

University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

UNIVERSITY OF SOUTHAMPTON

Structuring the world's knowledge:
Socio-technical processes and data quality
in Wikidata

by

Alessandro Piscopo

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

October 2019

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

STRUCTURING THE WORLD'S KNOWLEDGE:
SOCIO-TECHNICAL PROCESSES AND DATA QUALITY IN WIKIDATA

by **Alessandro Piscopo**

Wikidata is a collaborative knowledge graph by the Wikimedia Foundation which has undergone an impressive growth since its launch in 2012: it has gathered a user pool of almost two hundred thousand editors, who have contributed data about more than 50 million entities. In the fashion of other Wikimedia projects, it is completely bottom-up, i.e. everything within the knowledge graph is created and maintained by its users.

These features have drawn the attention of a growing number of researchers and practitioners from several fields. Nevertheless, research about collaboration processes in Wikidata is still scarce. This thesis addresses this gap by analysing the socio-technical fabric of Wikidata and how that affects the quality of its data. In particular, it makes a threefold contribution: (*i.*) it evaluates two previously uncovered aspects of the quality of Wikidata, i.e. provenance and its ontology; (*ii.*) it is the first to investigate the effects of algorithmic contributions, i.e. bots, on Wikidata quality; (*iii.*) it looks at emerging editor activity patterns in Wikidata and their effects on outcome quality.

Our findings show that bots are important for the quality of the knowledge graph, albeit their work needs to be continuously controlled since they are potentially able to introduce different sorts of errors at a large scale. Regarding human editors, a more diverse user pool—in terms of tenure and focus of activity—seems to be associated to higher quality. Finally, two roles emerge from the editing patterns of Wikidata users, leaders and contributors. Leaders perform more edits and have a more prominent role within the community. They are also more involved in the maintenance of the Wikidata schema, their activity being positively related to the growth of its taxonomy.

This thesis contributes to the understanding of collaborative processes and data quality in Wikidata. Further studies should be carried out in order to confirm whether and to what extent its insights are generalisable to other collaborative knowledge engineering platforms.

Contents

Declaration of Authorship	xv
Acknowledgements	xvii
Nomenclature	xix
1 Introduction	1
1.1 Motivation	2
1.2 Research questions and contributions	5
1.3 Thesis outline	7
1.4 Previous publications by the author	7
1.5 Data	8
I Background work	9
2 Online knowledge collaboration	11
2.1 Online collaboration	11
2.2 Online collaboration frameworks	13
2.3 Social and organisational characteristics	13
2.4 User motivation	17
2.5 Tools and technologies	17
2.6 Success of online communities and outcome quality	18
2.7 Wikipedia	19
3 Filling the knowledge gap	23
3.1 The Semantic Web	23
3.2 Ontologies and ontology engineering	25
3.3 Linked Data	26
3.4 Knowledge graphs	27
3.5 The social Semantic Web	31
4 What we talk about when we talk about quality	35
4.1 About data quality	35
4.2 Quality dimensions	37

II Wikidata	43
5 Wikidata as a knowledge graph	45
5.1 The data model of Wikidata	45
5.2 The Wikidata ontology	52
5.3 Accessing Wikidata	55
6 Wikidata as a collaborative system	57
6.1 Editing Wikidata	57
6.2 Wikidatians	60
7 The quality of Wikidata	69
7.1 Data quality in Wikidata	69
7.2 Wikidata quality from the eyes of Wikidatians	73
III Collaborative work and data quality	79
8 Research questions and methodology	81
8.1 RQ1: Quality of human and bot edits	82
8.2 RQ2: Group composition and data quality	84
8.3 RQ3: User profiles and ontology quality	85
8.4 Data	88
8.5 Summary	89
9 Back to the sources	91
9.1 Wikidata reference policy	92
9.2 Related work	93
9.3 Approach	95
9.4 Reference editing activity: bots vs humans	102
9.5 Reference evaluation	103
9.6 Discussion	111
9.7 Limitations	113
9.8 Summary	113
10 The right mix of users	115
10.1 Related work	116
10.2 Research hypotheses	120
10.3 Data and methods	121
10.4 Results	124
10.5 Discussion	126
10.6 Limitations	127
10.7 Summary	128
11 Who models the world?	129
11.1 The Wikidata ontology	130
11.2 Related work	130
11.3 Data and methods	133
11.4 Study 1 - Ontology quality	133

11.5 Study 2 - User roles	136
11.6 Study 3 - Relationship between user roles and ontology quality	137
11.7 Results	137
11.8 Discussion	142
11.9 Summary	146
12 Conclusions	151
12.1 Contributions and results	151
12.2 Wikidatians	154
12.3 Design suggestions	155
12.4 Limitations and future work	155
12.5 Final remarks and conclusions	157
A Properties not requiring a reference	159
B English translations of the epigraphs in Parts I and III	163
Bibliography	165

List of Figures

3.1	Example of RDF triple. Resources, including the property, are identified by URIs.	24
3.2	Example of triples of an ontology. Resources in grey are part of the ontology. Properties from RDFSchema and OWL are depicted as arrows, whereas rectangles are used for properties from other datasets.	26
5.1	Components of Wikidata’s data model. Image from https://www.wikidata.org/wiki/Wikidata:Introduction , consulted on 1 February 2019. . . .	46
5.2	Three claims from Item Q84, i.e. London. Both of them use the same property, i.e. population, but qualifiers specify that they refer to different points in time. Ranks are applied to indicate that the most up to date value should be preferred. An upper grey arrow in the claim rank icon indicates <i>preferred rank</i> , a middle grey square a <i>normal rank</i> and a lower grey arrow a <i>deprecated rank</i> . Image from https://www.wikidata.org/wiki/Q84 , consulted on 1 February 2019.	47
5.3	Meta-information in a selection of Knowledge Graphs.	47
5.4	Size of Wikidata (entities and triples), compared to four major projects. Figures about DBpedia, YAGO, Freebase, and OpenCyc from Färber et al. (2018); NELL from Ringler and Paulheim (2017).	48
5.5	Number of Items along the lifespan of Wikidata	48
5.6	Language capabilities of Wikidata, compared to DBpedia, YAGO, Freebase, OpenCyc, and NELL	49
5.7	Labels, descriptions, and aliases of Item Q84 (London). Human-readable labels, as well as descriptions and aliases, are added by users, therefore they may not be available in all the languages. A link below the main box allow to show all the languages available. On the right, the links to all the language versions of Wikipedia articles describing the Item. Image from https://www.wikidata.org/wiki/Q84 , consulted on 1 August 2018.	49
5.8	Number of classes and Properties over the lifespan of Wikidata	54
5.9	Ontology size of Wikidata, compared to DBpedia, YAGO, Freebase, OpenCyc, and NELL	54
6.1	An example of Wikidata Item in editing mode. According to the text typed in the box, the system suggests a number of Items, showing their description in the user’s language of choice for disambiguation purposes.	58

6.2	Graphical example of heavyweight edit. User <i>P.</i> edits Item Q515 (i.e. city). The meaning of Q515 is determined by its super-classes. Furthermore, <i>P.</i> 's revisions affect several other Items, namely all those down the sub-class hierarchy of Q515 and their instances, e.g. London, New York, Amsterdam, etc. Edits on Items with characteristics similar to Q515 may potentially affect a very large number of other Items in the graph. Opaque Items and relations are not seen directly by the user. To give an example of the possible effects of <i>P.</i> 's edits on the Item city, if she removed the statement city::subclass of::human settlement, a query for all Items that are instances of human settlement (and of its subclasses) would not retrieve any instance of city anymore.	59
6.3	Number and percentage of edits per user type	61
6.4	Number and percentage of Property edits per user type. The peak in bot edits after June 2017 is due to the activity of a number of bots importing Property constraints from Discussion pages to the related Property definitions. Please note that the scale differ from that in Figure 6.3. . . .	61
6.5	Number of registered users and monthly active registered users along the Wikidata lifespan	64
6.6	Gini coefficient over time. Above, all Items and Properties; below, only Property edits. Property editors refers to users who have ever performed any revision on Properties.	65
6.7	Percentage of edits (above) and Property edits (below) per yearly cohort and user type	66
6.8	Percentage of edits to Discussion pages per yearly cohort and user type .	66
7.1	Examples of Property constraints violations. Figure a. is taken from Item Q7259 (Ada Lovelace). The <i>format</i> constraint checks whether the value used as an object matches a regular expression, whereas the <i>property scope</i> constraint refers to a specificity of Wikidata's knowledge representation model, i.e. the type of statement where a Property can be used. Figure b. shows a violation for Q84 (London), suggesting that the information in a statement may be incomplete.	76
8.1	Community features and aspects of quality addressed by each research question	82
8.2	Pipeline of the two-stage approach adopted to address RQ1.a and RQ1.b	84
8.3	Variables available from Wikidata historical dumps. Please note that everything in Wikidata is a web page; the metadata provided with each revision specify whether it is e.g. an item/property, a community page, etc. and its format.	89
9.1	A microtask from T1	97
9.2	A microtask from T2	97
9.3	A microtask from T3.A	98
9.4	A microtask from T3.B	98
9.5	Occurrence of properties within the sample. The graph includes only properties with more than 5 occurrences to increase readability.	101
9.6	Number of sources added by type of user and reference. Lighter colours indicate sources added by the same author of the related statement. . . .	103

9.7	Time difference between creation of statements and addition of their related reference, by reference author. The scale has been adapted to increase readability.	104
9.8	Percentage of sources by relevance. Please note the small percentage of pages not in English.	107
9.9	Percentage of sources by authoritativeness. Sources added by sources are more commonly authoritative than those added by human editors.	108
9.10	Percentage of sources by relevance and authoritativeness	109
9.11	Relevance and authoritativeness by Property, ordered by number of references within the sample evaluated. The graph includes only Properties with more than 5 occurrences to increase readability.	110
10.1	Tenure distribution in Wikidata, in number of weeks since the first edit. The vertical line marks one year (52 weeks). Tenure is computed by counting the number of weeks between the first edit of a user and the last day in our dataset.	118
11.1	Evolution of number of entities, classes, and Properties in Wikidata over time	138
11.2	Wikidata quality assessment	139
11.3	Ontology depth	139
11.4	Number of editors by months of activity on Wikidata.	140
11.5	Proportion of contributions per user type and by yearly cohort over time and percentage of users per type. The count of anonymous users considers unique IP addresses, as these users are only known through them. Nothing prevents editors to connect from different addresses, though. Years in (c.) refer to the period between October of the previous year and September of the following (e.g. 2013 means Oct. 2012–Sep. 2013).	141

List of Tables

4.1	Data quality dimensions used by Färber et al. (2018). In italics the dimensions not originally in Wang and Strong (1996).	39
5.1	Properties used in Wikidata references (instances of Q18608359) at 1st October 2017. Usage has been calculated by counting the number of references in which each Property appears.	52
6.1	Number of users and edits per type. Please note that anonymous users are estimated by means of unique IP addresses. The same person may connect from different devices, meaning that different IP addresses may refer to the same user.	60
6.2	Breakdown of users by yearly cohort	65
7.1	Wikidata data quality studies from the literature	72
7.2	Showcase Item criteria	73
7.3	Types of publisher derived from Wikidata (2018f). On the right column, sub-types or, when these are missing, definitions of higher-level types.	75
8.1	Wikidata history database tables. <i>page id</i> may be an Item or Property QID or a page title, depending on the page type. The semi-automated tool is a boolean variable we created, which tells whether an edit has been made through such tools.	89
9.1	Authoritativeness of sources (ticks indicate authoritative)	93
9.2	Crowdsourcing experiment design	99
9.3	Sample characteristics. Humans include registered and anonymous users.	102
9.4	Task statistics (includes test questions)	106
9.5	Percentage of sources by type of author	107
9.6	Percentage of sources by type of publisher	108
9.7	Performance of prediction models for relevance and authoritativeness	111
10.1	Distribution of quality levels	122
10.2	Descriptive statistics and correlations among independent variables. Item age is expressed in days since Item creation.	125
10.3	Ordinal logistic regression of number of edits and group size, editor types, and diversity measures. Note: *** $p < 0.001$, ** $p < 0.01$.	125
10.4	Ordinal logistic regression of number of edits and group size, editor types, and diversity measures, trained on Items with at least one human edit. Note: *** $p < 0.001$, ** $p < 0.01$. Model 3 has been trained on the set of Items with at least one registered human edit.	126

11.1	Ontology metric frameworks evaluation against the requirements set in the present study	134
11.2	Wikidata quality indicators (Sicilia et al., 2012) used in the present analysis	135
11.3	Features used to cluster users	136
11.4	Ontology metrics figures at 1 October 2017. In brackets, 25 th and 75 th percentiles.	137
11.5	Role transition counts	140
11.6	Breakdown of users and leaders by yearly cohort. By leader, we refer to anyone who has taken on a leader role in at least one time frame.	140
11.7	Lagged regression analysis of proportion of activity of each user type on <i>noc</i> , <i>norc</i> , <i>nolc</i> , <i>ir</i> , <i>ap</i> , and <i>ad</i>	142

Declaration of Authorship

I, Alessandro Piscopo, declare that the thesis entitled *Structuring the world's knowledge: Socio-technical processes and data quality in Wikidata* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
8. Either none of this work has been published before submission, or parts of this work have been published as: [Piscopo et al. \(2017a\)](#), [Piscopo et al. \(2017d\)](#), [Piscopo et al. \(2017b\)](#), [Piscopo and Simperl \(2018\)](#)

Signature:

Date:

Acknowledgements

This thesis would have hardly been possible without the involvement and support of many people—family, friends, colleagues. I would like to thank (some of) those who have been close to me and helped me through these years of otherwise solitary research work.

First of all, Pauline. She put up bravely with more than three years of last-minute paper submissions, working weekends, and intermittent grumpiness. She knows how grateful I am for that and happy to have her by my side (and be by hers), but this is the chance to say it once more. Thanks to Sonny (aka Picci). Although he sometimes tried to sabotage my work by walking on my keyboard, the long winter days spent writing this thesis and the papers on which it is based would have been much longer without a cat sleeping on my lap. Then, who would not thank their parents? Many thanks to them and to Francesca, and Leo. A very special mention for my sister Maria Emanuela: so young, yet so wise and funny.

I would also like to thank my supervisor, Elena Simperl, for showing me what a great researcher looks like. I hope I have been able to follow her steps at least a little bit. Big thanks you also to Chris Phethean, who supported and worked with me for a big part of the studies presented here.

Finally, I am sincerely grateful to the EU Marie Curie ITN project WDAqua. Its contribution to my PhD could hardly be overstated. Among many things, it gave me the opportunity to get exceptional training, take part in several conferences, and collaborate with a number of institutions across Europe. I want to mention especially Wikimedia Germany and those who work there: thanks for your support and for making me feel at home during my secondment. Above all, WDAqua allowed me to meet a group of great people from all around the world. Research has the ability to join people across borders, valuing their contributions regardless of where they are from. We should all do our best to make this message resonate well beyond academia.

Nomenclature

Peer-production Form of online collaboration characterised by decentralised approach and participatory governance.

Semantic Web Expansion on the web that allows to share and reuse data across applications by giving it a well-defined meaning.

Linked Data Set of best practices to share and publish data on the Semantic Web.

Linked Open Data cloud Project that identifies Linked Data sources openly available on the web to facilitate their publication.

Knowledge Graphs Graph-based knowledge representations which describe real world entities and the relations between them.

Wikidata terminology

Item An Item describe a concrete or abstract entity, or a class of entities.

Property Properties describe relations between Items.

QIDs Alphanumeric IDs used to identify Items (Qx) and Properties (Px).

Label Besides QIDs, Items and Properties have human-readable labels.

Claim A property-value pair that connects an Item to a literal or to another Item.

Statement Statements are composed of a claim and possibly a reference or a qualifier.

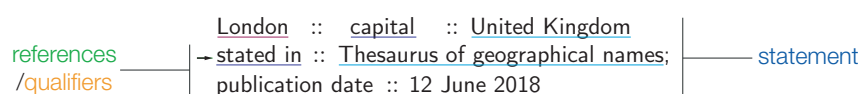
Reference References specify a source for a claim.

Qualifier Qualifiers add contextual information to a claim.

We used our own notation to represent Wikidata statements. Two colons represent a connection between a Property and its subject and value. These are colour-coded:



QIDs are expressed in **typewriter** font, whereas **sans** is used for labels. An arrow signals that references and qualifiers are attached to a claim. References/qualifiers are separated by a semicolon.



*Give every man thy ear but few thy voice;
Take each man's censure but reserve thy judgment.*
Hamlet, I-3

*What do you do with your education now that you have it
– and now that it is beginning to become obsolete even as you sit here?*
Choose one of five, speech by Edith Sampson (<https://bit.ly/2LfhDHZ>)

Chapter 1

Introduction

In January 2001, Jimmy Wales and Larry Sanger launched Wikipedia. In a short time, the free encyclopaedia achieved a massive success, acquiring shared recognition as an accurate and reliable source of information (Giles, 2005). Only available in English at the beginning, Wikipedia now can be read in 302 languages, with a total of more than 45 million articles and 500 million monthly unique visitors on average (Wikipedia, 2018b). It is the 5th most visited website globally¹ and more than 6200 publications have been dedicated to it and to other projects within the Wikimedia ecosystem up to this day².

11 years later, the Wikimedia Foundation launched Wikidata: a collaborative knowledge graph that gathers structured knowledge about the world by leveraging the efforts of a community of users. The initial aim of Wikidata was to act as a structured backbone for Wikipedia, e.g. by providing centralised interlanguage links between different versions of Wikipedia (Vrandečić, 2013). From that first goal, the ambition was to become being a key source of structured data for the web at large (Vrandečić, 2013). Wikidata’s user pool has grown so far up to 190 thousand registered users, who have contributed facts about over 55 million entities.³ This project appears to be set to replicate Wikipedia’s success, thanks to a range of factors: its open and collaborative approach, the support from the Wikimedia Foundation and the community gathered around it, insights from prior analogue projects, such as DBpedia (Auer et al., 2007) and Freebase (Bollacker et al., 2007), have created a fertile ground to make Wikidata one of the largest knowledge resources ever available in only around 6 years (Färber et al., 2018).

Knowledge graphs are a technology used to add context and depth to anything from web search to product recommendations and intelligent assistants. They are therefore crucial information sources for any AI application. They describe real-world entities, their relationships, and attributes (Paulheim, 2017). A knowledge graph typically, but

¹<https://www.alexa.com/siteinfo/wikipedia.org>, consulted on 1 February 2019.

²http://wikipapers.referata.com/wiki/List_of_publications, consulted on 19 December 2018.

³https://www.wikidata.org/wiki/Wikidata:Main_Page, consulted on 1 May 2018. This figure refers to the last revision of this thesis and may have increased since.

not necessarily, spans across several domains and is built on top of a conceptual schema, or *ontology*, which defines what types of entities (*classes*) are allowed in the graph, alongside the types of *properties* they can have.

This chapter will introduce the reader to the thesis. Section 1.1 outlines the motivations for the work. Section 1.2 presents the research questions addressed and the contributions made by this work. Subsequently, Section 1.3 provides an outline of the thesis. This work is based on a number of previous publications by the author: they are listed and summarised in Section 1.4. Finally, the data used for our analyses is described in Section 1.5.

1.1 Motivation

Wikidata has drawn the attention of researchers and practitioners from several fields due to some of its salient features (Vrandečić, 2013).

- It is **free**, i.e. its data is released under an open licence and can be reused by anyone.
- It is **collaborative**, which means that a community of users is entirely responsible to add and maintain its data, both for what concerns the entities it describes and its schema.
- It is **multilingual**, i.e. data can be edited and browsed in any of the 358 languages available, thanks to language-independent identifiers;
- It is a **secondary database**, as Wikidata’s knowledge is conceived to be verifiable and every assertion must be supported by a primary source.

Moreover, Wikidata is already tightly interlinked to numerous other structured knowledge bases. This allows its data to be easily integrated with data from other sources, thus facilitating the discovery of new information. Moreover, it effectively makes Wikidata part of the Semantic Web and a key node of the Linked Open Data cloud. The Semantic Web is a framework which extends the original web with the capability to interconnect data in order to allow its reuse and sharing across machines and applications (Berners-Lee et al., 2001). The Linked Open Data cloud is the set of intelinked datasets published on the web following the best practices known of Linked Data (Bizer et al. (2009b) and Chapter 3). An indication of the success of Wikidata so far is its inclusion in the last version of the RDF Primer (Schreiber et al., 2014), a document which provides guidance about the key concepts regarding RDF, the data model used to express facts on the Semantic Web. The use of Wikidata to give examples of resource identifiers on the web suggests that these are considered to be persistent—in other words, the authors of RDF

Primer 1.1 seem to believe that Wikidata is here to stay. This is not the only sign of Wikidata's success though. Its graph has been already adopted in a number of research projects (e.g. in [Elsahar et al. \(2018\)](#)) and by many organisations, such as museums⁴ and broadcasters.⁵

Socio-technical fabric. From a collaboration point of view, Wikidata shares features with peer-production systems and collaborative ontology development projects. On the one hand, its users add single pieces of information and work together similarly to platforms such as Wikipedia. On the other hand, they are responsible for creating and maintaining the schema of the knowledge graph, approaching an activity that is typically carried out by teams of trained professionals according to carefully crafted methodologies—ontology development—using grassroots contributions by the community and minimal guidelines and instructions. The intersection of these features, already noted by [Müller-Birn et al. \(2015\)](#), is unlike any prior project, with the partial exception of Freebase ([Bollacker et al., 2007](#)).

These activities are mediated by a range of tools that constitute the technical substrate of the system. Users perform edits through various interfaces, each affording different types of actions and providing different types of support. A particular tool, which exists also on Wikipedia, is **bots**. These are pieces of software which are able to automatically carry out actions on the platform at a large rate. One of their main tasks is editing the graph, e.g. adding data or importing information from other resources. Bots are also responsible for quality-control tasks. They patrol the graph for inconsistencies, producing reports that users may subsequently follow to fix errors. Given their disruptive potential—they edit the graph at a large scale, making revisions in the order of thousands per minute, bots are regulated by a community-defined policy, which requires to set a scope of actions for each bot and a user responsible for its operate. Bots have been so far the authors of the majority of edits in Wikidata. Yet, only a few studies have covered their activity so far (e.g. [Müller-Birn et al. \(2015\)](#); [Steiner \(2014\)](#)). Bots are one of the key technical components of Wikidata, albeit understandably not the only one. Users add and modify data, as well as communicate among them, by means of a web wiki interface. Plugins are available, which warn editors when they are about to perform a revision that may introduce any errors in the data. Other tools assist users in finding existing quality issues, or allow them to edit in large batches. Chapter 6 provides more detail about the different interfaces and technical tools in Wikidata. Wikidata is a socio-technical system, whereby this term encompasses both the social and organisational aspects of user interactions that take place on the platform and the technological artefacts in place to facilitate that and through which they occur. We refer to this intertwinement of technical solutions and social and organisational elements as to the **socio-technical fabric** of

⁴https://www.wikidata.org/wiki/Wikidata:Flemish_art_collections,_Wikidata_and_Linked_Open_Data, consulted on 1 February 2019.

⁵<http://wikimedia.fi/2016/04/15/yle-3-wikidata/>, consulted on 1 February 2019.

Wikidata. We investigated this in prior work (Piscopo et al., 2017c), specifically focusing on how early forms of participation change across editors’ activity lifespan, in terms of their identity and motivations, use of the interface, and perception of the community. We followed a qualitative approach and carried out semi-structured interviews with committed members of Wikidata. In these, interviewees reported to acquire a feeling of identity with the project and to establish stronger relationships with the community over time, albeit their core motivation remains unchanged. Furthermore, experienced users tend to devote more efforts to maintenance tasks, using a broader range of tools to perform a larger number of revisions. Whereas this study sheds light on several aspects of the behaviour of Wikidata users and their interaction with the technological layer of the platform, its scope is limited and a quantitative follow-up may help address some of the questions it leaves still open.

Data quality. A major area of interest within the field of online collaboration projects concerns the outcome of their work, whose quality is often investigated as a proxy for the success of the project (see Chapter 2). When data is the main product of collaboration, such as in the case of Wikidata, we talk about data quality. This is commonly defined as ‘fitness for use’ (Juran, 1962), which implies that quality is a concept dependent on the actual use case (Färber et al., 2018). Assessing data quality is important not only to evaluate the success of a project, but also to enable data consumers to understand whether a data source is suitable for them. The literature typically identifies a number of aspects of data quality, grouped into dimensions—the foundational study of Wang et al. (1993) is an example of that approach, more details about this are in Chapter 4.

The data quality of Wikidata has received so far insufficient attention. Although a growing body of literature exists around the topic (e.g. the works of Thakkar et al. (2016) or Färber et al. (2018)), several aspects remain uncovered. Yet, an understanding of Wikidata quality across various dimensions is key for a widespread adoption of this knowledge graph (Färber et al., 2018). In particular, one of the gaps in the literature regards the processes leading to the creation of knowledge. In their work about collaborative ontology development efforts (Strohmaier et al., 2013) distinguish the approaches evaluating the quality of ontologies as a product from those from those investigating the processes leading to their creation. An analysis of the latter, especially if combined with an evaluation of an ontology as product, is beneficial to the understanding of its quality and helps single out the areas that are likely to be contentious or problematic.

This thesis addresses specifically this aspect, i.e. collaboration processes and their effect on Wikidata quality. Our contributions could be used to improve the design of Wikidata, in order to meet the characteristics of its community and optimise outcomes.

1.2 Research questions and contributions

This thesis looks at the **socio-technical fabric** of Wikidata and analyses its relationship with the **quality** of the outcome of its community work, posing the overarching research question of *how the socio-technical fabric of Wikidata influences the quality of its data*. By community, we refer to the set of users participating to the construction and maintenance of Wikidata.

With regard to socio-technical fabric, we aim at understanding how different types of users differ in what concerns their contribution to the knowledge graph and the effects of these differences on outcome quality. We considered both explicit and implicit user types. Explicit user types result from directly observable features, whereas implicit ones arise from emerging activity patterns. As regards data quality, we investigate aspects that are specific to Wikidata, notably its Items, ontology, and provenance. We broke down our overarching question into three research questions:

RQ1 How do references added by bots and by humans compare with respect to their quality?

RQ2 To what extent does editor group diversity affect outcome quality in Wikidata?

RQ3 What features of editing roles affect the quality of the Wikidata ontology?

In addressing these questions, this thesis makes three main contributions to the collaborative knowledge engineering field: *(i.)* it gauges the data quality of two aspects of Wikidata previously uncovered, i.e. provenance and its ontology; *(ii.)* it delves into algorithmic contribution patterns in Wikidata and their role in bolstering a constant growth of the knowledge graph and their influence on its quality; *(iii.)* it investigates emerging activity patterns of human editors in Wikidata and the effects of these patterns on outcome quality. Our findings with respect to these contributions are summarised below.

Data quality evaluation. We evaluated Wikidata external sources and its ontology over time. The possibility to provide provenance for any fact stated in the graph is one of the features that set Wikidata apart from other knowledge graphs (Chapter 5). We introduced a two-staged approach to perform a large-scale assessment of Wikidata external sources. The first stage employed microtask crowdsourcing to gauge authoritativeness and relevance of sources for the piece of information they had to support. The results trained a machine learning algorithm which predicts the quality of sources on a large-scale. The crowdsourced evaluation of Wikidata sources revealed that these are mostly of good quality: ~ 60% are both relevant and authoritative. The majority

of sources are from governmental agencies or academic institutions and cannot be attributed to a single author, but are rather the product of these organisations as a whole. Overall, sources lack diversity: a large number of sources come from the same website, which may be attributed to some users performing edits in batch and to bots adding importing large amounts of data from a source.

Bot contributions and outcome quality. According to our findings bots are key for quality. However, their contributions need be balanced with human edits, in order to improve quality further. Bots often perform revisions in large batches. In this thesis, we demonstrate how this behaviour can introduce quality issues affecting large swathes of the graph. We observed primarily two types of issues: first, external sources imported to provide provenance for a large number of statements which subsequently become obsolete or stop working; second, numerous erroneous taxonomies. More frequent controls on bot activity may be effective in tackling these problems. This could be achieved for example by means of tools alerting human editors when substantial parts of the graph are modified by bots.

An evaluation of the Wikidata schema was interesting for two main reasons. First, it is the first example of large-scale ontology created and maintained in a completely bottom-up manner, whereby ontologies development typically follows well-defined practices and assigns editors different roles (see [Simperl and Luczak-Rösch \(2014\)](#)). Second, quality issues affecting the Wikidata ontology may hamper the discovery and creation of new information using its data. We applied a number of structural metrics to understand the evolution of the schema. The Wikidata ontology is large and messy. It has grown over time since the launch of the platform, reaching around 1.5M classes at the end of the period considered in our analysis (October 2017). The growth in the number of classes has a similar pace to that of the whole graph. Furthermore, a large part of the ontology is flat, with classes without instances or sub-classes, whereas other parts have extended, deep hierarchies. This suggests that a core schema exists, which is maintained by the community. These characteristics seem to confirm prior observations about the misuse of taxonomic relations by Wikidata editors ([Piscopo et al., 2017c](#)), i.e. editors create classes without a clear understanding of what these are and the consequences of adding them to the graph.

Activity patterns of human users and outcome quality. We investigated the relationships between different features of the Wikidata community and the outcome of its work. The first feature we looked at was the impact of group composition on the quality of Items, i.e. each entity described within Wikidata. Whereas more diverse groups generally lead to better outcomes, a larger number of members does not necessarily affect quality in a positive manner ([Piscopo et al., 2017b](#)). Moreover, the roles of human editors within Wikidata are less articulated than in other platforms ([Piscopo](#)

and Simperl, 2018). According to our quantitative analysis (Section 11.7.2), Wikidata users fall into two profiles, *leaders* and *contributors*. The first are typically more active, making a substantial use of automated editing tools and a larger number of edits on parts of the schema. The latter are generally active for shorter periods of time, perform fewer revisions, and focus more on adding data, rather than on editing existing pieces of information.

1.3 Thesis outline

The thesis is divided in three parts. The first covers background and related work about the main areas of research covered by this work, namely online collaborative systems, the Semantic Web, and data quality. Part II is dedicated to Wikidata, describing its approach to knowledge representation, its features as a collaborative system, and the research published so far about the quality of its data. Finally, Part III first discusses the research questions posed and presents the methods followed to address them. Subsequently, it addresses the questions previously enunciated and discusses our findings (Chapters 9-11). The Conclusions sum up our contributions and proposes future directions for research and possible applications. Finally, Appendix A lists the Properties that were left out in the experiment reported in Chapter 9 and Appendix B provides English translations for the epigraphs at the beginning of Parts I and III.

1.4 Previous publications by the author

The core of this thesis is based upon and expands on previous work published by its author.

Besides the already mentioned work in Piscopo et al. (2017c), where we performed a qualitative analysis of how Wikidata users' perception of their role and identity within the platform evolves along their activity lifespan, we carried out a number of quantitative studies about collaboration dynamics and quality in Wikidata. We focused on Wikidata external references in Piscopo et al. (2017a) and Piscopo et al. (2017d). The first collects a range of descriptive statistics of the sources used in Wikidata, in order to compare them to those from Wikipedia. The analysis covers different aspects, such as geographic provenance of sources, type, and web domain. Piscopo et al. (2017d) gauge the quality of external references in Wikidata, in terms of their relevance and authoritativeness.

Piscopo et al. (2017b) explores collaborative production processes in Wikidata. In particular, it investigates how the contribution of different types of editors, i.e. bots and

human editors, registered or anonymous, influences outcome quality in Wikidata. Moreover, the paper looks at the effects of tenure and interest diversity among registered editors.

[Piscopo and Simperl \(2018\)](#) study the relationship between different Wikidata editor roles and the quality of the Wikidata ontology. The paper proposes a framework to evaluate the ontology as it evolves. In order to identify user roles, it clusters editing activities in monthly time frames. These roles are successively linked to measures of ontology quality in order to verify whether any relation exists between them.

1.5 Data

For the purpose of analysis, this thesis uses data from the Wikidata historical dumps, made freely available by the Wikimedia foundation. This data includes each revision to every page in Wikidata, consisting of both the actual data and the metadata. Unless specified otherwise, the dataset used is updated to 1st October 2017. Further information is provided in [Section 8.4](#).

BACKGROUND WORK

Marco Polo descrive un ponte, pietra per pietra.

‘Ma qual è la pietra che sostiene il ponte?’ chiede Kublai Kan.

‘Il ponte non è sostenuto da questa o quella pietra,’ risponde Marco, ‘ma dalla linea dell’arco che esse formano.’

Kublai Kan rimane silenzioso, riflettendo.

Poi soggiunge: ‘Perché mi parli delle pietre? È solo dell’arco che mi importa.’

Polo risponde: ‘Senza pietre non c’è arco.’

LE CITTÀ INVISIBILI, ITALO CALVINO

Chapter 2

Online knowledge collaboration

Socio-technical systems are ‘social systems sitting upon a technical base’ (Whitworth, 2009). This definition follows that initially introduced in the 1950’s by the scholars who coined the term and encompasses a large range of systems, e.g. from a plane and its crew to social network platforms (Whitworth, 2009). What characterises a socio-technical system is that the social component does not merely exist next to the technical one, but provides context to that, by means of a complex network of feed-back and feed-forward interactions (Baxter and Sommerville, 2011; Whitworth, 2009). As a conjunction of its social structures and connections and the technical solutions through which these take place, Wikidata is a socio-technical system. This chapter sets out background work in a sub-field of socio-technical systems, i.e. online communities and open knowledge collaboration, describing relevant terminology and outlining the main areas of study.

2.1 Online collaboration

Online collaborative systems are a type of socio-technical system that has been extremely successful in recent years. Spanning from the development of software to the creation of generalist or domain-specific knowledge bases, online collaboration systems have produced artefacts or provided services with quality comparable to that of expert-based systems (Surowiecki, 2005). Broadly speaking, they are systems that support coordination and cooperation of a number of people through the Internet in order to perform a determined task (Bafoutsou and Mentzas, 2002). They comprehend very heterogeneous systems, including e.g. from the e-mail to platforms for sharing user-generated content such as YouTube.

Different terms emphasise various aspects of online collaboration. The term **online communities** highlights the social and aggregative aspects of online collaboration. It

refers to virtual spaces where groups of people interact and develop relationships, possibly to achieve a common goal (Kraut and Resnick, 2012). Definitions vary (Malinen, 2015), especially in terms of their focus on different aspects, such as the technology adopted (Preece, 2001b), the cultural exchanges among people in a community (Rheingold, 1996), or the relationships between them (Haythornthwaite and Wellman, 1998; Lin and Lee, 2006). However, four elements, i.e. social interactions, shared purpose, common policies, and technology, are present in most of the definitions (Malinen, 2015; Preece, 2001a). In particular, technology is generally recognised as ‘the foundation and medium through which community members interact, it is one of the key determinants of the dynamics of the community’ (Ma and Agarwal, 2007).

Another related term is **peer-production**. Benkler et al. (2015) define it ‘as a form of open creation and sharing performed by groups online that’ adopt a decentralised approach to achieve a common goal, gather participants with diverse motivations, typically non-monetary, and follow participatory governance and managements strategies.

Online knowledge collaboration provides a further specification of collective work on the web. Based upon Keegan and Fiesler (2017), Preece (2001b) and Maloney-Krichmar and Preece (2005) we define it as the social aggregation of people to create knowledge-based goods, following distributed, non-hierarchical forms of organisation and whose interaction is mediated by computer technology. Wikidata falls into this definition

The three concepts described so far—online communities, peer-production systems, and open knowledge collaboration—clearly overlap. Online communities may be peer-production systems, e.g. GNU/Linux and Wikipedia are two of the most widely known examples, although not necessarily. Peer-production systems may be considered as a subset of online communities, characterised by decentralised work, participatory governance, and a pool of users with heterogeneous motivations (Benkler et al., 2015). The difference between these and online knowledge collaboration concerns primarily the focus of the latter on the product, i.e. knowledge based goods.

In the remainder of this chapter, we provide an overview of the main areas of study concerning open knowledge collaboration and online communities. Following the categorisations proposed by Benkler et al. (2015) and Malinen (2015) and the aspects of online collaboration highlighted by Strohmaier et al. (2013), we identified the following areas:

- Online collaboration frameworks;
- Social and organisational dimensions;
- User motivation;
- Tools and technologies;

- Success of the system and outcome quality.

This thesis builds upon and expand on the social and organisational dimension and the quality areas. The majority of the studies cited in the following sections derive from analyses of either FLOSS projects or Wikipedia (or both). However, the body of work about the free encyclopaedia is so vast that may be considered a genre on its own. Therefore, we provide a short account of that at the end of this chapter.

2.2 Online collaboration frameworks

Diverse nomenclatures have been developed to classify online collaborative systems. The construction of framework of online collaboration is outside the scope of this thesis. Nonetheless, we provide a brief overview of works in this area. A large number of studies categorise online collaborative systems according to the 3 Cs ([Cruz et al., 2012](#)): communication, coordination, and collaboration. Time and space where the activity of the community takes place are used by [Grudin \(1994\)](#) and [Bafoutsou and Mentzas \(2002\)](#). Others take into account the type of tasks carried out by participants of a system. [McKenzie et al. \(2012\)](#) follows this approach, distinguishing between collaborative and creative platforms, whilst [De Sanctis and Gallupe \(1987\)](#) combine it with collaboration features, including group size in their framework. [Ziaie \(2014\)](#) brings together different theories—Activity Theory ([Kuutti, 1996](#)) and the socio-technical model developed by [Whitworth \(2009\)](#)—to explain the social, organisational, and technological dimensions of online communities. Finally, the social machines framework ([Shadbolt et al., 2013](#)) emphasises the combination of human and machine-driven components in web-based socio-technical systems, rather than considering technology only as a mediator between people and policies within a system.

