational Systems Biology Laboratory, University of São Paulo, São Paulo, Brazil

[9] Institute of Child Health (ICH), Kolkata, West Bengal, India

[10] Medica Superspecialty Hospital, Kolkata, West Bengal, India

[11] Ronin Institute, Montclair, New Jersey, United States of America

[12] Department of Evolutionary and Integrative Ecology, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

[13] School of Data Science, University of Virginia, Charlottesville, Virginia, United States

[14] Institute for Globally Distributed Open Research and Education (IGDORE), Jena, Germany

Corresponding Author: Daniel Mietchen
Email address: daniel.mietchen@ronininstitute.org

Urgent global research demands real-time dissemination of precise data. Wikidata, a collaborative and openly licensed knowledge graph available in RDF format, provides an ideal forum for exchanging structured data that can be verified and consolidated using validation schemas and bot edits. In this research paper, we catalog an automatable task set necessary to assess and validate the portion of Wikidata relating to the COVID-19 epidemiology. These tasks assess statistical data and are implemented in SPARQL, a query language for semantic databases. We demonstrate the efficiency of our methods for evaluating structured non-relational information on COVID-19 in Wikidata, and its applicability in collaborative ontologies and knowledge graphs more broadly. We show the advantages and limitations of our proposed approach by comparing it to the features of other methods for the validation of linked web data as revealed by previous research.

1 **Using logical constraints to validate statistical**
2 **information about disease outbreaks in collaborative**
3 **knowledge graphs: the case of COVID-19**
4 **epidemiology in Wikidata**

6 Houcemeddine Turki[1], Dariusz Jemielniak[2], Mohamed A Hadj Taieb[1], Jose E Labra Gayo[3],
7 Mohamed Ben Aouicha[1], Mus'ab Banat[4], Thomas Shafee[5], Eric Prud'hommeaux[6], Tiago
8 Lubiana[7,8], Diptanshu Das[9], Daniel Mietchen[8,10,11,12]

10 [1] Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of
11 Sfax, Sfax, Tunisia
12 [2] Department of Management in Networked and Digital Societies, Kozminski University,
13 Warsaw, Poland
14 [3] Web Semantics Oviedo (WESO) Research Group, University of Oviedo, Oviedo, Spain
15 [4] Faculty of Medicine, Hashemite University, Zarqa, Jordan
16 [5] La Trobe University, Melbourne, Victoria, Australia
17 [5] Swinburne University of Technology, Melbourne, Victoria, Australia
18 [6] World Wide Web Consortium, Cambridge, Massachusetts, United States of America
19 [7] Computational Systems Biology Laboratory, University of São Paulo, São Paulo, Brazil
20 [8] Ronin Institute, Montclair, New Jersey, United States of America
21 [9] Institute of Child Health (ICH), Kolkata, India
22 [9] Medica Superspecialty Hospital, Kolkata, India
23 [10] School of Data Science, University of Virginia, Charlottesville, Virginia, United States of
24 America
25 [11] Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
26 [12] Institute for Globally Distributed Open Research and Education, Jena, Germany

28 Corresponding Author:
29 Daniel Mietchen[8,10,11,12]
30 Ronin Institute, 127 Haddon Pl, Montclair, New Jersey 07043, United States of America
31 Email address: daniel.mietchen@roninstitute.org

Abstract

Urgent global research demands real-time dissemination of precise data. Wikidata, a
collaborative and openly licensed knowledge graph available in RDF format, provides an ideal
forum for exchanging structured data that can be verified and consolidated using validation
schemas and bot edits. In this research paper, we catalog an automatable task set necessary to
assess and validate the portion of Wikidata relating to the COVID-19 epidemiology. These tasks
assess statistical data and are implemented in SPARQL, a query language for semantic databases.
We demonstrate the efficiency of our methods for evaluating structured non-relational
information on COVID-19 in Wikidata, and its applicability in collaborative ontologies and
knowledge graphs more broadly. We show the advantages and limitations of our proposed
approach by comparing it to the features of other methods for the validation of linked web data as
revealed by previous research.

Introduction

Emerging infectious diseases demand scalable efforts targeting data acquisition, curation, and
integration to drive evidence-based medicine, predictive modeling, and public health policy
(Dong, Du, & Gardner, 2020; Xu, Kraemer, & Data Curation Group, 2020). Of particular
importance are outbreaks designated as Public Health Emergencies of International Concern,
which is currently the case with Polio (Wilder-Smith & Osman, 2020), COVID-19 (Wilder-Smith
& Osman, 2020), and Monkeypox (Kozlov, et al., 2022). Building on previous work (cf. Turki et
al., 2022) that explored how COVID-19-related information can be collaboratively curated in a
knowledge graph, the research presented here zooms in on how statistical information about
pathogens, diseases and disease outbreaks can be validated using logical constraints. As before,
we will use Wikidata (Vrandečić & Krötzsch, 2014) as an example for such a knowledge graph,
while the *SARS-CoV-2* virus, the COVID-19 disease as well as the COVID-19 pandemic will
serve as examples for curating and validating epidemiological information in such a knowledge
graph.

Agile data sharing and computer-supported reasoning about the COVID-19 pandemic and
SARS-CoV-2 virus allow us to quickly understand more about the disease's epidemiology,
pathogenesis, and physiopathology. This understanding can then inform the required clinical,
scholarly, and public health measures to fight the condition and handle its nonmedical
ramifications (Heymann, 2020; Mietchen & Li, 2020; RDA COVID-19 Working Group, 2020).
Consequently, initiatives have rapidly emerged to create datasets, web services, and tools to
analyze and visualize COVID-19 data. Examples include Johns Hopkins University's COVID-19
dashboard (Dong, Du, & Gardner, 2020) and the Open COVID-19 Data Curation Group's
epidemiological data (Xu, Kraemer, & Data Curation Group, 2020). Some of these resources are
interactive and return their results based on combined clinical and epidemiological information,
scholarly information, and social network analysis (Cuan-Baltazar, et al., 2020; Ostaszewski, et
al., 2020; Kagan, Moran-Gilad, & Fire, 2020). However, a significant shortfall in interoperability

71   is common: although these dashboards facilitate examination of their slice of the data, most of
72   them lack general integration with other sites or datasets.

73   The lack of technical support for interoperability is exacerbated by legal restrictions: despite
74   being free to access, the majority of such dashboards are provided under *All Rights Reserved*
75   terms or licenses. Similarly, >84% of the 142,665 COVID-19-related projects on the GitHub
76   repository for computing projects are under *All Rights Reserved*[1] terms (as of 4 February 2022).
77   Restrictive licensing of data sets and applications severely impedes their dissemination and
78   integration, ultimately undermining their value for the community of users and re-users. For
79   complex and multifaceted phenomena such as the COVID-19 pandemic, there is a particular need
80   for a collaborative, free, machine-readable, interoperable, and open approach to knowledge
81   graphs that integrate the varied data.

82   Wikidata[2] just fits this need as a CC0[3] licensed, large-scale, multilingual knowledge graph
83   used to represent human knowledge in a structured format (Resource Description Framework or
84   RDF) (Vrandečić & Krötzsch, 2014; Turki, et al., 2019). It, therefore, has the advantage of being
85   inherently **f**indable, **a**ccessible, **i**nteroperable, and **r**eusable, i.e., FAIR (Wilkinson, et al., 2016).
86   It was initially developed in 2012 as an adjunct to Wikipedia, but has grown significantly beyond
87   its initial parameters. As of now, it is a centralized, cross-disciplinary meta-database and
88   knowledge base for storing structured information in a format optimized to be easily read and
89   edited by both machines and humans (Erxleben, Günther, Krötzsch, Mendez, & Vrandečić,
90   2014). Thanks to its flexible representation of facts, Wikidata can be automatically enriched
91   using information retrieved from multiple public domain sources or inferred from synthesized
92   data (Turki, et al., 2019). This database includes a wide variety of pandemic-related information,
93   including clinical knowledge, epidemiology, biomedical research, software development,
94   geographic, demographic, and genetics data. It can consequently be a vital large-scale reference
95   database to support research and medicine in relation to a pandemic like the still ongoing
96   COVID-19 one (Turki, et al., 2019; Waagmeester, et al., 2021).

97   The key hurdle to overcome for projects such as Wikidata is that several of their features can
98   put them at risk of inconsistent structure or coverage: 1) collaborative projects use decentralized
99   contributions rather than central oversight, 2) large-scale projects operate at a scale where manual
100   checking is not possible, and 3) interdisciplinary projects regulate the acquisition of data to
101   integrate a wide variety of data sources. To maximize the usability of the data, it is therefore
102   important to minimize inconsistencies in its structure and coverage. As a result, methods of
103   evaluating the existing knowledge graphs and ontologies, integral to knowledge graph

---

1   [1]   120,109 of 142,665 as of 4 February 2022: https://github.com/search?q=covid-
2   19+OR+covid19+OR+coronavirus+OR+cord19+OR+cord-19
3   [2]   https://www.wikidata.org/
4   [3]   CC0 is a rights waiver similar to Creative Commons licenses, used to publish material into the public
5   domain. It waives as much copyright as possible within a given jurisdiction. Further information can be found at
6   https://creativecommons.org/publicdomain/zero/1.0/ .

104   maintenance and development, are of crucial importance. Such an evaluation is particularly
105   relevant in the case of collaborative semantic databases, such as Wikidata.

106   Knowledge graph evaluation is, therefore, necessary to assess the quality, correctness, or
107   completeness of a given knowledge graph against a set of predetermined criteria (Amith, He,
108   Bian, Lossio-Ventura, & Tao, 2018). There are several possible approaches to evaluating a
109   knowledge graph based on external information (so-called extrinsic evaluation), including
110   comparing its structure to a paragon ontology, comparing its coverage to source data, applying it
111   to a test problem and judging the outcomes, and manual expert review of its ontology (Brank,
112   Grobelnik, & Mladenic, 2005). Different systematic approaches have been proposed for the
113   comparison of ontologies and knowledge graphs, including NLP techniques, machine learning,
114   association rule mining, and other methods (Lozano-Tello & Gomez-Perez, 2004; Degbelo, 2017;
115   Paulheim, 2017). The criteria for evaluating ontologies typically include *accuracy*, which
116   determines if definitions, classes, properties, and individual entries in the evaluated ontology are
117   correct; *completeness*, referring to the scope of coverage of a given knowledge domain in the
118   evaluated ontology; *adaptability*, determining the range of different anticipated uses of the
119   evaluated ontology (versatility); and *clarity*, determining the effectiveness of communication of
120   intended meanings of defined terms by the evaluated ontology (Vrandečić, 2009; Obrst, Ceusters,
121   Mani, Ray, & Smith, 2007; Raad & Cruz, 2015; Amith, et al., 2018). However, extrinsic methods
122   are not the only ones that are used for evaluating such a set of criteria. Knowledge graphs can be
123   also assessed through an intrinsic evaluation that assesses the structure of the analyzed
124   knowledge graph thanks to the inference of internal description logics and consistency rules (e.g.,
125   Amith et al., 2018).

126   In the research reported here, we emphasize the use of intrinsic methods to evaluate
127   knowledge graphs by presenting our approach to quality assurance checks and corrections of
128   statistical semantic data in Wikidata, mainly in the context of COVID-19 epidemiological
129   information. This consists of a catalog of automatable tasks based on logical constraints expected
130   of the knowledge graph. Most of these constraints were not explicitly available in the RDF
131   validation resources of Wikidata before the pandemic and are designed in this work to support
132   new types of COVID-19 information in the assessed knowledge graph, particularly
133   epidemiological data. Our approach is built upon the outcomes of previous outbreaks such as the
134   Zika epidemic (Ekins et al., 2015) and aims to pave the way towards handling future outbreaks.
135   We implement these constraints with SPARQL and test them on Wikidata using the public
136   SPARQL endpoint of this knowledge graph, available at https://query.wikidata.org. SPARQL[4] is
137   a query language to search, add, modify or delete RDF data available over the Internet without
138   having to retrieve and process the entirety of a given ontological database. We introduce the
139   value of Wikidata as a multipurpose collaborative knowledge graph for the flexible and reliable
140   representation (Section 2) and validation (Section 3) of COVID-19 knowledge. Furthermore, we
141   cover the use of SPARQL to query this knowledge graph (Section 4). Then, we demonstrate how

7   [4]   An open license SPARQL textbook available in multiple languages can be found at
8   https://en.wikibooks.org/wiki/SPARQL.

142 statistical constraints can be implemented using SPARQL and applied to verify epidemiological
143 data related to the COVID-19 pandemic (Section 5). Finally, we compare our constraint-based
144 approach with other RDF validation methods through the analysis of the main outcomes of
145 previous research papers related to knowledge graph validation (Section 6) and conclude future
146 directions (Section 7).

Wikidata as a collaborative knowledge graph

148 Wikidata currently serves as a semantic framework for a variety of scientific initiatives ranging
149 from genetics (Burgstaller-Muehlbacher, et al., 2016) to invasion biology (Jeschke et al., 2021)
150 and clinical trials (Rasberry et al., 2022), allowing different teams of scholars, volunteers and
151 others to integrate valuable academic data into a collective and standardized pool. Its versatility
152 and interconnectedness are making it an example for interdisciplinary data integration and
153 dissemination across fields as diverse as linguistics, information technology, film studies, and
154 medicine (Turki, et al., 2019; Mitraka, et al., 2015; Mietchen, et al., 2015; Waagmeester, Schriml,
155 & Su, 2019, Turki, Vrandečić, Hamdi, & Adel, 2017; Wasi, Sachan, & Darbari, 2020;
156 Heftberger, Höper, Müller-Birn, & Walkowski, 2020), including disease outbreaks like those
157 caused by the Zika virus (Ekins et al., 2015) or *SARS-CoV-2* (Turki et al., 2022). However,
158 Wikidata's popularity and recognition across fields still vary significantly (Mora-Cantallops, et
159 al., 2019). Its multilingual nature enables fast-updating dynamic data reuse across different
160 language versions of a resource such as Wikipedia (Müller-Birn, Karran, Lehmann, & Luczak-
161 Rösch, 2015), with fewer inconsistencies from local culture (Miquel-Ribé & Laniado, 2018) or
162 language biases (Kaffee, et al., 2017; Jemielniak & Wilamowski, 2017).

163 The data structure employed by Wikidata is intended to be highly standardized, whilst
164 maintaining the flexibility to be applied across highly diverse use-cases. There are mainly two
165 essential components: Items, which represent objects, concepts, or topics; and properties, which
166 describe how one item relates to another[5]. A statement, therefore, consists of a subject item (*S*), a
167 property that describes the nature of the statement (*P*), and an object (*O*) that can be an item, a
168 value, an external ID, or a string, etc. While items can be freely created, new properties require
169 community discussion and vote, with about 10000 properties[6] currently available. Statements can
170 be further modified by any number of qualifiers to make them more specific, and be supported by
171 references to indicate the source of the information. Thus, Wikidata forms a continuously
172 growing, single, unified network graph, with 99M items forming the nodes, and 1706M
173 statements forming the edges as of July 20, 2022. A live SPARQL endpoint and query service,
174 regular RDF/JSON dumps, as well as linked data APIs and visualization tools, establish a
175 backbone of Wikidata uses (Malyshev, Krötzsch, González, Gonsior, & Bielefeldt, 2018;
176 Nielsen, Mietchen, & Willighagen, 2017).