2.3 Social and organisational characteristics

We include in the social and organisational characteristics of online communities all aspects that concern participation, governance, collaboration, and distribution of work within a system ([Strohmaier et al., 2013](#)). An extensive body of literature has been dedicated to the different facets of the subject.

Lurkers Participation in online communities is commonly seen in the rather broad sense of visiting an online platforms and engaging with it in some way ([Malinen, 2015](#)). Coherently with that, the most common distinction between forms of participation in online communities is between active and passive participation ([Malinen, 2015](#)). A large body of literature has been dedicated to the latter, investigating the phenomenon of

lurkers, i.e. users who take advantage of the content produced by a community, whilst not producing any or doing that only occasionally (Nonnecke and Preece, 2000). Lurkers often account for a large part of the user pool of online communities (Nonnecke and Preece, 2000; Preece et al., 2004; Halfaker et al., 2013). They have similar demographics and attitudes about online communities to active participants. Lurking is seen by active users as an acceptable form of participation, which may be beneficial for the community (Nonnecke and Preece, 2000) and is not simply amenable to ‘free-riding’ (Preece et al., 2004; Halfaker et al., 2013). Instead, it appears to be a form of legitimate peripheral participation through which users may move their first steps on a platform (Antin and Cheshire, 2010; Bryant et al., 2005; Panciera et al., 2010).

From newcomers to established members Numerous studies delve into the process through which newcomers become an integral part of a community. Bryant et al. (2005) observe how Wikipedia editors move to increasingly central roles, acquiring with time a feeling of identity with the project, gaining competence in using a larger number of tools, and becoming more involved in their community. This is consistent with the findings of Dennen (2014) regarding blog community users. The process may also be not linear though and members may move back and forth between the core and the periphery of the community, as Gray (2004) notes in their study of an online platform to support workplace learning.

Other works have taken into account the differences in editors’ behaviour by the year in which they joined a platform. Analysing user editing patterns by yearly cohort is crucial to avoid misrepresenting the underlying dynamics of the community (Barbosa et al., 2016). Aggregate analyses over the entire history of an online community can be subject to Simpson’s paradox. According to that, the comparison of two populations with respect to a determined attribute may lead to contrasting results, if the populations are separated into different categories (Wagner, 1982). This approach allowed Barbosa et al. (2016) and Keegan and Fiesler (2017) to highlight varying activity patterns in users of Reddit and Wikipedia respectively, depending on the year in which they joined their communities.

Core and periphery Quantitative descriptions of the amount of edit activities within online communities have consistently shown that a minority of participants does the lion’s share of the work, following roughly a power law, e.g. in Usenet (Fisher et al., 2006; Whittaker et al., 1998) or Wikipedia (Muchnik et al., 2013). This division of work, formalised mathematically by Borgatti and Everett (2000), has been described in terms of core-periphery (Long and Siau, 2007). The minority of more active contributors constitutes the core of the community. These participants take more responsibilities, are granted more privileges, and have stronger connections with other members of the community (Long and Siau, 2007). On the other hand, members at the periphery are

less interconnected among them and to the rest of the community, and perform a smaller number of actions. Several works have explored the articulation of user activities along the core-periphery axis, for instance in Wikipedia ([Arazy et al., 2011](#)) or in other FLOSS projects ([Amrit and van Hillegersberg, 2010](#); [Crowston et al., 2006](#)).

Governance and norms The articulation of labour discussed above is connected to the emergence of informal hierarchies and governance arrangements ([Benkler et al., 2015](#)). [De Laat \(2007\)](#) describes the evolution of large OSS projects from initial forms of spontaneous governance, stemming naturally from the varying amount and quality of the contributions made by different people, to more articulated ones, called internal governance and governance outside parties. Whereas the latter refers only to projects whose size requires to set a management structure in order to deal with relations with other organisations, internal governance concerns the rise of more or less formally defined structures within an online community. These involve the formalisation of procedures and norms, both in what concerns the activities within the community and its decision-making processes, breaking down a project into different sub-modules (horizontal differentiation), and the division of roles (vertical differentiation) ([De Laat, 2007](#)). These dynamics have been observed in many communities.

Extensive research has looked into norms and rules in online communities. Numerous works focus on the role of rules, policies, and guidelines in Wikipedia. [Butler et al. \(2008\)](#) and [Beschastnikh et al. \(2008\)](#) point out how rules help coordinate work in the online encyclopaedia, by assigning tasks and responsibilities and setting best practices to reach consensus and ensure quality. On the other hand, [Forte and Bruckman \(2008\)](#) and [Konieczny \(2009\)](#) analyse the influence of Wikipedia policies in creating a decentralised governance and preventing its community to become an oligarchy—this is the conclusion of [Konieczny \(2009\)](#), which is contested by [Shaw and Hill \(2014\)](#) though. [Keegan and Fiesler \(2017\)](#) investigate the creation and maintenance of rules in Wikipedia. Although a large number of users is active in rule-making practices along the lifespan of the project, this does not seem to divert them from the creation of content. Furthermore, their activity moves over time towards deliberating on the basis of existing rules, rather than creating new or revising old ones. Nevertheless, the system of Wikipedia rules and policies has been perceived as a burden by some community members, leading some scholars to argue that it might have slowed down its growth ([Suh et al., 2009](#)).

The division of larger projects into smaller, more manageable modules plays an important role in the organisation of FLOSS projects, as pointed out by [Narduzzo and Rossi \(2005\)](#) and shown by [Mockus et al. \(2002\)](#) with respect to Apache and Mozilla. In Wikipedia, this division of work is achieved through WikiProjects, which have the purpose of gather attention on and direct editors' efforts to areas of the encyclopaedia which need to be improved ([Tinati et al., 2015](#); [Warncke-Wang et al., 2015](#)).

Roles As part of the creation of an internal governance, formal roles may be set in what De Laat (2007) defines vertical differentiation. These roles may be needed to meet some requirements of the technology used, e.g. users with administrator privileges on a server, or to help manage a large community, as Butler et al. (2002) observed with regard to OSS projects. These formal roles are generally defined by structures and policies put in place by the community (Butler et al., 2002, 2008).

Besides formal roles, emergent, informal ones have been detected. These roles have been often placed on the core-periphery axis (Arazy et al., 2017; Marín et al., 2010). The fluid nature of online community allows ‘tensions among ideas, passion, time, and social ambiguity [to] ebb and flow’ (Faraj et al., 2011). Participants tend to take roles that are specific to the situation created by this movement (Faraj et al., 2011). As a consequence, roles may not be formally codified—they emerge from user activity patterns—and can change through the lifespan of a community project (Arazy et al., 2016; Faraj et al., 2011). Moreover, users tend to move across roles, often but not necessarily from the periphery to the centre (Dahlander and O’Mahony, 2011; O’Mahony and Ferraro, 2007), possibly responding to a change in the situation of the community they are part of. Although roles may coincide or overlap with formal or hierarchical roles, they are more generally determined by different levels of ‘expertise, identity, achievement, and community involvement’ (Arazy et al., 2015). In FLOSS projects, a concentric structure—the onion model (Crowston and Howison, 2005; Ducheneaut, 2005; Mockus et al., 2000; Scacchi, 2007)—has been employed to represent how roles are distributed along a core-periphery axis, with increasingly higher skills and reputation moving towards the centre. An analysis of functional roles in Wikipedia (Arazy et al., 2015), underpinned by the Reader-to-Leader framework (Preece and Shneiderman, 2009), has shown that users move in both directions between higher and lower responsibility roles, albeit with a lower number of editors getting to roles closer to the centre. Furthermore, the impact of core users on a system may change over time. Kittur et al. (2007a) note that the activity of ‘elite’ users progressively diminishes in Wikipedia and del.icio.us—a collaborative bookmarking site—giving place to the ‘rise of the bourgeoisie’, i.e. highly active users becoming less preponderant in the creation of content. This behaviour is observed whether elite users are defined as highly-active editors or as users with special rights (administrators) (Kittur et al., 2007a). The present work contributes to this area of study by looking at informal roles in Wikidata.

Relationships between communities Other lines of research have looked at the relationships between online communities. Jergensen et al. (2011) analysed several OSS projects within the same ecosystem, finding that participants do not behave consistently to their experience across projects. However, it appears that user with higher levels of experience are more likely to attain higher status when migrating to similar projects (Bird et al., 2007). Furthermore, a study of Wikia communities has shown

that membership overlap, especially of more experienced participants, is beneficial for the survival rate and success of a community (Zhu et al., 2014). Finally, Vincent et al. (2018) look at the relationship between communities in terms of influence on content creation. Their results indicate that whereas Wikipedia influences the creation of content in StackOverflow and Reddit, there is no evidence that this influence is reciprocal.

2.4 User motivation

The absence of clear incentives, e.g. monetary rewards, has led scholars to investigate the motivations behind participation in online communities and especially in peer-production systems (Butler et al., 2002). Works in this area have highlighted a broad range of reasons leading people to take part in online communities. Users can be driven by the desire to learn within a community of practice (Ye and Kishida, 2003). Other studies (Kollock, 1998) have found contributors of OSS projects to be led by a mixture of individual and social motivations, namely the expectation to receive something in return, to increase one's reputation and validate one's self image, being part of a group which benefits from the community work, and attachment to the community. Entertainment and information seeking are the main reasons for taking part in a user-generated online encyclopaedia, according to Lampe et al. (2010). Similar conclusions apply to Wikipedia. For Yang and Lai (2010), the feeling of self-efficacy deriving from meeting one's internal standard is one of the main drivers for sharing knowledge in the free encyclopaedia. Nov (2007) argue that fun and ideology are the main motivations of Wikipedia editors.

2.5 Tools and technologies

Literature about technological aspects of online collaboration systems has looked at how these facilitate participation and interaction between members (Malinen, 2015). In a study of an online health community, Maloney-Krichmar and Preece (2005) found that a sense of community may develop even in absence of specific virtual spaces designed for that purpose. Furthermore, community members seem to value reliable over cutting-edge technology. On the other hand, Gazan (2011) has shown that redesign in an online community may lead to disruption for the community itself, hampering communication and leading to a reduction of overall activity and an exodus of its members towards other platforms. Other studies, namely those following a socio-technical congruence approach (Cataldo et al., 2008), focus on production processes, looking at how social and technical structures within online communities support dependencies between tasks. Finally, a large body of work analyses the role of bots, i.e. software programmed to

perform revisions and various quality maintenance tasks, in Wikipedia. We report about this area of study in Section 2.7.

2.6 Success of online communities and outcome quality

Online communities have been able to create artefacts of quality comparable to that of expert-based systems despite a relative absence of traditional bureaucratic and management structure. This feature has raised substantial interest in both academia and the general public (Benkler et al., 2015; Surowiecki, 2005). A great deal of research relies on the open and distributed nature of FLOSS projects to explain the reasons of this success. For Weber (2004), these characteristics are inherently connected to high-quality outcomes. Raymond (2001) singles out the ability of the collective to effectively address any type of issue in his well-known claim that ‘given enough eyeballs, all bugs are shallow’ (the ‘Linus’s law’)—a concept formalised by Afuah and Tucci (2012), who has shown how distributed crowds are less likely to be stuck in local optima than localised teams. Faraj et al. (2011) argues that the fluidity of online collaborative systems may lead to reduced concerns of social conventions, ownership, and hierarchy, allowing more innovative knowledge collaboration to arise. Concerning the outcome of online communities, collaborative efforts have led to the development of largely successful products such as Apache and Linux. Wikipedia has notoriously been found to be comparable in terms of quality to the Encyclopaedia Britannica (Giles, 2005) and its editing model able to address quickly quality issues introduced by malicious users (Viégas et al., 2004). However, others have questioned these positive accounts. Keen (2011) maintains that peer-produced content is more likely to be of inferior quality, due to its openness to non-professional users. Raymond (2001)’s tenet is contested by Duguid (2006), who points out that the Linus’s law would only apply to software development contexts and warns against considering the quality of any peer-produced content as something fixed, whereas it changes over time.

A considerable amount of literature has analysed the relations between the organisational and social characteristics of online communities and the quality of their outcome. Roth et al. (2008) analysed the relation between structural and governance factors, e.g. the number of users and amount of contributions per user (structural) and the number of administrators (governance), on the growth of WikiProjects, finding a significant effect. Kittur and Kraut (2008) and Forte et al. (2012) focus on coordination within Wikipedia. Kittur and Kraut (2008) observe that higher levels of coordination—both implicit, through a few editors performing a larger share of the work, or explicit, through communication—are associated to higher quality. This is confirmed by Forte et al. (2012), who show that a small core of editors is responsible for structuring work in WikiProjects.

Group composition A substantial amount of literature has investigated the effects of group composition and diversity on performance. Arazy et al. (2011), Chen et al. (2010), Daniel et al. (2013), Lam et al. (2010), and Ren and Yan (2017) analyse how diversity affects outcome quality in online knowledge collaborations, obtaining similar results. Daniel et al. (2013) have examined the effects of three types of diversity—separation, variety, and disparity—on the outcome of OSS projects, noting that they have significant influence, either positive or negative, on community engagement and market success. Arazy et al. (2011) and Ren and Yan (2017) have looked at the relation between group diversity and task conflict and the quality of Wikipedia articles, finding that some types of diversity are positively associated to quality, namely contribution, cognitive, and group members’ orientation. Cognitive diversity refers to the mental models and interests of the members of a group and positively influences outcome performance. Lam et al. (2010) have looked at the effects of tenure diversity on the quality of the decisions to delete Wikipedia articles. Whereas the presence of newcomers by itself appears detrimental for outcome quality, in agreement with previous literature on offline settings (Moreland and Levine, 2014), a moderate tenure diversity is related to higher quality decisions. Chen et al. (2010) have studied how interest and tenure diversity influence productivity and withdrawal in Wikipedia projects. Interest diversity is a concept close to cognitive and functional diversity. It refers to the variety of members’ interests in a group. In collaborative projects such as Wikipedia or Wikidata, where users contribute voluntarily and generally choose which tasks to take on, an individual’s interests may actually determine their activity and function within the project. According to Chen et al. (2010), tenure diversity leads to higher productivity, but with diminishing results, while increasing member withdrawal, analogously to what noted by Lam et al. (2010). Interest diversity is linearly correlated to productivity, whereas no evidence is found about its influence on member withdrawal.

2.7 Wikipedia

Wikis are knowledge management systems that allow straightforward user contribution, typically by means of a web interface, and use markup systems to create link between or within pages (Breslin et al., 2009; Nalepa, 2010). Wikis record all revisions of a page, providing an effective version control functionality, and make use of discussion pages to enable asynchronous communication (Nalepa, 2010). Wikipedia is perhaps the most famous wiki. It counts with almost 49 million articles across 292 active language versions. The largest one, the English Wikipedia, has more than 5.7 million articles and have around 800 million monthly unique visitors¹, whilst 15 versions have more than 1 million articles.²

¹https://stats.wikimedia.org/v2/_#/en.wikipedia.org/reading/unique-devices/normal|bar|All|~total, consulted on 16 November 2018.

²https://en.wikipedia.org/wiki/List_of_Wikipedias, consulted on 16 November 2018.

The success of Wikipedia has generated considerable interest in the research community, leading to the creation of an area of study on its own. Literature reviews have covered research about different aspects of Wikipedia, e.g. the quality of its content (Mesgari et al., 2015) or its readers (Okoli et al., 2014). Jullien (2012) and Martin (2011) have attempted to compile a comprehensive review. A project exists—a wiki—to gather and organise all the literature about the free encyclopaedia, WikiLit.³

Most of the literature cited in the previous sections concerns Wikipedia, therefore we do not cover those areas again. However, we discuss some of the topics that are specific to research regarding the free encyclopaedia.

Content studies Several researchers have considered the quality of Wikipedia articles in terms of their accuracy and reliability (Mesgari et al., 2015). Besides the already cited article by Giles (2005), which assessed a selection of articles from the Natural Sciences field, a number of studies have looked at other areas. Rosenzweig (2006) and Holman Rector (2008) analysed samples of History-related articles, with varying results. Similar findings were found with regard to the medical field, with Devgan et al. (2007), Pender et al. (2008), and Rajagopalan et al. (2010) providing a positive evaluation, whereas Clauson et al. (2008) and Mercer (2007) found several flaws deriving from the lack of professional expertise on the subjects covered. Another important area of study looks at the differences between Wikipedia language versions and the lack of diversity. As regards the first, researchers such as Rogers (2013) have drawn attention upon the divergences that may exist between Wikipedias, for what concerns a number of aspects, such as depth of coverage and viewpoint. Other researchers have investigated the lack of diversity within language versions under several levels, caused by the unbalanced distribution of contributions across users, aggressive behaviour, and overly complex bureaucracy (Flöck et al., 2011). These aspects were taken into consideration in designing Wikidata, according to Vrandečić and Krötzsch (2014).

Bots Bots are pieces of software programmed to carry out repetitive tasks on Wikipedia. Over time, their activity has become key within the platform. They are the author of around ~ 15% of all edits over all Wikipedias (Steiner, 2014). Bots are key for quality control tasks, addressing and containing vandalism (Priedhorsky et al., 2007) and allowing to spot and correct low-quality edits in a few minutes (Geiger and Halfaker, 2013). In doing so, they enforce rules and policies established by the community, creating what Müller-Birn et al. (2013) have called algorithmic governance. Furthermore, they contribute to structuring new articles and—before the launch of Wikidata—maintained the links between different language versions (Niederer and van Dijck, 2010). On the other hand, bots may cause considerable harm, in case of malfunctioning (Stvilia et al., 2008). The activity of Wikipedia bots has been classified by Clément and Guitton (2015)

³http://wikilit.referata.com/wiki/Main_Page, consulted on 14 January 2018.

into different typical behaviour: ‘servant bots’, which carry out repetitive tasks replacing human editors, and ‘policing bots’, whose tasks focus on enforcing policies and norms.

Chapter 3

Filling the knowledge gap: the Semantic Web

In their seminal 2001 paper, [Berners-Lee et al. \(2001\)](#) presented their vision of a dense network of intelligent agents able to interact with each other. These agents were imagined to be able to exchange data and autonomously take decisions on the basis on the information extracted from the data. For instance, one agent could engage with another in order to book a medical appointment, negotiating the most suitable time for the patient and the GP. The paper argued that, in order to achieve the type of automated interactions they envisaged, an extension of the web would be needed. This extension is the Semantic Web.

Wikidata is a knowledge graph and is part of the Semantic Web, of which utilises several pieces of technology ([Malyshev et al., 2018](#)). This chapter describes the main concepts related to the Semantic Web, providing an overview of existing knowledge graphs. Furthermore, it explores collaborative projects in the Semantic Web area in order to provide some perspective to understand the characteristics of Wikidata.

3.1 The Semantic Web

A wealth of data is published on web, but that is not sufficient to build the type of intelligent agents described in [Berners-Lee et al. \(2001\)](#). This is because resources on the web are connected through semantically untyped hyperlinks ([Breslin et al., 2009](#)). Whereas humans can easily understand the relationship between two pages—e.g. a publication linked on a researcher’s page—the same task may be daunting for machines. The Semantic Web addresses this ‘knowledge gap’ ([Breslin et al., 2009](#)). It is an extension of the web where *entities*, i.e. individuals or types of things, rather than (or better, in addition to) documents, are connected by means of relations or *properties* ([Shadbolt](#)

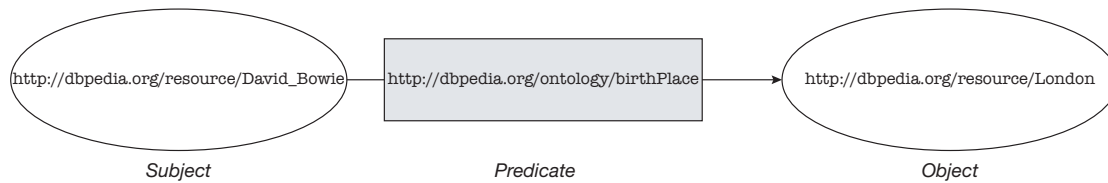


FIGURE 3.1: Example of RDF triple. Resources, including the property, are identified by URIs.

et al., 2006). The ultimate goal of the Semantic Web is to deploy semantics and connect information on a global scale (Shadbolt et al., 2006). The realisation of the Semantic Web is based on four requirements (Berners-Lee et al., 2001; Breslin et al., 2009):

- **Entity identity:** entities on the web must be unambiguous, each having a unique identifier.
- **Relationships:** in order to enable computers to gather knowledge about entities, different types of relationships must be expressed and defined.
- **Extensibility:** the Semantic Web is an extension of the web; as such, it aims to be at a global scale, thus requiring to be adaptable and able to evolve.
- **Ontologies:** different actors on the web may use different terms; in order to enable communication and represent knowledge, vocabularies are needed to provide an agreed upon definition of terms and relations.

RDF The *Resource Description Framework* (RDF) is a key component of the Semantic Web. It models data in triples, or statements, where a subject is connected to an object by a relation or property (see Figure 3.1). Whereas the subject of the triple must be an entity, or resource, the object can be either an entity or a literal, i.e. a string or a numeric value. Each entity can have several relations, connecting several triples and forming a graph. Resources are referred to by means of a **URI** (Universal Resource Identifier). URIs are *unique* and use global naming conventions, allowing anyone to link, discover, or refer to a resource (Shadbolt et al., 2006). Prefixes are often used to shorten URIs, replacing a common domain, e.g. `http://dbpedia.org/resource/` can be swapped for `dbr:` in `http://dbpedia.org/resource/David.Bowie`. Several serialisations of RDF exist, providing the possibility to use different syntaxes, such as XML, N3, and turtle (Shadbolt et al., 2006).

SPARQL Repositories able to store RDF content, called *triple stores*, have been developed (Shadbolt et al., 2006). These can have different functionalities, e.g. reasoning over the data, large-scale storage, etc. SPARQL is a language used to query and access these repositories (Harris et al., 2013). It is a graph-querying language (Breslin et al.,

2009), which enables querying and manipulating RDF graphs both on the web and in triple stores (Harris et al., 2013).

3.2 Ontologies and ontology engineering

Ontologies, or *schemas*, are essential components of the Semantic Web. They provide an ‘abstract, simplified view of the world’ (Gruber, 1993), capturing the classes of entities of interest in a given domain and their properties (Sicilia et al., 2012). They facilitate communication among people, as they standardise terminology and provide guidance for classification. They are equally useful to add context to algorithms for anything from web search to recommender systems and conversational agents. Different communities have developed their own ontologies, e.g. in health and life sciences (Ashburner et al., 2000; Bard and Rhee, 2004). Some ontologies included notable features which were subsequently implemented in Wikidata. For example, the cultural heritage ontology CIDOC-CRM (Doerr, 2003) affords the expression of contrasting pieces of information, a possibility existing also in Wikidata, as we see in Chapter 5. This is important to reflect diverse point of views and uncertain information.

The literature identifies two essential primitives for ontology modelling (Breslin et al., 2009):

- **Distinction between classes and instances.** Classes define sets of instances and can be ordered using subset relationships, i.e. asserting that a class is a subclass of another.
- **Properties.** These are required to assert relationships between entities, or between entities and literals.

Languages have been developed to express these and other modelling primitives. The Resource Description Framework Schema (Brickley and Guha, 2004) (rdfs) is used to state subclass relationships (`rdfs:subClassOf`), the sets of entities allowed as a subject (`rdfs:domain`) or object (`rdfs:range`) of a property, and the human-readable label of a resource (`rdfs:label`). The Web Ontology Language (OWL) (Hitzler et al., 2009) has been developed to add expressiveness to RDFSchema. It can be used to provide rules and constraints for properties and to link identifiers referring to the same resource, e.g. `db:David.Bowie` and `yago-res:David.Bowie` both refer to the musician David Bowie (Figure 3.2). The use of formal knowledge representation languages, which have a rigorous mathematical underpinning, allows one to use automatic reasoning to detect inconsistencies, classify new entities into classes based on their properties and derive new knowledge. For example, by inferencing over the resources in the graph in Figure 3.2, computers can deduce that every Musical Artist is a person and that every object of the birth place property is a place.

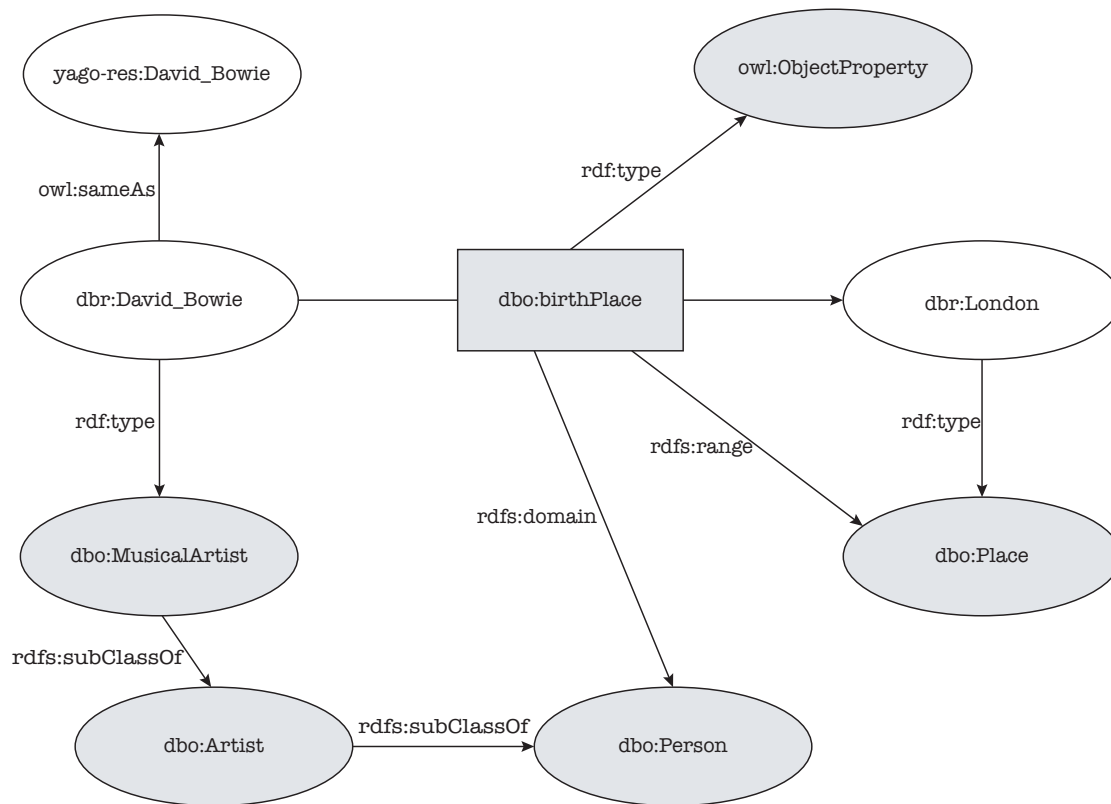


FIGURE 3.2: Example of triples of an ontology. Resources in grey are part of the ontology. Properties from RDFSchema and OWL are depicted as arrows, whereas rectangles are used for properties from other datasets.

3.3 Linked Data

Berners-Lee et al. (2001)’s paper sets out the goal to extend the web with an additional layer, the Semantic Web or Web of Data, that would allow connecting data and interweaving them to the document web on a global scale (Bizer et al., 2009a). The means to reach this goal are provided by Linked Data, ‘a set of best practices for publishing and connecting structured data on the web’ (Bizer et al., 2009a).

Linked Data is based on four principles, which prescribe the adoption of some of the technologies described in the previous sections to facilitate interconnection and discovery of the data. These are (Berners-Lee, 2006):

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

A subsequent revision of [Berners-Lee \(2006\)](#) adds a fifth principle:

5. All the above, plus: Link your data to other people’s data to provide context.

Data published following these standards—i.e. using the RDF data model to connect entities represented by URIs and making it available on the web—is called Linked Data itself. The Linking Open Data project has worked on identifying Linked Data sources openly available the web and facilitating their publication according to the principles mentioned above. So far, the project has gathered 1234 datasets¹, connected in what is called the Linked Open Data cloud.

3.4 Knowledge graphs

Knowledge graphs (KG) are essential components of the Semantic Web. The term was coined by Google in 2012 ([Singhal, 2012](#)), referring to the use of semantics to support web search. Whilst several definitions of KG are somewhat vague—‘a collection of relational facts that are often represented in the form of a triplet’ ([Wang et al., 2014](#)); ‘any graph-based knowledge repository’ ([Färber et al., 2018](#))—[Paulheim \(2017\)](#) pointed out some of the most common features of KGs:

1. KGs are **graph-based knowledge representations**, which describe real world entities and the relations between them.
2. These relations, together with the classes, are defined by a more or less formally-encoded schema, which nevertheless represents only a minor part of the graph. **The definition of KGs does not include ontologies** ([Paulheim, 2017](#)).
3. Moreover, **entities in a KG can be linked** between them and **to other entities in other graphs**.
4. Finally, KGs typically, but not necessarily, span across **several domains** ([Paulheim, 2017](#)).

Although the majority of KGs adopts Semantic Web and Linked Data standards ([Bizer et al., 2009a](#)), this is not a necessary condition and there exist examples that do not follow them, such as The World Factbook.² Furthermore, several KGs are openly accessible and their data released under an open licence. However, others are released under a proprietary licence and are not accessible, e.g. Google’s Knowledge Vault ([Dong et al., 2014](#)), or can be only accessed through an API, e.g. Wolfram Alpha or the Facebook Graph.

¹<https://lod-cloud.net/>, consulted on 18 January 2019.

²<https://www.cia.gov/library/publications/the-world-factbook/>

Creating KGs is not trivial and several approaches have been adopted, which vary in the way they combine human input and automated solutions. Some projects, like Cyc, are completely curated; other KG platforms rely on experts for specific types of activities (Freebase, Yago), define rules for how and by whom some activities should be carried out (DBpedia, Yago), or provide tools to facilitate collaboration (DBpedia). Sometimes the crowd takes care of adding and editing the data in the KG, either supported (Freebase) or unsupported by experts (Wikidata). Other solutions may rely more heavily on information extraction techniques (Knowledge Vault (Dong et al., 2014), PROSPERA (Nakashole et al., 2011)) and use human contributions only to improve their extraction algorithms (Open Mind Common Sense (Singh et al., 2002), NELL (Mitchell et al., 2015)).

In the following sections, we provide an overview of the most relevant projects in the field, pointing out the features that are of interest to understand the peculiarities of Wikidata. Far from being a complete overview of existing KGs, our selection includes projects that are openly accessible and are cross-domain, similarly to the selection from Färber et al. (2018).

DBpedia DBpedia was launched in 2006 and has since become one of the central nodes of the Linked Open Data cloud. DBpedia extracts information from Wikipedia to create a multilingual and multi-domain KG, which is made available following Linked Data best practices (Lehmann et al., 2015). The KG is generated by creating a URI for every article in Wikipedia and populating the resource with information from each of the article’s structured elements, such as infoboxes, links, images, and geo-coordinates. The property `prov:wasDerivedFrom` is used to connect each entity to its Wikipedia source. No other sources are referenced though (Färber et al., 2018).

To this day, 125 localised versions of DBpedia are available, and their maintenance is spread across a number of organisations (Lehmann et al., 2015). New releases of DBpedia have been issued approximately every two years. The last includes more than 4.5 million entities in the English version, its largest³, and contains a large number of incoming (~ 4M) and outgoing links (> 27M) (Lehmann et al., 2015). It can be accessed as RDF dumps, by retrieving the URIs of its entities on the web, and through a SPARQL endpoint. All the data in DBpedia is openly shareable and reusable. In order to keep the content of the KG updated between a release and the following one, the developers of DBpedia have created a live version, which monitors Wikipedia for changes to its pages and process them in order to keep the graph updated (Bizer et al., 2009b). This means that the live version of DBpedia reflects in only a few minutes any edits made to Wikipedia articles. Nevertheless, it cannot be edited directly.

³See <http://wiki.dbpedia.org/about>, consulted on 1 February 2019.

Whereas the extraction of information from Wikipedia is completely automatic, the ontology schema and the mappings between that and the Wikipedia infoboxes have been provided since 2010 by the efforts of a community of users (Lehmann et al., 2015). As of April 2013, 23 language mapping communities were active (Lehmann et al., 2015).⁴ Following the intentions of its creators, the DBpedia ontology is rather small: it contains 320 classes and 1650 properties, and has a maximal depth of 5 (Lehmann et al., 2015).

YAGO YAGO (acronym of Yet Another Great Ontology) exploits taxonomic knowledge within Wikipedia in combination with concept hierarchies from WordNet and information from GeoNames to generate a broad-covering KG (Suchanek et al., 2007). It covers over 10 million entities, with more than 120 million facts stated about them⁵. All the data is released under an open licence, which allows reuse and sharing, provided that attribution is given. The whole extraction process is automatic, with the exception of the mappings between English infobox attributes and YAGO relations (Mahdisoltani et al., 2015). The last version of YAGO is multilingual, drawing information from 10 language versions of Wikipedia (Mahdisoltani et al., 2015). YAGO3 relies upon the interlanguage links in Wikipedia to connect articles covering the same topic across different Wikipedia language versions. YAGO provides provenance for each statement by using its own vocabulary, indicating both the source and the extraction method (Färber et al., 2018). Moreover, it also allows to represent unknown and empty values.

YAGO’s classes are rather fine-grained, as they are generated from Wikipedia categories and their hierarchies built on the basis on Wordnet synset relations (Färber et al., 2018). As a consequence, YAGO has a much larger number of classes ($\sim 570,000$), compared to other KGs, such as DBpedia.

Freebase Launched in 2007 by Metaweb Technologies Inc., Freebase was acquired by Google Inc. in 2010 and subsequently shut down in June 2015 (Färber et al., 2018). It was a community-curated database, which aimed at covering the totality of human knowledge, making it accessible under an open licence. Freebase was the combined result of a massive data extraction from different sources, such as Wikipedia or MusicBrainz, and a community effort, which allowed it to cover up to around 49 million entities and to be continuously updated. Entities in Freebase used URIs, with localised human readable IDs in English and many other languages. Freebase enabled to express provenance and contextual information, as well as contrasting pieces of information, which could be ranked. Additionally, unknown and empty values for triples were accepted. Freebase did not rely on a rigid ontology, but on a ‘loose collection of structuring mechanism and conventions’ (Bollacker et al., 2008, p. 1247). Any user could add new relations,

⁴No information around the number of contributors in each community is available, to the best of our knowledge.

⁵<https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/>, consulted on 1 February 2019.

which had to be subsequently approved by editors with higher user rights, called admins (Färber et al., 2018). Freebase has had a considerable influence on Wikidata, not only with respect to the design of the KG, as we see in Chapter 5, but also in terms of data imported from the first into the latter. The migration of data from Freebase to Wikidata was carried out under the control of the community, using a tool called Primary Sources Tool (Tanon et al., 2016).

OpenCyc Cyc is a proprietary project launched in 1984, openly released in a smaller version called OpenCyc.⁶ It stores common sense facts, such as ‘every tree is a plant’, which have been gathered in the number of around two million, covering approximately 120 thousand entities (Ringler and Paulheim, 2017). Cyc/OpenCyc is completely contributed and maintained by experts, who manually crafted each of the axioms contained in the KG. Contrasting facts can be expressed and are all assumed true by default, although meta-level assertions can be added, stating that an assertion is more likely than another one (Lenat, 1995). Conversely to the projects described above Cyc/OpenCyc lacks the capability to express provenance (Färber et al., 2018). Moreover, Cyc/OpenCyc only provides labels in English.

ConceptNet ConceptNet is an openly released KG centred on word definitions (Speer et al., 2017). It contains around 28 million statements, providing support for 304 languages (Speer et al., 2017). Although ConceptNet previously accepted direct contributions from human users, its last version (5.5) does not anymore and has been built by extracting knowledge from different sources, e.g. Wiktionary⁷ and OpenCyc, and a subset of DBpedia. ConceptNet is available under Linked Open Data standards.

NELL The Never-Ending Language Learning (NELL) project continuously crawls the web to extract facts in order to build a KG of diverse, confidence-weighted beliefs (Mitchell et al., 2015). NELL has been running since 2010, generating triples about around 2.8 million entities so far.⁸ The extraction process relies on a manually developed ontology, which counts 293 categories. In order to improve the outcome of the automatic extraction tasks, NELL has tried to engage web users to improve the triples generated by its algorithms. NELL extracts information only from English language websites. However, versions in Portuguese (Duarte and Jr., 2014) and French (Duarte and Maret, 2017) have been developed independently.

⁶The distribution of OpenCyc has been discontinued early 2017. See www.cyc.com/opencyc/, consulted on 1 February 2019.

⁷https://en.wiktionary.org/wiki/Wiktionary:Main_Page, consulted on 1 February 2019.

⁸<http://rtw.ml.cmu.edu/rtw/overview>, consulted 1 February 2019.

3.5 The social Semantic Web

Although most of the efforts to create Linked Data resources have been driven primarily by researchers and academia (Bizer et al., 2009b), there have been various attempts to adopt collaborative approaches. Besides some of the KGs mentioned above, which rely either on a direct (i.e. Freebase) or indirect (i.e. DBpedia) contribution from a community of users, other projects that are relevant for the topic of this thesis come from the field of ontology engineering. Moreover, semantic wikis have tried to bring a wiki approach into the creation of semantic content. We provide an account of these last two types of systems in the next sections.

3.5.1 Collaborative ontology engineering

Ontology engineering encompasses not only the development process, but also the ontology life cycle, as well as the methodology, language, and tools used to build it (Gómez-Pérez et al., 2004). The actors involved in the ontology engineering process must agree upon several aspects. They need to share a common view or domain of interest, but also to find an agreed-upon way to structure the knowledge encoded in the ontology, in terms of concepts and relationships among these (Simplerl and Luczak-Rösch, 2014). Ontology engineering is thus a consensus-building process, for which collaborative approaches have been proven to be both useful and economically feasible (Simplerl and Luczak-Rösch, 2014). Collaboratively produced ontologies may reach a significant size with regard to the concepts defined (Strohmaier et al., 2013). Examples of collaboratively engineered ontologies are the International Classification of Diseases (ICD-11), whose 11th revision has more than 50,000 classes (Walk et al., 2015), SNOMED-CT, which includes over 300,000 classes around clinical health terminology (Walk et al., 2015), or the already mentioned Gene Ontology (Ashburner et al., 2000), which includes around 40,000 classes.

Whether collaboratively or not, engineering an ontology typically goes through a series of steps (Simplerl and Luczak-Rösch, 2014): first, an analysis of the domain covered by the ontology and of the requirements imposed by its application; following, the stakeholders need to agree upon which conceptual model to use and how to implement it. Once the first revision of the ontology has been published, it must be kept up to date with to changes in the domain of reference. Different releases and versions can be made public and, in order to keep clarity around each step and allow new stakeholders to join, detailed documentation must be kept about each decision. Members of a team working on an ontology engineering project may be assigned fixed roles and can act as knowledge engineers, ontology engineers, or domain experts (Gómez-Pérez et al., 2004). The latter inform the work of knowledge engineers to create a conceptual model of the domain, which is subsequently translated into a knowledge representation language by ontology

engineers (Simperl and Luczak-Rösch, 2014). Roles are generally less fixed in collaborative projects, albeit Simperl and Luczak-Rösch (2014) identify two that are commonly taken by users: ontology editors and ontology contributors. Whilst these two roles may overlap, contributors are usually limited to providing feedback and suggesting changes, which may then be approved and implemented by editors. Furthermore, each community member can take different roles, according to the requirements of the engineering process and the types of contributions he/she is allowed to make. An analysis of the tasks performed by community members has shown that more fine-grained differentiations can be made (Falconer et al., 2011; Walk et al., 2015). Strohmaier et al. (2013) examined the activity logs from five projects⁹ plus Wikipedia, in order to understand how editing behaviour varies across time, across users, and across concepts. Their findings suggest some similarities between collaborative ontology projects and the online encyclopaedia in terms of organisation of work—e.g. a power law distribution of work across users and across entities—whilst others characterise the projects considered, such as the changes happening in bursts, rather than uniformly. Falconer et al. (2011) analysed contributors’ activity patterns in three large-scale ontology development projects. The features considered included the type of change editors made to ontology, e.g. addition, modification, or deletion, and their centrality within the contributors’ network. The findings highlighted the existence of five distinct roles, each focused on a different type of activity, e.g. organisation, content creation, editing of existing content, and maintenance of the ontology hierarchies. Wang et al. (2015) examined user activity patterns in two collaborative ontology projects with the aim of predicting user behaviour. Their approach, association rule mining, is able to predict the action of ontology contributors on the basis of which parts of the ontology they previously contributed. Another study, carried out by Walk et al. (2015), uses an approach based on Hidden Markov Model called HypTrail to investigate sequential edit trails of collaborative ontology editors. They test a number of hypotheses over four large-scale ontology projects in order to understand the most likely transitions between contributors’ edits. According to their results, the most likely editing sequences of contributors of collaborative ontology projects follow hierarchical relations (e.g. super-, sub-class).

Another aspect that differentiates collaborative from traditional ontology engineering projects—i.e. carried out by small teams of highly trained ontology and knowledge engineers according to fixed roles (Chklovski and Gil, 2005)—is the need to synchronise contributions in order to avoid discrepancies and mediate between different points of views. Versioning software helps reach an agreement between participants publishing different types of revisions (Simperl and Luczak-Rösch, 2014). Alternatively, wiki-based approaches allow users to propose changes, add annotations, and make comments in a textual form. Approved changes are implemented by ontology editors at a later time.

⁹The National Cancer Institute’s Thesaurus (NCI Thesaurus) (Sioutos et al., 2007), The International Classification of Disease (ICD) revision 11 (ICD-11), The International Classification of Traditional Medicine (ICTM), The Ontology for Parasite Lifecycle (OPL), and the The Biomedical Resource Ontology (BRO)

Wiki-based platforms are designed with openness in mind (see Section 2.7). They allow anyone to edit, support better discussions between users, and their web-based interface represent a lower entry barrier than traditional ontology engineering platforms. To that end, researchers have designed semantic wikis that allow users to create structured knowledge by using a mainly textual interface. We discuss them in the following section.

3.5.2 Semantic wikis

Semantic wikis inherit the open and collaborative nature of wiki projects including community discourse, i.e. the possibility to have virtual spaces for discussion between members (Tempich et al., 2007), whilst incorporating semantics and data in formats that can be interpreted by machines (Krötzsch et al., 2007). Semantic wikis have been seen as a way to allow the collective creation of large scale collections of structured data, due to their ease of use and immediateness (Nalepa, 2010). Structured data can be added manually, in form of user annotations—although an extensive study of 230 Semantic Wikis showed that user participation to this task is generally low (Gil and Ratnakar, 2013), or automatically imported from content within the wiki itself, e.g. links or images’ metadata, or external sources that have already a structured format (RDF). Moreover, semantic wikis may support the collaborative creation of ontologies, by enhancing communication between users through the creation of textual descriptions and discussion pages, which facilitates cooperation. Semantic data may also be used to support textual content, adding search and reasoning capabilities (Krötzsch et al., 2007). For example, if we have a collection of texts in which all animals are tagged and an ontology which specifies the family to which each animal belongs, we can query the collection for all mentions *Felidae*, to find texts about cats, lions, etc. Schaffert et al. (2008) identifies various types of semantic wikis, according to their use of RDF, their way of expressing semantic relationships, and the reasoning capabilities they offer. Earlier examples of semantic wikis represent semantic relationships through typed links, intertwined in the text within the pages, using RDF, but with limited or no reasoning features (Schaffert et al., 2008). Furthermore, they typically have simple interfaces that allow users to edit separately text and metadata. Semperwiki (Oren, 2005) and Kawawiki (Kawamoto et al., 2006) belong to this typology. Later examples of semantic wikis, e.g. Semantic MediaWiki (Krötzsch et al., 2006) or OntoWiki (Frischmuth et al., 2015), are provided with ontology development capabilities, i.e. they do not require a predefined ontology and users can create new annotations, which could be used to structure knowledge within the wikis’ pages (Schaffert et al., 2008). In particular, the Semantic MediaWiki project aimed to support Wikimedia projects with semantic annotations (Krötzsch et al., 2006), anticipating to some extent one of the aims of Wikidata (Chapter 5). Finally, a third type of semantic wikis provides more advanced reasoning features, supporting usage and editing of OWL ontologies (Schaffert et al., 2008). IkeWiki (Schaffert, 2006) is an example of this type of semantic wikis.

While the semantic wiki approach lower the entry barrier for contributing structured data, opening up the possibility to involve larger user pools to build semantic knowledge, some researchers (e.g. [Nalepa \(2010\)](#)) lament their lack of expressive knowledge representation mechanisms. Furthermore, communities behind semantic wikis have so far achieved limited sizes, i.e. a few thousand users max ([Gil and Ratnakar, 2013](#)).

Chapter 4

What we talk about when we talk about quality

This thesis explores the relation between collaborative production processes and their outcomes in Wikidata, in terms of the quality of the data produced. This chapter introduces background work around data quality, which will be referred to when discussing previous studies about Wikidata quality in Chapter 7.

4.1 About data quality

The existing literature on data quality is extensive and commonly follows [Juran \(1962\)](#)’s definition of quality as ‘fitness for use’ (see [Batini et al. \(2009\)](#); [Fürber and Hepp \(2011\)](#); [Pipino et al. \(2002\)](#); [Wang and Strong \(1996\)](#); [Zaveri et al. \(2016\)](#)). Different perspectives follow this definition. Some stress data consumers’ needs ([Pipino et al., 2002](#); [Wang and Strong, 1996](#)), whereas others are centred on the suitability of data for the task to be performed ([Redman, 2001](#)). Both points of view stem from an empirical approach which has two underlying aspects: (*i.*) data quality is **task-dependent**, i.e. the same piece of data may be considered of sufficient quality for one task, but insufficient for another; (*ii.*) it is **subjective**, meaning that whereas a user may find a piece of data appropriate for a task, another may not deem the same piece of data suitable for the same task.

A drawback of this approach is that it ‘assumes that potential use of information is known and stable’ ([Lukyanenko et al., 2014](#)). In collaborative information systems, data is contributed by a possibly large number of users, who may have different levels of expertise and skills. Quality oversight of the content and shape of the data may become more difficult as a result of that. Moreover, contributors may be unaware or may disagree over the intended use of the data and have different motivations. To address this issue, [Lukyanenko et al. \(2014\)](#) incorporate the point of view of data contributors into their

definition of data quality. According to them, whilst this is determined by the fitness of data to the needs of data consumers, these needs are to be considered as perceived by data contributors. In large collaborative projects, such as Wikipedia or Wikidata, these needs are often defined through policies set by their contributors.

Regardless of the perspective adopted, data quality is generally considered a multi-dimensional construct, each dimension being a set of attributes that measure a single aspect of quality (Wang and Strong, 1996). Different sets of dimensions may be relevant for a user, depending on the task at hand (Bizer et al., 2009a). The literature diverges with respect to the dimensions included in each quality framework. The widely cited work of Wang and Strong (1996) classifies eight dimensions into four categories: *intrinsic*, *contextual*, *representational*, and *accessibility*. Intrinsic dimensions refer to those that are ‘independent of the user’s context’ (Zaveri et al., 2016), which implies that data has quality in its own right (Wang and Strong, 1996). Contextual dimensions are dependent on the task at hand and on the context of the data consumer (Wang and Strong, 1996). Representational and accessibility dimensions refer to the form in which the data is available and to how it can be accessed (Färber et al., 2018).

Other research adopts a different slant, depending on the type of data and system they focus on. For example, Pernici and Scannapieco (2003) propose a quality framework for web information systems. They emphasise the ephemeral nature of data publication on the web, compared to print. Web information systems may evolve continuously, shifting between different processes of information production. This leads Pernici and Scannapieco (2003) to include four dimensions in their framework: *expiration*, i.e. the time until which the data remains current; *completeness*, i.e. the extent to which the elements in a set are covered by an information source; *source reliability*, which indicates how credible the source on which an information system is; and *correctness*, i.e. the distance between a data value v and the correct value v' (Pernici and Scannapieco, 2003). Temporally modelling the evolution of these dimensions is important to provide measurable metrics, because of the contrast between the need of publishing high quality data and the stringent publication times on the web.

A large body of literature focuses on Linked Data quality. Hogan et al. (2010) address the subject by looking at systematic issues in RDF publishing, identifying four categories of symptoms that can occur when a software agent tries to retrieve a piece of data: *incomplete*, when relevant data cannot be retrieved; *incoherent*, i.e. a local piece of data cannot be interpreted according to the expectations of the data publisher; *hijack*, similar to incoherent, but referred to some remote piece of data; and *inconsistent*, when the interpretation returns an inconsistency in the data (Hogan et al., 2010). For each of these symptoms, the authors discuss and provide recommendations to data producers and consumers. The work of Bizer et al. (2009a) focuses on providing an approach to express quality meta-information about data, rather than on producing a new framework. Both studies are included in the survey of Zaveri et al. (2016), who reviewed more

than 100 articles within the field of Linked Data quality to compare the different perspectives in the field and produce an actionable framework for evaluating Linked Data sources. One of the outcomes of Zaveri et al. (2016)’s literature review is a set of quality dimensions specific for Linked Open Data. These dimensions are partially based upon the framework developed by Wang and Strong (1996), with the addition of a number of dimension that are specific to Linked Open Data, namely interlinking, versatility, licensing, and performance. Färber et al. (2018) compiled a data quality framework for their comparative evaluation of a number of knowledge graphs, which included also Wikidata. This selection rely primarily on the works of Wang and Strong (1996) and Zaveri et al. (2016). We provide a definition of the dimensions included in Färber et al. (2018)’s framework below.

4.2 Quality dimensions

4.2.1 Intrinsic dimensions

Accuracy Several definitions have been given to accuracy. For Wang and Strong (1996) it is the extent to which data is accepted as true and free of error, whilst Pernici and Scannapieco (2003) define it as the extent to which the data value v reflects the correct value v' . Others, such as Ballou and Pazer (1985) and Färber and Hepp (2011), assert that in order to be accurate data values must correspond to a state of things in the real-world, i.e. a reality existing objectively and independently from the observer. Some researchers (Batini et al., 2009) distinguish between *syntactic* and *semantic* accuracy. Whereas semantic accuracy corresponds to the definitions of accuracy mentioned above, syntactic accuracy refers to the closeness of a value v to any of the possible values in a definition domain D . Unless specified otherwise, this work uses the term accuracy in the sense of semantic accuracy.

Trustworthiness It indicates the extent to which the user deems data as ‘true’ (Pipino et al., 2002) and depends on both the trustworthiness of the data producers and the judgement of the data consumer (Dezani-Ciancaglini et al., 2012). Färber et al. (2018) note that this dimension subsumes other four, i.e. believability, reputation, objectivity, and verifiability. Hence, in order to be trustworthy data must be accepted as real and credible (believable) (Wang and Strong, 1996); its source and content must be highly regarded (reputable) (Wang and Strong, 1996); it must be impartial and free of bias (objective) (Wang and Strong, 1996); and its correctness must be easy to check (verifiable) (Naumann, 2002).

Consistency The definitions of consistency take into account various characteristics. According to Batini et al. (2009), a consistent dataset is free from ‘violations of semantic

rules defined over a set of data items.’ Zaveri et al. (2016) focus instead on aspects related to the Semantic Web and see consistency as the conformity with a particular knowledge representation and inference model. Finally, Mendes et al. (2012) argue that ‘a dataset is consistent if it is free of conflicting information.’ We adopt this definition in the current work.

4.2.2 Contextual dimensions

Relevancy This dimension concerns how useful and important data is for the task at hand (Wang and Strong, 1996; Zaveri et al., 2016).

Completeness Batini et al. (2009) include completeness among the set of basic data quality dimensions, defining it as the extent to which a dataset represent a corresponding collection of real-world objects. Other points of view take into account the context in which data is used. For Wang and Strong (1996) completeness is ‘the extent to which data are of sufficient breadth, depth, and scope for the task at hand.’ These three features correspond in some frameworks to sub-dimensions, i.e. schema, column, and population completeness (Färber et al., 2018).

Timeliness According to Zaveri et al. (2016), ‘timeliness measures how up-to-date data is relative to a specific task.’ Whereas data sources may vary and be updated at different times, these changes may not always reflect those occurring to the objects they represent. As a result, data may lose currency and become outdated for the task at hand of data consumers.

4.2.3 Representation dimensions

Ease of understanding In order to facilitate use, data must be unambiguous and understandable by its consumers (Wang and Strong, 1996). As regards Linked Data, whereas software agents rely on URIs (see Chapter 3) to unambiguously communicate between them, humans require labels and descriptions to visualise and browse RDF data (Hogan et al., 2010).

Interoperability The previous dimension refers to the representational characteristics of data from the point of view of human users. Interoperability concerns instead representation under a technical perspective, referring to the extent to which machines can obtain a consistent and clear interpretation of data which allows them to exchange and process information without ambiguities (Färber et al., 2018). The definition we

CATEGORY	Intrinsic	Contextual	Representational	Accessibility
DIMENSIONS	Accuracy	Relevancy	Ease of understanding	Accessibility
	Trustworthiness	Completeness	Interoperability	<i>Interlinking</i>
	<i>Consistency</i>	Timeliness		<i>Licence</i>

TABLE 4.1: Data quality dimensions used by [Färber et al. \(2018\)](#). In italics the dimensions not originally in [Wang and Strong \(1996\)](#).

follow here has been formulated by [Zaveri et al. \(2016\)](#): interoperability is the extent to which the data conforms with previous sources in terms of format and structure.

4.2.4 Accessibility dimensions

Accessibility Data sources on the web need to be timely available, in order to be integrated with other sources to produce tailored information for users ([Naumann, 2002](#)). Accessibility concerns this aspect and is defined by [Wang and Strong \(1996\)](#) as ‘the extent to which data is available or easily and quickly retrievable.’

Interlinking On the Linked Data web, datasets need to be interconnected to enable data integration. The interlinking dimension refers to that. It is the ‘degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources’ ([Zaveri et al., 2016](#)).

Licence The links between datasets on the Linked Data web may be useful to discover new information. However, some data sources may not be suitable for reuse for determined tasks. Therefore, it is important for consumers to provide datasets with a licence clearly expressing the terms for reuse and sharing ([Hogan et al., 2010](#)).

Part I — Summary

- Socio-technical systems are systems whose social component is strictly interconnected with their technical component. **Online knowledge collaborations** are a specific type of socio-technical system, which aims to create knowledge and follows distributed forms of organisation. Wikidata is an online knowledge collaboration.
- Online knowledge collaborations, and online communities in general, have been extensively studied. Areas of research include their social and organisational dimensions, the motivations of their participants, the tools and technologies they rely upon, and their quality. A large body of literature is dedicated to Wikipedia, representing an area of study on its own.
- Participation in online communities is often structured along a core-periphery axis, where more active users constitute the core. Several studies have observed roles emerging from user activity patterns situated along this axis.
- Numerous studies about success of online communities have delved into the relationship between group composition and outcome quality. More diverse groups may lead to various effects, depending on the features considered and on the type of outcome analysed.
- The **Semantic Web** is an extension of the web that aims to create a Web of Data that is processable by machines. Data on the Semantic Web is encoded in a graph structure, with facts represented as *subject-property-object* triples using the RDF data model. Ontologies are often used to facilitate communication, providing standardised terminologies. **Knowledge graphs** are graph-based knowledge representations that describe real world entities and the relations between them. Wikidata is a knowledge graph and part of the Semantic Web.
- **Data quality** is commonly defined as “fitness for purpose”. This definition is centred on data consumers’ needs, referring to the suitability of data for a task at hand. Data quality encompassed several dimensions. In the current work, we use a quality framework that includes: accuracy, trustworthiness, consistency, relevancy, completeness, timeliness, ease of understanding, interoperability, accessibility, interlinking, licence.

II

WIKIDATA

‘To begin with,’ said the Cat, ‘a dog’s not mad. You grant that?’

‘I suppose so,’ said Alice.

‘Well, then,’ the Cat went on, ‘you see, a dog growls when it’s angry, and wags its tail when it’s pleased. Now I growl when I’m pleased, and wag my tail when I’m angry. Therefore I’m mad.’

‘I call it purring, not growling,’ said Alice.

‘Call it what you like,’ said the Cat.

ALICE IN WONDERLAND, LEWIS CARROLL

Chapter 5

Wikidata as a knowledge graph

Part II (i.e. Chapters 5, 6, and 7) delves deeper into Wikidata, presenting its main features and discussing them in the context of prior literature and similar projects. Specifically, the following sections focus on how Wikidata represents knowledge: its data model, its approach to expressing instance and conceptual knowledge, and how its data can be accessed. The next chapter describes the community that edits and maintains Wikidata. Chapter 7 present background work about data quality in Wikidata.

5.1 The data model of Wikidata

Items and *Properties* are the building blocks of Wikidata’s knowledge. Items refer to concrete or abstract entities, e.g. London¹, Ada Lovelace, or love, or to classes of entities, such as human or music genre. Properties are used to state relationships between any two entities or between an entity and a literal. Items and Properties are identified by so-called *QIDs*, i.e. alphanumeric codes in which a letter—Q for Items, P for Properties—is followed by a number (e.g. Q5 or P31). QIDs are unique and can be considered as Uniform Resource Identifiers (URIs) (Vrandečić, 2013).

Statements assert facts about Items and Properties, i.e. their attributes and relationships with other entities. The core of a statement is the *claim*, a property-value pair that connects an Item to another Item or to a literal. The set of all statements, in which Items are linked to each other by means of Properties, is the knowledge graph. Besides Items and literals, Wikidata allows two special values as an object of a statement: if the value of a Property object is unknown or does not exist, the values *somevalue* or *novalue* can be used. For instance, Elizabeth I of England never married, thus the statement

¹We use a `sans` font when referring to the human-readable label of Wikidata Items and Properties and a `typewriter` font for their QIDs. Please see the Nomenclature at the beginning of this thesis for further detail concerning the conventions adopted in the current work.

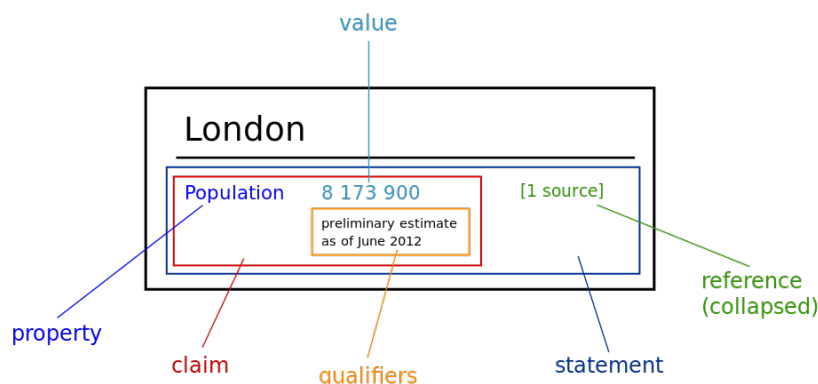


FIGURE 5.1: Components of Wikidata’s data model. Image from <https://www.wikidata.org/wiki/Wikidata:Introduction>, consulted on 1 February 2019.

regarding her marital status would be `spouse::novalue` (Erxleben et al., 2014). Statements can be further enriched: *qualifiers* can be added to a claim to add context, such as temporal validity, *references* can add provenance—i.e. a specification of where a piece of information is derived from—and *ranks* determine which claim must be preferred if multiple exist within the same statement. The capability to add meta-information to triples enables Wikidata to express a more diverse and verifiable knowledge (see sections 5.1.3 and 5.1.2) and is absent, or limited, in many of the KGs described in Chapter 3 (Figure 5.3). For example, population figures at different points in time can be stated, together with a reference pointing to the respective information sources (Figure 5.2), thus allowing users to choose the value they find most reliable. Figures 5.1 and 5.2 shows the components of Wikidata’s data model.

Following the wiki practice, Wikidata is organised as pages, grouped by common prefixes (a practice generally referred to as namespaces). Items and Properties have each their own namespaces. Other namespaces are dedicated to user pages, where editors can provide information about themselves, and other page types, such as policy, guidelines, and help pages. Each namespace has a related Talk namespace, where users can discuss and leave comments. Following the wiki approach, all revisions of each page can be consulted and restored, if a user has the necessary permissions.

Wikidata has reached a larger size, in terms of number of entities and possible relationships described, than many of the most commonly used KGs (Figure 5.4), counting at the end of September 2017 a total of almost 40 million Items and over 3500 Properties (Figure 5.5).

5.1.1 A multilingual knowledge graph

Whereas QIDs are handy for machines to use, they may be hard to remember for humans. Who would remember what Q84 refers to? Why is that related to Q145? Human-readable

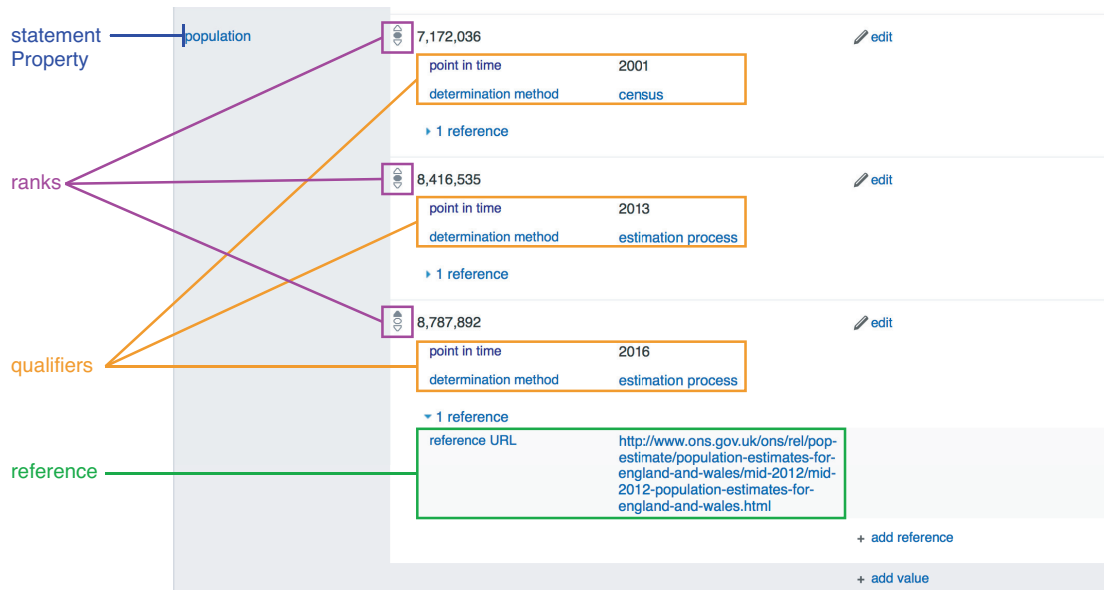


FIGURE 5.2: Three claims from Item Q84, i.e. London. Both of them use the same property, i.e. `population`, but qualifiers specify that they refer to different points in time. Ranks are applied to indicate that the most up to date value should be preferred. An upper grey arrow in the claim rank icon indicates *preferred rank*, a middle grey square a *normal rank* and a lower grey arrow a *deprecated rank*. Image from <https://www.wikidata.org/wiki/Q84>, consulted on 1 February 2019.

	WIKIDATA (October 2017)	DBpedia (April 2016)	YAGO (YAGO3)	Freebase (March 2015)	OpenCyc (May 2016)	NELL (08m.995)
Meta-information	✓	✗	✓	✓	✓	✓
Context	✓	✗	✓	✓	✗	✗
Provenance	✓	✗	✓	✓	✗	✗
Ranking	✓	✗	✓	✗	✓	✗

FIGURE 5.3: Meta-information in a selection of Knowledge Graphs.

labels, *descriptions*, or *aliases* are used to address this issue. Editors can add them in any of the 358 languages available in Wikidata (Vrandečić and Krötzsch, 2014) to help people identify which concept is represented by an Item or Property. A label is the ‘most common name that an Item [or Property] would be known by’ (Wikidata, 2018d) and it is not unique. Several Items/Properties may use the same label. Descriptions are short texts that provide information about an Item or a Property, with the aim to help disambiguate it (Wikidata, 2018c). Aliases are alternative names with which an Item or Property is also known (Wikidata, 2018b). Adopting QIDs to unambiguously identify entities enables Wikidata to be inherently **multilingual**. Language-independent data, e.g. numeric data, dates, or relationships between entities, entered in a language is immediately available in all other featured languages and can be (virtually) simultaneously modified by any user. The combination of QIDs and human-readable labels makes the information expressed by statements consistent across languages—a problem that has

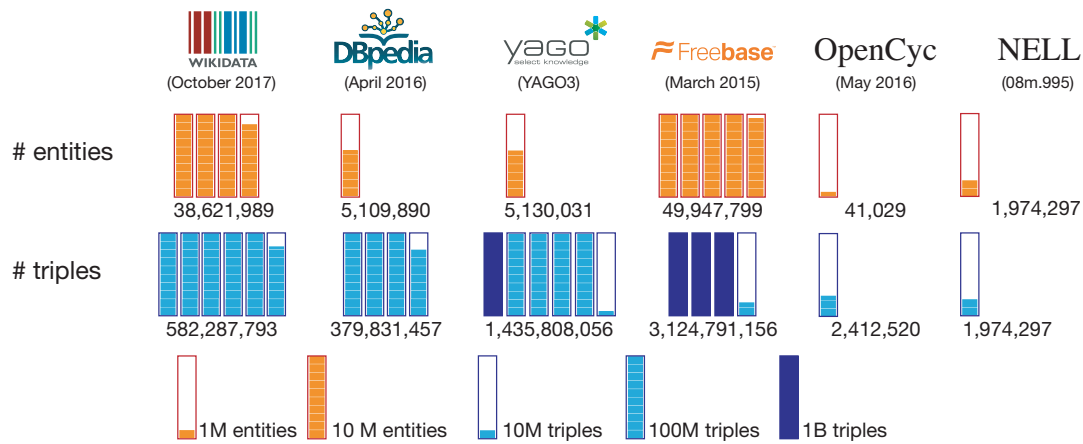


FIGURE 5.4: Size of Wikidata (entities and triples), compared to four major projects. Figures about DBpedia, YAGO, Freebase, and OpenCyc from [Färber et al. \(2018\)](#); NELL from [Ringler and Paulheim \(2017\)](#).

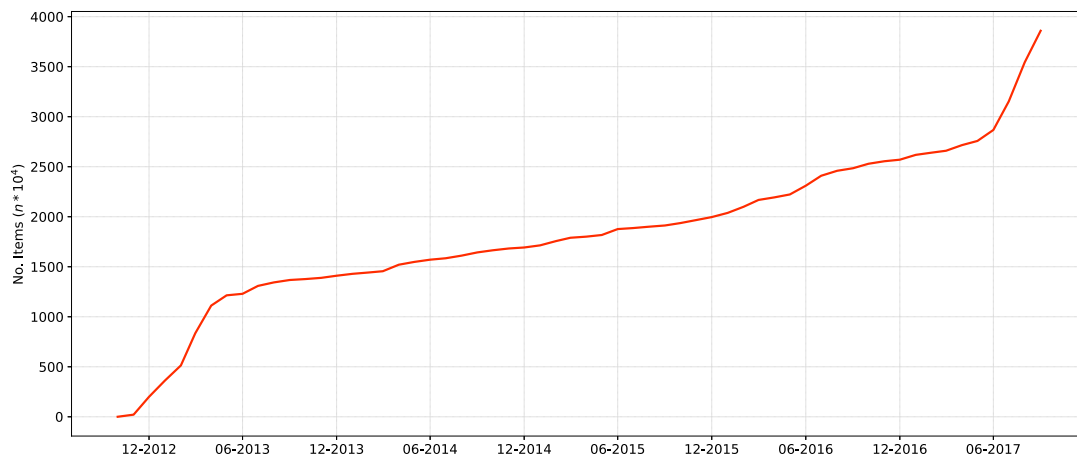


FIGURE 5.5: Number of Items along the lifespan of Wikidata

affected Wikipedia ([Kittur et al., 2007b](#); [Vrandečić, 2013](#)). For example, Wikidata would state the country where London is located through the statement







London :: country :: United Kingdom²

These are the English labels of the Items and Property included in this statement, which using Wikidata's QIDs would look like

Q84 :: P17 :: Q145

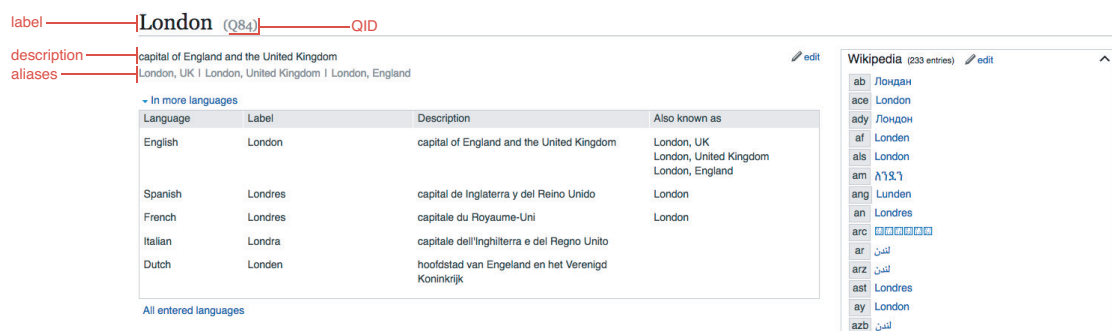
However, Dutch editors could render this statement in their language as

²The notation used to represent Wikidata statements is explained in the Nomenclature section at the beginning of this thesis.

	 WIKIDATA (October 2017)	 DBpedia (2016-04)	 YAGO (YAGO3)	 Freebase (March 2015)	 OpenCyc (2016-09-05)	 NELL (08m.995)
Multilingual support	✓	✓	✓	✓	✗	✗*
Multi-language labels	✓	✓	✓	✓	✗	✗
Language-independent URIs	✓	✗	✗	✗	✗	✗
# languages	358	125	>10 [†]	244	1	1*

* Other language versions have been developed independently from the original NELL.
[†] Labels available in 326 languages.

FIGURE 5.6: Language capabilities of Wikidata, compared to DBpedia, YAGO, Freebase, OpenCyc, and NELL



The screenshot shows the Wikidata interface for Item Q84 (London). On the left, there are sections for 'label', 'description', and 'aliases'. The main area displays a table of labels and descriptions in various languages (English, Spanish, French, Italian, Dutch). On the right, there is a list of Wikipedia articles in different languages, each with a link to the article.

FIGURE 5.7: Labels, descriptions, and aliases of Item Q84 (London). Human-readable labels, as well as descriptions and aliases, are added by users, therefore they may not be available in all the languages. A link below the main box allow to show all the languages available. On the right, the links to all the language versions of Wikipedia articles describing the Item. Image from <https://www.wikidata.org/wiki/Q84>, consulted on 1 August 2018.

Londen :: land :: Verenigd Koninkrijk

Whilst in Italian it would be

Londra :: Paese³ :: Regno Unito

The adoption of language-independent identifiers allows any content in Wikidata to be modified by editors, regardless of the languages they use. This is an obvious advantage, compared to other KGs which provide much less support to multiple languages (Figure 5.6). Another effect of this design choice has been that for every Wikipedia article, a page—or Item—was initially created on Wikidata to manage links to all related Wikipedia articles for each language. Each of their translations were added as labels, allowing a single Item to be re-used for each language version of Wikipedia (see example in Figure 5.7) (Vrandečić and Krötzsch, 2014).

³Capitalised in the original, see <https://www.wikidata.org/wiki/Q84> in Italian, consulted on 18 August 2018.

5.1.2 References

Wikidata has been conceived as a secondary database, meaning that it contains information stated by other primary sources, rather than claims about the real world (Vrandečić and Krötzsch, 2014). The possibility to add provenance, which is discussed in detail in this session, is thus one of its characterising features.

Wikidata’s provenance model is simple and relies on the above discussed data model. Prior literature on the topic identifies different aspects related to provenance: what it describes, its level of detail, and how to add it (Glavic and Dittrich, 2007). According to the object described, two types of provenance have been identified by Hartig (2009): *data-oriented* and *process-oriented*. Whereas the latter focuses on the method through which a piece of information is created, data-oriented provenance considers the source from which a data item is derived. This is the approach followed by Wikidata, where all content must be verifiable. Furthermore, the statement is the level of detail of the data item for which a source need to be specified, borrowing the terminology used by Glavic and Dittrich (2007). All statements in Wikidata must be supported by a source. Those that do not provide a reference are thus deemed unverified and should theoretically be removed (Wikidata, 2018j). Provenance can be either recorded at the moment of data creation (*eager* approach) or computed upon request (*lazy* approach), i.e. when data is used, following the categorisation made by Moreau et al. (2008). Wikidata adopts the former approach and editors are asked to add sources to the statements that they create or to those that miss one.

Community-generated rules define which types of statements are exempt from the requirement of being supported by a source (Wikidata, 2018j): undisputed claims representing common knowledge, e.g. `Earth :: instance of :: planet`; when a statement connects an Item to an external source of information, e.g. to an ID in a database; and when the source for a statement is an Item itself, e.g. a book and its author (Wikidata, 2018f).

References in Wikidata consist of a Property and a value attached to a claim. Various Properties can be used in references, the two main ones are P248 (stated in) and P854 (reference URL). P248 links a statement to another Item already existing in Wikidata. For example, this Property is used in the statement

London (Q84) :: capital of (P1376) :: United Kingdom (Q145)
→ stated in (P248) :: Thesaurus of Geographic Names (Q1520117)

P854 connects instead to an external URL that supports the statements, for instance

London (Q84) :: population (P1082) :: 8,787,892
→ reference URL (P854) :: http://www.ons.gov.uk/ons/...

This solution, which allows to include both internal and external references, stands in between those defined by [Glavic and Dittrich \(2007\)](#) as open and closed world approaches. Open world models have no control over the referenced data, which can therefore be modified and even deleted without notification ([Glavic and Dittrich, 2007](#)). On the other hand, closed world models do control the data used to provide provenance.