177 Importantly, Wikidata is based on free and open-source philosophy and software and is a
178 database that anyone can edit, similarly to the online encyclopedia, Wikipedia (Jemielniak,
179 2014). As a result, the emerging ontologies are created entirely collaboratively, without formal
180 pre-publication peer-review (Piscopo & Simperl, 2018), and developed in a community-driven
181 fashion (Samuel, 2017). This approach allows for the dynamic development of areas of interest

---

9 [5] Detailed information about the data structure of Wikidata can be found in Turki et al. (2022).
10 [6] For an updated list of available Wikidata properties, please see https://tools.wmflabs.org/hay/propbrowse/.

182 for the user community but poses challenges, e.g., to systematize and apportion class
183 completeness across topics (Luggen, Difallah, Sarasua, Demartini, & Cudré-Mauroux, 2019).
184 Also, since the edit history is available to anyone, tracing human and non-human contributions,
185 as well as detecting and reverting vandalism is available by design and relies on community
186 management (Pellissier Tanon & Suchanek, 2019) as well as on software tools like ORES
187 (Sarabadani, et al., 2017) or the Item Quality Evaluator[7]. Wikidata's quality is overall high, and
188 has been a topic of a number of studies already (e.g., Piscopo & Simperl, 2019; Shenoy et al,
189 2022).

190 Other ontological databases and knowledge graphs exist such as DBpedia, Freebase, and
191 OpenCyc (Färber, Bartscherer, Menne, & Rettinger, 2018; Pillai, Soon, & Haw, 2019). However,
192 much like the factors that led Wikipedia to rise to be a dominant encyclopedia (Shafee et al.,
193 2017; Jemielniak & Wilamowski, 2017), Wikidata's close connection to Wikimedia volunteer
194 communities and wide readership provided by Wikipedia have quickly given it a competitive
195 edge. The system, therefore, aims to combine the wisdom of the crowds with advanced
196 algorithms. For instance, Wikidata editors are assisted by a property suggesting system,
197 proposing additional properties to be added to entries (Zangerle, Gassler, Pichl, Steinhauser, &
198 Specht, 2016). Wikidata has subsequently exhibited the highest growth rate of any Wikimedia
199 project and was the first amongst them to pass one billion contributions (Waagmeester, et al.,
200 2020).

201 As a collaborative venture, its governance model is similar to Wikipedia (Lanamäki &
202 Lindman, 2018), but with some important differences. Wide permissions to edit Wikidata are
203 manually granted to approved bots and to Wikimedia accounts that are at least 4 days old and
204 have made at least 50 edits using manual modifications or semi-automated tools for editing
205 Wikidata[8]. These accounts are supervised by a limited number of experienced administrators to
206 prevent misleading editing behaviors (such as vandalism, harassment, and abuse) and to ensure a
207 sustainable consistency of the information provided by Wikidata[9]. As such, Wikidata is highly
208 relevant to the computer-supported collaborative work (CSCW) field, yet the number of studies
209 of Wikidata from this perspective is still very limited (Sarasua et al., 2019). To understand the
210 value of using SPARQL to validate the usage of relation types in collaborative ontologies and
211 knowledge graphs, it is important to understand the main distinctive features of Wikidata as a
212 collaborative project. Much as Wikidata is developed collaboratively by an international
213 community of editors, it is also designed to be language-neutral. As a result, it is quite possible to
214 contribute to Wikidata with only a limited command of English and to effectively collaborate
215 whilst sharing no common human language - an aspect unique even in the already rich ecosystem
216 of collaborative projects[10] (Jemielniak & Przegalinska, 2020). It may well be a cornerstone
217 towards the creation of other language-independent cooperative knowledge creation initiatives,

---

11 [7]   https://item-quality-evaluator.toolforge.org/
12 [8]   For an overview of the semi-automated editing tools for Wikidata, please see
13 https://www.wikidata.org/wiki/Wikidata:Tools .
14 [9]   Further information about the rights and governance of users in Wikidata is shown at
15 https://www.wikidata.org/wiki/Wikidata:User_access_levels .
16 [10]   For further details about the language representation of COVID-19 knowledge in Wikidata, please refer to
17 Turki et al. (2022), which has a figure and multiple tables on the subject.

218   such as Wikifunctions, which is an abstract, language-agnostic Wikipedia currently developed
219   and based on Wikidata (Vrandečić, 2021).

220   It is also possible to build Wikipedia articles, especially in underrepresented languages, based
221   on Wikidata data only, and create article placeholders to stimulate encyclopedia articles' growth
222   (Kaffee et al., 2018). This stems from combining concepts that are relatively easily inter-
223   translatable between languages (e.g., professions, causes of death, and capitals) with language-
224   agnostic data (e.g., numbers, geographical coordinates, and dates). As a result, Wikidata is a
225   paragon example of not only cross-cultural cooperation but also human-bot collaborative efforts
226   (Piscopo, 2018; Farda-Sarbas, et al., 2019). Given the large-scale crowdsourcing efforts in
227   Wikidata and the use of bots and semi-automated tools to mass edit Wikidata, its current volume
228   is higher than what can be reviewed and curated by administrators manually. It is quite intuitive:
229   as the general number of edits created by bots grows, so grows the number of administrative tasks
230   to be automated. Automation may include simplifying alerts, fully and semi-automated reverts,
231   better user tracking, or automated corrections or suggestions. However, the creation of automated
232   methods for the verification and validation of the ontological statements it contains is required
233   most.

234

235   Knowledge graph validation of Wikidata

236   As Wikidata properties are assigned labels, descriptions, and aliases in multiple languages (Red
237   in Fig. 2), multilingual information of these properties can be used alongside the labels,
238   descriptions, and aliases of Wikidata items to verify and find sentences supporting biomedical
239   statements in scholarly outputs (Zhang, et al., 2019). Such a process can be based on various
240   natural language processing techniques, including word embeddings (Zhang, et al., 2019; Chen,
241   et. al., 2020) and semantic similarity (Ben Aouicha & Hadj Taieb, 2016). These techniques are
242   robust enough to achieve an interesting level of accuracy, and some of them can achieve better
243   accuracy when the Wikidata classes of the subject and object of semantic relations are given as
244   inputs (Lastra-Díaz, et al., 2019; Hadj Taieb, Zesch, & Ben Aouicha, 2020). The subjects and
245   objects of Wikidata relations can likewise be aligned to other biomedical semantic resources such
246   as MeSH and UMLS Metathesaurus (Turki, et al., 2019). Thus, benchmarks for relation
247   extraction based on one of the major biomedical ontologies can be converted into a Wikidata-
248   friendly format[11] and used to automatically enrich Wikidata with novel biomedical relations or to
249   automatically find statements supporting existing biomedical Wikidata relations (Zhang, et al.,
250   2018). Furthermore, MeSH keywords of scholarly publications can be converted into their
251   Wikidata equivalents, refined using citation and co-citation analysis (Turki, 2018), and used to
252   verify and add biomedical Wikidata relations, e.g., by applying deep learning-based bibliometric-
253   enhanced information retrieval techniques (Mayr, Scharnhorst, Larsen, Schaer, & Mutschke,
254   2014; Turki, Hadj Taieb, & Ben Aouicha, 2018).

255   Another option of validating biomedical statements based on the labels and external identifiers
256   of their subjects, predicates, and objects in Wikidata can be the use of these labels and external
257   IDs to find whether the assessed Wikidata statements are available in other knowledge resources
258   (e.g., Disease Ontology) and in open bibliographic databases (e.g., PubMed). Several tools have

18   [11]   A Wikidata-friendly format of a database is an edition of that resource where items and predicates of triples
19   are replaced by their equivalents in Wikidata or in ontologies integrated with it.

259 been successfully built using this principle such as the *Wikidata Integrator*[12] that compares the
260 Wikidata statements of a given gene, protein or cell line with their equivalents in other structured
261 databases like NCBI's Gene resources, UniProt or Cellosaurus and adjusts them if needed.
262 Complementing this approach, *Mismatch Finder*[13] identifies Wikidata statements that are not
263 available in external databases, while *Structured Categories*[14] uses SPARQL to identify how the
264 members of a Wikipedia Category are described using Wikidata statements and to reveal whether
265 a statement is missing or mistakenly edited for the definition of category items (Turki, Hadj
266 Taieb, & Ben Aouicha, 2021), and *RefB*[15] (Fig. 1) extracts biomedical Wikidata statements not
267 supported by references using SPARQL and identifies the sentences supporting them in scholarly
268 publications using the PubMed Central search engine and a variety of techniques such as concept
269 proximity analysis.
270
271   *<place Figure 1 near here>*
272
273 In addition to their multilingual set of labels and descriptions, Wikidata properties are assigned
274 object types using wikibase:propertyType relations (Blue in Fig. 2). These relations allow the
275 assignment of appropriate objects to statements, so that non-relational statements cannot have a
276 Wikidata item as an object, while objects of relational statements are not allowed to have data
277 types like a value or a URL (Vrandečić & Krötzsch, 2014).
278
279   *<place Figure 2 near here>*
280
281 Just like a Wikidata item, a property can be described by statements (Green in Fig. 2). The
282 predicates of these statements link a property to its class (*instance of* [P31]), to its corresponding
283 Wikidata item (*subject item of this property* [P1629]), to example usages (*Wikidata property*
284 *example* [P1855]), to equivalents in other IRIs[16] (*equivalent property* [P1628] and *exact match*
285 [P2888]), to Wikimedia categories that track its usage on a given wiki (*property usage tracking*
286 *category* [P2875]), to its inverse property (*inverse property* [P1696]), or to its proposal
287 discussion (*property proposal discussion* [P3254]), etc. These statements can be interesting for
288 various knowledge graph validation purposes. The class, the usage examples, and the proposal
289 discussion of a Wikidata property can be useful through the use of several natural language
290 processing techniques, particularly semantic similarity, to provide several features of the use of
291 the property such as its domain of application (e.g., the subject or object of a statement using a

---

20   [12]   Wikidata Integrator is a bot framework for automatically curating genetic information provided by
21   Wikidata (https://github.com/SuLab/WikidataIntegrator). For Wikidata bots using this framework, refer to
22   https://www.wikidata.org/wiki/Wikidata:WikiProject_Gene_Wiki#Bot_accounts. The framework has been adapted
23   to various specific contexts, e.g., the curation of cell lines indexed in Cellosaurus, as per https://github.com/calipho-
24   sib/cellosaurus-wikidata-bot .
25   [13]   https://www.wikidata.org/wiki/Wikidata:Mismatch_Finder
26   [14]   https://www.wikidata.org/wiki/Wikidata:Structured_Categories
27   [15]   RefB: *Description* at https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/
28   RefB_(WikiCred), *Source code* at https://github.com/Data-Engineering-and-Semantics/refb/, *Wikidata edits* at
29   https://www.wikidata.org/wiki/Special:Contributions/RefB_(WikiCred).
30   [16]   Internationalized Resource Identifier (IRI) is a standardized character string (e.g., a URL) that recognizes a
31   given item in a semantic resource

292   Wikidata property related to medicine should be a medical item) and consequently to eliminate
293   some of the erroneous use by screening the property usage tracking category. The class of the
294   Wikidata item corresponding to the property can be used to identify the field of work of the
295   property and thus flag some inappropriate applications. In addition, the external identifiers of
296   such an item can be used for the verification of biomedical relations by their identification within
297   the semantic annotations of scholarly publications built using the *SAT+R* (Subject, Action,
298   Target, and Relations) model (Piad-Morffis, Gutiérrez, & Muñoz, 2019). The inverse property
299   relations can identify missing statements, which are implied by the presence of inverse statements
300   in Wikidata. However, using inverse properties has the downside that it causes redundancies in
301   the underlying knowledge graph.
302   Despite the importance of these statements defining properties, *property constraint* [P2302]
303   relations (Brown in Fig. 2) are the semantic relations that are primarily used for the validation of
304   the usage of a property. In essence, they define a set of conditions for the use of a property,
305   including several heuristics for the type and format of the subject or the object, information about
306   the characteristics of the property, and several description logics for the usage of the property as
307   shown in Table 1. Property constraints are either manually added by Wikidata users or inferred
308   with high accuracy from the knowledge graph of Wikidata or the history of human changes to
309   Wikidata statements (Pellissier Tanon, Bourgaux, & Suchanek, 2019; Hanika, et al., 2019).
310

311   *<place Table 1 near here>*

312

313   As shown in Fig. 2, a property constraint is defined as a relation where the property type is
314   featured as an object, and the detailed conditions of the constraint to be applied on Wikidata
315   statements are integrated as qualifiers to the relation. When a statement uses a property in a way
316   that does not conform to its corresponding constraint, the statement is automatically included in
317   the property constraint report[17] and is marked by an exclamation mark on the page of the subject
318   item (Fig. 3), so that either the item can be repaired by the community or by Wikidata bots, or the
319   property constraint can be renegotiated.
320

321   *<place Figure 3 near here>*

322

323   Although these methods are important to verify and validate Wikidata, they are not the only ones
324   that are used for these purposes. Various MediaWiki templates, Lua modules or bots can be used
325   to check, flag and in some cases fix inconsistencies. For instance, the Autofix template[18] allows
326   to specify regex patterns that then trigger bot edits, e.g., to enforce case normalization of values
327   for a given property.
328

329   In 2019, Wikidata announced the adoption of the Shape Expressions language (ShEx) as part of
330   the MediaWiki entity schemas extension[19]. ShEx was proposed following an RDF validation

---

32   [17]   https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations
33   [18]   https://www.wikidata.org/wiki/Template:Autofix
34   [19]   https://www.mediawiki.org/wiki/Extension:EntitySchema

331   workshop that was organized by W3C[20] in 2014 as a concise, high-level language to describe and
332   validate RDF data (Prud'hommeaux, Labra Gayo, & Solbrig, 2014). This Mediawiki extension
333   uses ShEx to store structure definitions (EntitySchemas or Shapes) for sets of Wikidata entities
334   that are selected by some query pattern (frequently the involvement of said entities in a Wikidata
335   class). This provides collaborative quality control where the community can iteratively develop a
336   schema and refine the data to conform to that schema. For those familiar with XML, ShEx is
337   analogous to XML Schema or RelaxNG. *SHACL* (Shapes Constraint Language), another
338   language used to constraint RDF data models, uses a flat list of constraints, analogous to XML's
339   Schematron. *SHACL* was adapted from SPIN (SPARQL Inference Notation) by the W3C Data
340   Shapes working group in 2014 and became a W3C recommendation in 2017 (Knublauch &
341   Kontokostas, 2017). However, ShEx was chosen to represent EntitySchemas in Wikidata, as it
342   has a compact syntax that makes it more human-friendly, supports recursion, and is designed to
343   support distributed networks of reusable schemas (Labra Gayo, Prud'hommeaux, Boneva, &
344   Kontokostas, 2017). Besides the possibility to infer ShEx expressions from the screening of a
345   large set of concerned items, they can be intuitively written by humans.

346   In Wikidata, ShEx-based EntitySchemas are assigned an identifier (a number beginning with
347   an E) as well as labels, descriptions, and aliases in multiple languages, so that they can be
348   identified by users. Entity schemas are defined using the ShEx-compact syntax[21], which is a
349   concise, human-readable syntax. A schema usually begins with some prefix declarations similar
350   to those in SPARQL. An optional start definition declares the shape which will be used by
351   default. In the example (Fig. 4), the shape `<app>` will be used, and its declaration contains a list
352   of properties, possible values, and cardinalities. By default, shapes are open, which means that
353   other properties apart from the ones declared are allowed. In this example, the values of property
354   `wdt:P31` are declared to be either a COVID-19 dashboard (`wd:Q90790055`), a search engine
355   (`wd:Q91136116`), or a dataset (`wd:Q91137337`). The `EXTRA` directive indicates that there can be
356   additional values for property `wdt:P31` that differ from the specified ones. The value for property
357   `wdt:P1476` is declared to be zero or more literals. The cardinality indicators come from regular
358   expressions, where '`?`' means zero or one, '`*`'; means zero or more, and '`+`' means one or more.
359   While the values for the other properties are declared to be anything (the dot indicates no
360   constraint) zero or more times, except for the properties `wdt:P577` and `wdt:P7103` that are
361   marked as optional using the question mark. Further documentation about ShEx can be found at
362   http://shex.io/ and in Labra Gayo et al. (2017).