Besides P248 and P854, other Properties can be used to specify sources, such as P143 (imported from), P1343 (described by source), or P887 (based on heuristic). P143 (imported from) is used as a pointer for bots importing data from Wikipedia and is not considered to be a valid reference, since Wikipedia is itself a secondary source ([Wikidata, 2018f](#)). Other Properties add further details to a reference, e.g. P792 (chapter).

```

London (Q84) :: capital (P1376) :: United Kingdom (Q145)
→ stated in (P248) :: Thesaurus of Geographic Names (Q1520117);
publication date (P577) :: 12 June 2018;
retrieved (P813) :: 18 June 2018

```

The number of references has constantly increased since Wikidata's launch. The percentage of referenced statements has surpassed that of unreferenced ones in April 2016. Table 5.1 contains statistics about Properties used to indicate a source. P248 is by far the most used Property to add references.

5.1.3 Qualifiers

Similarly to references, qualifiers consist of property-value pairs attached to a claim. Statements can be expanded through qualifiers, modifying the information they contain or adding context ([Wikidata, 2018e](#)). For example, in Figure 5.2 they are used to specify the point in time at which population was estimated and the method used, thus allowing to provide different values for London population over time.

Beyond adding contextual information, qualifiers and references enable the presence of contradicting statements. This affords the representation of alternative points of views and it is part to the overall approach of Wikidata to knowledge representation, which aims to facilitate the expression of diverse knowledge ([Vrandečić and Krötzsch, 2014](#)). One example of that is Item Q1218, i.e. Jerusalem, which is reported both as the capital of Israel and of the State of Palestine⁴. Qualifiers also allow the KG to convey uncertain information: the year of birth of the Greek philosopher Socrates is not known with certainty and has been indicated in either 469BC and 470BC; therefore, both dates have been added to the philosopher's page on Wikidata⁵. Additionally, qualifiers can be applied to multiple claims using the same Property, but related to

⁴<https://www.wikidata.org/wiki/Q1218>, consulted on 1 February 2019.

⁵<https://www.wikidata.org/wiki/Q913>, consulted on 19 January 2019.

Property	Property label	Description	No. of uses
P143	imported from	Used for data imported from other knowledge bases.	44,292,397
P248	stated in	Used with references pointing to items within Wikidata.	57,245,036
P854	reference URL	Points to external web page.	13,814,962
P887	based on heuristic	Points to some heuristics.	9551
P1343	described by source	Points to an Item representing a reference work.	170

Properties used to add contextual information to references

Property	Property label	Description	No. of uses
P792	chapter	Used to specify the number of chapter when the source is a book.	694
P813	retrieved	Date of retrieval for URLs.	62,841,549
P1065	archive URL	Link to an archived version of the web page used as a reference.	26,363
P1683	quote	Reports a quote providing evidence for a statement.	1604
P2960	archive date	Date in which a website was archived.	431
P3452	inferred from	Statement added based on related inverse statement found on the object Item.	1075
P1480	sourcing circumstances	Qualification of the truth or accuracy of a source.	709

TABLE 5.1: Properties used in Wikidata references (instances of Q18608359) at 1st October 2017. Usage has been calculated by counting the number of references in which each Property appears.

different points in time, as it may be in cases such as the population of a place or someone’s occupation. If users judge that a claim should be preferred against others in the same statement, they can assign to it a rank amongst *normal*, *preferred* or *deprecated*. Figure 5.2 shows an example of multiple claims using the same Property and enriched by qualifiers, references, and ranks. In that case, the most recent figure has been given a preferred rank. Since the inception of the project, the Wikidata community has added around 48 million qualifiers to a total of almost 600 million statements.

5.2 The Wikidata ontology

As we have seen in Chapter 3, KGs are often built on top of schemata, which set the possible relationships and attributes of entities within the graph. These schemata, or ontologies, can be either formally-defined, such as in the case of DBpedia, or emerge from the relationships between entities in the graph, e.g. as it is done in Freebase (Section 3.4). Wikidata does not have a predefined or formal ontology and adopts the

latter approach. This is part of an overall liberal approach to knowledge engineering, which trades off knowledge expressiveness for ease of use, in order to lower entry barriers and enable users with varying levels of experience and skills to contribute (Vrandečić, 2013; Vrandečić and Krötzsch, 2014). We discuss the practical implications of this approach in the next paragraphs.

Wikidata does not formally distinguish between Items that are classes, for example *city*, and Items that are entities, for example *London*. This distinction is important for any AI that would use the knowledge graph to understand whether in a given context the entity *London* would refer to the largest city in the United Kingdom or would rather stand for the British government, which is an entity of a different class, *executive body*. Taxonomic relations are described in Wikidata mainly through two Properties, P31 (instance of) and P279 (subclass of). Together with others, e.g. P1647 (subproperty of) or P1709 (equivalent class), these Properties have been created by the community having in mind analogue OWL/RDF relations, specifically `rdf:type` (P31) and `rdfs:subClassOf` (P279). Properties are defined similarly to Items. Structured data about them is added through statements, while human-readable information is expressed through labels, descriptions, and aliases.

Ontologies prescribe how Properties must be used to reduce the risk of inconsistencies. The domain and range of a Property can be specified, to define which classes of entities can take respectively as a subject and as an object. For example, the Property *head of government* must link to an entity that refers to a human and not, say, a musical instrument. Constraints may determine other attributes, e.g. the Property *spouse* is symmetric, meaning that the relationship *A::spouse::B* entails that *B::spouse::A*. Wikidata does not enforce any restriction on Properties. Any Property can take any value, with the exception of the data type, i.e. Item or literal, which is set by the system. Furthermore, there are no formally declared property types such as in OWL, where properties may be classified as either `owl:ObjectProperty` or `owl:DatatypeProperty`. Wikidata editors appear to value this design choice (Piscopo et al., 2017c), similarly to what noted in other online platforms (Hall et al., 2017). Initially, constraints were added to the Talk page of a Property, therefore being documented in a free text form. However, around mid-2016 the community started to express them as statements, which means they are effectively part of the graph and can be used by machines. Up to 1st October 2017, 20 constraints have been added to the graph. Whereas some have OWL equivalents, and have been previously expressed in RDF (Erxleben et al., 2014), others have not.

A consequence of Wikidata’s approach to knowledge engineering is that its ontological knowledge may change over time. Classes and entities can be added and modified by anyone. Properties cannot be added arbitrarily by community members, a practice allowed in Freebase (Section 3.4). They need to be proposed by editors who have

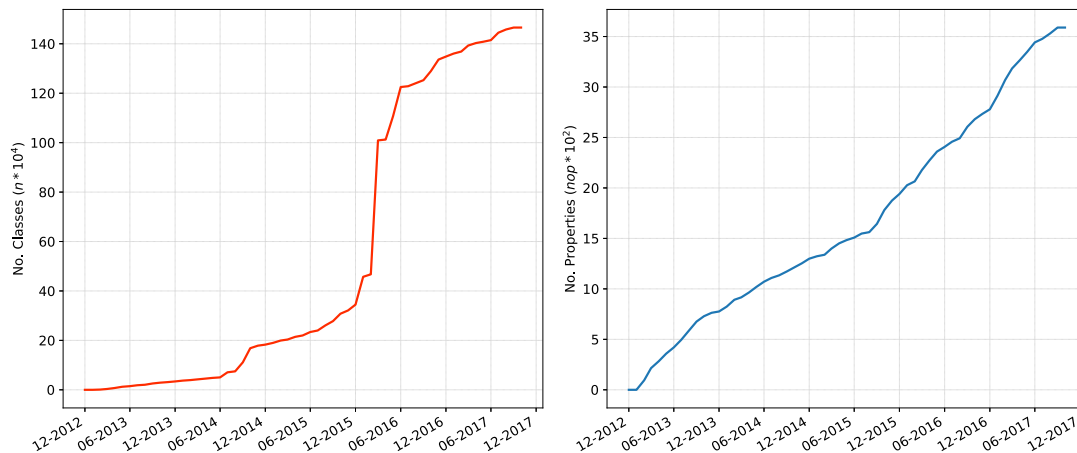


FIGURE 5.8: Number of classes and Properties over the lifespan of Wikidata













	 WIKIDATA (October 2017)	 DBpedia (April 2016)	 YAGO (YAGO3)	 Freebase (March 2015)	 OpenCyc (May 2016)	 NELL (08m.995)
Formally-defined ontology						
# classes	1,465,689	736	569,751	53,092	116,822	290
# Properties	3589	3555	106	70,902	18,028	1334

FIGURE 5.9: Ontology size of Wikidata, compared to DBpedia, YAGO, Freebase, OpenCyc, and NELL

previously been granted Property proposal rights, then reviewed and accepted by the community.

Several works (e.g. [Brasileiro et al. \(2016\)](#); [Erxleben et al. \(2014\)](#)) rely upon P31 and P279 to define classes in Wikidata and study its ontology. We follow the same approach in the current study. We define the Wikidata ontology as the set of all Properties and the Items that are used as classes, i.e. those that are subject or object of **subclass of** (P279), or object of instance of (P31). For example, in the statement

Ada Lovelace :: place of birth :: London

the ontology would include the Property **place of birth** alongside all classes to which the two entities, **Ada Lovelace** and **London** are linked to: humans, cities, capitals, financial centres etc. According to these criteria, the Wikidata ontology has constantly grown throughout its lifespan, including around 1.4 million classes and more than 3500 Properties in October 2017 (Figure 5.8). This is larger than many of the KGs previously discussed, as Figure 5.9 shows.

5.3 Accessing Wikidata

The data model of Wikidata was initially encoded in JSON. Only later an RDF version was made available. The conversion was not completely straightforward and different solutions have been proposed. These are discussed by [Erxleben et al. \(2014\)](#), who opted for translating the original Wikidata data model into RDF by using reification. This means that for every statement a corresponding resource is created, to which qualifiers, references, and/or ranks are attached by means of Properties. This approach, with few modifications, has been officially adopted by Wikidata. As regards data format and access, JSON and RDF dumps are released by the Wikimedia Foundation around twice a month and are freely downloadable. Besides the JSON and RDF encodings, Wikidata Items and Properties can be accessed and edited through a web interface. Finally, Wikidata can be queried through a SPARQL endpoint.⁶

⁶<https://query.wikidata.org/>, consulted on 15 January 2019.

Chapter 6

Wikidata as a collaborative system

Wikidata is a free, community-driven project. *Free* because its philosophy allows anyone to contribute: anybody can contribute anything, with minimal restrictions posed on the type of edits allowed, and without registration required. *Community-driven* because it is entirely edited and maintained by its community of users.

This chapter discusses these two features and explores various aspects of collaboration in Wikidata. Moreover, we provide a description of the editing activities carried out by Wikidata users.

6.1 Editing Wikidata

Wikidata can be edited through several interfaces. The easiest and probably the most straightforward one is the web interface. Every entity in Wikidata is represented as a web page, as described in Chapter 5. Any user can retrieve the entity he/she is interested in from the Wikidata main page. From there, he/she can land on the Item page and add or modify any content. The interface facilitates edits by showing suggestions of a number of Items with a label matching the text entered by the user (Figure 6.1).

Semi-automated editing tools, such as *QuickStatements*¹ or *The Wikidata Game*², are another commonly used interface. These allow users to edit at a much higher rate than it would be possible through the web interface, e.g. QuickStatements accepts csv files with a list of statements to be added, or (in The Wikidata Game) to check the quality of statements suggested by an algorithm.

¹<https://tools.wmflabs.org/quickstatements/#/>, consulted on 15 January 2019.

²<https://tools.wmflabs.org/wikidata-game/distributed/>, consulted on 15 January 2019.

Ada Lovelace (Q7259)

English mathematician, considered the first computer programmer

[edit](#)

Augusta Ada Byron | Lady Ada | Augusta Ada King, Countess of Lovelace | Ada Byron | Augusta Ada King | Ada King

[In more languages](#)

Statements

place of birth	<input type="text" value="London"/>	✓ publish ✖ remove ✕ cancel ⓘ
	London capital of England and the United Kingdom	✖ remove
	London city in Southwestern Ontario, Canada	+ add qualifier
	London city in Madison County, Ohio, United States	+ add reference
	London city in Kentucky, United States	+ add value

FIGURE 6.1: An example of Wikidata Item in editing mode. According to the text typed in the box, the system suggests a number of Items, showing their description in the user's language of choice for disambiguation purposes.

Revisions in Wikidata may differ greatly in terms of the effort they require from users. The majority of edits require little skills or knowledge other than the factual information that is added to the graph. For example, if the piece of information regarding Ada Lovelace's birthplace is missing, a user can just add a claim, using the **place of birth** Property and relating it to the **Q84 (London)** Item. Existing statements do not need to be adjusted to add this new piece of information. The same applies to labels: if a user wants to add the Italian label for London, he/she will only need to add it in the appropriate place.

Other tasks may be less trivial, though. Adding statements using the Properties **instance of** or **subclass of** theoretically requires the user to be familiar enough with knowledge engineering concepts to understand the difference between the two, and use them accordingly. Similar skills are required when modifying Properties, whereby understanding their relationship with other Properties, the constraints that may apply to them, and their intended use is essential. Based on these differences, tasks can be classified into **lightweight** and **heavyweight**, according to the definition given by [Haythornthwaite \(2009\)](#). Lightweight edits, such as adding claims and labels, require little specialised knowledge and are largely independent from other revisions. On the other hand, carrying out revisions on taxonomic hierarchies, modifying Properties, or working on claims entailing particular types of relations (e.g. meronymic relations) demands at least some experience of knowledge engineering principles. These types of edits may impact a larger portion of the graph and can be defined as heavyweight. For instance, a change in the definition of the Item *city* may influence all the Items that are instances or subclasses, thereby possibly preventing users from finding the information they need from Wikidata. Figure 6.2 shows an example of heavyweight edit and what it entails for other Items in the graph. The distinction between lightweight and heavyweight revisions, whilst not made by editors as such, is somehow recognised by them. Some editors reported in [Piscopo](#)

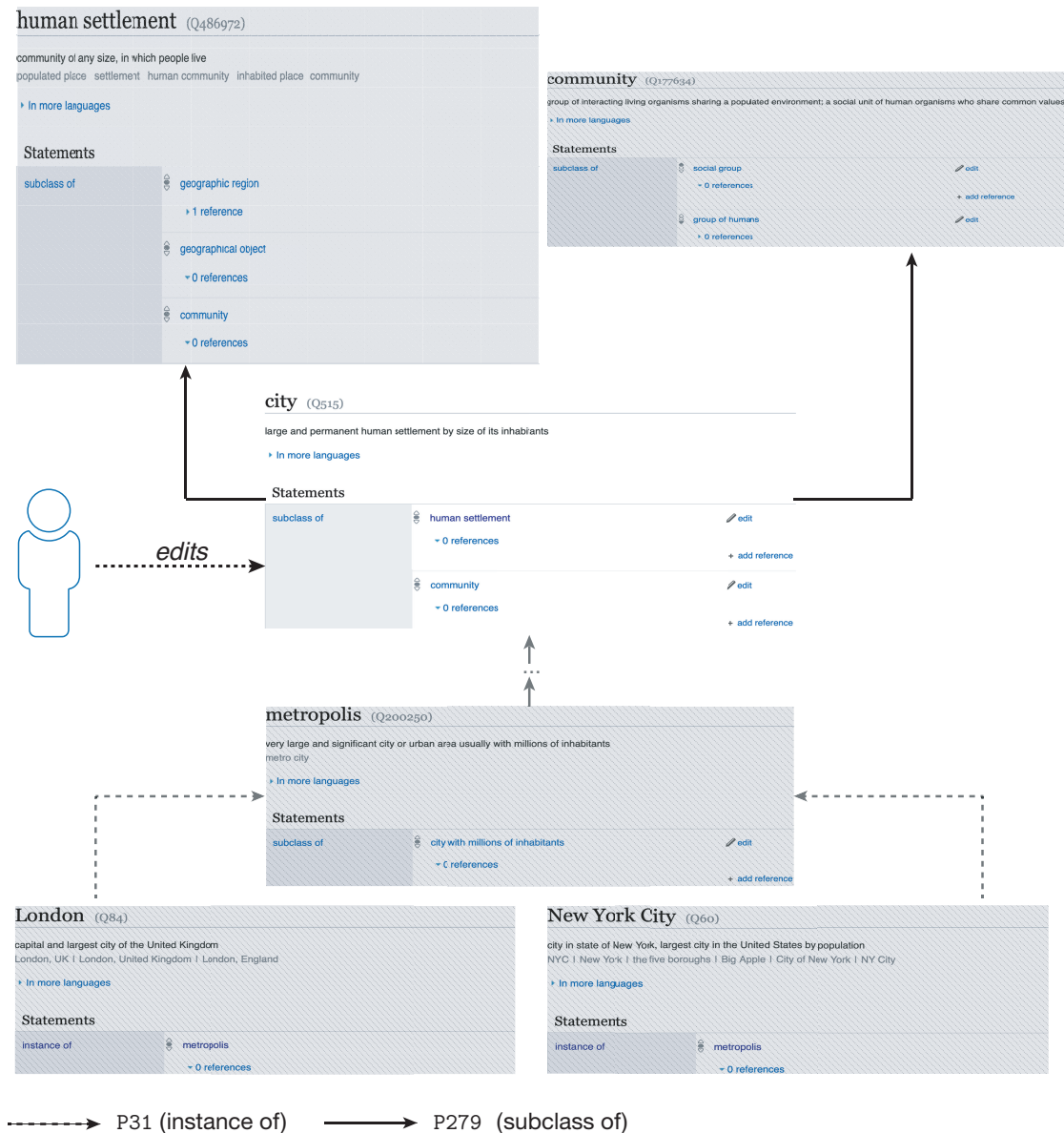


FIGURE 6.2: Graphical example of heavyweight edit. User *P*. edits Item Q515 (i.e. city). The meaning of Q515 is determined by its super-classes. Furthermore, *P*.’s revisions affect several other Items, namely all those down the sub-class hierarchy of Q515 and their instances, e.g. London, New York, Amsterdam, etc. Edits on Items with characteristics similar to Q515 may potentially affect a very large number of other Items in the graph. Opaque Items and relations are not seen directly by the user. To give an example of the possible effects of *P*.’s edits on the Item city, if she removed the statement city::subclass of::human settlement, a query for all Items that are instances of human settlement (and of its subclasses) would not retrieve any instance of city anymore.

et al. (2017c) to perform edits on the taxonomic structure of the graph, in opposition to revisions adding simple pieces of information.

Finally, another class of edits made by users concerns discussion, or Talk, pages. The wiki approach underpinning Wikidata includes communal spaces where editors can exchange opinions, discuss policies and norms, and reach consensus around various matters

User type	# users	# edits
Registered Human	190,765	171,824,150
Anonymous Human	548,956	2,329,109
Bots	407	384,660,528

TABLE 6.1: Number of users and edits per type. Please note that anonymous users are estimated by means of unique IP addresses. The same person may connect from different devices, meaning that different IP addresses may refer to the same user.

regarding the graph. Each page on Wikidata has a corresponding Talk page, including Items and Properties. Nevertheless, only a very small number of Items have active Talk pages (10,787 over around 40 million), i.e. pages in which a discussion has been initiated. Properties with active Discussion pages are in proportion more numerous (2569 over 3589 Properties), probably because constraints previously used to be added to Talk pages.

6.2 Wikidatians

The Wikidata community has continuously grown along the whole lifespan of the project, reaching a total of more than 190 thousand unique users. Users do not need to register to contribute. Besides humans, pieces of software called *bots* (Section 6.2.2) are active on Wikidata. They are programmed to carry out various types of tasks on the system, such as editing Items and Properties or patrolling the graph for quality control checks.

Not all users contribute equally and differences in terms of edit volume exist along various axes. The main difference is by user type (Table 6.1). Bots are the authors of the majority of edits on Wikidata. In the early years of the project, their edits reached around 90%, being one of the highest shares of automated contribution within the Wikimedia ecosystem (Steiner, 2014). Over the years, the percentage of bot revisions has declined though, albeit remaining over 50% (Figure 6.3). This picture is reversed when it comes to Properties, where registered editors account for almost the totality of edits (Figure 6.4). Anonymous editors carry out the smallest number of revisions, although their contribution is comparable to bots when it comes to Properties (Figure 6.4). Other differences exist among registered human users (*registered users* or *humans* from now on). We look at these in Section 6.2.4.

6.2.1 User rights

Anybody can contribute anything in Wikidata, with no restrictions on what editors can do, with minimal exceptions. These are typically features that are deemed to be crucial for the graph, such as creating Properties and changing data types, or are particularly debated, such as blocked Items. Wikidata has a system of user rights that determines

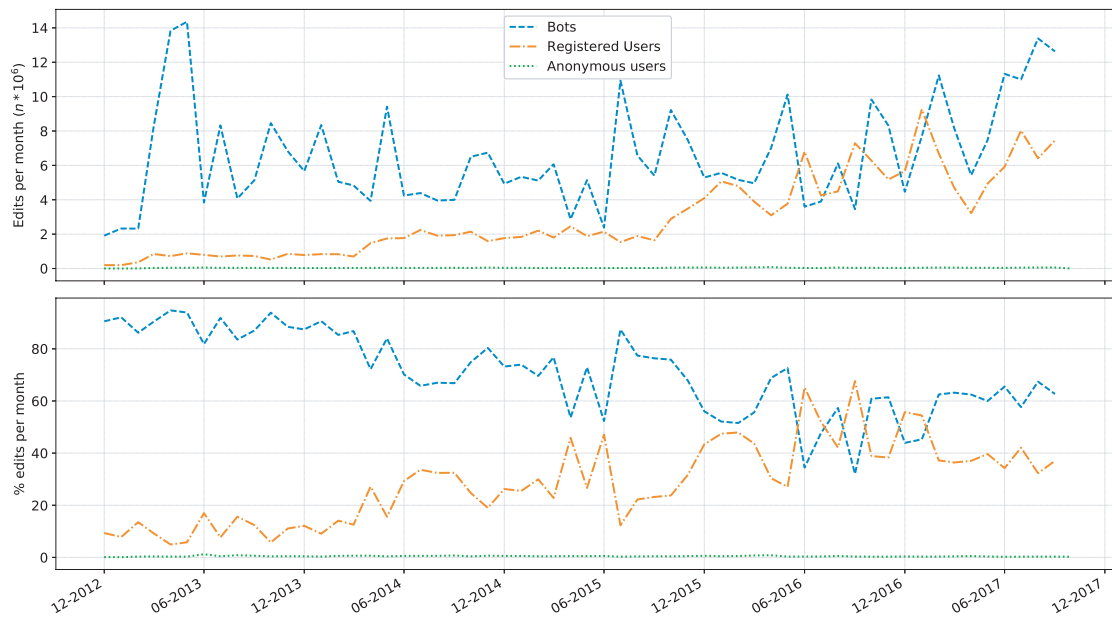


FIGURE 6.3: Number and percentage of edits per user type

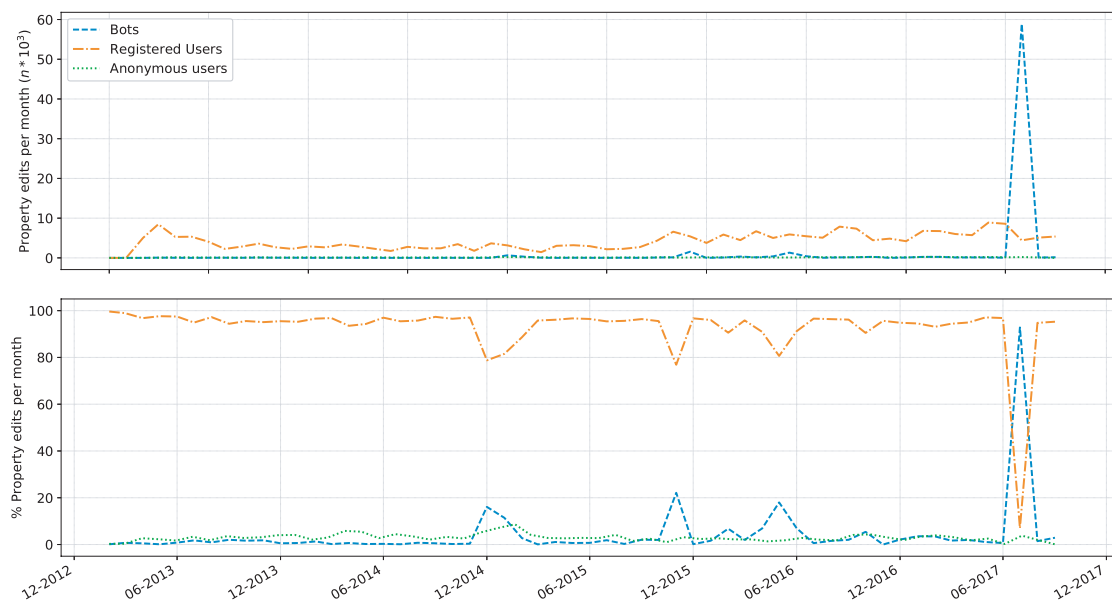


FIGURE 6.4: Number and percentage of Property edits per user type. The peak in bot edits after June 2017 is due to the activity of a number of bots importing Property constraints from Discussion pages to the related Property definitions. Please note that the scale differ from that in Figure 6.3.

the type of actions permitted to each user, similarly to Wikipedia. Rights are granted by assigning each user to one or multiple groups. One of these is the *bot* group, which we discuss in the following section. The other groups are mainly reserved for human users and form the organisational structure of the Wikidata community.

Some groups, i.e. *unregistered users*, *new users*, and *autoconfirmed users*, are automatically assigned to editors as they perform any type of action on the platform. Autoconfirmed—and confirmed—users are authorised to edit semi-blocked pages and

read abuse logs after a certain number of days (4) from registering, a minimum of 50 edits, and/or having manually confirmed their account (Wikidata, 2018i). Editors in this group can revise Items and create new ones, and use various types of tools to perform revisions in batch.

Community approval is required to become member of the *administrator*, *steward*, *bureaucrat*, *translation administrator*, and *bot* groups. These grant different sets of rights, which may go from performing edits at a very high rate (i.e. hundreds of revisions per minute), to deleting pages, blocking and unblocking other users, modifying page templates, and creating new Properties, etc. *Administrators* (or *sysops*) and *stewards* are the groups with the largest sets of rights. These users oversight the work on Wikidata and are highly involved in the social administration of the community. Other groups, which we call *lower administrator groups*, need only the approval of an administrator to accept new members. Each of these groups is granted a subset of the administrators' rights. For example, *Property creators* are the only ones allowed to create new Properties, besides administrators; *rollbackers* can revert large number of edits if they deem those as vandalism; *ombudsmen* are in charge of investigating possible violations of privacy, etc.

6.2.2 Bots

Bots have a prominent role in the production of Wikidata's knowledge. We have already seen that they author the majority of edits on the platform. In this section, we look in more detail at how bots are run and what their role on the platform is.

In order to operate freely, i.e. without any editing rate cap, bots must be approved by the Wikidata community. Each bot is proposed by a user, who commits to continuously check its work and to suspend it in cases it causes any harm to the graph. The procedure to approve a bot is described by a set of Wikidata policies (Wikidata, 2018a). Editors must open a *Request for permission*³, in which they need to provide a detailed description of the activities that their bot is planned to do. After a test run (between 50 and 250 edits), the community can leave comments and vote in favour or against the activation of the bot. To revoke the authorisation, any user can open a Talk page on a dedicated section of the Wikidata platform, asking to suspend the bot's activities.

One of the first functions for Wikidata Items has been to act as inter-language hubs for different language versions of Wikipedia articles (Vrandečić and Krötzsch, 2014). Therefore, the first Wikidata bots harvested inter-language links from all Wikipedias and added them over to Wikidata, where each Item was connected to the corresponding articles in several language versions of the free encyclopaedia. Besides this first task,

³https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot, consulted on 6 August 2018.

bot activities have focused on importing new statements and enriching the knowledge graph. A 2015 study (Müller-Birn et al., 2015) shows that almost 90% of automated editing in Wikidata concerns the addition or modification of Item statements (58%) or labels/descriptions/aliases (30%). Nonetheless, bots perform other types of tasks as well, similar to human editors in constructing Wikidata’s knowledge. According to Müller-Birn et al. (2015), bots are active as reference editors, Item creators, Item editors, Item experts, and Property editors, five of the six user roles identified for Wikidata editors.

Given the key role of bots in editing Wikidata, it is important to study how they affect its quality. Their contributions have been studied so far either in terms of their percentage over the total of edits, compared to different Wikipedia language versions (Steiner, 2014), or to identify emerging activity patterns, in the above mentioned study by Müller-Birn et al. (2015). We will look at how they impact quality of references (Chapter 9) and Items (Chapter 10).

6.2.3 Anonymous users

Anonymous editors perform a minimal part of edits on Wikidata. The effects of anonymous users on quality have been studied in other online platforms. Research has looked at the behaviour of this type of users as a whole group. Since the same person can use different IPs in various editing session, it is hard to follow the same user through different sessions. Findings are contrasting. Whereas users who edit anonymously may have lower levels of attachment and are often responsible for spam and vandalism in Wikipedia (Adler and de Alfaro, 2007), other studies have found that some anonymous users produce high quality revision on the same platform (Anthony et al., 2009). To the best of our knowledge, no study has looked at the effects of anonymous users’ contributions on quality in Wikidata as of yet.

6.2.4 Registered users

Registered human editors are the largest part of the Wikidata community. Their number has continuously grown, both in terms of unique users, who have reached a total of around 190,000 in October 2017, and of monthly active users, topping 17,000 in the same period (Figure 6.5).

The level of engagement of registered users in Wikidata varies greatly, following a pattern already observed in other online communities (Ortega et al., 2008). A core of editors does the lion’s share of the work, whereas a long tail of users contributes only marginally. Ortega et al. (2008) have computed the Gini coefficient of the number of edits per users in several Wikipedia language versions. This coefficient, introduced by Gini (1936) to estimated income inequality, provides the level of inequality in a set of values. It is computed as

$$\frac{\sum_{i=1}^n (2i - n - 1)x_i}{n \sum_{i=1}^n x_i} \quad (6.1)$$

where the observed values x_i are ranked in ascending order and n is the number of values observed. It ranges from 0 to 1, with higher values meaning higher inequality. We followed the same approach to gain some insights about the inequality of contributions in Wikidata. At the most recent date of our dataset (October 2017), the Gini coefficient of Wikidata edits per user was 0.97 (0.99 taking into consideration also bots). This is higher than any of the Wikipedia versions investigated by [Ortega et al. \(2008\)](#) (English Wikipedia 0.93 Gini). However, the evolution of the Gini coefficient in Wikidata shows a pattern similar to those of the Wikipedias analysed by [Ortega et al. \(2008\)](#), whereby it shows initial lower values, to subsequently grow and stabilise on values closer to 1 (Figure 6.6).

We have noted in Chapter 2 how analysing editor behaviour by the year in which they joined the platform may uncover dynamics that are likely to be under- or misrepresented by taking into account the entire history of an online community. Hence, we divided editors in five yearly cohorts, from Wikidata's launch (October 2012) to the end of our dataset (September 2017), each including users joining from October of one year to September of the following (e.g. October 2012 to September 2013 etc.). Users who joined earlier in the project (before September 2013) are more active over the whole lifespan of Wikidata, considering different sets of activities. Early participants perform a higher number of revisions than any other user group, with the exception of bots, looking at the whole of Wikidata (Figure 6.7 and Table 6.2). They are also by far the most active cohort when considering only Property edits (Figure 6.7) or Discussion pages (Figure 6.8). This

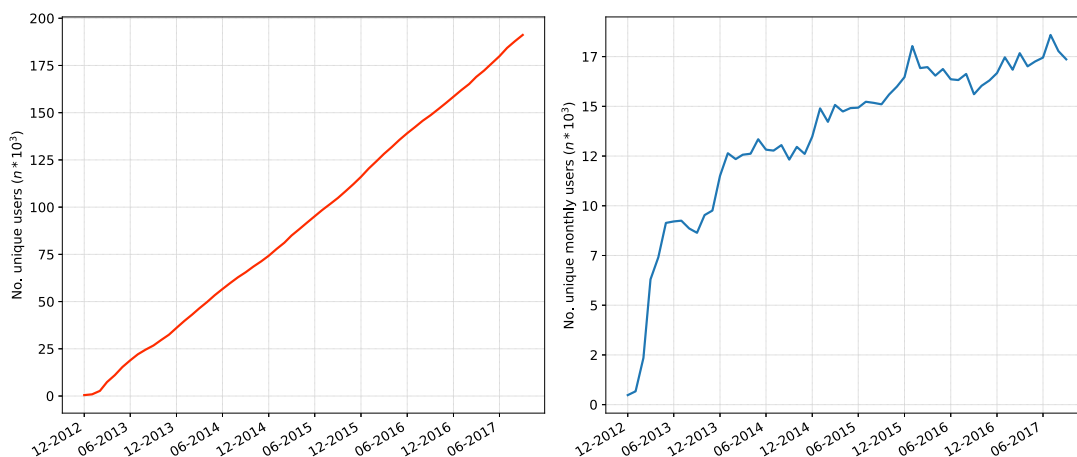


FIGURE 6.5: Number of registered users and monthly active registered users along the Wikidata lifespan

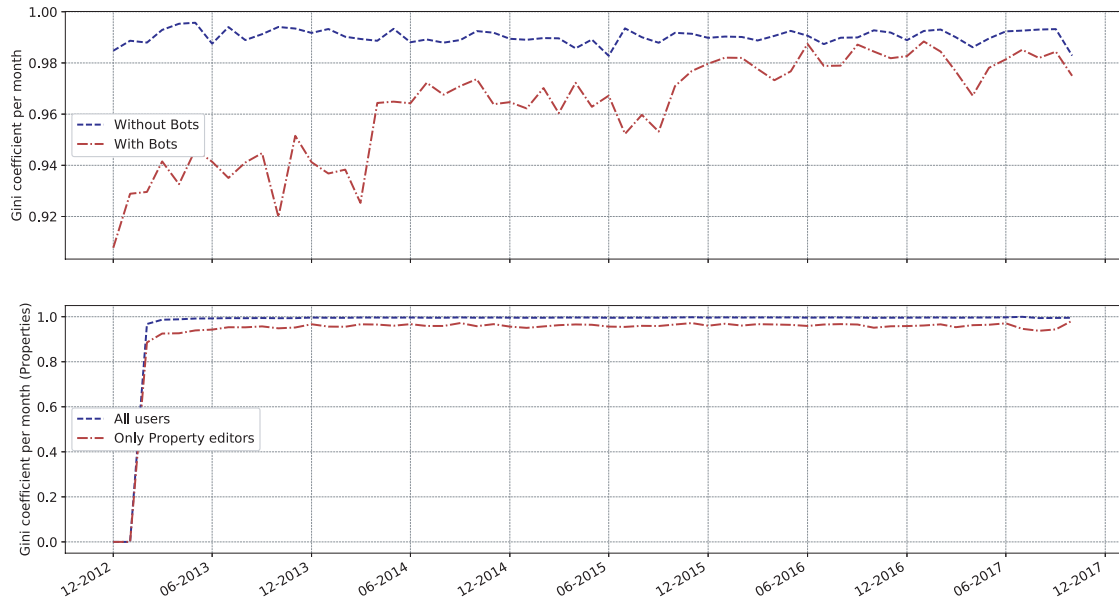


FIGURE 6.6: Gini coefficient over time. Above, all Items and Properties; below, only Property edits. Property editors refers to users who have ever performed any revision on Properties.

first joined	# users	# edits	Avg. edits per month
Oct. 2012–Sep. 2013	26,492	115,014,652	2,053,833
Oct. 2013–Sep. 2014	38,679	34,890,801	712,057
Oct. 2014–Sep. 2015	39,318	10,557,617	285,341
Oct. 2015–Sep. 2016	43,714	6,961,678	278,467
Oct. 2016–Sep. 2017	42,562	4,399,402	338,415
Total	190,765	171,824,150	3,124,075

TABLE 6.2: Breakdown of users by yearly cohort

may be explained by a self-selection bias, i.e. users arriving to Wikidata in its early stage may be on average those who are more inclined to like it, a behaviour that has been noted also among users of other platforms (e.g. Reddit ([Barbosa et al., 2016](#))). The majority of the most active Wikidata editors have been also established Wikipedians, as discussed by [Piscopo et al. \(2017c\)](#), therefore having the opportunity to participate in early discussions about the launch of the new Wikidata project. This would be in agreement with the self-selection bias hypothesis.

In [Piscopo et al. \(2017c\)](#), we have carried out a qualitative study in which we analysed how user activities change as they gain experience in Wikidata, shedding some light on how these differences may be connected. We interviewed seven experienced Wikidata users about their perception of themselves within the platform, of the community, and of the interface, asking them how this changed compared to their time as novices. Interviewees' responses were examined through the lenses of activity theory ([Kuutti, 1996](#)) and legitimate peripheral participation ([Wenger and Lave, 1991](#)). Established users seem to perform a larger number of revisions and are more active within the community, taking

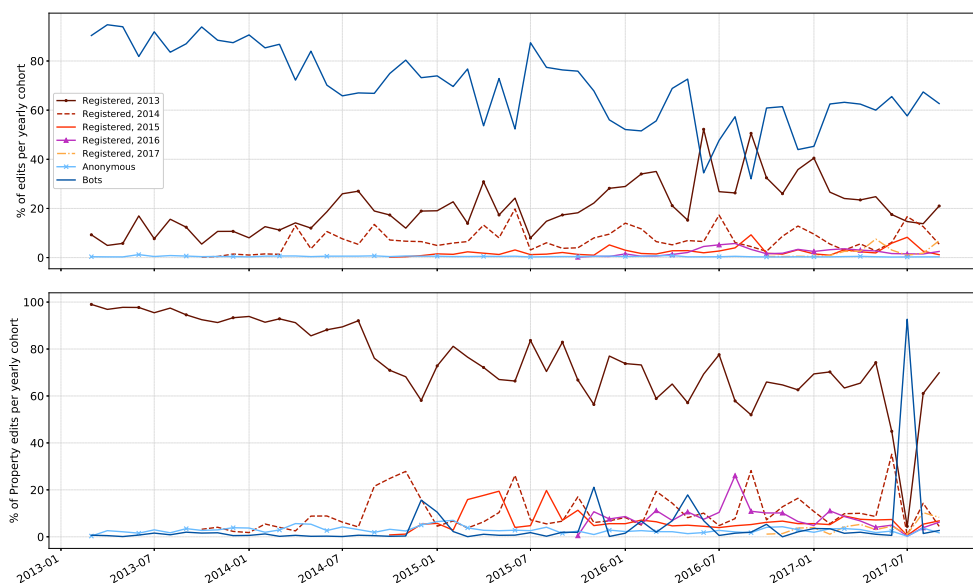


FIGURE 6.7: Percentage of edits (above) and Property edits (below) per yearly cohort and user type

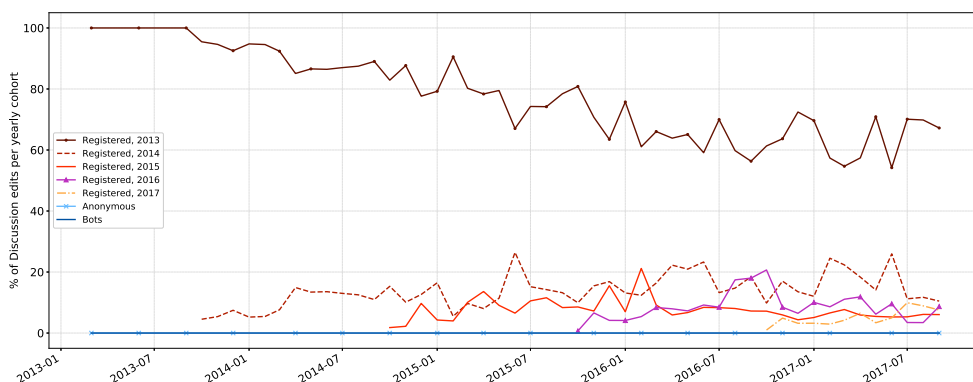


FIGURE 6.8: Percentage of edits to Discussion pages per yearly cohort and user type

part more often in discussions with other members and mentoring new ones. Moreover, whilst as novices they focus on a narrower number of Items of their interest at the beginning, as they become more experienced they are more likely to edit a broader range of Items, e.g. adding or correcting a type of statements. This is connected to their augmented sense of responsibility, which leads them to gain a higher level look on the quality of the graph, and their increased use of semi-automated tools. Additionally, established users report to carry out design and maintenance tasks on the ontology of Wikidata. Some of them voice their concerns about common errors in the construction of taxonomic hierarchies (i.e. involving Properties P31 and P279) done by novices, echoing what was reported in [Brasileiro et al. \(2016\)](#).

Some considerations may be drawn about the effects of the features observed so far on the outcome of community work. First, Wikidata users participate in different ways

and to a varying extent to the activities that build the knowledge graph. Prior literature has shown that diversity in groups may lead to different outcomes. We discuss this in Section 6.2.4.1 and suggest some implications for Wikidata. Second, user roles may emerge from the editing patterns of Wikidata editors. This is suggested by the findings from [Piscopo et al. \(2017c\)](#), but has not been supported quantitatively as of yet. Section 6.2.4.2 is dedicated to earlier studies about user roles in Wikidata and other collaborative platforms.

6.2.4.1 Group composition and diversity

Group diversity may be a double edged sword, as discussed in Chapter 2. In the previous sections, we have seen that the behaviour of Wikidata editors varies along different lines. Some users are highly active, whereas others perform only a few edits. Furthermore, only a small number of editors has even done any revision on a Property. User behaviour is not static and may change over time. It has been noted that Wikidata editors change the focus and scope of their activity as they gain experience within the system ([Piscopo et al., 2017c](#)). This suggests that users with different levels of experience may bring complementary skills in building high-quality Items. Editor activity may also vary along their edit scope, i.e. the object and type of revisions made ([Piscopo et al., 2017c](#)). Some users carry out similar tasks, i.e. adding references, on a broad spectrum of Items, whereas others narrow down their activity on a smaller selection, specialising on a single domain. The contribution of these two types of users may thus be complementary and necessary to create good quality content. Verifying this hypotheses—the complementarity of editors with different levels of experience and different edit scope—would allow to harness the contribution of various types of users of the platform, e.g. facilitating the collaboration between editors with diversified skills.

6.2.4.2 User roles

Prior research has looked into emergent user roles (see Chapter 2) in Wikidata. [Müller-Birn et al. \(2015\)](#) have analysed Wikidata edit logs in order to identify roles emerging from activity patterns related to the type of information produced by editors. Their work draws upon peer-production systems and collaborative ontology engineering literature to profile editors based on their revisions’ scope (e.g. claims, labels, references) and type (e.g. addition or deletion). Human users are categorised into six profiles. Two of them are primarily engaged in modifications to the ontology (*Property editors* and *Property engineers*), whereas other profiles focus to adding and revising other elements, such as references and common Items. Bots have similar profiles, with the exception of Property engineers. The work of [Müller-Birn et al. \(2015\)](#) analyses profiles emerging by editors focusing on different types and scopes of edits. However, their research overlooks other dimensions, such as tenure, number of edits, participation in the discussions,

contribution to the ontology. These aspects are brought up in the study of [Piscopo et al. \(2017c\)](#), which sheds some light on the tasks undertaken by users with different levels of experience in Wikidata and hints at the existence of various user roles within the platform. Nevertheless the research is qualitative and based on a small sample of interviews. This thesis includes a quantitative follow-up of that study, to identify Wikidata user profiles over time and understand their impact on the part of Wikidata that seems to rely on experienced contributors, its ontology.

Chapter 7

The quality of Wikidata

The previous two chapters have described the main features of Wikidata and discussed some of its specificities. Drawing from these and from the background work in Chapter 4, this chapter presents prior studies about the data quality of Wikidata, both from the research literature and from the Wikidata community. The work presented in this chapter has been subsequently expanded in [Piscopo and Simperl \(2019\)](#).

7.1 Data quality in Wikidata

A growing body of literature looks at data quality in Wikidata. In the following, we provide an overview of works in the area, drawing attention to the aspects that have not been covered yet.

Several studies compare Wikidata to other resources. [Thakkar et al. \(2016\)](#) adopt a data consumer viewpoint, relying upon the framework developed by [Zaveri et al. \(2016\)](#). They choose a task—i.e. question answering around a number of domains—and compare the quality of Wikidata and DBpedia over several dimensions. Their results show that Wikidata outperforms DBpedia in what concerns completeness of the data (in some of the domains analysed), diversity (i.e. the availability of different formats), and trustworthiness. With regard to the latter, although provenance richness varies in Wikidata, i.e. the extent to which sources are specified for the piece of data in the graph, [Thakkar et al. \(2016\)](#) note that this feature—i.e. references—is simply missing in DBpedia.

[Färber et al. \(2018\)](#) propose a quality framework to select the most suitable knowledge graph for a given task. Their work compares five KGs—DBpedia, Wikidata, YAGO, OpenCyc, and Freebase—along a set of dimensions derived from [Wang and Strong \(1996\)](#), [Bizer \(2007\)](#) and [Zaveri et al. \(2016\)](#) (see Table 4.1). They perform a high-level evaluation, i.e. providing for each KG an overall score in each dimension. Some of their metrics are expressed as a ratio between the number of correct instances to

the total of instances, e.g. syntactic accuracy of literals is measured as the proportion of literal values matching an expected pattern. Other metrics simply gauge whether a certain feature is present in a KG. For example, the *trustworthiness on statement level* metric allows three values, depending on whether provenance is used at statement level, at resource level, or not used. The evaluation of Färber et al. (2018) is useful to provide an overview of the capabilities of Wikidata. Concerning accuracy, 100% of the RDF triples and the literals evaluated in Färber et al. (2018) were syntactically correct (see Section 4.2 for definitions of syntactic and semantic accuracy), whereas > 90% of the triples assessed matched those in the gold standard—the second best performer within the set of KGs studied. Wikidata achieves the highest scores in the trustworthiness dimension, due to being manually curated, to the possibility to add references for each statement, and to the support of empty and unknown values (Färber et al., 2018). With respect to consistency, the evaluation looks at the existence of schema restrictions checks at the time of statement creation, a feature implemented in Wikidata editing interface. However, it gauges other aspects of consistency that are hardly measurable in Wikidata, because of its approach to expressing ontological knowledge—e.g. contrarily to the other KGs evaluated, Wikidata does not use OWL; Färber et al. (2018) evaluate the number of inconsistent axioms by checking disjoint statements via `owl:disjointWith`, which is not used in Wikidata, therefore no inconsistent axioms are found. Moreover, Wikidata scores well in other dimensions, such as relevancy, because of the possibility to rank statements; timeliness, for being continuously updatable by editors and for qualifiers, which can be used to specify the temporal validity of a statement; accessibility, where Wikidata is one of the KGs to provide full RDF exports, have a public SPARQL endpoint, and support content negotiation to provide an HTML representation of its resources; and licensing, as all the data in Wikidata can be openly reused and shared with no restrictions. On the other hand, the results for completeness and interlinking are somewhat mixed. Compared to a gold standard schema, Wikidata contains the highest degree of classes and properties, compared to the other KGs in the study. However, several properties are not used to create statements for the entities of the classes the should describe. With respect to interlinking, Wikidata is connected to several external resources by functioning links. Nevertheless, it does not use the standard approach used on the Linked Data web to connect resources describing the same entity, i.e. via the `owl:sameAs` relation.

Whilst Färber et al. (2018) have a comprehensive look at various quality dimensions, their study gives little insights about how good the data in the graph is. This is especially true with respect to dimensions such as trustworthiness/verifiability, consistency, or relevance, whereby the metrics used in their study provide only discrete values about the presence or absence of a determined feature. Other approaches focus only on a single dimension. Prasojo et al. (2016) addresses completeness in Wikidata with a tool called COOL-WD¹. Combining crowdsourced work, information extraction techniques,

¹The tool can be found at <https://cool-wd.inf.unibz.it/>, consulted on 8 August 2018.

and entailments from the Wikidata RDF graph (Darari et al., 2016), COOL-WD can create completeness statements, describing whether an Item, a statement, or parts of the graph are complete. Human users can manage and create completeness statements, therefore adding a further manual check to the system. Kaffee and Simperl (2018) have examined the availability of labels in different languages for seven datasets, including samples from the Linked Open Data cloud, government datasets, datasets published by museums, and Wikidata. This was found to have the most comprehensive coverage in terms of proportion of entities with human-readable labels. Furthermore, it was the most diverse, supporting the largest variety of languages and having the least unequal distribution of coverage across languages.

7.1.1 Ontology quality

Other works have addressed the quality of the Wikidata ontology. Erxleben et al. (2014) exploited Property constraints, and the relations P31 (instance of) and P279 (subclass of) to extract an OWL ontology from Wikidata. The study of Brasileiro et al. (2016) used the same Properties to explore common issues in the Wikidata taxonomy, highlighting three main anti-patterns, attributable mainly to the misuse of P31 and P279. This generally consists of using a type or subclass relation in a statement that would require a different one. For example, in

Ada Lovelace :: instance of :: computer scientist

the correct Property would be *occupation*. Whereas in

MSF Canada :: subclass of :: Médecins Sans Frontières

the correct Property would be *part of*. Other quality issues involve an incorrect object Item or cause redundancies (rather than inconsistencies). For example, these occur when an Item is a sub-class of two Items, one of which is an instance of the other (Brasileiro et al., 2016).

Wikidata allows anyone to edit (virtually) any part of the graph, in a completely bottom-up fashion. There is no editorial oversight on its ontology, which can change quickly, as a results of edits on Properties and on the use of P31 and P279. An application looking up information in Wikidata via an API would receive very different results at different points in time, as these results very much depend on the ontological structure of the knowledge graph. While previous studies have started investigating how users edit the ontology, none of them has looked at editing activities over time as of yet.

Paper	Dimensions	Comparative	Results
Thakkar et al. (2016)	Availability, Completeness, Timeliness, Interlinking, Data diversity, Semantic accuracy, Consistency, Trust and Provenance, Conciseness, Coverage, Licensing.	Yes (DBpedia)	Task-based evaluation which looks at the fitness of two KGs for open domain question-answering using the Luzzu (Debattista et al., 2016) framework. Wikidata outperforms DBpedia in most of the dimensions considered. However, only slices of the two KGs are used for the evaluation.
Färber et al. (2018)	All those in Table 4.1.	Yes (DBpedia, YAGO, OpenCyc, and Freebase)	High-level comparative evaluation of KGs. Dimensions are evaluated over the whole graph, either as a ratio of valid instances over a total (e.g. completeness or accuracy), or as a variable representing the degree to which a feature is supported by the KG (e.g. trustworthiness).
Prasojo et al. (2016)	Completeness.	No	Tool which enables to generate completeness information regarding Wikidata statements.
Brasileiro et al. (2016)	Consistency.	No	Evaluation of taxonomy hierarchies. Three anti-patterns are found, which represent possible misuses of P31 and P279.
Kaffee and Simperl (2018)	Ease of understanding.	Yes (Billion Triple Challenge 2010 and 2014, UK and Taiwan governmental datasets, Swiss National Library data, Linked Open British National Bibliography)	Comparative evaluation of language coverage and diversity. Wikidata outperforms the other datasets, with respect to coverage and diversity of languages used.

TABLE 7.1: Wikidata data quality studies from the literature

Component	Showcase Item criteria
Statements	Minimum 10 statements with: <ul style="list-style-type: none"> – Non-Wikimedia Sources for non-trivial statements; – Appropriate ranks; – Qualifiers when applicable; – An image associated (optional).
Human-readable labels & descriptions	Labels, descriptions, and properties in ≥ 4 languages; When appropriate, aliases in each language.
Wikimedia links	Sitelinks to a complete and correct set of applicable pages on Wikimedia projects.

TABLE 7.2: Showcase Item criteria

7.2 Wikidata quality from the eyes of Wikidatians

As part of their efforts to define governance and norms regulating their communities, on-line knowledge collaborations have often developed policies and put in place a number of strategies to uphold quality (see Chapter 2). The Wikidata community has followed this practice as well, adopting consensus-based strategies from its elder sister Wikipedia and inheriting some of its policies. In the following, we describe the Wikidata community-based initiatives to control and assess quality.

7.2.1 Item quality

Items represent entities in the real world and are seen by editors as ‘unitary topics’ (Piscopo et al., 2017c). The community has undertaken several initiatives to measure quality of Items. Showcase Items (Wikidata, 2018h) are a set of Items selected by the community as outstanding examples of the capabilities of the system. The number of Showcase Items varies, but has been so far in the order of the few dozens. Showcase Items must meet a number of criteria (see Table 7.2) covering the different elements composing Items, i.e. statements, human-readable labels, and links to other Wikimedia projects.

Yapinus et al. (2017) relied upon the Showcase Item’s criteria to devise, in close-collaboration with the community of Wikidata, a single-grading scale which assigns labels to Items from A (the highest) to E. The grading scale covers the completeness of an Item, described as the number of relevant statements; the number of the sources used to support the statements; the labels and descriptions in an appropriate number of languages; links to other wiki projects; and possibly whether media files are attached². Quality criteria of the single-grading scheme are reviewed through discussions with the community and have subsequently been used to run an evaluation campaign, to which the Wikidata community could take place. The result of this campaign is a sample of

²https://www.wikidata.org/wiki/Wikidata:Item_quality, consulted on 1 February 2019.

5000 Items, each evaluated by one or more editors. A pilot campaign was run prior to the main one, in order to refine the quality labels.

Another attempt to address Wikidata quality is ORES, acronym of Objective Revision Evaluation Service³. ORES is not properly a community-based initiative, but has been built by developers working at the Wikimedia Foundation (Sarabadani et al., 2017). This tool, which has an API that can be queried by other applications, uses machine learning algorithms to detect damaging revisions on Wikipedia and Wikidata. It is also able to provide a quality score for the Item or Property at the moment in which the revision evaluate has been made, i.e. it can provide a quality assessment over time. The labels and the criteria used are those from the Item quality experiment mentioned above.

7.2.2 Reference quality

The **verifiability policy** (Wikidata, 2018j) specifies which statements need to be supported by a reference and sets the quality requirements for that. Statements must be verifiable by consulting a referenceable primary source. This must be **accessible** ‘by at least some’ Wikidata contributors to confirm the source firsthand (Wikidata, 2018j). A good reference must also be **relevant**—it must provide evidence for the claim it is linked to. Additionally, good references must be **authoritative** or ‘deemed trustworthy, up-to-date, and free of bias for supporting a particular statement’ (Wikidata, 2018j).

Wikidata defines authoritative sources by describing suitable types of publishers and authors. This is also the approach of Wikipedia, whose policy Wikidata refers to. Specifically, the term ‘source’ has three meanings in Wikipedia (Wikipedia, 2018a): the *type of work* itself, the *author* of the work, and the *publisher* of the work. The Wikidata policy specifies types of sources that are authoritative: books; academic, scientific and industry publications; policy and legislation documents; news and media sources. These must have a corresponding entity in Wikidata, linked to claims through P248 (**stated in**). Databases and web pages may also be authoritative. Databases require a corresponding property already defined in the knowledge graph, pointing to an entry in the database. Authoritativeness of web pages, referenced through P854 (**reference URL**), depends on their author and publisher type. Authors may be *individuals* (one or more identifiable persons), *organisations*, or *collective* (a number of individuals who often utilise a username and whose contribution is voluntary). Sources whose author is unknown should be avoided, as well as user-generated sources, e.g. forums or social review sites. Regarding publishers, sources with no editorial oversight and relying on rumours and personal opinions are generally not considered authoritative. Government agencies, companies and organisations, and academic institutions are authoritative publishers (Wikidata, 2018j). Self-published sources are generally not accepted, nor are websites with promotional

³<https://ores.wikimedia.org/>, consulted on 1 October 2018.

Type of publisher	Sub-types (when applicable) / <i>Definition</i>
Academic and scientific organisations	Academic and research institutions (e.g. universities and research centres, but not museums and libraries); Academic publishers; Other academic organisations.
Companies or organisations	Vendors and e-commerce companies; Political or religious organisations; Cultural institutions; Other types of company.
Government agencies	<i>Any governmental institution, national or supranational.</i>
News and media outlets	Traditional news and media (e.g. news agencies, broadcasters); Non-traditional news and media (e.g. online magazines, platforms to collaboratively create news).
Self-published sources	<i>Any sources that do not belong to any organisation/company, maintained by the authors themselves.</i>

TABLE 7.3: Types of publisher derived from Wikidata (2018f). On the right column, sub-types or, when these are missing, definitions of higher-level types.

purposes or those affected by political, financial, or religious bias. Wikipedia pages are not good references because they are not primary sources and are collectively created. Table 7.3 shows publisher types.

References are among the features that set Wikidata apart from similar projects. Provenance facilitates the reuse of data by improving error-detection and the selection of pieces of information based on their source (Lehmann et al., 2012). The lack of provenance information or the use of poor sources may affect trustworthiness of the data and hinder the its reuse for business and other purposes (Hartig and Zhao, 2009). Additionally, the availability of provenance information can increase trust in the project, as noted in Wikipedia (Lucassen and Schraagen, 2010). On a practical side, a method to detect bad external sources would support editors in maintaining Wikidata knowledge graph. Yet, no evaluation of Wikidata references has been carried out so far.

7.2.3 Constraint violations

Section 5.2 introduced Property constraints in Wikidata, observing that although they are not enforced, they are used to scan the graph for possible quality issues. This task is carried out by *KrBot*, a bot which has regularly scanned since April 2013 the whole knowledge graph and reported constraint violations for each Property. Constraints are used to define how Properties should be used and the relations that should exist—or not exist—for the classes they apply to. For instance, Property P26 (spouse) has the *symmetric* constraints (see Section 5.2). Other constraint violations may indicate potential errors, such as those related to the format of literals or stating that the Item

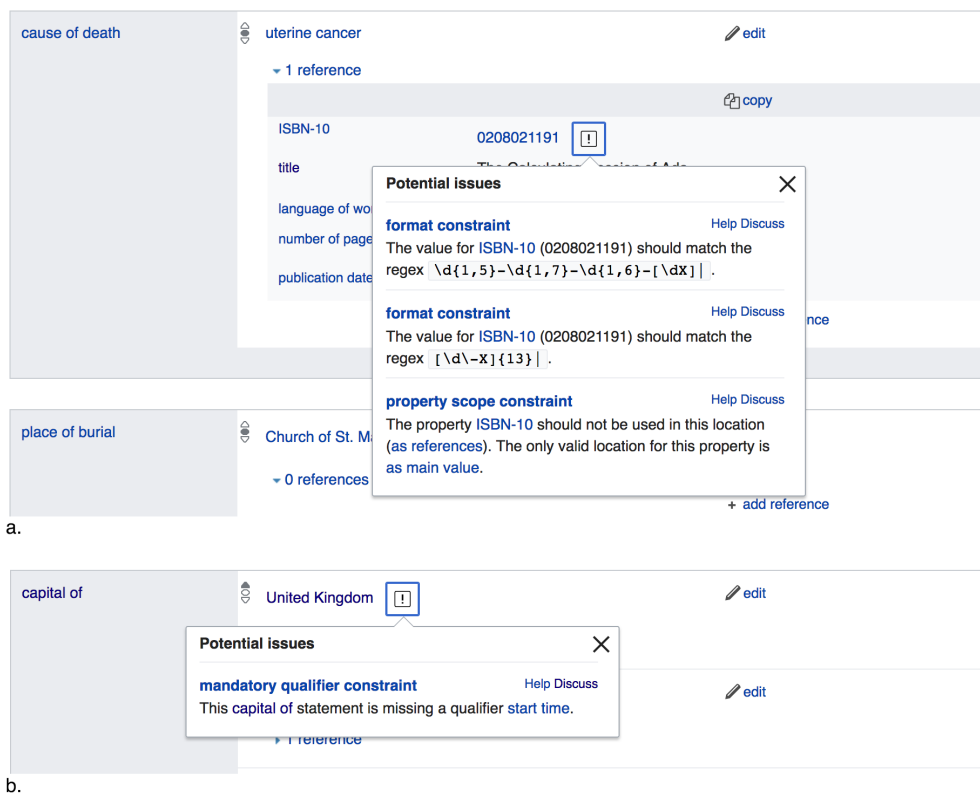


FIGURE 7.1: Examples of Property constraints violations. Figure a. is taken from Item Q7259 (Ada Lovelace). The *format* constraint checks whether the value used as an object matches a regular expression, whereas the *property scope* constraint refers to a specificity of Wikidata’s knowledge representation model, i.e. the type of statement where a Property can be used. Figure b. shows a violation for Q84 (London), suggesting that the information in a statement may be incomplete.

used as an object must be of a determined type. Figure 7.1 shows two examples of notifications of constraint violations.

Part II — Summary

- Wikidata is a collaborative knowledge graph. Its community is entirely responsible for adding content, maintaining its quality, and defining its policies and norms.
- **Items** and **Properties** are Wikidata’s building blocks. The first describe any type of entity, either abstract or concrete. The latter are used to express any relation between Items or between Items and literals. Wikidata’s knowledge graph consists in the set of all relations expressed by its Properties.
- Wikidata gives users the possibility to add provenance to every fact stored in the knowledge graph, via **references**. These can be either *internal*, i.e. pointing to another Item within Wikidata, or *external*, linking to a web page without the graph.
- Differently from other widely used knowledge graphs, Wikidata does not encode its schema in a formal ontology. Instead, it relies on a loose set of relations, which treat conceptual knowledge in the same fashion as any other type of information on Wikidata. Taxonomic hierarchies are expressed through the Properties P31 (instance of) and P279 (subclass of). Classes are not formally differentiated from any other Item and are by convention defined as the Items that are object of P31, or subject or object of P279.
- Wikidata editors can be either **humans** or **bots**. Although the latter numerically represent only a small part of Wikidata’s user pool, they carry out by far the majority of edits. A similar inequality in terms of number of contributions is observed for human editors, where a core of users does the lion’s share of work.
- Prior work has pointed out the different characteristics of novices and experienced users. The former perform a smaller number of edits and are less active in the community.

III

COLLABORATIVE WORK AND DATA QUALITY

Ya se sabe: por una línea razonable o una recta noticia hay leguas de insensatas cacofonías, de fárragos verbales y de incoherencias.

LA BIBLIOTECA DE BABEL, JORGE LUIS BORGES

Perché la ruota giri, perché la vita viva, ci vogliono le impurezze, e le impurezze delle impurezze: anche nel terreno, come è noto, se ha da essere fertile. Ci vuole il dissenso, il diverso, il grano di sale e di senape.

IL SISTEMA PERIODICO, PRIMO LEVI

Chapter 8

Research questions and methodology

In Part I and II, we have described the features that set Wikidata apart from prior projects in the fields of open knowledge collaboration and of the the Semantic Web. Wikidata’s large user pool has created in a few years a large size knowledge graph following a completely bottom-up approach, whereby prior analogue projects relied on the contribution of knowledge engineering experts. Because of that, some researchers have argued that Wikidata may represent a novel type of collaborative platform ([Müller-Birn et al., 2015](#)). Furthermore, Knowledge graphs such as Wikidata are crucial resources for AI applications. Nevertheless, research has devoted little attention to the collaborative processes occurring within the Wikidata community and to how these processes affect the quality of its outcome, i.e. the data in the graph. Our work builds upon and expands on prior knowledge in the online collaboration and knowledge representation fields, posing the following **overarching research question**:

How does the socio-technical fabric of Wikidata influence the quality of its data?

Investigating the socio-technical processes behind the production of the data will shed a light on the issues affecting the data itself ([Strohmaier et al., 2013](#)). Moreover, our study will contribute to understanding whether Wikidata actually represents a different paradigm of online collaboration, as suggested by some. Additionally, gaining insights into which community processes determine quality can contribute to design appropriate tools to improve them and address quality issues.

We address the overarching question by looking specifically at three aspects: *i.* the quality of references, as a function of their author type, i.e. human or bot; *ii.* the influence of group composition on Item quality; *iii.* the influence of Wikidata emerging user roles on the quality of its ontology. Corresponding to these three aspects, we pose the following research questions:

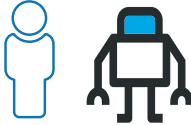
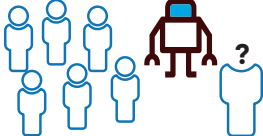
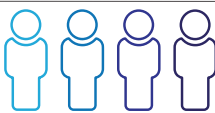
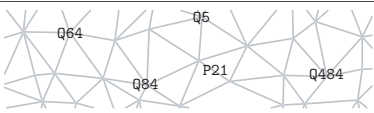
	Community features	Quality of
RQ1	 Humans vs bots	$\underline{Qx} :: \underline{Py} :: \underline{Qz}$ $\rightarrow \underline{Pj} :: \underline{Qk}$ Statements (references)
RQ2	 Groups of humans, bots, and anonymous	\underline{Qx} Items
RQ3	 Profiles of registered editors	 Sets of Items and Properties (<i>ontology</i>)

FIGURE 8.1: Community features and aspects of quality addressed by each research question

RQ1 How do references added by bots and by humans compare with respect to their quality?

RQ2 To what extent does editor group diversity affect Item quality in Wikidata?

RQ3 What features of editing roles affect the quality of the Wikidata ontology?

This chapter builds upon the background literature described in Parts I and II and provides further details about the research questions outlined above. Moreover, it describes the methods followed to address each question.

8.1 RQ1: Quality of human and bot edits

The community of Wikidata is entirely responsible for adding and maintaining data. Activity on the system is carried out either by people, registered or anonymous, or by pieces of software, called bots. These have a substantial influence on the quality of the graph. Although their share of contributions has declined since the first years of Wikidata, bots still perform the majority of revisions and are able to change quickly the shape of the graph by making large numbers of edits in a very short time (Chapter 6). The quality of bot contributions to Wikidata has not been investigated as of yet. RQ1 seeks to fill this gap with regard to one of the fields that constitute Wikidata statements, i.e. references. The possibility to add these, specifying the provenance of a single piece of information, is a characterising feature of Wikidata. In spite of that, no study has evaluated their quality so far. Furthermore, whereas references are added by bots and

human alike, we know from what discussed in Chapter 6 that editors work differently from bots. These generally import statements in batches from other knowledge resources and add a link to those as a reference. On the other hands, humans show a range of behaviours, which span from adding single statements, to using semi-automated tools. It is thus important to understand the effects of these different types of contributors on reference quality.

In order to answer RQ1 we restricted our scope on external references, i.e. those linking to resources outside Wikidata by means of property P854, as these provide a direct indication of the sources Wikidata is derived from. This is important to understand how diverse and reliable the knowledge in Wikidata is. We defined the quality of references in terms of relevance and authoritativeness (Chapter 7). In addition to gauging reference quality, we wanted our approach to be applicable on a large scale to the whole of Wikidata. This in order to be potentially utilised as a tool to monitor quality on the platform and predict problematic references over the whole of Wikidata. To that end, we evaluated the extent to which non-relevant and non-authoritative references can be predicted by our approach. Hence, we broke down RQ1 into two sub-questions:

RQ1.a How do references added by bot and human editors compare with respect to their relevance and authoritativeness?

RQ1.b To what extent can non-relevant and non-authoritative references be predicted in Wikidata?

The contribution of the work carried out to address RQ1 is threefold. First, it is the first evaluation of Wikidata references so far. Second, it proposes an approach to evaluate Wikidata references on a large scale. Third, it helps understand the influence of human-made and automated contributions on the quality of the graph and the types of errors they introduce.

8.1.1 RQ1: Methods

In order to address the two sub-questions into which we have broken down RQ1, we developed an approach that evaluates Wikidata references in terms of relevance and authoritativeness. Whereas RQ1.a aims to compare references contributed by bots and by people, the objective of RQ1.b is to gauge the performance of a large-scale scale evaluation approach of Wikidata provenance. In order to tackle these two aims, we adopted a two-staged approach relying on two complementary methods: microtask crowdsourcing and machine learning. Because of the advantages outlined above, we performed a crowdsourced evaluation of references, which was used to train a machine learning model to predict their quality. Machine learning can be easily applied on a large-scale and is

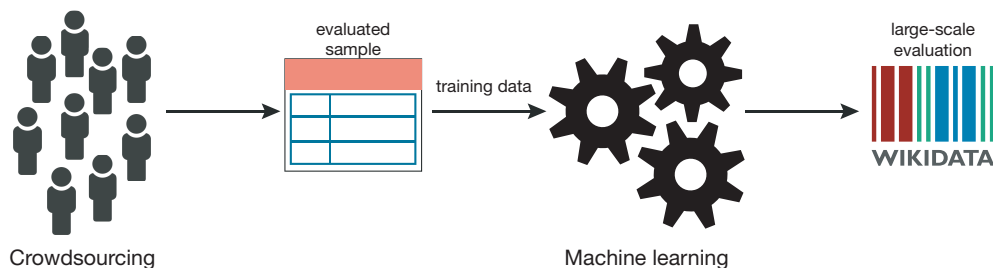


FIGURE 8.2: Pipeline of the two-stage approach adopted to address RQ1.a and RQ1.b

virtually costless. This two-stage approach is illustrated in Figure 8.2. Further details are provided in Chapter 9.

8.2 RQ2: Group composition and data quality

People of different backgrounds, skills, and perspectives put their efforts together to build Wikidata’s knowledge. Editing activity levels widely vary among editors, as we have seen in Chapter 6. Users who joined earlier are likely to be perform more revisions and take part more frequently in community discussions (Section 6.2.4.1). Furthermore, we observed in [Piscopo et al. \(2017c\)](#) that editors’ behaviour may change along their activity lifespan. Experienced editors may focus on different types of Items than novices and use semi-automated tools to perform revisions at higher rates that would be possible through the web interface. The effects of group composition on quality in online collaborative systems has been covered by prior research. However, this aspect has not yet been investigated with regard to Wikidata. With **RQ2**, we want to find the ‘right mix’ of users that leads to good quality in Wikidata. We scope out Items to evaluate quality. Items are the building blocks of Wikidata (Chapter 5) and are recognisable as unitary entities around which users may coalesce in groups.

We consider two aspects of the relationship between group composition and Item quality. First, we analyse the influence of the share of contributions of bots, registered, and anonymous human users. Second, we investigate the effects of the distribution of two group features: length of activity and task knowledge, i.e. tenure and interest diversity. Previous work has shown that these two variables characterise a large part of the variation in Wikidata editors’ behaviour ([Piscopo et al., 2017c](#)). Moreover, we look at the influence of different proportions of bot and human contributions on an Item quality—we have noted in the prior sections the importance of automated editing in Wikidata.

The contribution of RQ2 is twofold. First, it expands previous knowledge about the effects of group composition on quality in online knowledge collaborations, by gaining insights about these in a system, Wikidata, which presents unique characteristics in the field. Second, the findings of our analysis may be used in the future to inform tools to

monitor the composition of editor groups contributing to Item and involve other users, if the group composition is likely to lead to low-quality outcomes.

8.2.1 RQ2: Methods

In order to answer RQ2, we performed a regression analysis, Ordinal Logistic Regression (OLR), which was suitable to the grading scale used to measure Item quality. OLR takes into account the ordering of discrete response variables, such as the Item quality labels used in this study, compared to other models which are either suitable to binary responses (standard logistic regression) or do not make any assumptions about the ordering of outcome discrete variables (multinomial logistic regression) (Bender and Grouven, 1997). OLR splits the distribution of the data corresponding to each rank in the response variable. It relies on the assumptions that independent variables have the same effect across different responses (*proportional odds* assumption) (Brant, 1990). The ordering of these is modelled by considering cumulative probabilities for all different response categories, rather than by single category. As a consequence, the output of OLR provides an intercept value for each threshold between categories in the outcome variable.

The regression analysis predicted the quality of an Item, given a number of independent variables describing the proportion of human and bot contributions and the level of tenure and interesting diversity in the group of editors that ever performed a revision on that Item. We formulated five hypotheses about the relationships of the variables taken into consideration with Item quality. Further details about the regression approach, the variables used, and the hypotheses are provided in Chapter 10.

8.3 RQ3: User profiles and ontology quality

Wikidata editors show different behaviours, both for what concerns the type and the volume of tasks they carry out. These differences have been discussed in Chapter 6. Numerous users perform only a small number of revisions, whilst a minority carries out in the order of thousands of edits. Discussion pages are used by Wikidata community members to communicate among them and to reach consensus about a range of matters. Only a limited number of users has ever been active on these pages. Similar variations have been observed with respect to the object of the revisions made by users. The same core-periphery distribution of edit numbers noted for the whole graph applies when looking at Properties only. Furthermore, prior studies, both quantitative (Müller-Birn et al., 2015) and qualitative (Piscopo et al., 2017c), have shown that a part of the community focuses on creating and maintaining the structure of Wikidata’s knowledge, i.e. the ontology.

As we have seen in chapters 2 and 3, emergent user roles, or profiles, have been identified by analysing activity patterns in online communities and collaborative ontology development projects. Modelling user roles is important to understand participation and production processes across online collaboration platforms (Preece and Shneiderman, 2009). Furthermore, they can be used to improve the design of tools to support editors (Falconer et al., 2011).

RQ3 looks at Wikidata under a collaborative ontology engineering perspective. It aims to identify editor roles in Wikidata and understand their influence on ontology quality. A high-quality ontology is key to enable the discovery of information in Wikidata, e.g. a correct taxonomy is necessary to find all instances of a determined class. This research question links socio-technical processes and their outcome. First, it performs an evaluation of the Wikidata ontology over time; second, it seeks to identify Wikidata user profiles emerging from activity patterns not analysed by prior literature; and third, it investigates how users in each of the profiles found influence the quality of the ontology. Addressing RQ3 contributes to the understanding of collaborative processes in large-scale knowledge engineering projects. Furthermore, the approaches devised to tackle RQ3 can be used to support the diagnosis of quality issues and design appropriate solutions.

8.3.1 RQ3: Methods

Research Question 3 investigates roles emerging from user activity patterns and the influence of these on the quality of the ontology. In order to address all the aspects related to this question, we broke it down in three studies:

Study 1. We defined a suitable quality framework for the Wikidata ontology and subsequently applied it to perform an evaluation;

Study 2. We determined user profiles on the basis of their editing patterns;

Study 3. We linked the findings about user profiles to ontology quality evaluation results.