363

364   *<place Figure 4 near here>*

365

366   Due to the ease of using ShEx to define EntitySchemas, it has been used successfully to validate
367   Danish lexemes in Wikidata (Nielsen, Thornton, & Labra-Gayo, 2019) and biomedical Wikidata
368   statements (Thornton, et al., 2019). During the COVID-19 pandemic, Wikidata's data model of
369   every COVID-19-related class as well as of all major biomedical classes has been converted to an
370   EntitySchema, so that it can be used to validate the representation of COVID-19 Wikidata
371   statements (Waagmeester, et al., 2021). These EntitySchemas were successfully used to enhance

---

35   [20]   https://www.w3.org/2012/12/rdf-val/report
36   [21]   ShEx schemas can also be defined in RDF-based representations like Turtle or JSON-LD.

372 the development and the robustness of the semantic structure of the data model underlying the
373 COVID-19 knowledge graph in Wikidata and are accordingly made available at a subpage of
374 Wikidata's WikiProject COVID-19[22]. Significant efforts are currently underway to further
375 simplify the definition of EntitySchemas by making them more intuitive and concise, enabling an
376 increase of the usage of ShEx to validate semantic knowledge in Wikidata (Samuel, 2021).

377   Beyond these interesting methods, validation constraints can be inferred and used to verify
378 semantic statements in a knowledge graph through the use of the full screening of RDF dumps or
379 the use of SPARQL queries. RDF dumps are particularly used for screening Wikidata items in a
380 class to identify common features of the assessed entities based on a set of formal rules (Marx &
381 Krötzsch, 2017; Hanika et al. 2019). These features involve common characteristics of the data
382 model of the concerned class with patterns of used Wikidata properties such as symmetry and are
383 later used to verify the completeness of the class and validate the statements related to the
384 evaluated class. The analysis of RDF dumps for Wikidata can be coupled with the federated
385 screening of the RDF dumps of other knowledge graphs such as DBpedia, allowing to evaluate
386 Wikidata shapes based on aligned external structures for the same domain (Ahmadi & Papotti,
387 2021). Nowadays, efforts are provided to extend inference-based methods for the validation of
388 Wikidata through the development of probabilistic approaches to identify when a statement is
389 unlikely to be defined for an item allowing to enhance the evaluation of the completeness of
390 Wikidata as an open knowledge graph (Arnaout, et al., 2021). As SPARQL has been designed to
391 extract a searched pattern from a semantic graph (Pérez, Arenas, & Gutierrez, 2009), it has been
392 used to query and harmonize competency questions[23], and to evaluate ontologies and knowledge
393 graphs in a context-sensitive way (Vasanthapriyan, Tian, & Xiang, 2017; Bansal & Chawla,
394 2016; Martin, 2018). Indeed, a sister project presents how SPARQL can be used to generate data
395 visualizations[24] (Nielsen, Mietchen & Willighagen 2017; Shorland, Mietchen & Willighagen,
396 2020). Validating RDF data portals using SPARQL queries has been regularly proposed as an
397 approach that gives great flexibility and expressiveness (Labra Gayo & Alvarez Rodríguez,
398 2013). However, academic literature is still far from revealing a consensus on methods and
399 approaches to evaluate ontologies using this query language (Walisadeera, Ginige, &
400 Wikramanayake, 2016), and other approaches have been proposed for validation (Thornton, et
401 al., 2019; Labra-Gayo, et al., 2019). Currently, there is mostly an effort to normalize how to
402 define SPARQL queries, particularly for knowledge graph validation purposes, to save runtime
403 and ameliorate the completeness of the output of a query using a set of heuristics and axioms
404 (Salas & Hogan, 2022).

405   In Wikidata, the Wikidata Query Service (https://query.wikidata.org) allows querying the
406 knowledge graph using SPARQL (Malyshev, et al., 2018; Turki, et al., 2019). The query service
407 includes a specific endpoint (https://query.wikidata.org/sparql) that allows programmatic access

---

37 [22]   The data models for WikiProject COVID-19 are accessible via
38 https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Data_models.
39 [23]   Competency questions: A set of requirements ensuring consistency of a knowledge graph, constraints
40 determining what knowledge to be involved in a knowledge graph (Wiśniewski, Potoniec, Ławrynowicz, & Keet,
41 2019).
42 [24]   For SPARQL-based visualizations of COVID-19 information in Wikidata, see https://speed.ieee.tn/, https://
43 egonw.github.io/SARS-CoV-2-Queries/, https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Queries,
44 and https://scholia.toolforge.org/topic/Q84263196.

408 to SPARQL queries in programming languages. The required Wikidata prefixes are already
409 supported in the backend of the service and do not need to be defined (Malyshev, et al., 2018).
410 What the user needs to do is to formulate their SPARQL query (Black in Fig. 5) and click on the
411 Run button (Blue in Fig. 5). After an execution period of up to sixty seconds, the results will
412 appear (Green in Fig. 5) and can be downloaded in different formats (Brown in Fig. 5), including
413 JSON, TSV, CSV, HTML, and SVG. Different modes for the visualization of the query results
414 can be chosen (Purple in Fig. 5), particularly table, charts (line, scatter, area, bubble), image grid,
415 map, tree, timeline, and graph. The query service also allows users to use a query helper (Red in
416 Fig. 5) that can generate basic SPARQL queries, and to get inspired by sample queries (Yellow in
417 Fig. 5), especially when they lack experience. It also allows users to generate a short link for the
418 query (Pink in Fig. 5) and code snippets to embed the query results in web pages and computer
419 programs (Brown in Fig. 5) (Malyshev, et al., 2018).

420

421   *<place Figure 5 near here>*

422

423 Constraint-driven heuristics-based validation of epidemiological data

424 The characterization of epidemiological data is possible using a variety of statistical measures
425 that show the acuteness, the dynamics, and the prognosis of a given disease outbreak. These
426 measures include the cumulative count[25] of cases (P1603 [199569 statements, Orange in Fig. 6],
427 noted *c*, as defined before), deaths (P1120 [243250 statements[26], Black in Fig. 6], noted *d*),
428 recoveries (P8010 [36119 statements, Green in Fig. 6], noted *r*), clinical tests (P8011 [21249
429 statements, Blue in Fig. 6], noted *t*), and hospitalized cases (P8049 [5755 statements, Grey in Fig.
430 6], noted *h*) as well as several measurements done by the synthesis of the values of
431 epidemiological counts such as case fatality rate (P3457 [51504 statements, Red in Fig. 6], noted
432 *m*), basic reproduction number (P3492, noted R_0), minimal incubation period in humans (P3488,
433 noted *mn*), and maximal incubation period in humans (P3487, noted *mx*) (Rothman, Greenland, &
434 Lash, 2008). For all these statistical data, every information should be coupled by a *point in time*
435 (P585, noted *Z*) qualifier defining the date of the stated measurement and by a *Determination*
436 *method* (P459, noted *Q*) qualifier identifying the measurement method of the given information
437 as these variables are subject to change over days or according to used methods of computation.

438

439   *<place Figure 6 near here>*

440

441 From count statistics (*c*, *t*, *d*, *h*, and *r* statements), it is possible to compare regional
442 epidemiological variables and their variance for a given date (*Z*) or date range, and relate these to
443 the general disease outbreak (each component defined as a *part of* [P361] of the general outbreak)
444 as shown in Table 2. Such comparisons are enabled using statistical conditions that are
445 commonly used in epidemiology (Zu, et al., 2020). Tasks V1 and V2 have been generated from
446 the evidence that COVID-19 started in late 2019 and that its clinical discovery can only be done

45 [25]   We found the Wikidata properties reflecting epidemiological data about COVID-19 outbreaks using a
46 specific SPARQL query available at https://w.wiki/5UsE. Please note that current results can return new properties
47 that did not exist as of August 8, 2020 such as Number of vaccinations [P9107].
48 [26]   As of August 8, 2020. For updated statistics, see https://w.wiki/Z5m.

447 through medical diagnosis techniques (Zu, et al., 2020). Tasks V3 and V4 have been derived
448 from the fact that *c, d, r,* and *t* are cumulative counts. Consequently, these variables are only
449 subjects to remain constant or increase over days. Task V5 is motivated by the fact that an
450 epidemiological count cannot return negative values. Tasks V6, V7, V8, and V9 are due to the
451 evidence that *d, r,* and *h* cannot be superior to *c* as COVID-19 deaths are the consequence of
452 severe infections by SARS-CoV-2 that can only be managed in hospitals (Rothman, Greenland,
453 & Lash, 2008) and as a patient needs to undergo COVID-19 testing to be confirmed as a case of
454 the disease (Zu, et al., 2020). V10 is built upon the assumption that *c, d, r, h,* and *t* values can be
455 geographically aggregated (Rothman, Greenland, & Lash, 2008).
456

457   *<place Table 2 near here>*
458

459 This task set has been applied using ten simple SPARQL queries that can be found in Appendix
460 A where <PropertyID> is the Wikidata property to be analyzed and has returned 5496
461 inconsistencies in the COVID-19 epidemiological information (as of August 8, 2020) as shown in
462 Table 3. Among these potentially inaccurate statements, 2856 were *number of cases* statements,
463 2467 were *number of deaths* statements, 189 were *number of recoveries* statements, 9 were
464 *number of clinical tests* statements, and 10 were *number of hospitalized cases* statements. This
465 distribution of the inconsistencies among epidemiological properties is explained by the
466 dominance of *number of cases* and *number of deaths* statements on the COVID-19
467 epidemiological information. Most of these inconsistencies are linked to a violation of the
468 cumulative pattern of major variables. These issues can be resolved using tools for the automatic
469 enrichment of Wikidata like QuickStatements (cf. Turki, et al., 2019) or adjusted one by one by
470 active members of WikiProject COVID-19.
471

472   *<place Table 3 near here>*
473

474 Concerning the variables issued from the integration of basic epidemiological counts (*m, R_0, mn,*
475 and *mx* statements), they give a summary overview of the statistical behavior of the studied
476 infectious pandemic and that is why they can be useful to identify whether the stated evolution of
477 the morbidity and mortality caused by the outbreak is reasonable (Delamater, et al., 2019).
478 However, the validation of these variables is more complicated due to the complexity of their
479 definition (Delamater, et al., 2019; Backer, Klinkenberg, & Wallinga, 2020; Li, et al., 2020). The
480 basic reproduction number (*R_0*) is meant to be a constant that characterizes the dissemination
481 power of an infectious agent. It is defined as the expected number of people (within a community
482 with no prior exposure to the disease) that can contract a disease via the same infected individual.
483 This variable should exceed the threshold of 1 to define a contagious disease (Delamater, et al.,
484 2019). Although *R_0* can give an idea about the general behavior of an outbreak of a given disease,
485 any calculated value depends on the model used for its computation (e.g., *SIR* Model) as well as
486 the underlying data and is consequently a bit imprecise and variable from one study to another

487   (Delamater, et al., 2019). That is why it is not reliable to use this variable to evaluate the accuracy
488   of epidemiological counts for a given pandemic. The only heuristic that can be applied to this
489   variable is to verify if its value exceeds 1 for diseases causing large outbreaks. The incubation
490   period of a disease gives an overview of the silent time required by an infectious agent to become
491   active in the host organism and cause notable symptoms (Backer, Klinkenberg, & Wallinga,
492   2020; Li, et al., 2020). This variable is very important, as it reveals how many days an inactive
493   case can spread the disease in the host's environment before the host is being symptomatically
494   identified. As a result, it can give an idea about the contagiousness of the infectious disease and
495   its basic reproduction number (R_0). However, the determination of the incubation period -
496   especially for a novel pathogen - is challenging, as a patient often cannot identify with precision
497   the day when they had been exposed to the disease, at least if they did not travel to an endemic
498   region or had not been in contact with a person they knew to be infected. This factor was behind
499   the measurement of falsely small incubation periods for COVID-19 at the beginning of the
500   COVID-19 epidemic in China (Backer, Klinkenberg, & Wallinga, 2020). Furthermore, the use of
501   minimal (*mn*) and maximal (*mx*) incubation periods in Wikidata to epidemiologically describe a
502   disease instead of the median incubation period is a source of a lack of accuracy of the extracted
503   values (Backer, Klinkenberg, & Wallinga, 2020; Li, et al., 2020). Minimal and maximal
504   incubation periods for a given disease are obtained in the function of the mean ($\bar{x}$) and standard
505   deviation (σ) of the measures of the confidence interval of observed incubation periods in

506   patients. Effectively, *mn* is equal to $\bar{x} - \frac{z * \sigma}{\sqrt{n}}$ and *mx* is equal to $\bar{x} + \frac{z * \sigma}{\sqrt{n}}$ where *n* is the number of

507   analyzed observations and *z* is a characteristic of the hypothetical statistical distribution and of
508   the statistical confidence level adopted for the estimation (Altman, et al., 2013). As a
509   consequence, *mn* and *mx* variables are modified according to the number of observations (*n*) with
510   a smaller difference between the two variables for higher values of *n*. The two measures also vary
511   according to the used statistical distribution and that is why different values of *mn* and *mx* were
512   reported for COVID-19 when applying different distributions (Weibull, gamma, and log-normal
513   distribution) using a confidence level of 0.95 on the same set of observed cases (Backer,
514   Klinkenberg, & Wallinga, 2020). Similarly, the two variables can change according to the
515   adopted confidence level (*p* - 1) when using the same statistical distribution where a higher
516   confidence level is correlated with a higher difference between the calculated *mn* and *mx* values,
517   as shown in Fig. 7 (Ward & Murray-Ward, 1999; Altman, et al., 2013). Given these reasons and
518   despite the significant importance of the two measures, these two statistical variables cannot be
519   used to evaluate statistical epidemiological counts for COVID-19 due to their lack of precision
520   and difficulty of determination.
521
522   *<place Figure 7 near here>*
523
524   As for the reported case fatality rate (*m*), it is the quotient of the cumulative number of deaths (*d*)
525   and the cumulative number of cases (*c*) as stated in official reports. It is consequently
526   straightforward to validate for a given disease by comparing its values with reported counts of
527   cases and deaths (Rothman, Greenland, & Lash, 2008). Here, two simple heuristics can be