8.3.1.1 Study 1: Ontology evaluation

The choice of an ontology evaluation approach is dependent on context, which includes the purpose of the assessment and the available data (Brank et al., 2005). RQ3 aims to gain a quantitative understanding of the effects of different editing patterns on the quality of the Wikidata ontology. Therefore, we sought an approach that

- R1** considers primarily factors that editors could potentially influence (as opposed to externalities around the use of the ontology by developers, its suitability for a particular task, etc.);
- R2** is able to assess the ontology over time to observe its evolution;
- R3** uses only the ontology for the assessment and does not require additional task-based evaluations (a counterexample would be aspects such as completeness or coverage, which need to take into account a reference model or a gold standard);
- R4** includes indicators that could be implemented unambiguously and be computed automatically (for example, aspects such as the understandability of an ontology can be assessed in various ways).

We followed an ontology validation approach, which investigates whether the ontology is fit for purpose (see Chapter 11). The processes used by the Wikidata community are defined and assessed by the community itself, according to their principles and models of governance, hence an ontology verification approach would not have been suitable to our case. Moreover, our analysis covered structural aspects of quality, which are influenced directly by the activity of Wikidata editors (R1), can be observed over time without requiring external tools (R2 and R3), and can be evaluated by means of metrics computed automatically (R4). Other aspects, e.g. vocabulary, syntax, context are related to applications using Wikidata or to its data model and as such not connected to editing activities. Semantics is an aspect that could possibly meet the requirements set. However, state-of-the-art tools to check e.g. the consistency of the ontology may not work, given its large size.

We surveyed the extensive research around structural ontology metrics to inform the design of our framework. In order to identify relevant papers, we crossed the results from queries to widely used academic literature search engines (i.e. Google Scholar and Web of Science) with the references found in a number of ontology evaluation surveys (Brank et al., 2005; Navarro et al., 2010; Hlomani and Stacey, 2014; Vrandečić, 2010). We used the following keywords: ‘ontology metrics’, ‘ontology evaluation framework’, and ‘ontology evaluation’. From the results, we selected only papers including primarily structural metrics. We finally evaluated the degree to which each metric within the frameworks in our selection met the requirements set above and, if suitable, added it to our framework. Subsequently, we applied the selected metrics to evaluate the Wikidata ontology at monthly intervals since the creation of the property P279, in March 2013.

8.3.1.2 Study 2: User roles

In order to identify emergent user roles, we clustered editors according to a number of features, described in Chapter 11, using the k -means algorithm. The choice of features

was informed by prior studies about community dynamics in Wikidata and other platforms. We assumed that users' activity patterns may vary over time. Therefore, we divided our datasets into monthly timeframes and, for each timeframe i , we created an activity vector for each user u . Since the number of contributors may change across timeframes, the total number of vectors was equal to $\sum_{i=1}^n u_i$. All registered human users were included. We used the gap statistic to estimate the most suitable number of clusters.

8.3.1.3 Study 3: From user profiles to quality

We formulated two research hypotheses linking user role features to the ontology quality metrics based on the findings from the first two studies. We looked at this relation in terms of how the activity of a determined user type, specifically leaders, results in changes in various ontology metrics. Our approach follows Kittur and Kraut (2008) and consists in applying a lagged multiple regression model to predict changes in an ontology metric considered between two points in time $metricT_n - metricT_{n-1}$, holding it constant at $metricT_{n-1}$. This approach has the advantage to control for regression towards the mean and to remove the influence of $metricT_{n-1}$ on the relation between predictors and the dependent variable (Kittur and Kraut, 2008).

8.4 Data

The Wikimedia Foundation releases dumps containing the history of each page within Wikidata. Everything within the platform, including items and properties, has a corresponding web page, which is assigned to a particular namespace. For each version of a page, called *revision*, several metadata fields are available: the revision id, the parent revision id, the name and identifier of the user responsible for the revision, the timestamp of the revision, and a comment. All variables available with each revision are shown in Figure 8.3. In order to make the dumps processable for our experiments, we extracted the metadata related to each revision and the data about every statement, reference, or qualifier revision into a PostgreSQL database. Deleted statements were appropriately tagged. Table 8.1 shows the outcome of this extraction. The ontology graph used to address RQ3 was built on the basis of the P31 (instance of) and P279 (subclass of) Properties. Besides revision data, we extracted from various sources information regarding Wikidata users. Administrators and bots were identified through the lists of users by group made available by the Wikimedia Foundation, cross-checking these with data extracted from the *Request for Permission* pages in Wikidata. Edits made through semi-automated tools were identified by looking at the comments and tags attached to each revision (Sarasua et al., 2019).

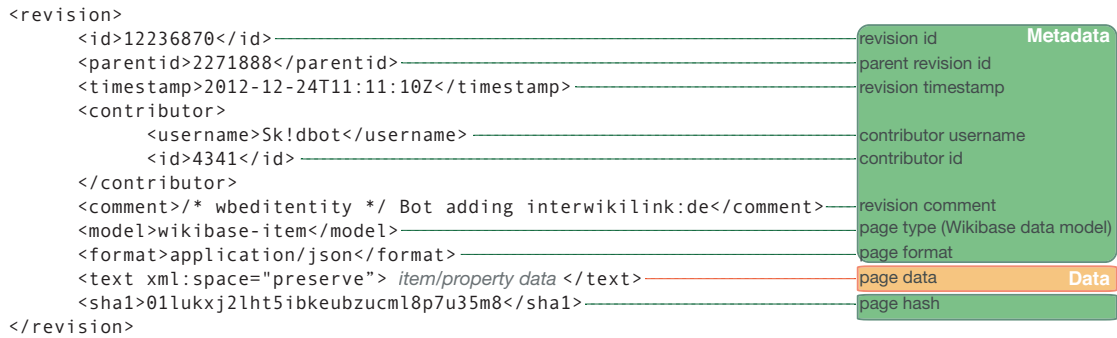


FIGURE 8.3: Variables available from Wikidata historical dumps. Please note that everything in Wikidata is a web page; the metadata provided with each revision specify whether it is e.g. an item/property, a community page, etc. and its format.

Table	Revision history	Statement history	Reference history	Qualifier history
Columns	revision id	revision id	revision id	revision id
	parent id	statement id	statement id	statement id
	page id	statement property	reference id	qualifier id
	timestamp	statement value	reference	qualifier
			property	property
	comment		reference value	qualifier value
	semi-automated tool			

TABLE 8.1: Wikidata history database tables. *page id* may be an Item or Property QID or a page title, depending on the page type. The semi-automated tool is a boolean variable we created, which tells whether an edit has been made through such tools.

Rome was not built in a day, nor was this thesis. The studies reported in the current have been carried out along a time span of approximately three years. Because of that, the experiments reported rely on datasets covering different timespans. The analysis in Chapter 9 (RQ1) employs data up to 1st October 2016; Chapter 10 utilises data updated to 1st April 2017; and Chapter 11 relies on datasets updated to 1st October 2017. We discuss the limitations of that in Chapter 11. All the code and part of the datasets produced in this thesis are available at https://github.com/Aliossandro/structuring_the_world_knowledge_phd_thesis.

8.5 Summary

Our work examines quality as a function of various aspects of Wikidata community fabric. We structured our analysis along three research questions.



RQ1 looks at the quality of a component of Wikidata statements—i.e. reference—as a function of the type of author, i.e. human or bot, that authored them. We evaluated external references following a two-staged approach, which (*i.*) allows us to gain insights about their quality and (*ii.*) is suitable to perform a large-scale evaluation.

RQ2 explores how editors mingle into groups and what the effects on groups' outcomes are. In particular, we look at the relation between group composition and Item quality. In this step, RQ2 expands the horizon of RQ1 with respect to the editors' features analysed, by taking into account different characteristics of human users, namely their tenure and activity focus.

RQ3 examines the influence of roles emerging from editor activity patterns on the quality of the Wikidata ontology, being the first study so far to embrace both features concerning revisions on the graph and user interactions within the community.

Chapter 9

Back to the sources: quality of human and bot edits

	Community features	Quality of
RQ1	  Humans vs bots	$\underline{Qx} :: \underline{Py} :: \underline{Qz}$ $\rightarrow \underline{Pj} :: \underline{Qk}$ Statements (references)

This chapter addresses RQ1, by examining and comparing the quality of bot and human contributions for what concerns references. This research question was broken down into two sub-questions:

RQ1.a How do references added by bot and human editors compare with respect to their relevance and authoritativeness?

RQ1.b To what extent can non-relevant and non-authoritative references be predicted in Wikidata?

We define quality following the community-developed policy concerning references in Wikidata. This has been discussed in Chapter 7 and is recapitulated below in Section 9.1. Bots and humans present different editing behaviours, as seen in Chapter 6. Bots generally commit massive number of revisions at a very large rate, whereas the activity of humans may vary according to the tools they use to perform their edits. We dig down into what these differences entail in what concerns the creation and modification of references in Section 9.4, where we characterise the contributions of each of these two user groups, e.g. looking at the amount of reference edits and to whether they add sources contextually to the creation of a statement or add them to previously created

ones. Section 9.3 provides further details about the approach followed to address RQ1.a (i.e. comparing relevance and authoritativeness of references added by human editors and bots) and RQ1.b (i.e. the evaluation of an automated approach to predict quality of Wikidata references). The remainder of the chapter presents and discusses the results of the different parts of the experiment.

9.1 Wikidata reference policy

We described the Wikidata policy on references in Section 7.2.2. In this section we summarise the salient points of this policy, integrating them with a few details. The content below is based on Wikidata (2018f), Wikidata (2018g), Wikidata (2018j), and Wikipedia (2018a).

When is a reference required? All statements must be supported by a source, with three notable exceptions (Wikidata, 2018g): (*i.*) statements expressing common knowledge, e.g. Ada Lovelace – instance of – human; (*ii.*) statements that refer to external sources, such as those linking to external identifiers like Ada Lovelace – GND ID – 119232022; (*iii.*) statements referring to an item that is a source itself, this is the case of books, films, and LP records, such as Cosmicomics – author – Italo Calvino. The list of the properties for which a reference is not needed can be found in Appendix A.

Types of references. Different Properties can be used to source a statement in Wikidata. The main ones are P248 and P854. P248 (stated in) connects a statements to an item that supports it. P854 (reference URL) links to an external resource. Other Properties may be used, e.g. P143 (imported from Wikimedia project). However, references of this type are not deemed to be reliable and are supposed to be replaced by others using either P248 or P854 (Wikidata, 2018f).

Reference quality. The Wikidata policy requires sources to be relevant and authoritative. Relevance refers to *whether a source supports a determined statement, either directly or by entailing the information conveyed by the latter*. Authoritative sources must be ‘deemed trustworthy, up-to-date, and free of bias’ (Wikidata, 2018j). Recognising that this definition is fairly subjective, Wikidata identifies some features that are likely to be associated to authoritative sources, namely their type of author and publisher. We described more in detail these in Chapter 7. Here we provide a table with the combination of author and publisher that are associated to authoritative references (Table 9.1).

Publisher \ Author	Individual	Organisation	Collective
Academic and research institution	✓	✓	✗
Academic publisher	✓	✓	✗
Other academic	✓	✓	✗
Government agency	✓	✓	✗
Vendor or e-commerce company	✗	✗	✗
Political or religious organisation	✗	✗	✗
Cultural institution	✓	✓	✗
Other type of company	✓	✓	✗
Traditional news and media	✓	✓	✗
Non-traditional news and media	✓	✗	✗
Self-published source	✗	✗	✗

TABLE 9.1: Authoritativeness of sources (ticks indicate authoritative)

9.2 Related work

Relevance and authoritativeness are variously defined in the literature. As regards authoritativeness, the definition is less straightforward and refers to aspects that are often mentioned by the literature on the topic, i.e. trustworthiness, currency, and objectivity. This section discusses previous work of relevance regarding approaches for the evaluation of sources in the web. Research on information quality in the web generally connects authoritativeness to credibility and trustworthiness (see Section 7.2.2), which means that we took into consideration also evaluation methods originally developed for those contexts. We discuss in the following approaches to evaluate relevance and authoritativeness, which we divided into automated and manual, according to the degree of human participation they involve.

9.2.1 Automated approaches

The field of citation recommendation provides a large body of research about reference evaluation. DeFacto (Lehmann et al., 2012) is a system which finds relevant and trustworthy pages to enrich knowledge graphs with provenance information. For each triple to be enriched with a source, the algorithm extracts a number of keywords, which are then used to issue a query. The pages resulting from the query are analysed through a supervised machine learning algorithm, combined to NLP techniques, to produce a confidence score of the probability that the page contains the information from a triple. Additionally, DeFacto evaluates the trustworthiness of resulting pages by taking into consideration how much the topic of a page is covered on the web and in the search results, and the PageRank (Page et al., 1999) of the page. In spite of its potential as a system for discovering new sources for Wikidata and assessing existing ones, some

drawbacks hamper DeFacto’s application to address RQ1. First, its measure of trustworthiness would need to be tested, in order to understand whether and how it matches the definition of authoritativeness used by Wikidata. Second, after a preliminary analysis of Wikidata external sources, we found that their format varies greatly, e.g. tables, free text, files to be downloaded, etc., which would likely make the framework used by DeFacto inadequate for many of them.

[Fetahu et al. \(2016\)](#) evaluate a supervised machine learning approach to find appropriate news citations on the web for statements in Wikipedia articles requiring to be supported by a source. The algorithm first classifies statements according to the type of citation they require. Subsequently, it issues a query to a search engine and uses basic textual entailment and topic modelling features to select the resulting pages with relevant citations. These must come from authoritative sources and entail the statement they support. The measure of authoritativeness is generated by computing the probability of a type, i.e. the subject of the statements such as *politician* or *athlete*, and an article section to use a determined web domain. Their model performs variously depending on the article type. Authoritativeness and entailment of the source match the requirements set by the Wikidata verifiability policy. However, compared to [Fetahu et al. \(2016\)](#), RQ1 addresses the evaluation of Wikidata external references, rather than the discovery of new ones. We leave the evaluation of automated methods to discover new sources for Wikidata statements to future work.

Quantitative approaches to measure authoritativeness typically rely on the analysis of inter-links between pages. The PageRank algorithm is a well-known example of this approach ([Page et al., 1999](#)). [Kleinberg \(1999\)](#) presents an algorithm that generates authority measures using inter-links within sub-graphs of the web. These approaches diverge from that followed by Wikidata, which focuses instead on identifying principles to define credible and authoritative web sources.

9.2.2 Manual approaches

Other approaches identify a number of criteria to help users evaluate credibility on the web. Normative criteria are designed to provide advice to users in discerning credible sources, whereas descriptive ones aim instead to understand user behaviour in information seeking practices ([Pattanaphanchai et al., 2013](#)). These models may assess credibility on various levels and include different indicators, which can comprise relevance and authoritativeness, among others ([Wathen and Burkell, 2002](#)). The models are often applied to guide users in assessing web content through checklist approaches, i.e. sets of questions that can be answered in order to assess the credibility of an information source ([Metzger et al., 2010](#)). The pieces of information required for this type of evaluation, e.g. the author’s qualification or credentials or the last time a page was updated ([Flanagin and Metzger, 2007](#)), may be difficult to gather though. However, the

use of proxies, i.e. measures of features that characterise trustworthy or credible sources, may be suitable to our purpose, namely for what concerns authoritativeness, a highly subjective concept whose unbiased evaluation may be problematic (Metzger et al., 2010). Prior studies suggest that credibility, which is very close to authoritativeness, is consistently assessed under a positive bias by web users (Kakol et al., 2013). An evaluation approach relying on proxy features has been experimented with success in Wikipedia by Ford et al. (2013), who evaluated sources by classifying their type, author, and publisher. Since Wikidata verifiability policy explicitly refers to Wikipedia, we based our authoritativeness evaluation on Ford et al. (2013) to devise publisher and author types. We assessed only these two aspects because Wikidata external references already include only one type of sources, i.e. web pages, which may be authoritative depending on their author and publisher. Moreover, we followed Wikidata and Wikipedia verifiability policies to identify which combinations of author and publisher types corresponded to authoritative sources.

9.2.3 Crowdsourcing

Crowdsourcing is a problem solving approach in which tasks are outsourced to a large group of people through an open call (Simperl, 2015). Tasks may be carried out either in exchange of a monetary reward or for free. This technique has become a popular means to create scientific resources and perform data curation tasks, as its results may reach higher levels of quality than fully-automated approaches, whilst being still being cost-efficient and timely if correctly planned (Kittur et al., 2008). It is based on the principle that several cheap non-expert judgements can reach a performance comparable to that of few expensive experts (Eickhoff and de Vries, 2011). Crowdsourcing has been successfully applied to several types of task, including the assessment of Web sources' relevance. Alonso et al. (2008) outline the strengths of a crowdsourced approach to assess the relevance of web pages with regard to a determined subject. This technique has faster completion times, compared to experiments involving experts or online surveys, and low cost, whilst yielding high quality results. Additionally, it is flexible, in the sense that it can be used for a large range of tasks, as Simperl (2015) points out.

The characteristics of crowdsourcing make it suitable to understand the quality of external references of Wikidata with respect to the above described criteria of relevance and authoritativeness. This approach has previously shown to be accurate and efficient in similar contexts (e.g. in Alonso et al. (2008)).

9.3 Approach

We evaluated only external references, which were 13% of the total. With the purpose of addressing RQ1.a, we investigated (*i.*) the extent to which Wikidata external references

are **relevant**; (ii.) the extent to which Wikidata external references are **authoritative**; In addition, (iii.) we performed an evaluation of the extent to which non-relevant and non-authoritative Wikidata references can be predicted on a large scale (RQ1.b).

9.3.1 Source Evaluation

We designed three crowdsourcing tasks to assess reference quality, which were carried out on CrowdFlower ¹. All tasks included one type of microtask, except one, which included two. In order to increase the clarity of microtasks, we refined their design by launching test runs of small samples (between 50–100) of references to be evaluated. User behaviour (number of missed questions and completion time) was observed to understand microtask clarity.

Relevance

The first task (**T1**) was designed to assess relevance by asking users to find the pieces of information composing a statement within its source. Each microtask in T1 evaluated a reference, i.e. a statement with its attached source. In order to decrease the cognitive burden on workers we structured microtasks along three questions, one for each element of a statement (subject, property, object). For each of these, we asked whether the source provided information about it. Users were prompted the successive questions only if they responded positively to the prior one (e.g. we asked about the Property of a statement only if evidence about its subject was found in the source). English labels were shown for every part of each statement, instead of their URIs. In the case of pages not working or requiring a log in, or for pages not in English, users could select the appropriate responses. Figure 9.1 illustrates an example of T1 microtask.

Authoritativeness

A similar concept to authoritativeness—credibility—is consistently assessed under a positive bias by web users (Kakol et al., 2013). Hence, instead of directly questioning users about the authoritativeness of a source, which would have likely given overly subjective responses, we tested whether sources matched the types specified by Wikidata policy and asked the crowd to classify them, similar to the approach followed for Wikipedia’s sources by Ford et al. (2013). Namely, we determined sources’ author and publisher types.

Author type was assessed in **T2**. Microtasks in T2 (Figure 9.2) asked users to indicate the most appropriate author type for a source. The typology of authors was based upon

¹<https://www.crowdfunder.com/>. The platform has changed its name into Figure Eight since the time of our experiment. See <https://www.figure-eight.com/>.

Please read the following statement

The Water Diviner → cast member → Megan Gale

and examine carefully this page <http://www.imdb.com/title/tt3007512/fullcredits>

Does the page contain information about 'The Water Diviner'?

- ☒ Yes
- ☐ No
- ☐ The page is not in English
- ☐ The link does not work/requires to log in

● Please examine the source carefully, including any document or table associated.

Does the page contain information about 'cast member' that is related to 'The Water Diviner'?

Please note that 'cast member' may be expressed also as 'film starring,actor,actress,starring'.

- ☒ Yes
- ☐ No

Does the page contain the information 'Megan Gale' in relation to 'The Water Diviner' and 'cast member'?

- ☐ Yes
- ☒ No

Please specify.

- ☐ There is a value related to 'The Water Diviner' and 'cast member', but it is different from 'Megan Gale'
- ☐ I couldn't find any information related to 'The Water Diviner' and 'cast member'

FIGURE 9.1: A microtask from T1

Please examine carefully the following web page

<http://americanart.si.edu/collections/search/artwork/?id=21513>

Which of the following types best describes the author of the page? (required)

- ☐ Individual
- ☒ Organisation
- ☐ Collective
- ☐ The page is not in English
- ☐ The link does not work/requires to log in

If you are unsure about your answer, please look again at the Instructions above or at the [example page](#).

FIGURE 9.2: A microtask from T2

what discussed above (Chapter 7) and included Individual, Organisation, and Collective. Users were shown only the source, rather than the whole reference. Therefore, T2 evaluated only unique web pages, meaning that a lower number of microtasks than T1 was required.

Task 3 (T3) evaluated publisher type. We assumed that pages belonging to the same domain had the same publisher. Hence, we collected judgements for unique domains, rather than for each single reference. **T3.A** (Figure 9.3) included only higher-level types of publisher: *academic and scientific organisations*, *companies and organisations*, *government institutions*, *news and media*, and *self-published sources*. It consisted of a multiple choice question to select the most appropriate type of publisher. **T3.B** (Figure 9.4) collected judgements related to the sub-types in Table 7.3. In T3.B users were asked whether the publisher type obtained from the previous task was appropriate for the source, in order to test contributors' performance and verify the results of T3.A.

Please examine carefully the following web page
<http://addons.mozilla.org>

Which of the following types best describes the publisher of the page?

- ☐ Academic or scientific organisation
- ☐ Governmental institution, national or supranational
- ☐ Company or organisation
- ☐ News and media outlet
- ☐ Self-published source
- ☐ Other
- ☐ The page is not in English
- ☐ The link does not work/requires to log in

❗ By publisher we mean the organisation or group who is responsible, i.e. hosts and maintains, the content of a page.

If you are unsure about your answer, please look again at the Instructions above or at the [example](#) page.

FIGURE 9.3: A microtask from T3.A

Please examine carefully the following web page
<http://americanart.si.edu>

Is 'company or organisation' the most appropriate type for the publisher of the page provided?

- ☒ Yes
- ☐ No
- ☐ The page is not in English
- ☐ The link does not work/requires to log in

Which type of company or organisation applies best?

- ☐ Vendor or e-commerce company
- ☐ Political or religious association or foundation
- ☐ Cultural institution, e.g. museum, library, or cultural heritage foundation
- ☐ Other

If you are unsure about your answer, please look again at the Instructions above or at the [example](#) page.

FIGURE 9.4: A microtask from T3.B

If users answered positively, they were asked to classify the sub-type of the source publisher. User pools of T3.A and T3.B were independent from each other. Appropriate options were given for pages not working, requiring to log in, or pages not in English in T2, T3.A, and T3.B—in these cases, no further evaluation was required.

Quality Assurance

Crowdsourcing is vulnerable to users who perform poorly due to lack of skills, malicious behaviour, or distraction ([Eickhoff and de Vries, 2013](#)). We adopted various strategies to tackle this issue. We added gold standard questions to tasks and excluded workers whose performance fell under a certain threshold, which we set to 80% in all tasks. Tasks were structured in pages, each containing a number of microtasks which varied depending on the task. Workers were first required to pass a test consisting of a page of test questions with an accuracy above or equal to the threshold set. Additionally, a test question

was included in each page of work. Users had to keep an accuracy above the minimum threshold throughout their contribution. We followed previous research regarding the experimental design of workers' qualification, granularity of task, and monetary rewards (see Table 9.2). Based on observations collected during test runs of the tasks, we accepted only workers with a previous accuracy rate of 85%—the highest allowed by CrowdFlower.²—to select highly performing users (Eickhoff and de Vries, 2013). Payments per microtask were determined according to (Snow et al., 2008). Correct answers were selected by majority voting over five assignments per microtask, following Acosta et al. (2013). Information on how to complete the task and links to clarifying examples³ were available on each page.

	T1	T2	T3.A	T3.B
Worker qualification	≥ 85%	≥ 85%	≥ 85%	≥ 85%
Granularity (microtasks per page)	10	8	8	8
Monetary reward (per microtasks)	\$0.08	\$0.06	\$0.05	\$0.05
Assignments	5	5	5	5
Min. worker accuracy	80%	80%	80%	80%

TABLE 9.2: Crowdsourcing experiment design

9.3.2 Automatic Evaluation Model

We used a machine learning classifier to identify not relevant or not authoritative sources. We trained a supervised algorithm for each outcome variable, using the labels obtained through the crowdsourcing experiment. Both relevance and authoritativeness models included the three types of features list below.

1. Features concerning the source itself:

URL reference uses. Number of times a URL has been used as a reference.

Domain reference uses. Number of times a domain has been used.

Source HTTP code. HTTP response code given by the source link.

2. Features related to the semantics of the statements the source is referred to:

Statement property. The property used in the statement.

Statement item. The item subject of the statement, represented as a vector of its structured components, i.e. labels and aliases were excluded.

Statement object. The object of the statement represented as a vector.

Subject parent class. Item parent class, i.e. object of property P279 (subclass of) or P31 (instance of).

Property parent class. Property parent class, i.e. object of P279 or P31.

Object parent class. Item parent class, i.e. object of P279 or P31.

²<http://crowdflowercommunity.tumblr.com/post/108559336035/new-performance-level-badge-requirements>

³Examples were provided for T2, T3.A, and T3.B: <https://wdref-author-evaluation.000webhostapp.com/>, <https://wdref-evaluation.000webhostapp.com/>.

3. Features editing activity on the statement:

Author type. Anonymous, bot, or registered human.

Author activity. Total number of revisions carried out by the reference creator, prior to adding it.

Author activity concerning references. Proportion between number of reference edits and total number of edits carried out by the author of the reference. Editors who are more active on references are more likely to add good sources.

The rationale for adding the first class of features was that more frequently used sources are more likely to be checked by several users and therefore to be trusted. Regarding statement semantics, we assumed that if a reference is good for a statement, it might be good for similar statements as well. Activity metrics were added as users with a larger number of edits may be more trustworthy, according to previous findings (Potthast et al., 2008). We included the same features in both models, as they could contribute to various extents to their accuracy.

We tested three algorithms that previously performed well in different tasks, Naive Bayes, SVM, and Random Forests. Models were trained using the Python library `scikit-learn` (Pedregosa et al., 2011). We used the appropriate options provided by that library to account for the imbalance in the outcome variables.

9.3.3 Evaluation data

For the purpose of evaluation, we extracted a sample of references from the Wikidata historic dumps, pre-processed as described in Chapter 8. The experiment was carried out using data updated to 1st October 2016. We selected all statements containing external references, excluding those pointing to a Wikimedia link and those not requiring any reference.

We counted a total of 30,959,834 unique references at the 1st October 2016. External references (P854) accounted for around 13.2% of the total, i.e. 4,056,576. Around 46% of these (1,855,487) pointed to two domains (`uniprot.org` and `ebi.ac.uk`) and were added by two bots, MicrobeBot⁴ and ProteinBoxBot⁵. Most of the references linking to these domains have been subsequently moved to qualifiers or claims using a specific database Property (e.g. P352), therefore we removed them from the sample. Furthermore, we left out all references referring to statements using Properties which do not require to be sourced according to the Wikidata policy (Wikidata, 2018f)⁶. After these

⁴<https://www.wikidata.org/wiki/User:MicrobeBot>, consulted on 4 December 2018.

⁵<https://www.wikidata.org/wiki/User:ProteinBoxBot>, consulted on 4 December 2018.

⁶The selection of Properties which do not required any reference was performed manually by the authors and it is provided in Appendix A. Although some of the Properties excluded may actually need a reference in some cases, we opted to leave them out anyway.

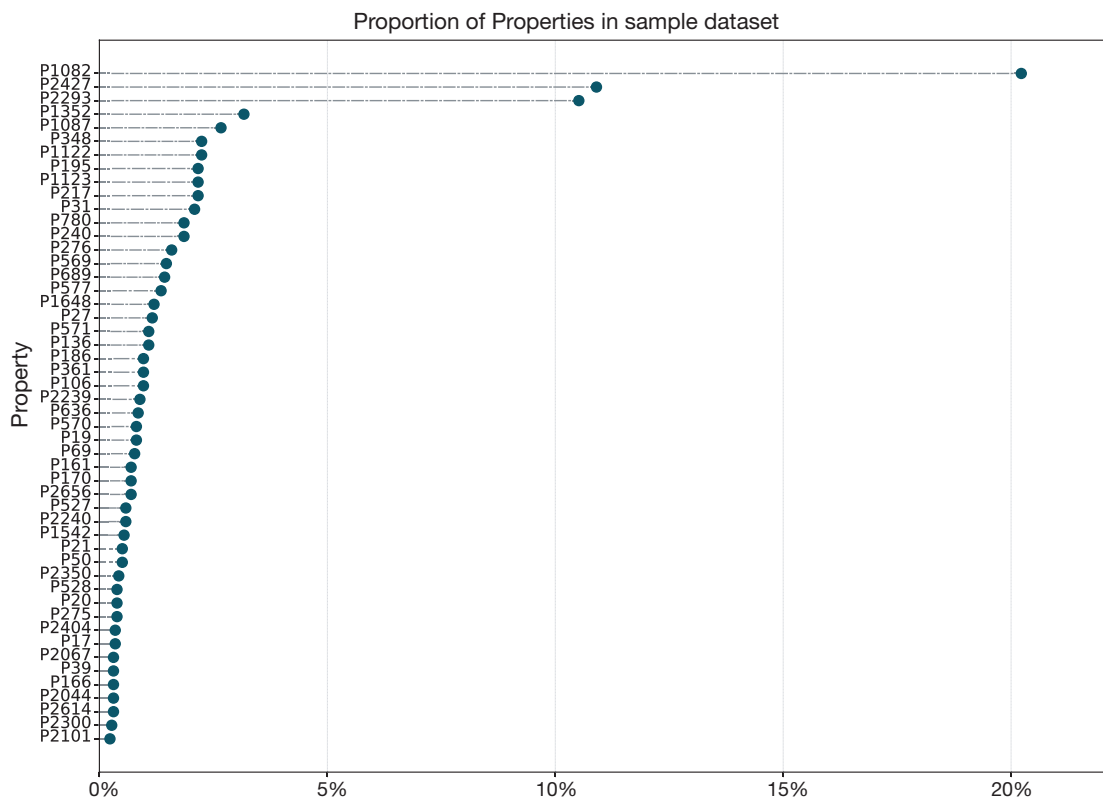


FIGURE 9.5: Occurrence of properties within the sample. The graph includes only properties with more than 5 occurrences to increase readability.

first steps, 1,429,993 references were left. Only references in English were selected; non-English references were identified by dropping all those whose source did not have an international top-level domain or one from an English-speaking country⁷. This left a total of 83,215 references, from which we sampled 2586 (99% confidence level, 2.5% margin of error; further details in Table 9.3.3). In order to reflect the different subject-object relations supported by references, we drew the sample so that it would proportionally include Properties according to their proportion in our dataset. 182 Properties in total were included. Figure 9.5 shows the proportion of the most common Properties within the sample.

We automatically tested the validity of each link by querying its HTTP code with the python library `requests`. Pages that returned a 404 code or timed out were flagged as not working and not submitted to the crowd. One link⁸, used in several references (512, 19.8%), redirected to another page which did not contain the data initially hosted by the link and was judged as not relevant. Two more links⁹ (282 uses, 11% of the total) pointed to csv files that were automatically checked. Both links were classified as relevant and not submitted to the crowd. Other pages belonged to research projects

⁷We kept the following top-level domains: tv, au, gov, com, net, org, info, edu, uk, mt, eu, ca, mil, wales, nz, ph, euweb, ie, id, info, ac, za, int, london, museum.

⁸<http://www.census.gov/popest/data/counties/totals/2013/files/C0-EST2013-Alldata.csv>

⁹https://figshare.com/articles/GRID_release_2015_12_14/2010108, https://figshare.com/articles/GRID_release_2016-05-31/3409414.

	Total instances	Total Items	Total statements	Total Properties	Total URLs	Total domains	Avg. domains per Property	Avg. edits per reference
All	2586	2372	2583	182	1674	345	3	1.03
Added by bots	1175	1108	1,175	30	486	38	3	1
Added by humans	1411	1269	1408	173	1189	325	2.7	1.2

TABLE 9.3: Sample characteristics. Humans include registered and anonymous users.

which explicitly stated the names of their authors. We labelled their author type as ‘individual’ and did not submit them for evaluation. After this filtering, the datasets submitted to the crowd included 1701 references (T1), 1178 unique URLs (T2), and 335 unique domains (T3.A and T3.B).

9.4 Reference editing activity: bots vs humans

References are seldom edited: the average number of revisions is 1.3 (median 1, interquartile range 1, 1; external references 1.2 revisions on average, median 1, interquartile range 1, 1), only 22% of them are ever modified after they are created (0.5% are edited more than five times). These percentages are lower for external references: only 15.2% of these are edited at least once after their creation and 0.03 more than five times. However, references added by bots are more edited on average (2.9 times vs. 1.5, median 2 vs. 1, interquartile range 1, 4 vs 1, 2). A *t-test* confirmed that the difference is significant ($p=0$). This might indicate issues with bot-added references, which therefore need to be subsequently revised by human editors.

Regarding edits volume, bots are unsurprisingly the most active authors of references by far, both external and internal (Figure 9.6). This is in line with what we observed in Chapter 6 concerning algorithmic editing activity in Wikidata—bots perform the majority of revisions of the platform. However, the ratio between references added by bots and by human editors is 9.2 to 1, whereas the ratio for all revisions is 2 to 1. According to the Wikidata policy (Wikidata, 2018a), bots must complement all statements they add with a link to the source they are imported from—content from Wikipedia uses P143. This is done in a fairly consistent way, as we observed that around 88% of references added by bots are created by the same bot that created the statement. This is suggested also by the median time between statement and reference creation (1 second) for statements created by bots (Figure 9.7). Human editors take longer to add references, with median=3s—the difference, tested by means of a Mann-Whitney *U* test is significant ($p < 0.05$)—and a much wider interquartile range (1 second-3.5 months), which leads to suppose that adding statements without a supporting source

is a much more common behaviour among human editors. We also analysed the time elapsed between the creation of a statement and of its related reference when this is added by someone else than the statement creator. When this is a bot, a much longer time is needed before a source is attached, with a median of 404.9 days against 136.8 for statements created by humans (Figure 9.7 on the right side).

References are seldom edited: the average number of revisions is 1.3 (median 1, interquartile range 1, 1; external references 1.2 revisions on average, median 1, interquartile range 1, 1), only 22% of them are ever modified after they are created (0.5% are edited more than five times). These percentages are lower for external references: only 15.2% of these are edited at least once after their creation and 0.03 more than five times. However, references added by bots are more edited on average (2.9 times vs. 1.5, median 2 vs. 1, interquartile range 1, 4 vs 1, 2). A *t-test* confirmed that the difference is significant ($p=0$). This might indicate issues with bot-added references, which therefore need to be subsequently revised by human editors. The remaining sections of this chapter provide further insights on whether this is the case.

9.5 Reference evaluation

This section presents the results of the experiment carried out to address RQ1.a and RQ1.b. First, we evaluate the crowdsourcing experiment, detailing each step of the process. Second, we provide the findings of the reference evaluation, breaking down the results by editor type and feature evaluated. Finally, we describe the performance of the machine learning algorithms tested.

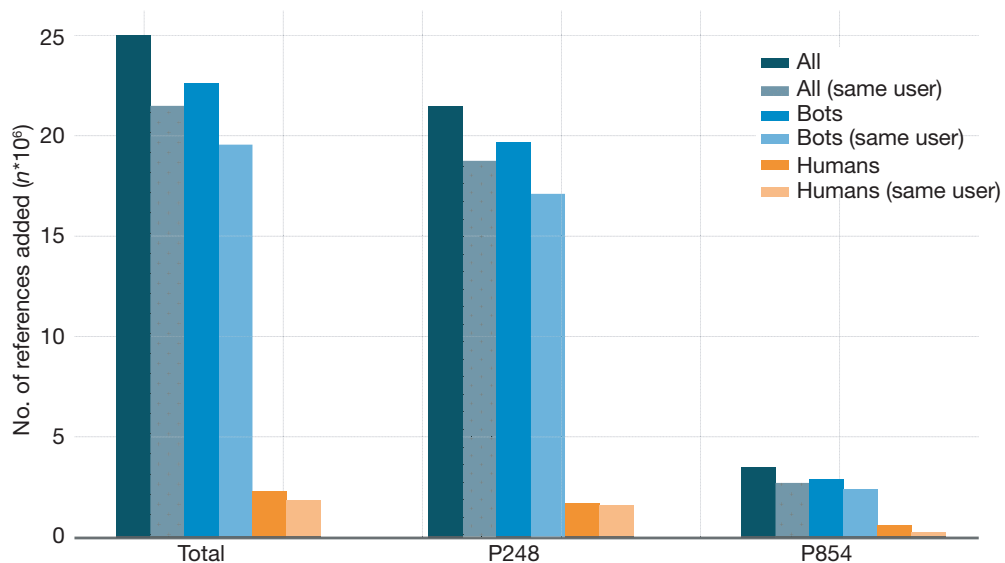


FIGURE 9.6: Number of sources added by type of user and reference. Lighter colours indicate sources added by the same author of the related statement.