528   applied using SPARQL queries as shown in Appendix B. As the number of deaths is less than or
529   equal to the number of cases of a given disease, m values should be set between 0 and 1. That is
530   why Task M1 is defined to extract m statements where $m > 1$ or $m < 0$. Also, as $m = d / c$ for a
531   date Z, m values that are not close to the corresponding quotients of deaths by disease cases
532   should be identified as deficient and m values should be stated for a given date Z if mortality and
533   morbidity counts exist. Thus, Task M2 is created to extract m values where the absolute value of
534   ($m - d/c$) is superior to 0.001, and Task M3 is developed to identify (item, date) pairs where m
535   statements are missing and c and d statements are available in Wikidata. Absolute values for Task
536   M2 are obtained using SPARQL's ABS function, and deficient (item, date) pairs are eliminated
537   in Task M3 where $m > 1$ and $c < d$.
538   As a result of these three tasks, we interestingly identified 143 problematic m statements and
539   7116 missing m statements. 133 of the problematic statements are identified thanks to Task M2
540   and concern 25 Wikidata items and 31 distinct dates, and only 10 deficient statements related to 3
541   Wikidata items and 8 distinct dates are found using Task M1. These statements should be
542   checked against reference datasets to verify their values and to determine the reason behind their
543   deficiency. Such a reason can be the integration of the wrong case and death counts in Wikidata,
544   or a bug or inaccuracy within the source code of the bot making or updating such statements. The
545   verification process can be automatically done using an algorithm that compares Wikidata values
546   (c, d, and m statements) with their corresponding ones in other databases (using file or API
547   reading libraries) and subsequently adjusts statements using the Wikidata API directly or via
548   tools like QuickStatements (Turki et al., 2019). As for the missing m statements returned by M3,
549   they are linked to 395 disease outbreak items and to 205 distinct dates and concern 70%
550   (7116/10168) of the (case count, death count) pairs available in Wikidata. The outcome of M3
551   proves the efficiency of comparative constraints to enrich and assess the completeness of
552   epidemiological data available in a knowledge graph, particularly Wikidata, based on existing
553   information. Consequently, derivatives of Task M3 can build to infer d values based on c and m
554   statements or to find c values based on d and m statements. The missing statements found by such
555   tasks can be integrated in Wikidata using a bot based on Wikidata API and Wikidata Query
556   Service to ameliorate the completeness and integrity of available mortality data for epidemics,
557   mainly the COVID-19 pandemic (Turki, et al., 2019).
558

Discussion

560   The results presented here demonstrate the value of our statistical constraints-based validation
561   approach for knowledge graphs like Wikidata across a range of features (Tables 2 and 3). These
562   tasks successfully address most of the competency questions, particularly conceptual orientation
563   (*clarity*), coherence (*consistency*), strength (*precision*), and full coverage (*completeness*).
564   Combined with previous findings in the context of bioinformatics (Bolleman, et al., 2020; Marx
565   & Krötzsch, 2017; Darari, et al., 2020), this proves that the efficiency of rule-based approaches to
566   evaluate semantic information from scratch displays a similar accuracy as other available
567   ontology evaluation algorithms (Amith, et al., 2019; Zhang & Bodenreider 2010). The efficiency
568   of these constraint-based assessment methods can be further enhanced by using machine learning
569   techniques to perform imputations and adjustments on deficient data (Bischof, et al., 2020). The

570   scope of rule-based methods can be similarly expanded to cover other competency questions such
571   as non-redundancy (*conciseness*) through the proposal of other logical constraints to tackle them,
572   such as a condition to find taxonomic relations to trim in a knowledge graph (examples can be
573   found at https://www.wikidata.org/wiki/Wikidata:Database_evaluation). The main limitation of
574   applying the logical constraints using SPARQL in the context of Wikidata is that the runtime of a
575   query that infers or verifies a complex condition or that analyzes a huge amount of class items or
576   property use cases can exceed the timeout limit of the used endpoint (Malyshev, et al., 2018;
577   Chah & Andritsos, 2021). Here, the inference of logical constraints and the identification of
578   inconsistent semantic information through the analysis of full dumps of Wikidata can be more
579   efficient, although this comes with advanced storage and processing requirements (Chah &
580   Andritsos, 2021). Another option can be either the loading of a Wikidata dump and the running
581   of those queries on a designated SPARQL endpoint with more permissive timeout settings or the
582   use of one of the publicly available clones, though their data is usually less complete[27].
583   These evaluation assignments covered by our approach can be done by other rule-based
584   (*structure-based* and *semantic-based*) ontology evaluation methods. Structure-based methods
585   verify whether a knowledge graph is defined according to a set of formatting constraints, and
586   semantic-based methods check whether concepts and statements of a knowledge graph meet
587   logical conditions (Amith, et al., 2018). Some of these methods are software tools, particularly
588   Protégé extensions such as OWLET (Lampoltshammer & Heistracher, 2014) and OntoCheck
589   (Schober, et al., 2012). OWLET infers the JSON schema logics of a given knowledge graph,
590   converts them into OWL-DL axioms, and uses the semantic rules to validate the assessed
591   ontological data (Lampoltshammer & Heistracher, 2014). OntoCheck screens an ontology to
592   identify structural conventions and constraints for the definition of the analyzed relational
593   information and consequently to homogenize the data structure and quality of the ontology by
594   eliminating typos and pattern violations (Schober, et al., 2012). Here, the advantage of applying
595   constraints using SPARQL is that its runtime is faster, as it does not require the download of the
596   full dumps of the evaluated knowledge graph (Malyshev, et al., 2018). The benefit of our method
597   and other structure-based and semantic-based web-based tools for knowledge graph validation
598   like OntoKeeper (Amith, et al., 2019) and adviseEditor (Geller, et al., 2013), when compared to
599   software tools, is that the maximal size of the knowledge graphs that can be assessed by web
600   services is larger than the one that can be evaluated by software tools because the latter depends
601   on the requirements and capacities of the host computer (Lampoltshammer & Heistracher, 2014;
602   Schober, et al., 2012). These drawbacks of other structure-based tools can indeed be solved
603   through the simplification of the knowledge graph by reducing redundancies using techniques
604   like ontology trimming (Jantzen, et al., 2011) or through the construction of an abstraction
605   network to decrease the complexity of the analyzed knowledge graph (Amith, et al., 2018;
606   Halper, et al., 2015). However, knowledge graph simplification processes are time-consuming,
607   and resulting time gain can consequently be insignificant (Jantzen, et al., 2011; Amith, et al.,
608   2018; Halper, et al., 2015).

---

49   [27]For instance, the query SELECT (COUNT(*) AS ?c) WHERE {?s ?p ?o} currently gives 11857528152 results on
50   the clone at https://wikidata.demo.openlinksw.com/sparql that was set up by Chalupsky et al. (2021), while the live
51   Wikidata result as of 23 July 2022 is 14040950269.

609   Such tasks can also be solved using data-driven ontology evaluation methods. These
610   techniques process texts in natural languages to validate the concepts and statements of a
611   knowledge graph and currently include intrinsic (*lexical-based*) and extrinsic (*cross-validation,*
612   *big data-based,* and *corpus-based*) methods (Amith, et al., 2018). Lexical-based methods use
613   rules implemented in SQL or SPARQL to retrieve items and glosses corresponding to a concept
614   and their semantic relations (mostly *subclass of* statements) (Rector & Iannone, 2012; Luo,
615   Mejino Jr, & Zhang, 2013). These items are then compared against a second set of rules to
616   identify inconsistencies in their labels, descriptions, or semantic relations (Amith, et al., 2018).
617   The output can then be analyzed using natural language processing techniques such as hamming
618   distance measures (Luo, Mejino Jr, & Zhang, 2013), semantic annotation tools (Rector &
619   Iannone, 2012), and semantic similarity measures (Amith, et al., 2018) to comparatively identify
620   deficiencies in the semantic representation, labelling, and symmetry of the assessed knowledge
621   graph. Conversely, extrinsic data-based methods extract the usage and linguistic patterns from
622   raw text corpuses such as bibliographic databases and clinical records (*Corpus-based methods*) or
623   from gold standard semantic resources like large ontologies and knowledge graphs (*Cross-*
624   *validation methods*) or social media posts and interactions, Internet of Things data or web service
625   statistics (*Big data-based methods*) (Amith, et al., 2018; Sebei, Hadj Taieb, & Ben Aouicha,
626   2018; Rector, Brandt, & Schneider, 2011; Gangemi, et al., 2005) using structure-based and
627   semantic-based ontology evaluation methods as explained above (Rector, Brandt, & Schneider,
628   2011) as well as a range of techniques including machine learning (Bean, et al., 2017; Zhang, et
629   al., 2018), topic modeling using Latent Dirichlet Analysis (Abd-Alrazaq, et al., 2020), word
630   embeddings (Zhang, et al., 2019), statistical correlations (Vanderkam, et al., 2013) and semantic
631   annotation methods (Li, et al., 2016). The returned features of the analyzed resources are
632   compared to the ones of the analyzed knowledge graph to assess the accuracy and completeness
633   of the definition and use of concepts and properties (Amith, et al., 2018).

634   When compared to our proposed approach, lexical-based methods have the advantage to
635   identify and adjust characteristics of a knowledge graph item based on its natural language
636   information of a knowledge graph item, particularly terms and glosses (Rector & Iannone, 2012;
637   Luo, Mejino Jr, & Zhang, 2013). The drawback of using semantic similarity, word embeddings,
638   and topic modeling techniques in such approaches is that these techniques are sensitive to the
639   used parameters, to input characteristics, and to the chosen models of computation and can
640   consequently give different results according to the context of determination (Lastra-Díaz, et al.,
641   2019; Hadj Taieb, Zesch, & Ben Aouicha, 2020). The current role of constraints in the extraction
642   of lexical information and respective semantic relations (Rector & Iannone, 2012; Luo, Mejino
643   Jr, & Zhang, 2013) proves that the scope of constraint-based validation should not only be
644   restricted to rule-based evaluation but also to lexical-based evaluation. Yet, the function of
645   logical conditions should be expanded to refine the list of pairs (lexical information, semantic
646   relation) to more accurately identify deficient and missing semantic relations and defective
647   lexical data and to support multilingual lexical-based methods. This would build on the many
648   SPARQL functions that analyze strings in knowledge graphs[28] such as STRLEN (length of a
649   string), STRSTARTS (verification of a substring beginning a given string), STRENDS

---

52   [28]   Detailed information about string functions in SPARQL can be found at https://www.w3.org/TR/sparql11-
53   query/#func-strings.

650   (verification of a substring finishing a given string), and CONTAINS (verification of a substring
651   included in a given string) (DuCharme, 2013; Harris, Seaborne, & Prud'hommeaux, 2013).

652   As for the extrinsic data-driven methods, they are mainly based on large-scale resources that
653   are regularly curated and enriched. Raw-text corpora are mainly composed of scholarly
654   publications (Raad & Cruz, 2015) and blog posts (Park, et al., 2016). Information in scholarly
655   publications is ever-changing according to the dynamic advances in scholarly knowledge,
656   particularly medical data (Jalalifard, Norouzi, & Isfandyari-Moghaddam, 2013). This expansion
657   of scientific information in scholarly publications is highly recognized in the context of COVID-
658   19 where detailed information about COVID-19 disease and the SARS-CoV-2 virus is published
659   within less than six months (Kagan, Moran-Gilad, & Fire, 2020). Big data is the set of real-time
660   statistical and textual information that is generated by web services including search engines and
661   social media and by the Internet of Things objects including sensors (Sebei, Hadj Taieb, & Ben
662   Aouicha, 2018). This data is characterized by its value, variety, variability, velocity, veracity, and
663   volume (Sebei, Hadj Taieb, & Ben Aouicha, 2018) and can be consequently used to track the
664   changes of the community knowledge and consciousness over time (Abd-Alrazaq, et al., 2020;
665   Turki, et al., 2020). Large semantic resources are ontologies and knowledge graphs that are built
666   and curated by a community of specialists and that are regularly verified, updated, and enriched
667   using human efforts and computer programs (Lee, et al., 2013). These resources represent broad
668   and reliable information about a given specialty through machine learning techniques (Zhang, et
669   al., 2018) and the crowdsourcing of scientific efforts (Mortensen, et al., 2014) and can be
670   consequently compared to other semantic databases for validation purposes. Examples of these
671   resources are the COVID-19 Disease Map (Ostaszewski, et al., 2020) and SNOMED-CT[29] (Lee,
672   et al., 2013).

673   Large-scale knowledge graphs are dynamic corpora. Changes in the logical and semantic
674   conditions for the definition of knowledge in a particular domain need to be identified to adjust
675   the assessed knowledge graph accordingly. Rule-based and lexical-based approaches (especially
676   constraints-based methods) are therefore less simple to apply than extrinsic data-driven methods
677   (Amith, et al., 2018). Nonetheless, the growing and changing nature of gold-standard resources
678   require continuous human efforts and an advanced software architecture to maintain (e.g.,
679   structure-based and semantic-based methods), process (e.g., *word embeddings* and *latent
680   Dirichlet analysis*), and store (e.g., *Hadoop* and *MapReduce*) these reference resources
681   (Mortensen, et al., 2014; Le, et al., 2013; Sebei, Hadj Taieb, & Ben Aouicha, 2018). This
682   architecture has advanced hardware requirements and its results are subject to change according
683   to the used parameters (Sebei, Hadj Taieb, & Ben Aouicha, 2018).

684   These tasks are in line with the usage of Shape Expressions as well as property constraints and
685   relations for the validation of data quality and completeness of the semantic information of class
686   items in knowledge graphs as shown in the "Knowledge graph validation of Wikidata" section. A
687   ShEx ShapeMap is a pair of a triple pattern for selecting entities to validate and a shape against
688   which to validate them. This allows for the definition of the properties to be used for the items of
689   a given class (Prud'hommeaux, Labra Gayo, & Solbrig, 2014; Waagmeester, et al., 2021) and
690   property constraints and relations based on the meta-ontology (i.e., data skeleton) of Wikidata.
691   Expressions written in shape-based property usage validation languages for RDF (e.g., *SHACL*)

54   [29]   Systematized Nomenclature Of Medicine - Clinical Terms

692   can be used to state conditions and formatting restrictions for the usage of relational and non-
693   relational properties (Erxleben, et al., 2014; Thornton, et al., 2019; Gangemi, et al., 2005).
694   SPARQL can be more efficient in inferring such information than the currently existing
695   techniques that screen all the items and statements of a knowledge graph one by one to identify
696   the conditions for the usage of properties (e.g., *SQID*) mainly because SPARQL is meant to
697   directly extract information according to a pattern without having to evaluate all the conditions
698   against all items of a knowledge graph (Marx & Krötzsch, 2017; Hanika, et al., 2019; Pérez,
699   Arenas, & Gutierrez, 2009).

700   The separate execution of value-based constraints is common in the quality control of XML
701   data. Typically, structural constraints are managed by RelaxNG or XML Schemas, while value-
702   based constraints are captured as Schematron. Much as Schematron rules are typically embedded
703   in RelaxNG, the consistency constraints presented above can be embedded in Shape Expressions
704   Semantic Actions or in SHACL-SPARQL as shown in Fig. 8 (Melo & Paulheim, 2020). These
705   supplement structural schema languages with mechanisms to capture value-based constraints and
706   in doing so, provide context for the enforcement of those constraints. The implementation of
707   value-based constraints shown in the "Constraint-driven heuristics-based validation of
708   epidemiological data" section can likewise be implemented in a shapes language (Labra-Gayo, et
709   al., 2019). Parsing the rules in Table 2 would allow the mechanical generation or augmentation of
710   shapes, providing flexibility for how the rules are expressed while still exploiting the power of
711   shape languages for validation. More generally, ontology-based and knowledge graph-based
712   software tools have the potential to provide wide data and platform interoperability, and thus their
713   semantic interoperability is relevant for a range of downstream applications such as IoT and WoT
714   technologies (Gyrard, Datta, & Bonnet, 2018).

715
716   *<place Figure 8 near here>*
717

Conclusion

719   In this paper, we investigate how to best assess epidemiological knowledge in collaborative
720   ontologies and knowledge graphs using statistical constraints which we describe based on the
721   example of COVID-19 data in Wikidata. Collaborative databases produced through the
722   cumulative edits of thousands of users can generate huge amounts of structured information
723   (Turki, et al., 2019) but as a result of their rather uncoordinated development, they often lead to
724   uneven coverage of crucial information and inconsistent expression of that information. The
725   resulting gaps are a significant problem (conflicting values, reasoning deficiencies, and missing
726   statements). Avoiding, identifying, and closing these gaps is therefore of top importance. We
727   presented a standardized methodology for auditing key aspects of data quality and completeness
728   for these resources[30].

729   This approach complements and informs shape-based methods for data conformance to
730   community-decided schemas. The SPARQL execution does not require any pre-processing, and
731   is not only applicable to the validation of the representation of a given item according to a
732   reference data model but also to the comparison of the assessed statistical statements. Our method

---

55 [30]   This method can be adapted to meet the needs of the user. For instance, the SPARQL queries can be
56 slightly adjusted to assess other patterns in collaborative ontologies such as the usage of classes.

733   is demonstrated as useful for measuring the overall accuracy and data quality on a subset of
734   Wikidata and thus highlights a necessary first step in any pipeline for detecting and fixing issues
735   in collaborative ontologies and knowledge graphs.

736   This work has shown the state of the knowledge graph as a snapshot in time. Future work will
737   extend this to investigate how the knowledge base evolves as more biomedical knowledge is
738   integrated into it over time. This will require incorporating the edit history in the SPARQL
739   endpoint APIs of knowledge graphs (Pellissier Tanon & Suchanek, 2019, Dos Reis, Pruski, Da
740   Silveira, & Reynaud-Delaître, 2014) to dynamically visualize time-resolved SPARQL queries.
741   We will also couple the information inferred using this method[31] with Shape Expressions and the
742   explicit constraints of relation types to provide a more effective enrichment, refinement, and
743   adjustment of collaborative ontologies and knowledge graphs with statistical data. This will be an
744   excellent infrastructure to enable the support of non-relational information. Although our paper
745   focuses on COVID-19, applying the basic approach outlined here to any disease outbreak -
746   especially those designated as Public Health Emergencies of International Concern[32], which have
747   particularly urgent needs for efficient sharing and quality assessment of data - would require
748   minimal modification, since Wikidata is progressing quickly towards greater integration with
749   biomedical reference data. Consequently, frameworks similar to the one presented here can be
750   designed for validating other types of biomedical data as well as data from other knowledge
751   domains. We look forward to extending our proposed approach to allow knowledge graphs to
752   handle non-relational statements about future epidemics and other disasters such as earthquakes
753   as well as to clinical trials.

754

Author statements

756   **Data availability:** All the SPARQL queries used in this research work are provided in the
757   appendices. The Internet Archive links of the URLs cited by this paper are made available at
758   https://web.archive.org/save/https://www.wikidata.org/w/index.php?title=User:Daniel_Mietchen/
759   sandbox&oldid=1580603965.

760

Acknowledgements

762   We thank the Wikidata community, Olivier Corby (Université Côte d'Azur, France), Odile Papini
763   (Aix-Marseille Université, France), Egon Willighagen (Maastricht University, Netherlands), and
764   Mahir Morshed (University of Illinois at Urbana-Champaign, United States of America) for
765   useful comments and discussions about the topic of this research paper. This research paper is
766   published on behalf of the WikiProject COVID-19 members: Jan Ainali, Susanna Ånäs, Erica
767   Azzellini, Mus'ab Banat, Mohamed Ben Aouicha, Alessandra Boccone, Jane Darnell, Diptanshu
768   Das, Lena Denis, Rich Farmbrough, Daniel Fernández-Álvarez, Konrad Foerstner, Jose Emilio
769   Labra Gayo, Mauricio V. Genta, Mohamed Ali Hadj Taieb, James Hare, Alejandro González
770   Hevia, David Hicks, Toby Hudson, Netha Hussain, Jinoy Tom Jacob, Dariusz Jemielniak, Krupal
771   Kasyap, Will Kent, Samuel Klein, Jasper J. Koehorst, Martina Kutmon, Antoine Logean, Tiago
772   Lubiana, Andy Mabbett, Kimberli Mäkäräinen, Tania Maio, Bodhisattwa Mandal, Nandhini

---

57   [31]   This information can be represented in the form of RDF triples where the subject is the studied relation
58   type and integrated into Wikidata.
59   [32]   Epidemiological data about the monkeypox epidemic have begun to be tracked, e.g. via the item
60   Q112070734 for the 2022 monkeypox outbreak and similar entries with a more regional focus like Q112059351 for
61   the 2022 monkeypox outbreak in the United Kingdom.

773 Meenakshi, Daniel Mietchen, Nandana Mihindukulasooriya, Mahir Morshed, Peter Murray-Rust,
774 Minh Nguyễn, Finn Årup Nielsen, Mike Nolan, Shay Nowick, Julian Leonardo Paez, João
775 Alexandre Peschanski, Alexander Pico, Lane Rasberry, Mairelys Lemus-Rojas, Diego Saez-
776 Trumper, Magnus Sälgö, John Samuel, Peter J. Schaap, Jodi Schneider, Thomas Shafee, Nick
777 Sheppard, Adam Shorland, Ranjith Siji, Michal Josef Špaček, Ralf Stephan, Andrew I. Su, Hilary
778 Thorsen, Houcemeddine Turki, Lisa M. Verhagen, Denny Vrandečić, Andra Waagmeester, and
779 Egon Willighagen.
780

References

782 Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of tweeters
783   during the COVID-19 pandemic: infoveillance study. *Journal of medical Internet research*, *22*(4),
784   e19016. doi:10.2196/19016.
785 Ahmadi, N., & Papotti, P. (2021, April). Wikidata Logical Rules and Where to Find Them. In *Companion
786   Proceedings of the Web Conference 2021* (pp. 580-581). doi:10.1145/3442442.3452343.
787 Altman, D., Machin, D., Bryant, T., & Gardner, M. (Eds.). (2013). *Statistics with confidence: confidence
788   intervals and statistical guidelines*. John Wiley & Sons. ISBN:978-0-727-91375-3.
789 Amith, M., He, Z., Bian, J., Lossio-Ventura, J. A., & Tao, C. (2018). Assessing the practice of biomedical
790   ontology evaluation: Gaps and opportunities. *Journal of Biomedical Informatics*, *80*, 1-13.
791   doi:10.1016/j.jbi.2018.02.010.
792 Amith, M., Manion, F., Liang, C., Harris, M., Wang, D., He, Y., & Tao, C. (2019). Architecture and
793   usability of OntoKeeper, an ontology evaluation tool. *BMC medical informatics and decision
794   making*, *19*(4), 152. doi:10.1186/s12911-019-0859-z.
795 Arnaout, H., Razniewski, S., Weikum, G., & Pan, J. Z. (2021, April). Negative knowledge for open-world
796   Wikidata. In *Companion Proceedings of the Web Conference 2021* (pp. 544-551).
797   doi:10.1145/3442442.3452339.
798 Backer, J. A., Klinkenberg, D., & Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-
799   nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance*,
800   *25*(5), 2000062. doi:10.2807/1560-7917.ES.2020.25.5.2000062.
801 Bansal, R., & Chawla, S. (2016). Design and development of semantic web-based system for computer
802   science domain-specific information retrieval. *Perspectives in Science, 8*, 330–333.
803   doi:10.1016/j.pisc.2016.04.067.
804 Bean, D. M., Wu, H., Iqbal, E., Dzahini, O., Ibrahim, Z. M., Broadbent, M., et al. (2017). Knowledge
805   graph prediction of unknown adverse drug reactions and validation in electronic health records.
806   *Scientific reports*, *7*(1), 1-11. doi:10.1038/s41598-017-16674-x.
807 Ben Aouicha, M., & Hadj Taieb, M. A. (2016). Computing semantic similarity between biomedical
808   concepts using new information content approach. *Journal of biomedical informatics*, *59*, 258-
809   275. doi:10.1016/j.jbi.2015.12.007.
810 Bischof, S., Harth, A., Kämpgen, B., Polleres, A., & Schneider, P. (2018). Enriching integrated statistical
811   open city data by combining equational knowledge and missing value imputation. *Journal of Web
812   Semantics*, *48*, 22-47. doi:10.1016/j.websem.2017.09.003.
813 Bolleman, J., de Castro, E., Baratin, D., Gehant, S., Cuche, B. A., Auchincloss, A. H., et al. (2020).
814   HAMAP as SPARQL rules—A portable annotation pipeline for genomes and proteomes.
815   *GigaScience, 9*(2), giaa003. doi:10.1093/gigascience/giaa003.
816 Brank, J., Grobelnik, M., & Mladenic, D. (2005). A survey of ontology evaluation techniques.
817   *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)* (pp. 166–

818   170). Ljubljana, Slovenia: Citeseer. http://citeseerx.ist.psu.edu/viewdoc/summary?
819   doi=10.1.1.101.4788.
820   Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., et al.
821   (2016). Wikidata as a semantic framework for the Gene Wiki initiative. *Database, 2016*, baw015.
822   doi:10.1093/database/baw015.
823   Chah, N., & Andritsos, P. (2021). WikiMetaData Studio: Dashboards From Data Profiling the Languages,
824   Properties, and Items of Wikidata. In *Proceedings of the 2nd Wikidata Workshop*
825   *(Wikidata@ISWC 2021)* (pp. 13:1-13:8). http://ceur-ws.org/Vol-2982/paper-13.pdf.
826   Chalupsky, H., Szekely, P., Ilievski, F., Garijo, D., & Shenoy, K. (2021). Creating and Querying
827   Personalized Versions of Wikidata on a Laptop. In *Proceedings of the 2nd Wikidata Workshop*
828   *(Wikidata@ISWC 2021)* (pp. 4:1-4:13). http://ceur-ws.org/Vol-2982/paper-4.pdf.
829   Chen, Q., Lee, K., Yan, S., Kim, S., Wei, C. H., & Lu, Z. (2020). BioConceptVec: Creating and
830   evaluating literature-based biomedical concept embeddings on a large scale. *PLoS computational*
831   *biology*, *16*(4), e1007617. doi:10.1371/journal.pcbi.1007617.
832   Cuan-Baltazar, J. Y., Muñoz-Perez, M. J., Robledo-Vega, C., Pérez-Zepeda, M. F., & Soto-Vega, E.
833   (2020). Misinformation of COVID-19 on the internet: infodemiology study. *JMIR public health*
834   *and surveillance*, *6*(2), e18444. doi:10.2196/18444.
835   Darari, F., Nutt, W., Razniewski, S., & Rudolph, S. (2020). Completeness and soundness guarantees for
836   conjunctive SPARQL queries over RDF data sources with completeness statements. *Semantic*
837   *Web*, *11*(3), 441-482. doi:10.3233/SW-190344.
838   Degbelo, A. (2017). A Snapshot of Ontology Evaluation Criteria and Strategies. *Proceedings of the 13th*
839   *International Conference on Semantic Systems* (pp. 1–8). New York: ACM.
840   doi:10.1145/3132218.3132219.
841   Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T., & Jacobsen, K. H. (2019). Complexity of the
842   basic reproduction number (R0). *Emerging infectious diseases*, *25*(1), 1-4.
843   doi:10.3201/eid2501.171901.
844   Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real
845   time. *The Lancet infectious diseases*, *20*(5), 533-534. doi: 10.1016/S1473-3099(20)30120-1.
846   Dos Reis, J. C., Pruski, C., Da Silveira, M., & Reynaud-Delaître, C. (2014). Understanding semantic
847   mapping evolution by observing changes in biomedical ontologies. *Journal of biomedical*
848   *informatics*, *47*, 71-82. doi:10.1016/j.jbi.2013.09.006
849   DuCharme, B. (2013). *Learning SPARQL: querying and updating with SPARQL 1.1*. O'Reilly Media, Inc.
850   ISBN:978-1449306595.
851   Ekins, S., Mietchen, D., Coffee, M., Stratton, T. P., Freundlich, J. S., Freitas-Junior, L., et al. (2016).
852   Open drug discovery for the Zika virus. *F1000Research*, *5*, 150.
853   doi:10.12688/f1000research.8013.1.
854   Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. (2014). Introducing Wikidata to the
855   Linked Data Web. *The Semantic Web – ISWC 2014* (pp. 50–65). Springer International
856   Publishing. doi:10.1007/978-3-319-11964-9_4.
857   Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked data quality of DBpedia, Freebase,
858   OpenCyc, Wikidata, and YAGO. *Semantic Web, 9*(1), 77–129. doi:10.3233/SW-170275.
859   Farda-Sarbas, M., Zhu, H., Nest, M. F., & Müller-Birn, C. (2019). Approving automation: analyzing
860   requests for permissions of bots in wikidata. In *Proceedings of the 15th International Symposium*
861   *on Open Collaboration* (pp. 1-10). doi:10.1145/3306446.3340833.

862   Gangemi, A., Catenacci, C., Ciaramita, M., & Lehmann, J. (2005). A theoretical framework for ontology
863   evaluation and validation. In *SWAP* (Vol. 166, p. 16).
864   http://www.loa.istc.cnr.it/old/Papers/swap_final_v2.pdf.
865   Gyrard, A., Datta, S. K., & Bonnet, C. (2018). A survey and analysis of ontology-based software tools for
866   semantic interoperability in IoT and WoT landscapes. *2018 IEEE 4th World Forum on Internet of*
867   *Things (WF-IoT),* (pp. 86–91). doi:10.1109/WF-IoT.2018.8355091.
868   Hadj Taieb, M. A., Zesch, T., & Ben Aouicha, M. (2020). A survey of semantic relatedness evaluation
869   datasets and procedures. *Artificial Intelligence Review*, *53*(6), 4407-4448. doi:10.1007/s10462-
870   019-09796-3.
871   Halper, M., Gu, H., Perl, Y., & Ochs, C. (2015). Abstraction networks for terminologies: supporting
872   management of "big knowledge". *Artificial intelligence in medicine*, *64*(1), 1-16.
873   doi:10.1016/j.artmed.2015.03.005
874   Hanika, T., Marx, M., & Stumme, G. (2019). Discovering implicational knowledge in Wikidata. In
875   *International Conference on Formal Concept Analysis* (pp. 315-323). Springer, Cham.
876   doi:10.1007/978-3-030-21462-3_21.
877   Harris, S., Seaborne, A., & Prud'hommeaux, E. (2013). SPARQL 1.1 query language. *W3C*
878   *recommendation*, *21*(10), 778.
879   Heftberger, A., Höper, J., Müller-Birn, C., & Walkowski, N.-O. (2020). Opening up Research Data in
880   Film Studies by Using the Structured Knowledge Base Wikidata. In H. Kremers, *Digital Cultural*
881   *Heritage* (pp. 401–410). Springer International Publishing. doi:10.1007/978-3-030-15200-0_27.
882   Heymann, D. L. (2020). Data sharing and outbreaks: best practice exemplified. *The Lancet, 395* (10223),
883   469-470. doi: 10.1016/S0140-6736(20)30184-7.
884   Jalalifard, M., Norouzi, Y., & Isfandyari-Moghaddam, A. (2013). Analyzing web citations availability and
885   half-life in medical journals. *Aslib Proceedings*, *65*(3), 242. doi:10.1108/00012531311330638.
886   Jantzen, S. G., Sutherland, B. J., Minkley, D. R., & Koop, B. F. (2011). GO Trimming: Systematically
887   reducing redundancy in large Gene Ontology datasets. *BMC research notes*, *4*(1), 267.
888   doi:10.1186/1756-0500-4-267.
889   Jemielniak, D. (2014). *Common knowledge?: An ethnography of Wikipedia.* Stanford: Stanford University
890   Press. ISBN:978-0804789448
891   Jemielniak, D., & Wilamowski, M. (2017). Cultural diversity of quality of information on Wikipedias.
892   *Journal of the Association for Information Science and Technology, 68*(10), 2460-2470.
893   doi:10.1002/asi.23901.
894   Jemielniak, D., & Przegalinska, A. (2020) *Collaborative Society*, Cambridge, MA: MIT Press. ISBN:978-
895   0262537919.
896   Jeschke, J. M., Heger, T., Kraker, P., Schramm, M., Kittel, C., Mietchen, D. (2021) Towards an open,
897   zoomable atlas for invasion science and beyond. *NeoBiota, 68*: 5-18. doi:
898   10.3897/neobiota.68.66685.
899   Kaffee, L. A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., & Pintscher, L. (2017). A glimpse into
900   babel: An analysis of multilinguality in wikidata. *Proceedings of the 13th International*
901   *Symposium on Open Collaboration* (p. 14). ACM. doi:10.1145/3125433.3125465.
902   Kaffee, L.-A., & Simperl, E. (2018). Analysis of Editors' Languages in Wikidata. *Proceedings of the 14th*
903   *International Symposium on Open Collaboration* (p. 21). ACM. doi:10.1145/3233391.3233965
904   Kagan, D., Moran-Gilad, J., & Fire, M. (2020). Scientometric trends for coronaviruses and other emerging
905   viral infections. *GigaScience, 9*(8), giaa085. doi:10.1093/gigascience/giaa085.