9.5.1 Evaluation data

Crowdsourcing gold standard

A gold standard for each task was created by the author of this thesis in collaboration with a fellow researcher, manually labelling random samples of each of the datasets submitted to the crowd. The sizes of the annotated samples were determined to ensure that workers would not respond twice to the same question (sample size: T1:333; T2:116; T3.A and T3.B:67). Inter-rater agreement of gold standard questions (using Cohen's kappa) was between moderate and substantial for the four tasks: T1:0.447; T2:0.802; T3.A:0.587; T3.B:0.545. Divergent judgements were settled by mutual agreement. Furthermore, sources assessed in T1 had varying levels of difficulty. In some the information sought could be easily found, whereas others were very technical or contained long text. To better assess the crowd's performance, we labelled each reference in T1's gold standard as 'easy' or 'hard'. We found 239 easy and 94 hard references.

Machine learning data

We aimed to build binary classifiers to predict relevance and authoritative of sources. Hence, we converted the judgements collected into binary labels for each of these two

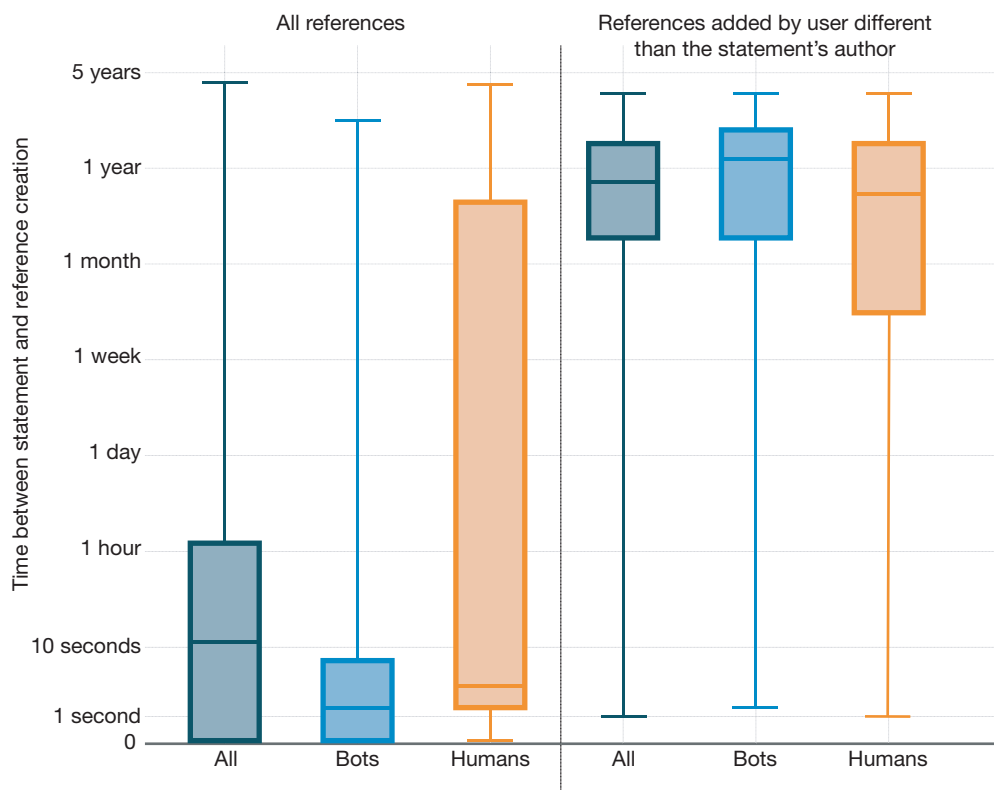


FIGURE 9.7: Time difference between creation of statements and addition of their related reference, by reference author. The scale has been adapted to increase readability.

outcome variables, i.e. relevant vs. non-relevant and authoritative vs. non-authoritative. We followed Wikidata and Wikipedia verifiability policies to identify the combinations of author and publisher types corresponding to authoritative sources (Table 9.1). Wikidata contemplates exceptions for sources generally considered as ‘bad’, e.g. self-published sources are acceptable in references regarding their author. For the purpose of analysis, we classified these types of sources as always not authoritative. We deemed not relevant, nor authoritative, sources with non-working links or that required log in as these were not accessible. We also excluded all references classified as not in English by crowdworkers. After this filtering, the dataset used to train the models had 2550 instances (1781 relevant vs. 769 non-relevant; 1610 authoritative vs. 940 non-authoritative).

9.5.2 Metrics

Crowdsourcing experiment

CrowdFlower provides a full report for each task, which includes every response, plus several details about workers, e.g. id, country, and their previous accuracy rate. We extracted from this data the metrics we used to evaluate the performance of crowdworkers. For each task, we measured the percentage of correct answers to test questions, inter-rater agreement (measured as Fleiss’ kappa (Acosta et al., 2013)), and completion time.

Predictive models

We evaluated the performance of the predictive models by comparing them to a baseline. For the relevance model, the baseline was generated by matching English labels of subject and object of a statement in the source text. Sources where both matched were labelled as relevant. In case of labels composed of several words, if any of them were found in a page, we considered that a match. For authoritativeness, a blacklist of deprecated domains has been compiled within the *primary sources tool* project (Tanon et al., 2016). This list is currently used to exclude non-authoritative sources, thus we judged it as a meaningful term of comparison for an approach assessing reference authoritativeness. We deemed not authoritative all sources whose domain was not included in this blacklist.

9.5.3 Crowdsourcing experiment evaluation

The accuracy of trusted workers, i.e. contributors whose accuracy did not drop under 80%, was around 90% and their responses had Fleiss’ kappa between 0.335 and 0.534, indicating fair to moderate agreement. These figures suggest that judgements collected had good quality (see Table 9.4).

	Microtasks	Total judgements	Trusted judgements	Total workers	Trusted workers	Fleiss' k	Trusted workers accuracy	Time	Cost
T1	1701	13,330	9671	457	218	0.335	90.4%	45h	\$858
T2	1178	14,340	9170	749	322	0.534	89.3%	90h	\$500
T3.A	345	4325	1950	322	60	0.435	89.9%	81h	\$116
T3.B	345	3622	2555	239	116	0.391	90.5%	24h	\$119

TABLE 9.4: Task statistics (includes test questions)

More than half of participants who worked on T1 were discarded due to a low accuracy rate. However, this was the task with the highest rate of microtasks completed per hour (37), i.e. the average number of microtasks successfully completed by the minimum number of workers (5) per hour. Furthermore, workers' accuracy was high on both easy (91.5%) and hard (89.7%) references.

T2 took longer to complete (90h), although not by microtasks/hour (13). The accuracy rate of all contributors to T2 was lower than T1 (72% vs. 75%). Task 3.A appeared to be the most difficult. The accuracy of its overall user pool (including trusted and non-trusted workers) was the lowest, with 66% of correct responses to test questions. Consequently, a high number of contributors were expelled from the task, leading to very long completion times. However, responses to T3.A had a moderate inter-rater agreement (0.435). 94.8% of the responses were confirmed by the first question of T3.B.

9.5.4 Relevance and authoritativeness evaluation

The next sections report the findings of the reference evaluation. The results presented include both references assessed through crowdsourcing and those previously evaluated by the researchers (see section 9.3.3).

The majority (68.9%) of sources evaluated in T1 were relevant. 7.5% were not working and only 1.4% of sources were found to not be in English, indicating that the approach followed to select only English-language pages likely had high precision.¹⁰ Of non-relevant sources, most of them (93.7%, 20.9% of the total) were not related to the subject of the statement. Please note that this includes redirected links pointing to a new, working page. Overall, human editors added more relevant references than bots (90% vs. 30.3%). In general, bot edits are associated with lower quality references: we found a moderate negative correlation between percentage of bot edits and relevance (-0.3). Evaluation results by type of user are shown in Figure 9.8.

¹⁰We make no assumptions about recall, though, therefore we might have missed some sources that were in English, but had different top-level domain than those selected.

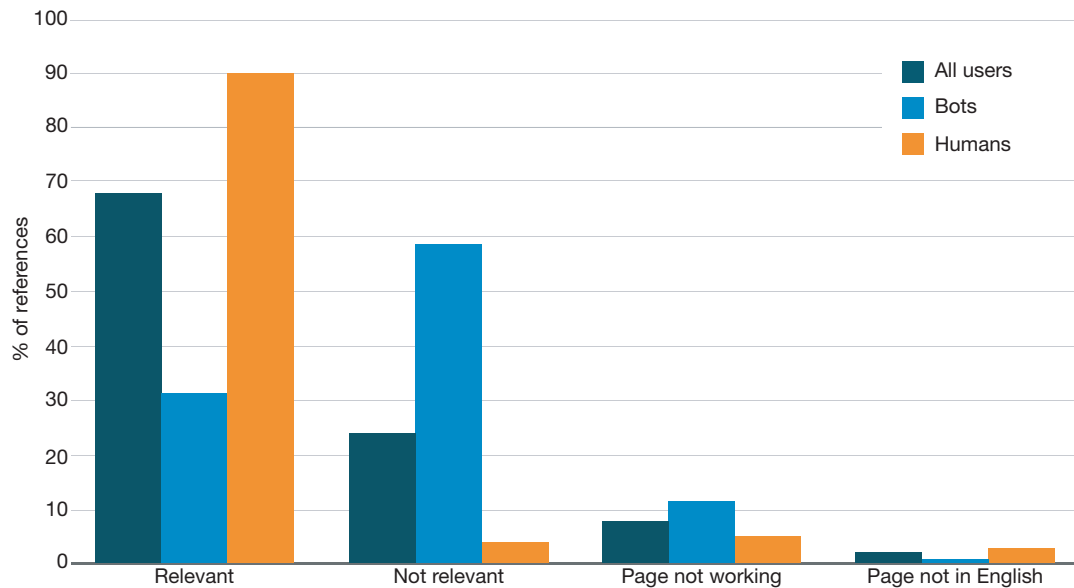


FIGURE 9.8: Percentage of sources by relevance. Please note the small percentage of pages not in English.

Source type	Unique sources			Unique domains		
	Humans	Bots	All Users	Humans	Bots	All Users
Organisation	75.7	81.4	78.2	72.4	50.5	65.8
Individual	10.8	4.5	7.9	11.8	13.1	12.5
Collective	5.3	0.2	2.9	6.1	0.6	4.5
Page not working	3.9	0.2	2.1	4.9	0.6	3.7
Page not in English	4.3	13.7	3.7	4.9	35.2	13.4

TABLE 9.5: Percentage of sources by type of author

Concerning publisher type, the majority of references pointed to sources published by government agencies (37.5%). Academic institutions were the second most common type (around 24%). This changes if we look at the occurrences of unique web domains. In this case, government agencies slip to 5.8%, whereas ‘other companies or organisations’ become the most used sources with 19.9%. Regarding editor types, governments were still the most common among both bots and humans. However, the situation differs depending on whether all references are considered or unique domains. This is common to other publisher types and affects especially bot-added sources. Table 9.5.4 shows percentages of publisher type by user type, for all references and unique domains.

Organisation staff were by far the most common author type (78%) overall, and both among bot- and human-added references (see Table 9.5.4). Sources created by identifiable individuals follow (7.9%) and appear to be reused less often than those authored by organisations (12.5% of unique URLs). Collectively-authored sources represented only 2.9% of our sample. Whereas these were only 0.2% of bot-added pages, they were 5.3%

Publisher type	Unique sources			Unique domains		
	Humans	Bots	All Users	Humans	Bots	All Users
Governmental agencies	32.7	44.4	37.5	34.2	1.5	5.8
Other companies & organisations	15.3	12.6	14.4	17.6	27	19.9
Academic & research institutions	13	12.6	12.4	15.3	28.2	7.8
Other academic organisations	10.3	12.6	11.2	0.4	1.2	1.2
Cultural institutions	7.7	11.9	15	8.6	28.8	15
Vendors & e-commerce companies	7.3	1.8	5.4	8.6	1	15.9
Non-traditional news & media	3.7	1.2	2.5	4.3	2.9	10.1
Self-published	3	0.2	1.6	2.5	0.1	5.4
Traditional news & media	2	0	1.1	2.4	0	5.2
Political or religious institutions	0.9	4.6	1.2	0.9	4.6	1.7
Academic publishers	0.4	0	0.2	0.5	0	1.1
Others	0.1	0	0.1	0.1	0	0.3

TABLE 9.6: Percentage of sources by type of publisher

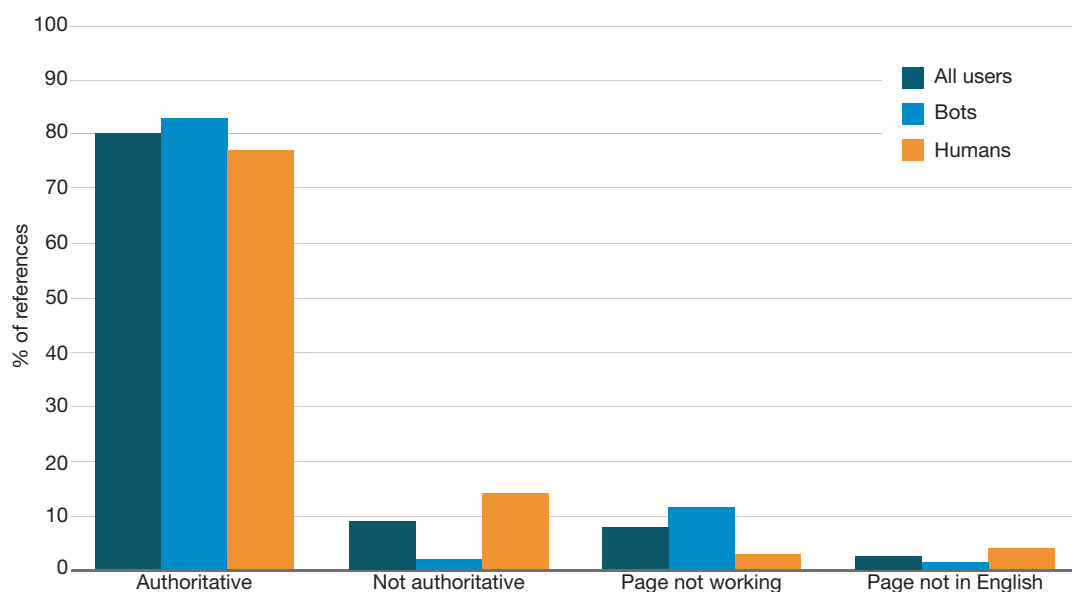


FIGURE 9.9: Percentage of sources by authoritativeness. Sources added by sources are more commonly authoritative than those added by human editors.

of those created by human users. Finally, applying the criteria in Table 9.1, 79.9% of the references were classified as authoritative (Figure 9.9). Conversely to relevance, sources added by bots resulted more authoritative (83.5% vs. 77%). Furthermore, the majority of the bot-added references that were not found to be authoritative was not working (12.1% of the total), whereas only a few were judged to be non-authoritative (2.4% of the total). The percentage of non-English language sources was around 2.4% higher than what found in T1. This was likely due to T2 and T3 evaluating unique links and web domains, contrarily to T1 where sources could be repeated if used in more than one reference.

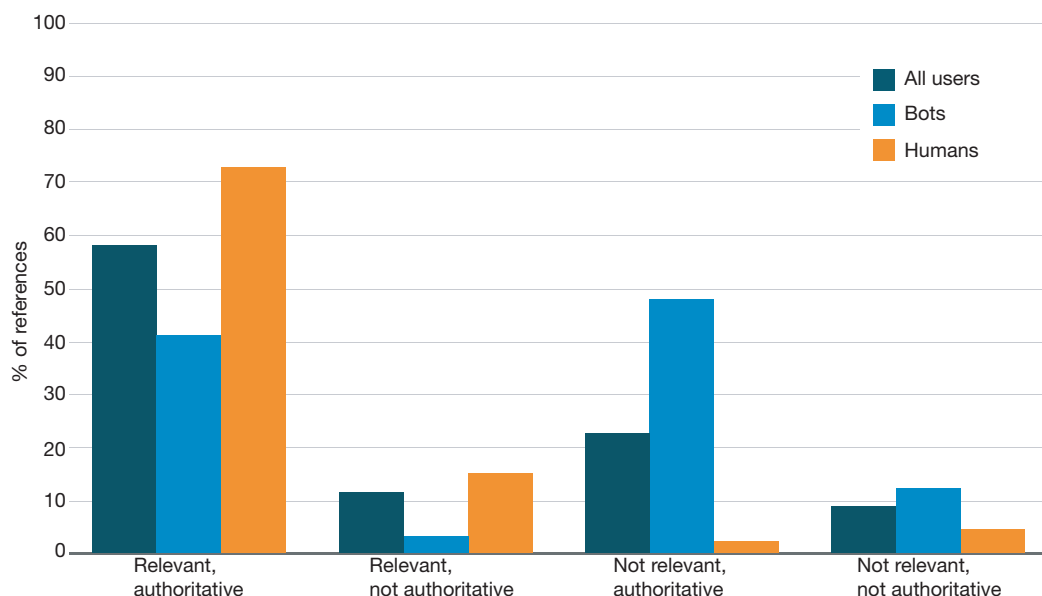


FIGURE 9.10: Percentage of sources by relevance and authoritativeness

Overall, sources are both relevant and authoritative (Figure 9.10). References created by human editors suffer more often from lack of authoritativeness. On the contrary, bot-contributed sources are generally authoritative, but may be not relevant for the statement they are attached to. References must be accessible, therefore several were classified as neither relevant nor authoritative because they were not working or required to log in. Some pages redirected to a new one, which often was not relevant. These were possibly valid at the time of addition, but subsequently changed.

Looking at relevance and authoritativeness by Property (Figure 9.11), references tend to be more relevant than authoritative—for over 182 properties in our sample, 153 have a greater or equal number of relevant references than authoritative. Properties with bad references vary with respect to their domain. Whereas some of the Properties ranking higher for number of not relevant references are connected to the medical domain (P636, route of administration, P689, afflicts, P2240, median lethal dose), other Properties among the most relevant vary, e.g. P21 (sex or gender—this may actually often be classified among those not requiring any reference), P1542 (has effect), and P170 (creator).

9.5.5 Quality prediction models

The trained models were binary classifiers aiming to predict non-relevant and non-authoritative references. We used stratified 10-folds cross-validation to estimate the algorithms' performance. Stratified cross-validation ensures outcome classes have the same distribution in the subsets selected in each fold and improves the comparability of different algorithms (Forman and Scholz, 2010). The F_1 measure was computed on

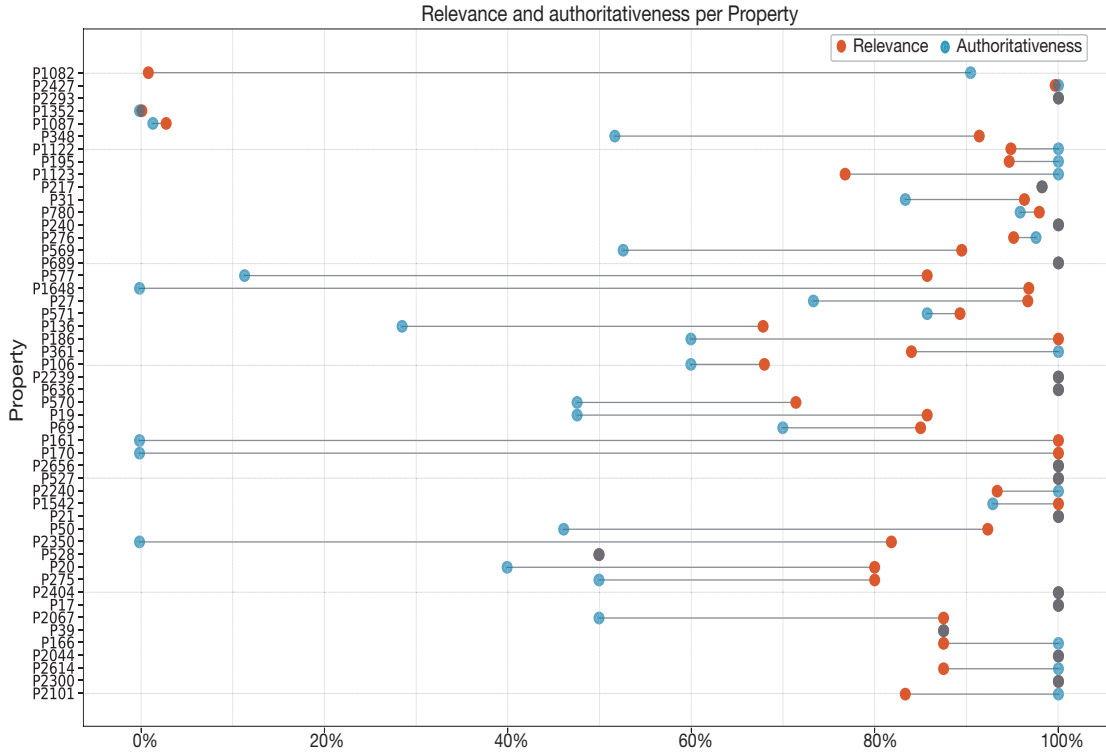


FIGURE 9.11: Relevance and authoritativeness by Property, ordered by number of references within the sample evaluated. The graph includes only Properties with more than 5 occurrences to increase readability.

true and false positive over all folds, providing a more unbiased estimate compared to other methods (Forman and Scholz, 2010). We used Matthews correlation coefficient (MCC) to estimate the level of agreement between predicted and observed labels. MCC has values between -1 and $+1$, with higher values indicating better agreement (Baldi et al., 2000). Class unbalance was addressed by adjusting prediction weights in SVM and Random Forest (Pedregosa et al., 2011). Further details about implementation and hyperparameters of the models are provided in this thesis’ GitHub repository.¹¹

The **relevance** baseline was good at predicting non-relevant sources ($F_1 = 0.84$, $MCC = 0.68$), although it was outperformed by all models. Random Forest provided the best scores. The **authoritativeness** baseline gave worse results ($F_1 = 0.53$, $MCC = 0.15$). All trained models outperformed the baseline, with Random Forest yielding the highest F_1 (0.89) and MCC (0.83). Results for both models are shown in Table 9.7.

¹¹Data and code available at https://github.com/Aliossandro/structuring_the_world_knowledge_phd_thesis.

	Model	P	R	F ₁	AUC-PR	MCC
Relevance	Baseline	0.88	0.83	0.84	0.81	0.68
	Naive Bayes	0.94	0.94	0.90	0.92	0.86
	Random Forest	0.95	0.95	0.92	0.94	0.89
	SVM	0.94	0.94	0.91	0.94	0.87
Authoritativeness	Baseline	0.71	0.65	0.53	0.62	0.16
	Naive Bayes	0.90	0.90	0.86	0.88	0.78
	Random Forest	0.93	0.92	0.89	0.93	0.83
	SVM	0.90	0.90	0.89	0.90	0.79

TABLE 9.7: Performance of prediction models for relevance and authoritativeness

9.6 Discussion

9.6.1 Crowdsourced evaluation

The crowdsourced experiment provided accurate results, as shown by the level of agreement between workers and the percentage of correct responses to test questions. Task completion times differed greatly, probably due to the task type. T1 asked users to find a piece of information within a web page and seemed to be straightforward. Conversely, the classification tasks T3.A and T3.B were harder. This may be due to the classification system used appearing unclear for workers, or clashing with their prior knowledge, leading to erroneous responses, similar to what has been noted before in taxonomy creation tasks (Karampinas and Triantafyllou, 2012). Nevertheless, the judgements collected in T3.B largely confirmed T3.A. The difference between the percentage of non-English language sources in T1 and T2, T3.A, and T3.B may be due to the specificities of the tasks: the information sought to complete T1 may have been easier to retrieve, which may have led workers to respond to the question, instead of flagging the page as not in English.

The majority of references examined included relevant sources, although those added by humans and bots diverged considerably. This (see Table 9.8) is caused by a link to a US census dataset that was redirected to another page, which did not contain relevant data anymore. By removing the references using that link, the total percentage of relevant sources goes up to 85.9%. Considering the breakdown by type of reference creator, sources added by human editors are still more relevant (90.7% vs. 77.2%) than those added by bots. The case of the malfunctioning US census page may not be an isolated case. In Section 9.4 we have seen that, whereas bots most commonly add sources at the same time or shortly after the creation of a statement, there seem to be some form of control on the references they add, considering the number of times their references are edited by other users. The results of our crowdsourced evaluation suggest that this control might not be enough to address issues such as invalid URLs becoming outdated or invalid. Collaborative production systems with no editorial oversight are inherently vulnerable to fluctuations in the quality of their content (Faraj et al., 2011). This consideration may

apply even more for Wikidata, whereby [Piscopo and Simperl \(2018\)](#) have shown that sudden burst of activity by single editors, either bot or human, may change significantly the data in the graph. More continuous control from the community are required to address issues like those highlighted above to occur—the eyeballs required to make all bugs shallow ([Raymond, 2001](#))—either manually or automatically, e.g. a frequent check of URL validity.

Government agencies are the most common publisher type, both among human- and bot-added references. Sources are generally authored by organisation staff and not by individuals. Two classes of publishers showed large differences between percentage of references and percentage of unique domains (Table 9.5.4). In both categories, the skewness is likely to be determined by the massive automatic generation of statements by bots. This led us to hypothesise that typical bot editing patterns may result in a lower degree of diversity of source types. The data confirmed this: in spite of similar numbers of references by bots and humans (46.3% vs 53.6%), bots used 36 web domains, compared to 295 by humans. This analysis should increase awareness about the current limitations of using bots to add references, and in turn help design bots that follow a more nuanced approach to reference selection. Nevertheless, this conclusion should be taken with a grain of salt; our sample includes only English-language sources—a minority of all Wikidata external sources [Piscopo et al. \(2017d\)](#)—and may fail to represent other aspects of the data, such as a more diverse selection of web domains in references. Future work may use different samples in order to address this issue.

The distribution of author and publisher types for references did not match that observed in Wikipedia ([Ford et al., 2013](#)), despite the partial overlap of the two communities ([Piscopo et al., 2017c](#)). This has been already noted also in [Piscopo et al. \(2017d\)](#). A smaller number of news sources is used in Wikidata references, compared to the online encyclopedia. Whereas Wikidata recommends primary sources as references, Wikipedia asks editors to use secondary sources and officially disapproves of primary ones, in line with the rule that the encyclopedia cannot contain original research.

References added by bots appear to be more authoritative on average. Bots generally import statements from a number of sources that are previously selected by the human user responsible for them. Requesting permission to run a bot on Wikidata requires a certain degree of cognition of the policies and practices of the platform. Hence, bots are likely to be maintained and run by more experienced users, who are aware of which types of references are considered authoritative by the Wikidata policy. The lower authoritativeness of human-contributed references is a topic for future studies. In particular, it would be important to understand what type of users add worse sources, e.g. examining the relation between the level of experience of users and the type of references they add.

9.6.2 Machine learning experiment

The predictive models for relevance and authoritativeness performed well, which may support our intuition that that sources from a website that are good for a type of statement, i.e. using a determined property with defined domain and range, are likely to be good for similar statements. Another explanation may regard the characteristics of references in Wikidata. From a total of around 2000 Properties at October 2016 (the date of our sample), only about 200 have references. Sources from the same web domain tend to have the same level of quality. On the other hand, the number of domains per Property is low. As a consequence, the algorithm may find ‘easy’ to assess combination of Properties and domains. If the number of Properties with references and the diversity of web domains used will increase, further research should evaluate how this affects the performance of predictive models of reference quality. It should also seek to understand how to adapt these models to be implemented in Wikidata, to help editors find bad references.

9.7 Limitations

The study described in the current chapter presents some limitations, which should be addressed in further studies.

Future work should validate whether our results hold true for non-English sources. Besides using outgoing links, Wikidata expresses provenance by means of internal connections, which were not examined in this study. These are a substantial part of Wikidata references and should be examined in the future, in order to achieve a comprehensive evaluation of provenance quality in Wikidata.

The Wikidata verifiability policy ([Wikidata, 2018j](#)) provides some example of source types which are likely to be authoritative. However, it also adds the caveat that authoritativeness is dependent on the combination between statement and source. For example, a sport newspaper may be an authoritative source with respect to the 2016 Olympic games, but less for what concerns other topics, such as literature or politics. Our approach does not take into account this relationship between source and statement. Future work should devise suitable approaches to address this limitation.

9.8 Summary

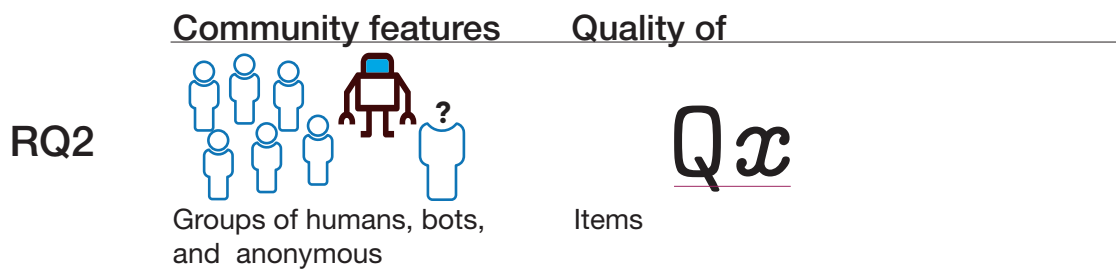
In this chapter we have looked at the quality of the revisions of human and bot editors, specifically for what concerns one of the most interesting features of Wikidata, i.e. references. Human-contributed sources appear to be overwhelmingly relevant, in contrast

to those added by bots. This is due primarily to sources previously added by bots and not working anymore, suggesting the need of more continuous and rigorous controls on automatically authored edits.

In the next chapter, we take into account a large number of features of human users, investigating how differences in terms of experience and activity focus come into play when editors work together on a Item.

Chapter 10

The right mix of users: Group diversity and Item quality



The Wikidata user pool is rather heterogeneous, as we have seen in Chapter 6. Editors may have different levels of experience, devote their attention to different tasks, use different tools. They may edit after registering or anonymously, and they may even be pieces of software (bots) programmed by other human users in order to carry out revisions on a large scale.

Research question 2 investigates the effects of this differences on quality. In particular, it looks at how the outcome of groups of editors is influenced by the diversity of their members. In other words, RQ2 seeks to understand the ‘right mix’ of users required in order to achieve good quality in Wikidata. We report RQ2 in the following, for convenience of the reader.

RQ2 To what extent does group diversity affect outcome quality in Wikidata?

The following sections present related work about prior approaches to study the effects of group composition and diversity on quality. After that, some of the observation made in the previous chapter regarding group diversity in the literature and in Wikidata are summed up. Subsequently, Section 10.2 outlines the hypotheses tested to address RQ2.

The subsequent section describes the data used and provides more depth about the methods. Finally, results are presented and described.

10.1 Related work

Online communities host heterogeneous pools of users. The effects of diversity on performance have been investigated by the literature, which has highlighted both positive and negative outcomes. Diversity appears to be both an opportunity as well as a challenge for work teams (Milliken and Martins, 1996) in offline contexts. Differences among group members may generate a “creative abrasion” that positively affects performance according to Arazy et al. (2011). On the other hand, diversity may hamper the identification of users within a group (Milliken and Martins, 1996). Researchers have tried to explain these mixed effects by categorising various types of diversity. In their review of studies about diversity in organisational groups, Milliken and Martins (1996) distinguish observable and underlying attributes. Dissimilarity with regard to observable attributes, such as age, gender, or race, lead to higher turnover and lower integration (Milliken and Martins, 1996). Underlying attributes may refer to personality characteristics, values, skills and knowledge, and functional background, among others. Skills- or knowledge-related diversity affects positively performance in top-management and project teams, whereas the effects of other types of underlying attributes are less clear. In a similar fashion, Arazy et al. (2011) identify surface- and deep-level diversity. Whereas the first encompasses demographic characteristics, the latter regards expertise, knowledge, and functional background. Deep-level diversity entails a higher variety of perspectives, which create better conditions for creativity and knowledge sharing. Other attempts to interpret different types of diversity in respect to outcome quality see two contrasting viewpoints, the *social category* and the *information/decision making* perspectives (Van Knippenberg et al., 2004). The social category perspective focuses more on relational aspects. Homogeneous groups benefit from higher cohesion and member commitment, thus being able to produce a better output. The information/decision making perspective is slanted instead towards job-related attributes and connected to less evident aspects of members, e.g. educational or functional background. Diversity also influences positively performance according to this perspective. Van Knippenberg et al. (2004) combine these two perspectives, by connecting them to the requirements and the elaboration of tasks. Diversity would lead to better performance in case of complex information-processing tasks, with respect to simple, repetitive ones.

A great deal of previous research has focused on demographic features of group members. Ancona and Caldwell (1992) explored direct and indirect effects on performance by the distribution of organisation tenure and functional speciality in the team, with mixed results. Whereas both types of diversity have a direct negative effect on team-

and managerial-rated performance, their indirect effects look more complex. More heterogeneous groups with regard to tenure, i.e. the length of activity within a team, are able to define better their goals and priorities. Higher functional diversity improves external communication. Both clarity of goals and priorities and improved external communication positively affect performance. These conflicting findings suggest a complex relationship between group diversity and outcomes, with effects that may change according to the context and the type of diversity studied. [Pelled et al. \(1999\)](#) draw similar conclusions about the complexity of the relationship between several types of diversity, conflict, and performance. Their study, carried out on corporation teams, finds a positive association between tenure and functional diversity and task conflict. In turn, this affects positively cognitive task performance, thus suggesting that differences in organisational tenure and functional background of group members may indirectly improve their outcome. Other variables, e.g. race and gender, do not seem to directly influence conflict.

In Chapter 8 we have singled out two traits along which we want to study diversity, i.e. tenure and interest. Tenure diversity refers to the distribution of the length of the activity lifespan within a community. Previous studies have found it to have a positive, curvilinear interaction with decision quality ([Lam et al., 2010](#); [Ren et al., 2016](#)). Interest diversity defines the variety of members' interests in a group ([Chen et al., 2010](#)). In collaborative projects such as Wikipedia or Wikidata, where users contribute voluntarily and generally choose which tasks to take on, an individual's interests may actually determine their activity and function within the project. Interest diversity is a concept close to cognitive diversity, which describes the mental models and interests of the members of a group ([Arazy et al., 2011](#)). Both interest and cognitive diversity have been found to positively affect performance ([Arazy et al., 2011](#); [Chen et al., 2010](#)).

We have seen in Chapter 6 that Wikidata editors are likely to show a different behaviour, depending on the year they joined the platform. Participants that started in the first year after the launch of the system are on average more active than users joining later. The majority of users in Wikidata (64%) have contributed for longer than a year (529 days on average, Figure 10.1). As discussed in chapters 2, 6, and 8 editors with different levels of experience and interests are likely to contribute to different aspects of the graph. This also suggested by the qualitative study we carried out in [Piscopo et al. \(2017c\)](#). That work analyses the evolution of user perception of their activity and their role within the Wikidata community along their lifespan on the platform. Data was collected through semi-structured interviews to highly-active Wikidata users, who were asked about their experiences as novices and subsequently as experienced users. Data was analysed under the frameworks of Legitimate Peripheral Participation ([Wenger and Lave, 1991](#)), which explains how members become fully participant in a community of practice, and Activity Theory ([Kuutti, 1996](#)), which formalises the roles and interaction between actors in socio-technical systems. Concerning user motivation and perception

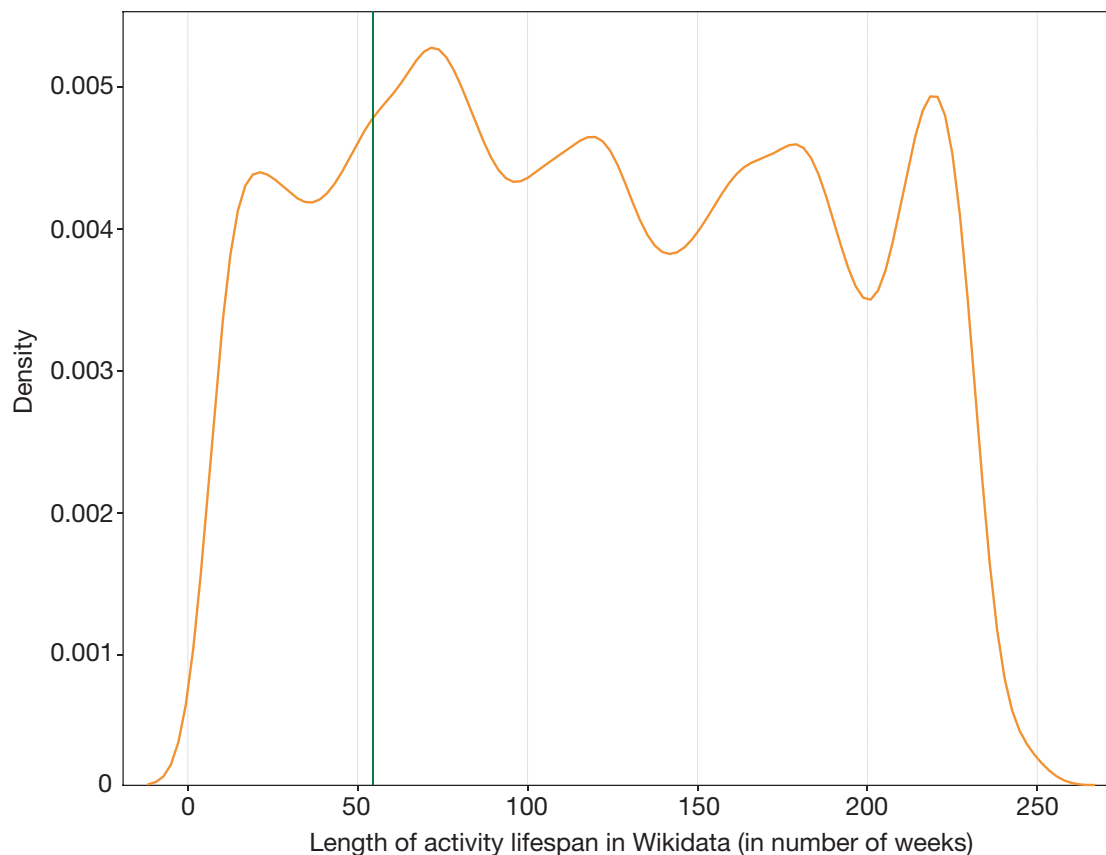


FIGURE 10.1: Tenure distribution in Wikidata, in number of weeks since the first edit. The vertical line marks one year (52 weeks). Tenure is computed by counting the number of weeks between the first edit of a user and the last day in our dataset.

of self and of the community, members of Wikidata are attracted by the simplicity of the project and by the possibility to spread knowledge by contributing and sharing structured data. This motivation does not seem to change as the level of user experience increases. With higher levels of responsibility, users feel more bound to the project and develop an identity as “Wikidatians”. They become more aware of different ways to use the Wikidata interface and are more likely to use a broader range of tools. Seasoned editors take on higher level tasks, contributing to the creation of conceptual knowledge and to quality control and maintenance. Furthermore, they focus on a particular type of edit on a larger number of Items at the same time, e.g. adding all missing references or statements with a determined Property, as opposed to novices, who often work on smaller sets of Items, generally related to a topic, and carry out revisions using the web interface, thus at a smaller pace. Experienced users also appear to have a more central role within the community, as they mentor newcomers and introduce them to Semantic Web concepts. These findings suggest that editors with different levels of experience may have diverse sets of skills, activity patterns, and community involvement. This chapter quantitatively analyses the effects of some aspects of this diversity on the quality of Items.