906 Knublauch, H., & Kontokostas, D. (2017, 6). *Shapes Constraint Language (SHACL), W3C*
907   *Recommendation 20 July 2017.* W3C Recommendation, #w3c#. Retrieved from
908   https://www.w3.org/TR/2017/REC-shacl-20170720/
909 Kozlov, M. (2022). Monkeypox declared a global emergency: will it help contain the outbreaks?. *Nature*.
910   doi:10.1038/d41586-022-02054-7.
911 Krishnan, L., Ogunwole, S. M., & Cooper, L. A. (2020). Historical Insights on Coronavirus Disease 2019
912   (COVID-19), the 1918 Influenza Pandemic, and Racial Disparities: Illuminating a Path Forward.
913   *Annals of Internal Medicine, 173*(6), 474-481. doi:10.7326/M20-2223.
914 Labra Gayo, J. E., & Alvarez Rodríguez, J. M. (2013). Validating statistical index data represented in
915   RDF using SPARQL queries. *RDF Validation Workshop. Practical Assurances for Quality RDF*
916   *Data.* Cambridge: http://www.w3.org/2012/12/rdf-val.
917 Labra Gayo, J. E., Prud'Hommeaux, E., Boneva, I., & Kontokostas, D. (2017). Validating RDF data.
918   *Synthesis Lectures on Semantic Web: Theory and Technology*, *7*(1), 1-328.
919   doi:10.2200/s00786ed1v01y201707wbe016.
920 Labra-Gayo, J. E., García-González, H., Fernández-Alvarez, D., & Prud'hommeaux, E. (2019).
921   Challenges in RDF validation. In Current Trends in Semantic Web Technologies: Theory and
922   Practice (pp. 121-151). Springer, Cham. doi:10.1007/978-3-030-06149-4_6.
923 Lastra-Díaz, J. J., Goikoetxea, J., Hadj Taieb, M. A., García-Serrano, A., Ben Aouicha, M., & Agirre, E.
924   (2019). A reproducible survey on word embeddings and ontology-based methods for word
925   similarity: linear combinations outperform the state of the art. *Engineering Applications of*
926   *Artificial Intelligence*, *85*, 645-665. doi:10.1016/j.engappai.2019.07.010.
927 Lampoltshammer, T. J., & Heistracher, T. (2014). Ontology evaluation with Protégé using OWLET.
928   *Infocommunications Journal*, *6*(2), 12-17. https://www.researchgate.net/profile/Thomas-
929   Lampoltshammer/publication/263692985_Ontology_evaluation_with_Protege_using_OWLET/
930   links/00b4953bced7997952000000/Ontology-evaluation-with-Protege-using-OWLET.pdf.
931 Lanamäki, A., & Lindman, J. (2018). Latent Groups in Online Communities: a Longitudinal Study in
932   Wikipedia. *Computer Supported Cooperative Work (CSCW)*, *27*(1), 77-106. doi:10.1007/s10606-
933   017-9295-8.
934 Lee, D., Cornet, R., Lau, F., & De Keizer, N. (2013). A survey of SNOMED CT implementations.
935   *Journal of biomedical informatics*, *46*(1), 87-96. doi:10.1016/j.jbi.2012.09.006.
936 Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., et al. (2016). BioCreative V CDR task
937   corpus: a resource for chemical disease relation extraction. *Database*, *2016*. doi:10.1093/database/
938   baw068.
939 Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., et al. (2020). Early transmission dynamics in
940   Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine,*
941   *382*, 1199-1207. doi:10.1056/NEJMoa2001316.
942 Lozano-Tello, A., & Gomez-Perez, A. (2004). Ontometric: A Method to Choose the Appropriate
943   Ontology. *Journal of Database Management (JDM)*, *15*(2), 1-18.
944   https://www.igi-global.com/article/ontometric-method-choose-appropriate-ontology/3308.
945 Luggen, M., Difallah, D., Sarasua, C., Demartini, G., & Cudré-Mauroux, P. (2019). Non-parametric Class
946   Completeness Estimators for Collaborative Knowledge Graphs—The Case of Wikidata. *The*
947   *Semantic Web – ISWC 2019* (pp. 453–469). Springer International Publishing. doi:10.1007/978-3-
948   030-30793-6_26.
949 Luo, L., Mejino Jr, J. L., & Zhang, G. Q. (2013). An analysis of FMA using structural self-bisimilarity.
950   *Journal of biomedical informatics*, *46*(3), 497-505. doi:10.1016/j.jbi.2013.03.005.

951 Malyshev, S., Krötzsch, M., González, L., Gonsior, J., & Bielefeldt, A. (2018). Getting the most out of
952   wikidata: Semantic technology usage in wikipedia's knowledge graph. *International Semantic*
953   *Web Conference* (pp. 376-394). Springer, Cham. doi:10.1007/978-3-030-00668-6_23.
954 Martin, P. A. (2018). Evaluating Ontology Completeness via SPARQL and Relations-between-Classes
955   Based Constraints. *11th International Conference on the Quality of Information and*
956   *Communications Technology (QUATIC),* (pp. 255–263). doi:10.1109/QUATIC.2018.00045.
957 Marx, M., & Krötzsch, M. (2017). SQID: Towards Ontological Reasoning for Wikidata. In *Proceedings*
958   *of the ISWC 2017 Posters & Demonstrations Track*. CEUR Workshop Proceedings.
959   https://iccl.inf.tu-dresden.de/web/Inproceedings3169/en.
960 Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., & Mutschke, P. (2014). Bibliometric-enhanced
961   information retrieval. *European Conference on Information Retrieval* (pp. 798-801). Springer,
962   Cham. doi:10.1007/978-3-319-06028-6_99.
963 Melo, A., & Paulheim, H. (2020). Automatic detection of relation assertion errors and induction of
964   relation constraints. *Semantic Web, 11*(5), 801-830. doi:10.3233/SW-200369.
965 Mietchen, D., Hagedorn, G., Willighagen, E., Rico, M., Gómez-Pérez, A., Aibar, E., Rafes, K,, Germain,
966   C., Dunning, A., Pintscher, L., & Kinzler, D. (2015). Enabling open science: Wikidata for
967   research (Wiki4R). *Research Ideas and Outcomes, 1*, e7573. doi: 10.3897/rio.1.e7573.
968 Mietchen, D., & Li, J. (2020). Quantifying the Impact of Data Sharing on Outbreak Dynamics (QIDSOD).
969   *Research Ideas and Outcomes, 6*, e54770. doi: 10.3897/rio.6.e54770.
970 Miquel-Ribé, M., & Laniado, D. (2018). Wikipedia Culture Gap: Quantifying Content Imbalances Across
971   40 Language Editions. *Frontiers in Physics, 6*, 54. doi:10.3389/fphy.2018.00054.
972 Mitraka, E., Waagmeester, A., Burgstaller-Muehlbacher, S., Schriml, L. M., Su, A. I., & Good, B. M.
973   (2015). Wikidata: A platform for data integration and dissemination for the life sciences and
974   beyond. *bioRxiv,* 031971. doi:10.1101/031971.
975 Mora-Cantallops, M., Sánchez-Alonso, S., & García-Barriocanal, E. (2019). A systematic literature
976   review on Wikidata. *Data Technologies and Applications, 53*, 250–268. doi:10.1108/DTA-12-
977   2018-0110.
978 Mortensen, J. M., Minty, E. P., Januszyk, M., Sweeney, T. E., Rector, A. L., Noy, N. F., & Musen, M. A.
979   (2014). Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of
980   SNOMED CT. *Journal of the American Medical Informatics Association, 22,* 640-648.
981   doi:10.1136/amiajnl-2014-002901.
982 Müller-Birn, C., Karran, B., Lehmann, J., & Luczak-Rösch, M. (2015). Peer-production System or
983   Collaborative Ontology Engineering Effort: What is Wikidata? *Proceedings of the 11th*
984   *International Symposium on Open Collaboration* (pp. 20:1–20:10). New York: ACM.
985   doi:10.1145/2788993.2789836.
986 Nielsen, F. Å., Mietchen, D., & Willighagen, E. (2017). Scholia, scientometrics and wikidata. In
987   *European Semantic Web Conference* (pp. 237-259). Springer, Cham. doi:10.1007/978-3-319-
988   70407-4_36.
989 Nielsen, F. Å., Thornton, K., & Labra-Gayo, J. E. (2019). Validating Danish Wikidata lexemes. In *15th*
990   *International Conference on Semantic Systems, SEMPDS 2019.* Karlsruhe: CEUR-WS.
991 Obrst, L., Ceusters, W., Mani, I., Ray, S., & Smith, B. (2007). The Evaluation of Ontologies. *Semantic*
992   *Web*, 139–158. doi:10.1007/978-0-387-48438-9_8.
993 Ostaszewski, M., Mazein, A., Gillespie, M. E., Kuperstein, I., Niarakis, A., Hermjakob, H., et al. (2020).
994   COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host
995   interaction mechanisms. *Scientific data, 7*(1), 136. doi:10.1038/s41597-020-0477-8.

996   Park, M. S., He, Z., Chen, Z., Oh, S., & Bian, J. (2016). Consumers' use of UMLS concepts on social
997   media: diabetes-related textual data analysis in blog and social Q&A sites. *JMIR medical*
998   *informatics*, *4*(4), e41. doi:10.2196/medinform.5748.

999   Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods.
1000   *Semantic Web, 8*(3), 489-508. doi:10.3233/SW-160218.

1001   Pellissier Tanon, T., & Suchanek, F. (2019). Querying the Edit History of Wikidata. *The Semantic Web:*
1002   *ESWC 2019 Satellite Events* (pp. 161–166). Springer International Publishing. doi:978-3-030-
1003   32327-1_32.

1004   Pellissier Tanon, T., Bourgaux, C., & Suchanek, F. (2019). Learning how to correct a knowledge base
1005   from the edit history. In *The World Wide Web Conference* (pp. 1465-1475).
1006   doi:10.1145/3308558.3313584.

1007   Pérez, J., Arenas, M., & Gutierrez, C. (2009). Semantics and complexity of SPARQL. *ACM Transactions*
1008   *on Database Systems (TODS), 34*(3), 16. doi:10.1145/1567274.1567278.

1009   Piad-Morffis, A., Gutiérrez, Y., & Muñoz, R. (2019). A corpus to support ehealth knowledge discovery
1010   technologies. *Journal of biomedical informatics*, *94*, 103172. doi:10.1016/j.jbi.2019.103172.

1011   Pillai, S., Soon, L.-K., & Haw, S.-C. (2019). Comparing DBpedia, Wikidata, and YAGO for Web
1012   Information Retrieval. *Intelligent and Interactive Computing* (pp. 525–535). Springer Singapore.
1013   doi:10.1007/978-981-13-6031-2_40.

1014   Piscopo, A., & Simperl, E. (2018). Who Models the World?: Collaborative Ontology Creation and User
1015   Roles in Wikidata. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 141:1–
1016   141:18. doi:10.1145/3274410.

1017   Piscopo, A., & Simperl, E. (2019). What we talk about when we talk about Wikidata quality: a literature
1018   survey. In *Proceedings of the 15th International Symposium on Open Collaboration* (pp. 17:1-
1019   17:11). doi:10.1145/3306446.3340822.

1020   Prud'hommeaux, E., Labra Gayo, J. E., & Solbrig, H. (2014). Shape Expressions: An RDF Validation and
1021   Transformation Language.In *Proceedings of the 10th International Conference on Semantic*
1022   *Systems* (pp. 32-40). doi:10.1145/2660517.2660523

1023   Raad, J., & Cruz, C. (2015). A survey on ontology evaluation methods. *Proceedings of the International*
1024   *Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*
1025   (pp. 179-186). ACM. doi:10.5220/0005591001790186.

1026   Rahman, S., Montero, M. T. V., Rowe, K., Kirton, R., & Kunik Jr, F. (2021). Epidemiology, pathogenesis,
1027   clinical presentations, diagnosis and treatment of COVID-19: a review of current
1028   evidence. *Expert review of clinical pharmacology*, *14*(5), 601-621.
1029   doi:10.1080/17512433.2021.1902303.

1030   Rasberry, L., Tibbs, S., Hoos, W., Westermann, A., Keefer, J., Baskauf, S. J., et al. (2022). WikiProject
1031   Clinical Trials for Wikidata. *medRxiv*. doi:10.1101/2022.04.01.22273328.

1032   RDA COVID-19 Working Group (2020). *RDA COVID-19; recommendations and guidelines*, *5th release*
1033   *28 May 2020*. Research Data Alliance. doi:10.15497/RDA00046.

1034   Rector, A. L., Brandt, S., & Schneider, T. (2011). Getting the foot out of the pelvis: modeling problems
1035   affecting use of SNOMED CT hierarchies in practical applications. *Journal of the American*
1036   *Medical Informatics Association*, *18*(4), 432-440. doi:10.1136/amiajnl-2010-000045.

1037   Rector, A., & Iannone, L. (2012). Lexically suggest, logically define: Quality assurance of the use of
1038   qualifiers and expected results of post-coordination in SNOMED CT. *Journal of biomedical*
1039   *informatics*, *45*(2), 199-209. doi:10.1016/j.jbi.2011.10.002.

1040   Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology*. Lippincott Williams &
1041   Wilkins. ISBN:978-1451190052.

1042   Salas, J., & Hogan, A. (2022). Semantics and Canonicalisation of SPARQL 1.1. *Semantic Web*.
1043   doi:10.3233/SW-212871.
1044   Samuel, J. (2017). Collaborative Approach to Developing a Multilingual Ontology: A Case Study of
1045   Wikidata. *Research Conference on Metadata and Semantics Research* (pp. 167–172). Springer.
1046   doi:10.1007/978-3-319-70863-8_16.
1047   Samuel, J. (2021, April). ShExStatements: Simplifying Shape Expressions for Wikidata. In *Companion*
1048   *Proceedings of the Web Conference 2021* (pp. 610-615). ACM. doi:10.1145/3442442.3452349.
1049   Sarabadani, A., Halfaker, A., & Taraborelli, D. (2017, April). Building automated vandalism detection
1050   tools for Wikidata. In *Proceedings of the 26th International Conference on World Wide Web*
1051   *Companion* (pp. 1647-1654). ACM. doi:10.1145/3041021.3053366.
1052   Sarasua, C., Checco, A., Demartini, G., Difallah, D., Feldman, M., & Pintscher, L. (2019). The evolution
1053   of power and standard Wikidata editors: comparing editing behavior over time to predict lifespan
1054   and volume of edits. *Computer Supported Cooperative Work (CSCW)*, *28*(5), 843-882.
1055   doi:10.1007/s10606-018-9344-y.
1056   Schober, D., Tudose, I., Svatek, V., & Boeker, M. (2012). OntoCheck: verifying ontology naming
1057   conventions and metadata completeness in Protégé 4. *Journal of Biomedical Semantics*, *3*(Suppl
1058   2), S4. doi:10.1186/2041-1480-3-S2-S4.
1059   Sebei, H., Taieb, M. A. H., & Aouicha, M. B. (2018). Review of social media analytics process and big
1060   data pipeline. *Social Network Analysis and Mining*, *8*(1), 30. doi:10.1007/s13278-018-0507-0.
1061   Shafee, T., Masukume, G., Kipersztok, L., Das, D., Häggström, M., & Heilman, J. (2017). Evolution of
1062   Wikipedia's medical content: past, present and future. *J Epidemiol Community Health*, *71*(11),
1063   1122-1129. doi:10.1136/jech-2016-208601.
1064   Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., & Szekely, P. (2022). A Study of the Quality of
1065   Wikidata. *Journal of Web Semantics*, *72*, 100679. doi:10.1016/j.websem.2021.100679.
1066   Shorland, A., Mietchen, D., & Willighagen, E. (2020). *Wikidata Queries around the SARS-CoV-2 virus*
1067   *and pandemic*. NL: Zenodo. doi:10.5281/zenodo.3977414.
1068   Thornton, K., Solbrig, H., Stupp, G. S., Labra Gayo, J. E., Mietchen, D., Prud'Hommeaux, E., &
1069   Waagmeester, A. (2019). Using Shape Expressions (ShEx) to share RDF data models and to guide
1070   curation with rigorous validation. *European Semantic Web Conference* (pp. 606-620). Springer.
1071   doi:10.1007/978-3-030-21348-0_39.
1072   Turki, H. (2018). Citation analysis is also useful to assess the eligibility of biomedical research works for
1073   inclusion in living systematic reviews. *Journal of clinical epidemiology*, *97*, 124-125.
1074   doi:10.1016/j.jclinepi.2017.11.002.
1075   Turki, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2018). MeSH qualifiers, publication types and relation
1076   occurrence frequency are also useful for a better sentence-level extraction of biomedical relations.
1077   *Journal of biomedical informatics*, *83*, 217-218. doi:10.1016/j.jbi.2018.05.011.
1078   Turki, H., Shafee, T., Hadj Taieb, M. A., Ben Aouicha, M., Vrandečić, D., Das, D., & Hamdi, H. (2019).
1079   Wikidata: A large-scale collaborative ontological medical database. *Journal of Biomedical*
1080   *Informatics*, *99*, 103292. doi:10.1016/j.jbi.2019.103292.
1081   Turki, H., Vrandečić, D., Hamdi, H., & Adel, I. (2017). Using WikiData as a Multi-lingual Multi-dialectal
1082   Dictionary for Arabic Dialects. *2017 IEEE/ACS 14th International Conference on Computer*
1083   *Systems and Applications (AICCSA)* (pp. 437–442). IEEE. doi:10.1109/AICCSA.2017.115.
1084   Turki, H., Hadj Taieb, M. A., Ben Aouicha, M., & Abraham, A. (2020). Nature or Science: what Google
1085   Trends says. *Scientometrics*, *124*(2), 1367-1385. doi:10.1007/s11192-020-03511-8.

1086   Turki, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2021). Coupling Wikipedia Categories with Wikidata
1087   Statements for Better Semantics. In *Proceedings of the 2nd Wikidata Workshop (Wikidata@ISWC*
1088   *2021)* (pp. 8:1-8:6). http://ceur-ws.org/Vol-2982/paper-8.pdf.
1089   Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Ben Aouicha, M., et al. (2022).
1090   Representing COVID-19 information in collaborative knowledge graphs: the case of
1091   Wikidata. *Semantic Web*, *13*(2), 233-264. doi:10.3233/SW-210444.
1092   Vanderkam, D., Schonberger, R., Rowley, H., & Kumar, S. (2013). *Nearest neighbor search in google*
1093   *correlate*. Google Inc. https://research.google/pubs/pub41694/.
1094   Vasanthapriyan, S., Tian, J., & Xiang, J. (2017). An Ontology-Based Knowledge Framework for Software
1095   Testing. *Knowledge and Systems Sciences* (pp. 212–226). Springer Singapore. doi:10.1007/978-
1096   981-10-6989-5_18.
1097   Vrandečić, D. (2009). Ontology Evaluation. In R. S. S. Staab, *Handbook on Ontologies* (pp. 293–313).
1098   Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-92673-3_13.
1099   Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of*
1100   *the ACM*, *57*(10), 78-85. doi:10.1145/2629489.
1101   Vrandečić, D. (2021). Building a multilingual Wikipedia. *Communications of the ACM*, *64*(4), 38-41.
1102   doi:10.1145/3425778.
1103   Waagmeester, A., Schriml, L., & Su, A. I. (2019). Wikidata as a linked-data hub for Biodiversity data.
1104   *Biodiversity Information Science and Standards*, *3*, e35206. doi:10.3897/biss.3.35206.
1105   Waagmeester, A., Willighagen, E. L., Su, A. I., Kutmon, M., Gayo, J. E. L., Fernández-Álvarez, D., et al.
1106   (2021). A protocol for adding knowledge to Wikidata: aligning resources on human
1107   coronaviruses. *BMC biology*, *19*(1), 12:1-12:14. doi:10.1186/s12915-020-00940-y
1108   Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Malachi, G., Griffith, O. L., et al.
1109   (2020b). Wikidata as a knowledge graph for the life sciences. *eLife*, *9*, e52614.
1110   doi:10.7554/eLife.52614.
1111   Ward, A., & Murray-Ward, M. (1999). *Assessment in the classroom*. Wadsworth Publishing Company.
1112   ISBN:978-0534527044.
1113   Walisadeera, A. I., Ginige, A., & Wikramanayake, G. N. (2016). Ontology Evaluation Approaches: A
1114   Case Study from Agriculture Domain. *Computational Science and Its Applications -- ICCSA 2016*
1115   (pp. 318–333). Springer International Publishing. doi:10.1007/978-3-319-42089-9_23.
1116   Wasi, S., Sachan, M., & Darbari, M. (2020). Document Classification Using Wikidata Properties.
1117   *Information and Communication Technology for Sustainable Development* (pp. 729–737).
1118   Singapore: Springer. doi:10.1007/978-981-13-7166-0_73.
1119   Wilder-Smith, A., & Osman, S. (2020). Public health emergencies of international concern: a historic
1120   overview. *Journal of travel medicine*, *27*(8), taaa227. doi:10.1093/jtm/taaa227.
1121   Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The
1122   FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1),
1123   160018:1-160018:9. doi:10.1038/sdata.2016.18.
1124   Wiśniewski, D., Potoniec, J., Ławrynowicz, A. & Keet, C. M. (2019). Analysis of Ontology Competency
1125   Questions and their formalizations in SPARQL-OWL. *Journal of Web Semantics, 59*, 100534.
1126   doi:10.1016/j.websem.2019.100534.
1127   Xu, B., Kraemer, M. U., & Data Curation Group (2020). Open access epidemiological data from the
1128   COVID-19 outbreak. *The Lancet Infectious Diseases*, *20*(5), 534. doi:10.1016/S1473-
1129   3099(20)30119-5.
1130   Zangerle, E., Gassler, W., Pichl, M., Steinhauser, S., & Specht, G. (2016). An Empirical Evaluation of
1131   Property Recommender Systems for Wikidata and Collaborative Knowledge Bases. *Proceedings*

1132   *of the 12th International Symposium on Open Collaboration* (pp. 18:1–18:8). New York: ACM.
1133   doi:10.1145/2957792.2957804.
1134   Zhang, G. Q., & Bodenreider, O. (2010). Large-scale, exhaustive lattice-based structural auditing of
1135   SNOMED CT. In *AMIA Annual Symposium Proceedings* (Vol. 2010, p. 922). American Medical
1136   Informatics Association. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041382/.
1137   Zhang, Y., Lin, H., Yang, Z., Wang, J., Zhang, S., Sun, Y., & Yang, L. (2018). A hybrid model based on
1138   neural networks for biomedical relation extraction. *Journal of biomedical informatics, 81*, 83-92.
1139   doi:10.1016/j.jbi.2018.03.011
1140   Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word
1141   embeddings with subword information and MeSH. *Scientific data, 6*(1), 52:1-52:9.
1142   doi:10.1038/s41597-019-0055-0.
1143   Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., & Zhang, L. J. (2020). Coronavirus
1144   disease 2019 (COVID-19): a perspective from China. *Radiology, 296*(2), E15-E25.
1145   doi:10.1148/radiol.2020200490.

Constraint types for the usage of Wikidata properties

Each property constraint is given with its Wikidata identifier, an English label and an English description

Wikidata ID	Constraint type	Description
Q19474404	single value constraint	Constraint used to specify that this property generally contains a single value per item
Q21502404	format constraint	Constraint used to specify that the value for this property has to correspond to a given pattern
Q21502408	mandatory constraint	status of a Wikidata property constraint: indicates that the specified constraint applies to the subject property without exception and must not be violated
Q21502410	distinct values constraint	Constraint used to specify that the value for this property is likely to be different from all other items
Q21510852	Commons link constraint	Constraint used to specify that the value must link to an existing Wikimedia Commons page
Q21510854	difference within range constraint	Constraint used to specify that the value of a given statement should only differ in the given way. Use with qualifiers minimum quantity/maximum quantity
Q21510856	mandatory qualifier constraint	Constraint used to specify that the listed qualifier has to be used
Q21510862	symmetric constraint	Constraint used to specify that the referenced entity should also link back to this entity
Q21510863	used as qualifier constraint	Constraint used to specify that a property must only be used as a qualifier
Q21510864	value requires statement constraint	Constraint used to specify that the referenced item should have a statement with a given property
Q21510495	relation of type constraint	relation establishing dependency between types/meta-levels of its members
Q21510851	allowed qualifiers constraint	Constraint used to specify that only the listed qualifiers should be used. Novalue disallows any qualifier
Q21510865	value type constraint	Constraint used to specify that the referenced item should be a subclass or instance of a given type
Q21514353	allowed units constraint	Constraint used to specify that only listed units may be used
Q21510857	multi-value constraint	Constraint used to specify that a property generally contains more than one value per item
Q21510859	one-of constraint	Constraint used to specify that the value for this property has to be one of a given set of items
Q21510860	range constraint	Constraint used to specify that the value must be between two given values
Q21528958	used for values only constraint	Constraint used to specify that a property can only be used as a property for values, not as a qualifier or reference
Q21528959	used as reference constraint	Constraint used to specify that a property must only be used in references or instances of citation (Q1713)
Q25796498	contemporary constraint	Constraint used to specify that the subject and the object have to coincide or coexist at some point in history
Q21502838	conflicts-with constraint	Constraint used to specify that an item must not have a given statement
Q21503247	item requires statement constraint	Constraint used to specify that an item with this statement should also have another given property
Q21503250	type constraint	Constraint used to specify that the item described by such properties should be a subclass or instance of a given type
Q54554025	citation needed constraint	Constraint specifies that a property must have at least one reference
Q62026391	suggestion constraint	status of a Wikidata property constraint: indicates that the specified constraint merely suggests additional improvements, and violations are not as severe as for regular or mandatory constraints
Q64006792	lexeme value requires lexical category constraint	Constraint used to specify that the referenced lexeme should have a given lexical category
Q42750658	value constraint	class of constraints on the value of a statement with a given property. For constraint: use specific items (e.g., "value type constraint", "value requires statement constraint", "format constraint", etc.)
Q51723761	no bounds constraint	Constraint used to specify that a property must only have values that do

		not have bounds
Q52004125	allowed entity types constraint	Constraint used to specify that only listed entity types are valid for this property
Q52060874	single best value constraint	Constraint used to specify that this property generally contains a single "best" value per item, though other values may be included as long as the "best" value is marked with a preferred rank
Q52558054	none of constraint	Constraint specifying values that should not be used for the given property
Q52712340	one-of qualifier value property constraint	Constraint used to specify which values can be used for a given qualifier when used on a specific property
Q52848401	integer constraint	Constraint used when values have to be integer only
Q53869507	property scope constraint	Constraint to define the scope of the property (main value, qualifier, references, or combination); only supported by KrBot currently

Table 1. Constraint types for the usage of Wikidata properties. Each property constraint is given with its Wikidata identifier, an English label and an English description.

Tasks for the heuristics-based evaluation of epidemiological data using the Wikidata SPARQL endpoint

Each validation task is given with its identifier, a brief description of the heuristic validation criteria and an example where the data does not fit them. See the section "Constraint-driven heuristics-based validation of epidemiological data" for definitions of the epidemiological variables.

Task	Description	Sample filtered deficient statement
Validating qualifiers of COVID-19 epidemiological statements		
V1	Verify Z as a date > November 01, 2019	*COVID-19 pandemic in X* <number of cases> *5* <point in time> **March 25, 20**
V2	Verify Q as any subclass of (P279*) of medical diagnosis (Q177719)	*COVID-19 pandemic in X* <number of cases> *5* <point in time> *March 25, 2020* <determination method> **COVID-19 Dashboard**
Ensuring the cumulative pattern of c, d, r, and t		
V3	Identify c, d, r and t statements having a value in date $Z+1$ not superior or equal to the one in date Z (Verify if $d_Z \leq d_{Z+1}$, $r_Z \leq r_{Z+1}$, $t_Z \leq t_{Z+1}$, and $c_Z \leq c_{Z+1}$)	(*COVID-19 pandemic in X* <number of cases> **5** <point in time> *March 25, 2020*) AND (*COVID-19 pandemic in X* <number of cases> **6** <point in time> *March 24, 2020*)
V4	Find missing values of c, d, r and t in date $Z+1$ where corresponding values in dates Z and $Z+2$ are equal	(*COVID-19 pandemic in X* <number of cases> **5** <point in time> *March 24, 2020*) AND (*COVID-19 pandemic in X* <number of cases> **6** <point in time> *March 26, 2020*) AND (*COVID-19 pandemic in X* <number of cases> **no value** <point in time> *March 25, 2020*)
Validating values of epidemiological data for a given date		
V5	Identifying c, d, r, h, and t statements with negative values	*COVID-19 pandemic in X* <number of cases> **-5** <point in time> *March 25, 2020*
V6	Identify h statements having a value superior to the number of cases for a date Z	(*COVID-19 pandemic in X* <number of hospitalized cases> **15** <point in time> *March 25, 2020*) AND (*COVID-19 pandemic in X* <number of cases> **5** <point in time> *March 25, 2020*)
V7	Identify c statements having a value superior or equal to the number of clinical tests for a date Z	(*COVID-19 pandemic in X* <number of clinical tests> **4** <point in time> *March 25, 2020*) AND (*COVID-19 pandemic in X* <number of cases> **5** <point in time> *March 25, 2020*)
V8	Identify c statements having a value inferior to the number of deaths for a date Z	(*COVID-19 pandemic in X* <number of deaths> **10** <point in time> *March 25, 2020*) AND (*COVID-19 pandemic in X* <number of cases> **5** <point in time> *March 25, 2020*)
V9	Identify c statements having a value inferior to the number of recoveries for a date Z	(*COVID-19 pandemic in X* <number of recoveries> **10** <point in time> *March 25, 2020*) AND (*COVID-19 pandemic in X* <number of cases> **5** <point in time> *March 25, 2020*)
V10	Comparing the epidemiological variables of a general outbreak with the ones of its components	(*COVID-19 pandemic in X* <number of cases> **10** <point in time> *March 25, 2020*) AND (*COVID-19 pandemic in Y* <number of cases> **5** <point in time> *March 25, 2020*) WHERE *X* is a district of *Y*

Table 2. Tasks for the heuristics-based evaluation of epidemiological data using the Wikidata SPARQL endpoint. Each validation task is given with its identifier, a brief description of the heuristic validation criteria and an example where the data does not fit them. Grey cells include human-readable categories of the considered tasks. SPARQL queries corresponding to V1-V10 are available in Appendix A. See the section "Constraint-driven heuristics-based validation of epidemiological data" for definitions of the epidemiological variables.

Matrix overview of data quality issues identified per validation task and epidemiological Wikidata property.

Rows represent validation tasks as defined in Table 2, columns the corresponding

epidemiological Wikidata properties, and the value in a given cell represents the number of

deficient statements identified by the row's specific task for the column's epidemiological

Wikidata property on a given date (August 8, 2020).

	c	d	r	t	h	Overall
V1	18	9	10	2	1	40
V2	2	91	6	0	0	99
V3	660	92	6	5		763
V4	2081	2247	149	1		4478
V5	0	0	0	0	0	0
V6	8				8	8
V7	1			1		1
V8	9	9				9
V9	17		17			17
V10	60	19	1	0	1	81
Overall	2856	2467	189	9	10	5496

Table 3. Matrix overview of data quality issues identified per validation task and epidemiological Wikidata property.
Rows represent validation tasks as defined in Table 2, columns the corresponding epidemiological Wikidata properties, and the value in a given cell represents the number of deficient statements identified by the row's specific task for the column's epidemiological Wikidata property on a given date (August 8, 2020).

Figure 1

RefB workflow

Process of RefB, a bot that adds scholarly references to biomedical Wikidata statements based on PubMed Central [Source: https://w.wiki/an$, License: CC BY 4.0]. The source code of RefB is available at https://github.com/Data-Engineering-and-Semantics/refb/ .
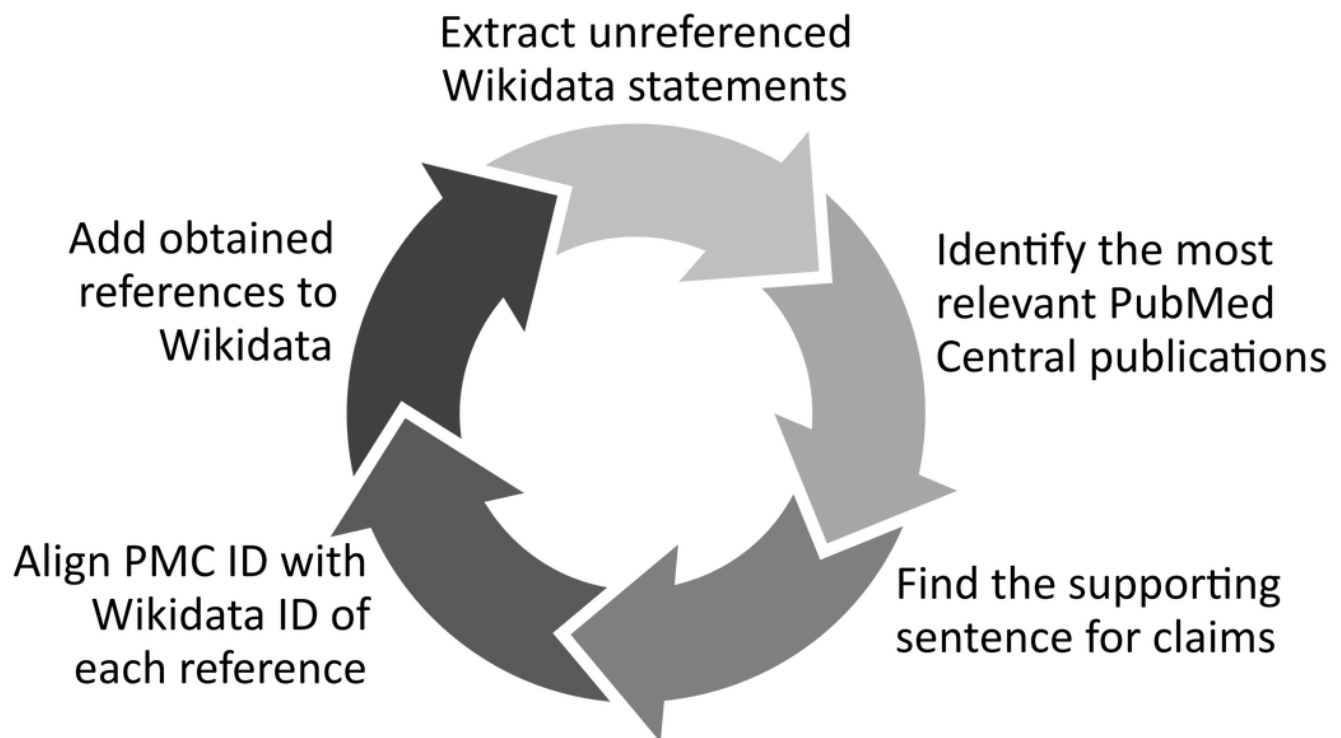
Figure 2

Example of a Wikidata property and its annotations

Wikidata page of a clinical property [Source: https://w.wiki/aeF , Derived from:
https://w.wiki/aeG , License: CC0]. It includes the labels, descriptions, and aliases of the
property in multiple languages (Red), the object data type (Blue), statements where the
property is the subject (Green) as well as property constraints (Brown).

symptoms (P780)

possible symptoms of a medical condition ✏ edit

▾ In more languages
Configure

Language	Label	Description	Also known as
English	symptoms	possible symptoms of a medical condition	
French	symptômes	manifestations ressenties par le patient atteint d'une maladie, plaintes exprimées par celui-ci	signes fonctionnels
Central Atlas Tamazight	No label defined	No description defined	
Arabic	الأعراض	No description defined	

All entered languages

Data type

Item

Statements

instance of	Wikidata property related to medicine	✏ edit
	▾ 0 references	
		+ add reference
		+ add value

subject item of this property	symptom	✏ edit
	▾ 0 references	
		+ add reference
		+ add value

Wikidata property example	meningitis	✏ edit
	symptoms headache	
	▾ 0 references	
		+ add reference
		+ add value

equivalent property	https://schema.org/signOrSymptom	✏ edit
	▾ 0 references	
		+ add reference
		+ add value

Constraints

property constraint	value type constraint	✏ edit
	class clinical sign	
	symptom	
	relation instance or subclass of	
	▾ 0 references	
		+ add reference
	type constraint	✏ edit
	class physiological condition	
	fictional medical condition	
	relation instance or subclass of	
	▾ 0 references	
		+ add reference
	citation needed constraint	✏ edit
	▾ 0 references	
		+ add reference
		+ add value

Figure 3

Example of a property constraint violation indicated via the Wikidata user interface

On the page of the Wikidata item Q3603152 (flash blindness), a constraint violation is indicated by the encircled exclamation mark. Clicking on it reveals the display of the popup with some further explanation. [File available on Wikimedia Commons: https://w.wiki/ZuJ , License: CC0].
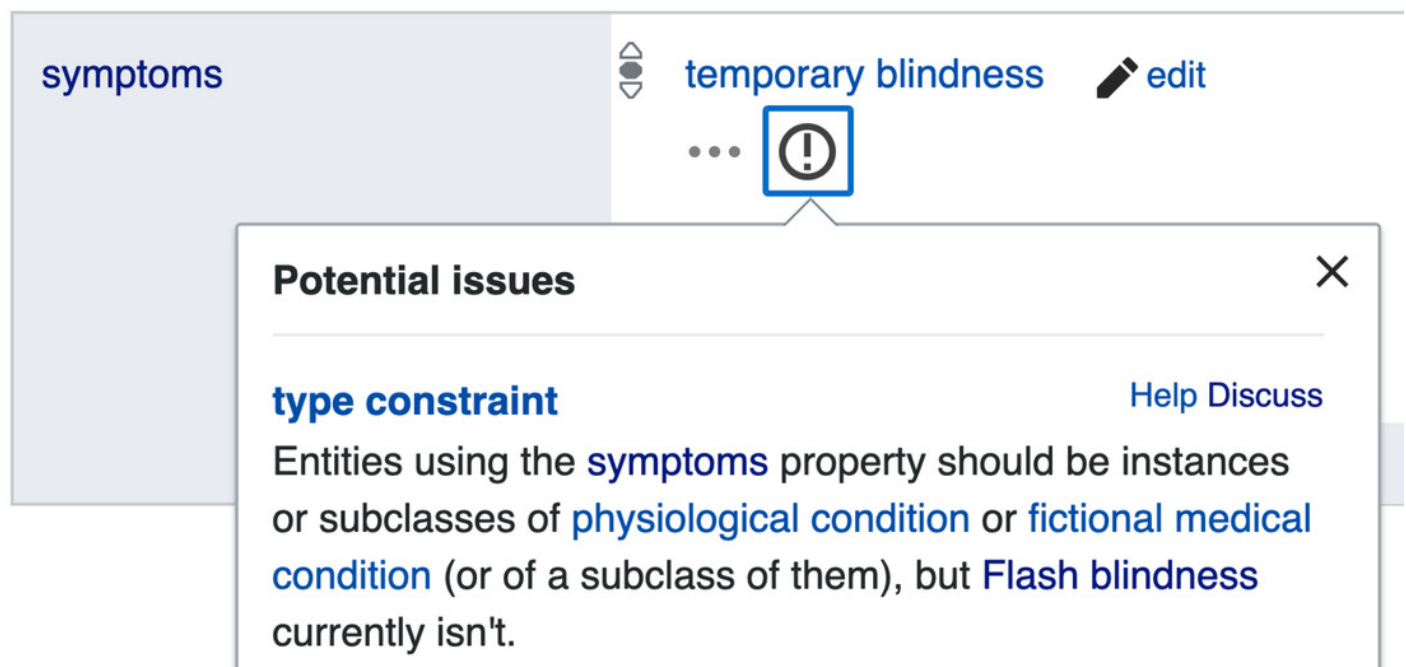
Figure 4

Entity Schema example

EntitySchema for COVID-19 dashboards, search engines and datasets [Source:

https://www.wikidata.org/wiki/EntitySchema:E205 . File available on Wikimedia Commons:

https://w.wiki/4rg5 , License: CC0.].

Figure 5

Web interface of the Wikidata Query Service

It involves a query field (Black), a query builder (Red), a short link button (Pink), a Run button (Blue), a visualization mode button (Purple), a download button (Brown), an embedding code generation button (Grey), a results field (green), and a sample query button (Yellow). [Source: https://w.wiki/aeH , Derived from: https://query.wikidata.org , License: CC0].

Figure 6

Sample statistical data available through Wikidata.

The item about the COVID-19 pandemic in Tunisia is shown. [Adapted from:

https://www.wikidata.org/wiki/Q87343682 , Source: https://w.wiki/uUr , License: CC0].

2020 COVID-19 pandemic in Tunisia (Q87343682)

viral outbreak in Tunisia   ✏ edit

2020 coronavirus outbreak in Tunisia

Statements

number of deaths	51	✏ edit
	point in time	8 August 2020
	▸ 1 reference	
	53	✏ edit
	point in time	14 August 2020
	▸ 1 reference	

case fatality rate	0.039	✏ edit
	point in time	8 April 2020
	▸ 1 reference	
	0.038	✏ edit
	point in time	4 April 2020
		7 April 2020
	▸ 1 reference	

number of cases	879	✏ edit
	point in time	18 April 2020
	▸ 1 reference	
	909	✏ edit
	point in time	21 April 2020
	▸ 2 references	

number of hospitalized cases	93	✏ edit
	point in time	22 April 2020
	▸ 1 reference	
	85	✏ edit
	point in time	20 April 2020
	▸ 1 reference	

number of recoveries	190	✏ edit
	point in time	21 April 2020
	▸ 2 references	
	170	✏ edit
	point in time	20 April 2020
	▸ 1 reference	

number of clinical tests	12,531	✏ edit
	point in time	13 April 2020
	▸ 1 reference	
	11,941	✏ edit
	point in time	12 April 2020
	▸ 1 reference	

Figure 7

Distribution statistics.

Confidence intervals for different p-values (*p*) when using a normal distribution [Source: https://w.wiki/aKT , License: Public Domain] (after Ward & Murray-Ward, 1999).
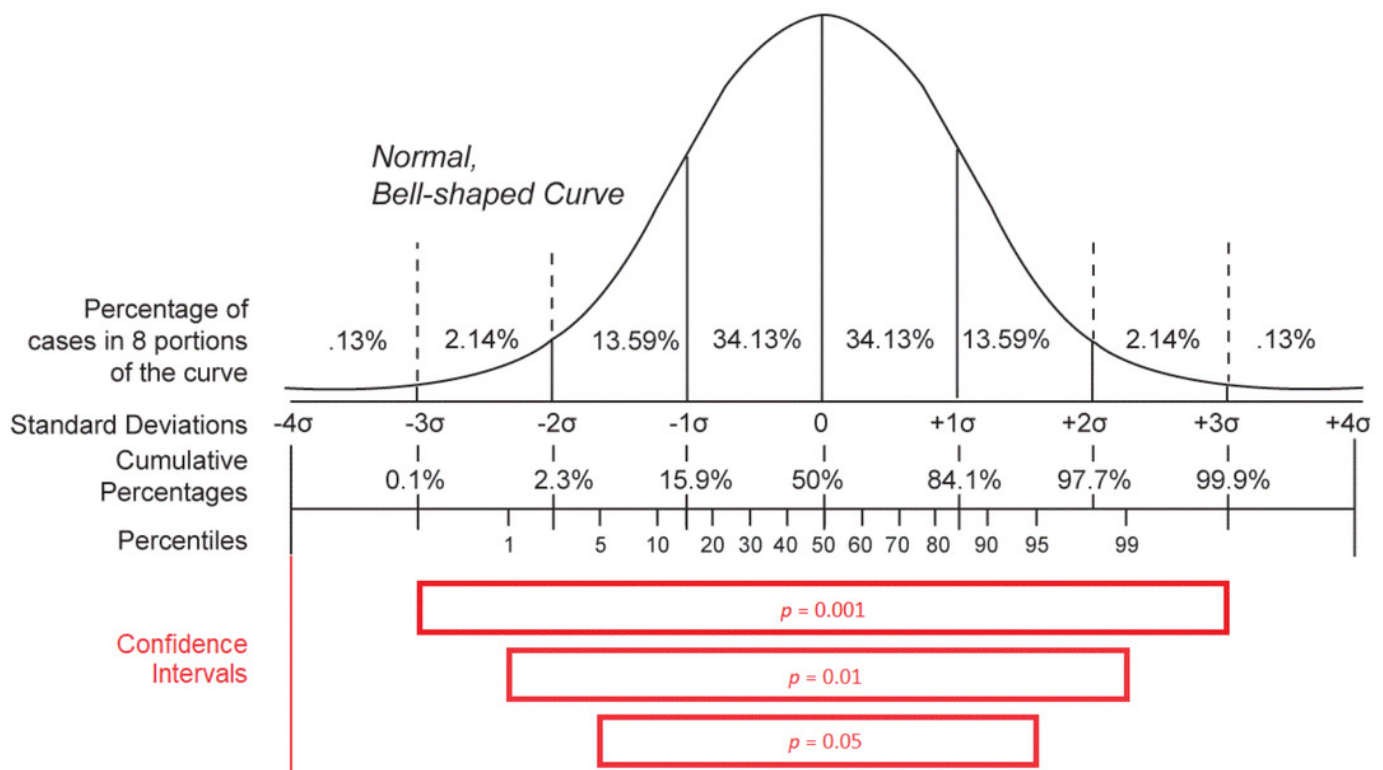
Figure 8

Key elements of data quality workflows on Wikidata.

Interactions between consistency rules, property statements, and RDF validation languages

[Source: https://w.wiki/ao5 , License: CC BY 4.0]