Previous approaches

[Forte et al. \(2012\)](#) analyse group work in Wikipedia, by focusing on nested organisational structures, specifically WikiProjects. In their work, they use McGrath's typology of group functions, which takes into account the activities required in order to maintain a group healthy, besides those related merely to production ([McGrath, 1991](#)). Underpinned by this theoretical framework, [Forte et al. \(2012\)](#) pursue a mixed-method approach to examine the mechanisms through which WikiProjects facilitate specialised work in the free encyclopaedia. They first carry out two rounds of interviews in order to gain an in-depth understanding of how members experience their activity within WikiProjects. The insights obtained from the interviews were subsequently utilised to inform the quantitative analysis of interactions across 379 project on the English Wikipedia.

Another approach is followed by [Brandes et al. \(2009\)](#), who investigate how group structure influence quality and level of conflict of Wikipedia articles. However, rather than looking at group composition—which is our aim—they focus on interactions, which they model by building an edit network from the revision history of each page.

A very common approach to research the effects of group characteristics on outcome quality is by applying a regression model to predict that. Small groups are the subject of [Lam et al. \(2010\)](#)'s study, which investigates the influence of group size, group bias, and tenure diversity on decision quality in Article for Deletion forums in Wikipedia. This approach relies on a logistic regression model to predict a binary variable describing whether a decision is reversed or not. [Ren et al. \(2016\)](#) investigate the evolution of diversity and its impact on group outcomes in Wikipedia. Using longitudinal data about WikiProjects, similarly to [Forte et al. \(2012\)](#), they apply Hierarchical Linear Models (HLMs) to understand the interaction between tenure diversity and interest variety and two dependent variables, i.e. group productivity and member withdrawal. HLMs are a type of Ordinary Least Squares (OLS) regression that are suitable to nested data ([Woltman et al., 2012](#)), such as in the case of the projects considered by [Ren et al. \(2016\)](#). Understanding the effects of diversity on group outcomes is also the aim of [Ren and Yan \(2017\)](#). The latter work focus on contribution diversity—i.e. the difference in editors' contribution within the same article—and experience diversity, a concept close to tenure diversity. The platform utilised for their study is again Wikipedia, namely articles within WikiProject. Conversely from the studies mentioned so far, [Ren and Yan \(2017\)](#) take into account features that act as mediators of user participation, specifically task conflict and task communication. Group outcome is measured by the peer-assigned score of the Wikipedia articles in the sample. The approach adopted to evaluate whether diversity features are significant predictors of the dependent variable is bootstrapped mediation analysis. This is a path analysis based on OLS that is able to estimate direct and indirect effects ([Ren and Yan, 2017](#)). [Arazy et al. \(2011\)](#) examine group

composition and task conflict and their effects on quality in English Wikipedia articles. Groups are modelled as the set of authors who edit an article. The analysis is two-staged: a quantitative one, which uses revision counts from Wikipedia logs, and a qualitative one, consisting of content analysis of the articles discussion pages. The latter is not suitable to Wikidata items, whose talk pages are seldom used (see Chapter 6). Group composition was operationalised by several variables, which included the percentage of administrators and the average number of edits on Wikipedia of group members. Furthermore, other independent variables comprehended cognitive diversity, expressed as a measure of sparsity of a matrix of all the articles edited by the group members, and task conflict, constructed using the article’s discussion page. Group size was used as a control variable, among others. The model applied was the partial least squares (PLS) algorithm, which is suitable for smaller samples and requires fewer assumptions about data distributions (Arazy et al., 2011). Finally, Daniel et al. (2013) examine the influence of three types of diversity, i.e. separation, variety, and disparity, on community engagement and market success of a number of Open Source Software (OSS) projects. Because the dependent variables are counts of occurrences, the model used is a Negative binomial moderated regression analysis, a particular case of Poisson regression (Daniel et al., 2013).

10.2 Research hypotheses

In order to address RQ2, we formulated and tested a number of hypotheses, which are presented in this section. Hypotheses 1-3 concern the proportion of contributions by different user types. Hypotheses 4-5 regard tenure and interest diversity.

The importance of bot contributions for outcome quality has been noted already with regard to Wikipedia (Niederer and van Dijck, 2010). In Wikidata, the amount of bot editing activity and its scope suggests that their contribution is a crucial factor for outcome quality. In the previous chapter, we have noted that bot activity has contrasting effects on the quality of references. On the one hand, bots add the majority of them and use more authoritative sources on average. On the other hand, the sources added by bots may become not relevant for the statements they are attached to, because of insufficient control over their work. With respect to Items, bots have added so far the largest part of their content, contributing to their completeness. Therefore, we formulate our first hypothesis:

Hypothesis 1: The percentage of bot edits is positively related to Item quality.

Although the contribution of bots is important to set the basic structure of Items—e.g. automatically adding Wikimedia links and labels in several languages—some tasks require human editors. These possess the knowledge and skills to add descriptions and aliases, and perform quality controls that are not routinely performed by bots. In their

analysis of the emergence of user roles connected to the division of labour in Wikidata, Müller-Birn et al. (2015) observe that bots and humans perform similar tasks, however with a different distribution. Bots' activities focus more on setting new statements, whereas human contributors primarily edit them and add references. Hence, bots and registered human editors may need to complement their efforts in order to achieve high-quality Items (see Chapter 6). On the other hand, users who edit anonymously may have lower levels of attachment and have shown to often generate spam and vandalism in other projects (Adler and de Alfaro, 2007). We refer to interaction between human and bot editors as the balance of the respective contributions to an Item. Higher interaction means a more equal contribution from each of these two user types.

Hypothesis 2: High levels of interaction between human and bot users are positively related to quality.

Hypothesis 3: The percentage of anonymous human edits is negatively related to Item quality.

As mentioned above, Wikidata editors take on different types of tasks along their evolution as part of the community (Piscopo et al., 2017c). Seasoned users focus on higher-level tasks, e.g. working on the conceptual structure of knowledge and on quality maintenance tasks, whereas newcomers tend to concentrate their efforts on adding and modifying statements. Items edited by users with various tenure levels may benefit from these different 'specialisations'. Additionally, more experienced users feel a sense of responsibility towards Wikidata. This might drive them to oversee the work done by other editors and help ensure quality.

Hypothesis 4: Tenure diversity is positively related to Item quality.

A similar process may be at play with regard to interest diversity. Editors working on a broader range of Items may lead to different perspectives to the creation of Items. One of the peculiarities of KGs is that the entities they contain are linked, allowing machines to perform inferences and reason following these connections. Users with heterogeneous interests may facilitate the creation of internal links.

Hypothesis 5: Interest diversity is positively related to Item quality.

10.3 Data and methods

In the following, we provide further details about the approach employed to test our research hypotheses, including the variables examined, the analysis strategy, and the data used.

10.3.1 Dependent variable

For the purpose of this study, we used as a dependent variable the quality measure generated by [Yapinus et al. \(2017\)](#) in close collaboration with the community of Wikidata. We described this quality scale in Section 7.2.1. Further details about this measure are provided below.

Item labels were collected for a sample of 5000 Wikidata Items, each evaluated by one or more Wikidata editors. A pilot campaign was previously run to verify and refine the quality of community-generated labels. The sample selection aimed to obtain a more balanced distribution of Item quality classes, compared to the entirety of Wikidata, where the majority of Items likely fall in classes C to E. Therefore, [Yapinus et al. \(2017\)](#) over-represented classes A and B by selecting a certain number of Items per size (in bytes), following the assumption that larger Items would more likely have higher quality. Additionally, they included a number of ‘special Items’, i.e. Items whose `QId` has a particular meaning or were created early in the project lifetime (thus having low `QId`), such as `Q2` (Earth). The distribution of Items per quality level is shown in Table 10.1.

Quality level	No. Items	No. Items (w/ at least 1 human edit)
A	322	322
B	438	419
C	1773	1671
D	986	702
E	1468	1010

TABLE 10.1: Distribution of quality levels

10.3.2 Independent variables

We present here the independent variables included in our analysis. Diversity measures referred only to registered human users—to which we refer from now on as human users—because anonymous users often cannot be tracked across different edit sessions. Bot users were not included as well.

Tenure diversity. This variable was computed for each Item by using the coefficient of variation ([Bedeian and Mossholder, 2000](#)) calculated on the number of days between each human user’s first edit and the last day in our dataset. This method has previously been applied to measure tenure diversity in [Ancona and Caldwell \(1992\)](#) and [Chen et al. \(2010\)](#).

Interest diversity. The closeness of the editing patterns of users working on the same Item has been used to estimate interest diversity, following the approach in [Arazy et al. \(2011\)](#). To build this metric, we generated a two-dimensional matrix, in which all members of the group—all human users that performed at least one edit on the Item considered—lie

on one dimension and all Items edited by anyone of them are on the other. Cells were assigned 0/1 values, according to whether a user had edited an Item. The sparsity of the matrix—the ratio between the number zeros and the total number of cells—reflects the extent of the overlap between group members’ editing patterns. Outcome values range from zero to one, with higher values indicating more diverse groups.

Proportion of bot edits. The proportion of edits made by bots over the total number of edits. This value was between 0 and 1.

Proportion of anonymous edits. The proportion of edits made by anonymous users over the total number of edits. This value was between 0 and 1.

Bot X Human edits. This variable captures the amount of interaction between bot and human editors. It was computed by multiplying the proportion of human edits by the proportion of bot edits. Considering the low amount of anonymous contributions, this variable can have values distributed in an inverted U shape, with higher values reflecting more balanced contributions from bots and humans.

10.3.2.1 Control variables

Number of edits. Items with a larger number of revisions are likely to have more statements and to have been reviewed and corrected more times.

Group size. The literature reports diverse effects of group size on outcome quality. Larger groups may negatively affect performance, because they reduce the likelihood of collaboration and increase the chance of conflicts (Levine and Moreland, 1990). On the other hand, more members likely entail a broader range of information sources (Surowiecki, 2005; Afuah and Tucci, 2012). We included group size as a control variable to account for these possible effects. Group size was measured by computing the number of unique editors for each Item.

Age of the Item. Older Items have likely been seen and reviewed more often. We used the number of days between the creation of an Item and the last day in our dataset as a control variable.

10.3.3 Analysis strategy

We performed an ordinal logistic regression (OLR) analysis to test the hypotheses. We trained four models to predict Item quality labels and verified the significance of the independent variables for prediction. The first model was the baseline and included only the control variables. Model 2 added variables related to the proportion of user type. Model 3 tested the influence of tenure and interest diversity, including only Items that have ever been edited by humans in order to reduce sparsity of the data. Model 4

tested all the independent variables together, using all the Items in our dataset. Tenure and diversity values were set to zero when no human users contributed to an Item.

10.3.4 Data

The data utilised in this experiment is updated on 1st of April 2017 and used the processed Wikidata dumps described in Section 8.4. We used the complete revision history of each Item in the labelled sample, including edit timestamp and user names.

Only 4987 Items over 5000 in the labelled sample were present in the dumps—the missing Items had been deleted by the Wikidata community. Of these 4124 were ever edited by human editors.

10.4 Results

Table 10.2 reports descriptive statistics of and correlations among the variables used in the analysis. The Items in the sample greatly vary in terms of number of edits, group size, and age. Both mean and median number of edits per Item were much larger than the figures for the whole of Wikidata (135.4 vs. 14.2 mean; 28 vs. 9 median). This was likely due to the method followed in drawing the sample (Section 10.3.1), which over-represented Items with higher quality. The proportion of human-made edits in the sample was higher than overall (0.46 vs. 0.28), which might be attributed to the fact that high-quality Items chosen were manually curated by editors.

The ratio between edit number and group size shows that each user in a group carried out on average four revisions. If we consider the median Item age (around four years) and number of edits, Items are seldom edited. The proportion of registered human edits was, not surprisingly, highly correlated to bot edits, therefore it was left out from the models. Regarding diversity, Items are edited by a population of editors which is moderately heterogeneous in terms of tenure. On the other hand, interest diversity was very high, indicating that on average editors focus on different sets of Items.

The baseline model (1, Table 11.7) shows a positive significant influence of Item age, edit number, and group size on Item quality. The increase in quality level is very low for all three variables though, with Item age having the smallest effect. Model 2 adds variables related to the contribution of different types of users to an Item. The proportion of bot edits has a positive significant interaction with the response variable, thus **supporting hypothesis 1**. The influence of bots on Item quality increases when these interact with human editors, as shown in Table 10.3, which **supports hypothesis 2**. The proportion of anonymous users is significant for prediction as well and influences negatively Item quality. This means that **hypothesis 3 was supported**.

	Mean	Median	Std	# Edits	p Bot edits	p Anonymous edits	p Human edits	Group size	Item age	Tenure div.
# Edits	135.4	28	239.19							
p Bot edits	0.53	0.50	0.35	-0.35						
p Anonymous edits	0.01	0	0.03	0.36	-0.27					
p Human edits	0.46	0.50	0.34	0.32	-0.99	0.18				
Group size	36.32	7	57.48	0.81	-0.47	0.49	0.43			
Item age	1182	1507	557.16	0.30	-0.15	0.22	0.13	0.47		
Tenure diversity	0.47	0.38	0.48	0.40	-0.49	0.27	0.48	0.56	0.62	
Interest diversity	0.89	0.98	0.19	0.11	0.01	0.12	-0.02	0.25	0.16	-0.12

TABLE 10.2: Descriptive statistics and correlations among independent variables. Item age is expressed in days since Item creation.

Model 3 was trained on Items with at least one human edit. The distribution of quality labels for this set of Items was more skewed towards higher levels, compared to the full dataset (Table 10.1). The results of model 3 show a significant positive interaction of tenure diversity with Item quality (Table 10.4), thus **supporting hypothesis 4**. Interest diversity was as well a significant predictor, albeit with a lower positive influence on quality. Hence, **hypothesis 5 was supported**. Finally, model 4 included all the dependent variables. Significant interactions did not change, with the exception of the proportion of anonymous edits, which ceased to be a predictor of quality. The effect of group size decreases, compared with model 2. Moreover, tenure diversity had a stronger positive influence on quality, whereas the effect of the interaction between bots and humans decreases.

	Model 1			Model 2		
	Coef.	SE	p	Coef.	SE	p
<i>Label</i> $\geq D$	-0.071	0.061		-1.302	0.104	***
<i>Label</i> $\geq C$	-1.255	0.064	***	-2.550	0.108	***
<i>Label</i> $\geq B$	-4.445	0.103	***	-5.767	0.136	***
<i>Label</i> $\geq A$	-6.217	0.132	***	-7.602	0.163	***
Item age	0.001	0.001	***	0.001	0.001	
Group size	0.028	0.001	***	0.033	0.001	***
# Edits	0.003	0.001	***	0.003	0.001	***
p Bot edits				1.400	0.102	***
Bot X Human				4.690	0.337	***
p Anonymous edits				-3.825	1.221	**

TABLE 10.3: Ordinal logistic regression of number of edits and group size, editor types, and diversity measures. Note: *** $p < 0.001$, ** $p < 0.01$.

	Model 3			Model 4		
	Coef.	SE	<i>p</i>	Coef.	SE	<i>p</i>
<i>Label</i> ≥ <i>D</i>	−1.174	0.178	***	−2.649	0.212	***
<i>Label</i> ≥ <i>C</i>	−2.387	0.181	***	−4.106	0.217	***
<i>Label</i> ≥ <i>B</i>	−5.890	0.214	***	−7.573	0.245	***
<i>Label</i> ≥ <i>A</i>	−7.484	0.226	***	−9.276	0.257	***
Item age	0.001	0.001		−0.001	0.001	***
Group size	0.015	0.001	***	0.025	0.002	***
# Edits	0.004	0.001	***	0.004	0.001	***
<i>p</i> Bot edits				2.469	0.123	***
Bot X Human				3.769	0.362	***
<i>p</i> Anonymous edits				−3.663	1.240	
Tenure diversity	1.550	0.110	***	2.804	0.117	***
Interest diversity	1.010	0.197	***	1.100	0.200	***

TABLE 10.4: Ordinal logistic regression of number of edits and group size, editor types, and diversity measures, trained on Items with at least one human edit. Note: *** $p < 0.001$, ** $p < 0.01$. Model 3 has been trained on the set of Items with at least one registered human edit.

10.5 Discussion

We have analysed the influence of group composition on outcome quality in Wikidata. First, we looked at how different proportions of bots, registered and anonymous human users affect quality. Second, we studied the effects of the distribution of two variables within groups of registered human users, tenure and members' interests.

The interaction between human editors and bots seems essential for the quality of Wikidata. It appears that the intertwinement of human and algorithmic contributions that led [Niederer and van Dijck \(2010\)](#) to define Wikipedia as a socio-technical system is also key for Wikidata quality. The division of work outlined by [Müller-Birn et al. \(2015\)](#) may explain the strong positive effect of bot–human interaction on Item quality. Each type of user contributes to Wikidata by carrying out the tasks in which they are specialised and require each other, in order to achieve good quality. Future work should investigate in detail this interaction at Item level, focusing on which share of light- and heavy-weight tasks (see Chapter 6 and [\(Haythornthwaite and Wellman, 1998\)](#)) need to take on each user type, in order to successfully build an Item. Furthermore, fewer than half of the Items in our datasets were ever edited by anonymous users. Although this reflects the overall edit distribution in Wikidata, this suggests that caution should be taken in interpreting results related to hypothesis 3 and that a more in-depth study should be conducted to draw clearer statements about that.

Heterogeneous groups in terms of tenure of their members are more likely to produce higher quality Items. This contradicts prior studies around tenure diversity in an offline context, such as [Ancona and Caldwell \(1992\)](#) and [Pelled et al. \(1999\)](#). On the contrary, it agrees with the observations around Wikipedia in [Lam et al. \(2010\)](#). An explanation

may be that in online contexts the importance of the relational aspect, which sees homogeneous groups perform better due to increased cohesion, decreases. More diverse groups would benefit from the different perspectives brought by their members, according to the information/decision making perspective (Van Knippenberg et al., 2004). This would apply specifically to Wikidata, where contrasting statements can coexist and editors do not need to discuss on talk pages to reach consensus, in contrast to Wikipedia, in which discussion pages are used to settle disputes. Another likely cause for the positive influence of tenure diversity on quality is the diversification of tasks carried out by users at different times of their activity within Wikidata (Piscopo et al., 2017c). The contributions of editors with various tenure levels may thus be complementary.

Our models show a significant interaction between interest diversity and quality. This finding is in agreement with previous research, which noted a linear correlation between this type of diversity and productivity (Chen et al., 2010) and between cognitive diversity and quality of decisions in Wikipedia (Lam et al., 2010). Varied editor interests may imply that these are more active over the whole KG and know its mechanisms better. Furthermore, group editors that are active over a wider portion of Wikidata may have increased chances to link an Item to others in the KG through statements. The interest diversity measure used does not take into account how conceptually distant the Items edited by members of a group are. For instance, two users may engage in adding content related to British musicians, while still working on different Items. Future work may rely on semantic similarity measures such as that presented in Ribón et al. (2016) in order to address this limitation.

Finally, this chapter addressed RQ2, aiming to shed light on the ‘right mix’ of users that leads to higher quality in Wikidata. According to the models trained, groups with higher levels of cooperation between bot and human editors (where tasks are more equally shared among these) are able to achieve better performance. ‘Ideal’ groups also benefit from including members with different tenure, which may address various quality issues. Group size has only a limited positive influence on performance, which partially contradicts previous observations around Wikipedia (Kittur and Kraut, 2008; Lam et al., 2010). The presence of anonymous users in these groups seems marginal and does not have any significant effect.

10.6 Limitations

Regarding the limitations of this work, cross-sectional approaches such as the one employed in our analysis may suffer from reverse causation and uncontrolled confounding factors (Kittur and Kraut, 2008). Longitudinal analyses are effective for addressing these issues. Nevertheless, no measures of quality over time are currently available for Wikidata, to the best of our knowledge. This is a relevant research topic for the future

of Wikidata and should be addressed by further studies. Several variables are at play in group work, such as the coordination among their members. Future research should explore how group diversity interact with other variables.

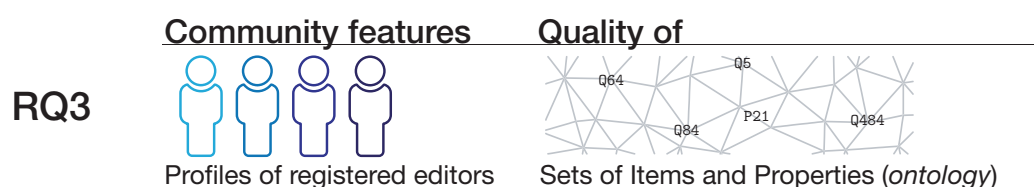
10.7 Summary

This chapter has addressed the relationship between group composition and outcome quality in Wikidata. We analysed how the contribution of these types of users and their interaction benefit Wikidata Item quality. Furthermore, we examined the effects of tenure and interest diversity across registered human users on outcome quality. Ordinal logistic regression analysis revealed that the interaction between human and algorithmic users is necessary to create high quality Items. Contributions from anonymous users are instead detrimental for quality. Concerning tenure and interest diversity, both these features have a positive influence on quality. More heterogeneous groups seem likely to benefit from the different experiences and skills of their members. One of our aims was to identify what are the characteristics of successful groups working on Wikidata Items. These groups are slightly larger than average. Their members are both human and bots and contribute in a balanced proportion. Human editors in these groups are likely to have diverse levels of experience and interests in Wikidata.

Chapter 11

Who models the world?

Ontology quality and user roles



In the previous chapter we analysed to what extent the combination of editors with different characteristics—bot vs. human, novices vs. experienced—influence outcome quality when it comes to building an Item. This chapter adopts another perspective, looking at the conceptual structure created by several Items, namely at what we called the ontology of Wikidata (see Section 5.2). In particular, we investigate the quality of the ontology and examine the impact of editors with various activity patterns on that. The research question we pose is

RQ3 What features of editing roles affect the quality of the Wikidata ontology?

In order to address this question, we carried out three related studies, each with their own methods: **Study 1**, to define a quality framework suitable for Wikidata based on existing literature, which we applied to evaluate the quality of the ontology over time; **Study 2**, to identify user roles based on their activity patterns, using the data from Section 10.3.4; the results of Study 1—a measure of the Wikidata ontology quality over time—and Study 2—roles and editing patterns of each user for every month—were linked together in **Study 3**, where we explored how specific user roles trigger changes in the quality of the Wikidata ontology. The findings of the first two studies were used to define hypotheses for the third. In the following, we sum up our prior observations

around the Wikidata ontology. After that, we describe the data and approaches used in the three studies. Finally, we present and discuss our findings.

11.1 The Wikidata ontology

As we have seen in Chapter 5, Wikidata does not have a predefined formal ontology. The ontology is loosely defined by the relationships between the Items and Properties in the graph and does not define classes as distinct from any other Item. Because of that, previous studies (e.g. [Erxleben et al. \(2014\)](#)) used Properties describing taxonomic relations (instance of (P31) and subclass of (P279)) to identify those Items that take the function of classes. A consequence of this approach is that, in an analogue fashion than the rest of the graph, Properties and classes within Wikidata can be edited by anyone. Taxonomic hierarchies may change quickly, depending on edits on Properties and on the use of P31 and P279, potentially influencing large swathes of the graph. Earlier work, e.g. [Brasileiro et al. \(2016\)](#), has already identified a number of issues commonly introduced by editors when using P31 and P279, which we have discussed in Chapter 7.

11.2 Related work

In the following we present previous approaches to identify user roles and to evaluate ontologies.

11.2.1 Ontology evaluation

Although some use the term “ontology” when talking about its conceptual structure, Wikidata does not rely on a formally-defined ontology. The work of the Wikidata users has been compared to a collaborative ontology development effort, since its loosely-defined schema is completely created and maintained bottom-up by the community ([Müller-Birn et al., 2015](#)). We include in the following an overview of ontology evaluation methods and approaches.

Assessing the quality of an ontology can follow various approaches ([Brank et al., 2005](#); [Hlomani and Stacey, 2014](#); [Vrandečić, 2010](#)). Choosing the most suitable one depends on context, including the purpose of the assessment and the available data ([Brank et al., 2005](#)). A basic distinction can be made in respect to the type of evaluation: *ontology validation* checks whether the ontology is fit for purpose and meets its requirements, whereas *ontology verification* focuses on the process (i.e. whether the ontology was built correctly according to some formalised process) [Vrandečić \(2010\)](#). Considering the informal, bottom-up fashion in which Wikidata is built, ontology validation is the

most relevant for our case. It can refer to various layers of an ontology: its labels (or vocabulary), its syntax (the correct use of a format), its semantics (whether the modelling is appropriate), and many more (Vrandečić, 2010).

Brank et al. (2005) identify four types of approaches to assess an ontology: (i.) comparing it to a ‘golden standard’; (ii.) gauging its fitness for a determined task; (iii.) comparing its knowledge representation to the information extracted from related data; (iv.) manually evaluating it. These approaches may be applied either to an ontology as a whole, or to one or more of the following six levels or layers (Brank et al., 2005): the *lexical, vocabulary, or data layer* refers to the concepts included in the ontology and how, i.e. with which terms, they have been represented. The *taxonomy layer* is related to the hierarchical structure of the concepts in the ontology, constructed by means of relationships such as ‘type of’ or ‘is-a’. The layer related to *other semantic relations*, i.e. all the relations other than those included in the previous layer. The *syntactic layer* focuses on the compliance of an ontology with the requirements of the language in which it is written. The *structure, architecture, and design layer*, concern the organisation of an ontology and its respect of some relevant criteria.

Other authors provide a different description of the layers that can be evaluated. For example, Vrandečić (2010) proposes methods to evaluate the vocabulary, the syntax, the structure, the semantics, the representation, or the context of an ontology. The *vocabulary* of an ontology refers to all the names contained in it, be it URIs or literals. The *syntax* concerns the format in which an ontology has been serialised and its conformance to its norms. As regards the *structure*, it is the most commonly approach to evaluate ontologies and it consists in measuring different aspects of an ontology graph. *Semantics* measure the knowledge representation models that are described by an ontology, whilst *representation* concern the relation between semantics and structure. *Context* refers to the applications accompanying an ontology (Vrandečić, 2010).

11.2.2 User Roles

A variety of approaches has been attempted to identify user roles in online platforms. Some studies focus on functional roles, determined by the organisational functions fulfilled by users. This is the case of Arazy et al. (2015), who analyse the activity profiles of functional roles in Wikipedia and their evolution over time. Access privileges (e.g. administrators) are manually mapped to functional roles. Patterns are identified for each of these activity by estimating the percentage of edits in each Wikipedia namespace made by users in that role. The transitions between roles are studied under the lens of the Reader-to-Leader framework (Preece and Shneiderman, 2009). Other works identify roles from emergent activity patterns within the platform examined. Walk et al. (2015) look at sequences of actions in ontology engineering projects with the aim of predicting what type of edit a user would perform next. The sequences are modelled by

means of the sequential pattern mining algorithm PrefixSpan (Pei et al., 2001), which finds sequences and their occurrences in a dataset. Moreover, in order to predict editors' actions Walk et al. (2015) use higher-order Markov chains. These models predict a state given k previous states, in contrast with first-order Markov chains, which predict the next state based only on the current. Two approaches are followed to determine k : Bayesian model selection, which penalises higher-order models and reduce the chances of overfitting (Claeskens et al., 2008), and cross-fold validation.

A more common method to identify emerging roles is by applying a clustering algorithm to user activity vectors. Typically, k -means is the algorithm of choice. This is the approach followed, among others, by Liu and Ram (2011), Arazy et al. (2016), Falconer et al. (2011), and Müller-Birn et al. (2015)—these in a study about Wikidata. In these studies, each user is represented as a vector which includes the counts of all the distinct types of actions he/she carried out. These types may be either derived by the researchers from an analysis of the editing activity on the platform, this is the approach followed by Liu and Ram (2011), Falconer et al. (2011), and Müller-Birn et al. (2015). Others rely on automated approaches, such as machine learning classifiers to categorise activities, e.g. Arazy et al. (2016). Some works build user vectors including actions across the whole lifetime of a project, Falconer et al. (2011) for instance, thus not considering possible variations in a contributor's activity patterns. On the other hand, other studies look at emergent roles over time or across different work units, e.g. pages or articles. Concerning the latter, Liu and Ram (2011) count edits by type for each contributor per Wikipedia page in their sample. Müller-Birn et al. (2015) analyse role transformations over time, hence collecting user activity vectors for each monthly timeframe in the lifespan of Wikidata. These perspective may be combined: this is done by Arazy et al. (2016), as they aim to understand role stability across different stages of the evolution of online platform. Hence, they divide Wikipedia's lifespan into two periods and build user activity vectors for each page in these. All the works cited above perform some sort of normalisation on the vectors, e.g. by dividing the count of each action type by the total count of all actions per user. Several studies, such as Liu and Ram (2011), leave out users with a number of contributions under a determined threshold, e.g. five, in order not to include so-called 'casual contributors' in their dataset. Other works adopt an inclusive approach, e.g. (Arazy et al., 2016), keeping all users regardless from their activity level, in order to model 'marginal profiles', such as vandals or occasional contributors.

Unsupervised clustering algorithm, such as k -means, require to determine the optimal number of clusters, i.e. k . For this purpose, a large number of strategies has been devised, each with different advantages and disadvantages. Liu and Ram (2011), Falconer et al. (2011), and Arazy et al. (2016) use measures of cluster *compactness* and *separation*, combined into an *Optimal Cluster Quality* (OCQ) measure, to find k . The rationale of this approach is that since each approach by itself is characterised by a different type

of bias, either towards larger or smaller k , two complementary measures such as within-cluster compactness and between-clusters separation can be combined to balance their respective biases and select an optimal k (Handl et al., 2005). However, other measures have outperformed these combined approaches. One of these is the *gap statistic*, which compares a measure of variance with a null reference distribution of the data (Tibshirani et al., 2001). We use this metric in Study 2.

11.3 Data and methods

11.3.1 Data

For the studies reported in this chapter, we used the Wikidata history dumps updated at 1st October 2017, processed as described in Section 8.4. The ontology graph was built on the basis of the P31 (instance of) and P279 (subclass of) Properties. Qualifiers were not included in the analysis, similarly to Erxleben et al. (2014). To examine the evolution of user activity and of ontology quality over time, we extracted monthly slices of the data and collected all variables for each slice. P31 was created in early February 2013, while P279 dates back from early March 2013. Hence, the first slice in our dataset is March 2013, the last is September 2017, for a total of 55 slices.

11.4 Study 1 - Ontology quality

Following the requirements outlined in Chapter 8, we set an evaluation framework to gauge the quality of the Wikidata ontology. The quality indicators included in the framework had to (i.) cover structural aspects of quality, which could be influenced by Wikidata editors (R1), (ii.) be able to be observed over time without requiring external tools (R2, R3), and (iii.) be able to be evaluated automatically (R4).

Based on these requirements, we evaluated a selection of previous ontology quality frameworks, collected through a survey of prior research in the field. We performed a selection of previous ontology quality frameworks through a survey of prior research in the field. The approaches in Orme et al. (2006) and Tello and Gómez-Pérez (2004) were not suitable for our purposes, as the first focused on measuring the level of connectedness between ontology pairs and the second used a system of subjective ratings which was not compatible with our requirements. The other frameworks shared a number of metrics related to breadth, depth, and fitness of use compared to the rest of the knowledge graph, e.g. number of classes, average sub-class hierarchy depth, and average number of instances per class. For the purpose of our analysis, we picked a set of 14 indicators to build our Wikidata ontology quality framework, primarily from Sicilia et al. (2012). The metrics in this work covered a wider range of aspects than those in Yao et al. (2005)

Framework	R1	R2	R3	R4
OntoMetric (Tello and Gómez-Pérez, 2004)	No	Partially	No	Partially
Gangemi et al. (Gangemi et al., 2006)	Partially	Yes	No	Partially
OntoQA (Tartir and Arpinar, 2007)	Yes	Yes	Partially	Yes
Orme et al. (Orme et al., 2006)	No	Yes	No	Yes
Sicilia et al. (Sicilia et al., 2012)	Yes	Yes	Partially	Yes
Yang et al. (Zhe et al., 2006)	Yes	Yes	Yes	Yes
Yao et al. (Yao et al., 2005)	Yes	Yes	Yes	Yes

TABLE 11.1: Ontology metric frameworks evaluation against the requirements set in the present study

and Yu et al. (2007), while being suitable to be implemented on Wikidata, which lacks the stringent logical definitions typical of formally-defined ontologies. For each of the metrics in our set, we adjusted them to capture the ontology model of Wikidata.

Our Wikidata ontology quality framework is shown in Table 11.2. The features counting the number of classes (*noc*, *norc*, *nolc*) and Properties (*nop*) measure the ontology size. *norc* is a count of the classes that have sub-classes, but no explicitly stated super-class. It can assume values between 1 and values very close to $|C|$, where higher values suggest more diverse knowledge in the ontology (Sicilia et al., 2012). *nolc* refers to the set of classes with no sub-classes. *noc* and *nolc* increasing at a pace comparable to that of the total of the Items in the graph would be a sign of a growing number of Items erroneously treated as classes, confirming prior findings around the misuse of P31 and P279. A slower growth would indicate that the taxonomy is actively and successfully maintained. Experienced users focus on this task according to Piscopo et al. (2017c). *noi* counts the Items for which type information is specified, i.e. which are subject of P31 statements, and ideally should be equal or close to the total of Items.

The average population of an ontology *ap* measures how instances are distributed across the classes of the ontology. Together with class richness (*cr*), it provides an indication of whether the information in the ontology is sufficient to describe the data (Sicilia et al., 2012). Low *cr* and *ap* may mean that the instances in the KG do not represent all the knowledge in the ontology (Tartir and Arpinar, 2007). This seems unlikely in Wikidata, where the ontology is not built separately from the rest of the graph, and could hence be attributed to other factors, such as the misuse of taxonomic Properties. Stable or increasing *cr* and *ap* may be a sign of successful efforts to maintain the ontology by a part of the Wikidata community, whereas decreasing values would point to empty classes continuously added without sufficient quality checks. Relationship Richness or *rr* is the ratio between the overall number of relations of the entities in an ontology, divided by the sum of the sub-class relations plus the overall relations. Values closer to 1 are characteristic of rich ontologies, whereas lower values indicate ontologies containing mostly

Indicator	Description	Feature
Number of instances ($ I $)	Items used as subject of P31 but not as an object and note connected to any other Items through P279.	<i>noi</i>
Number of classes ($ C $)	Items connected to at least another Item through P31 (as an object) or P279 (as a subject or object).	<i>noc</i>
Number of root classes	Classes for which no super-class exists ($ C_i , \neg \exists C_j C_i \not\subseteq C_j$).	<i>norc</i>
Number of leaf classes	Classes that have at least a super-class for which no sub-class exists ($ C_i , \neg \exists C_j C_j \not\subseteq C_i$).	<i>nolc</i>
Number of Properties ($ P $)	Possible relations between Items.	<i>nop</i>
Population	Number of instances per class.	<i>ap</i> (average)
		<i>mp</i> (median)
Class richness	Ratio between classes with instances and all instances ($\frac{ C' }{ C }$).	<i>cr</i>
Inheritance richness	Number of sub-classes per class.	<i>ir</i> (average)
		<i>mir</i> (median)
Relationship richness	Ratio between number of relations of classes except sub-class relations and all relations ($\frac{ P }{ SC + P }$).	<i>rr</i>
Class hierarchy depth	Explicit depth of class hierarchy. Class hierarchies are formed by chains of sub-class relations.	<i>ad</i> (average)
		<i>md</i> (median)
		<i>maxd</i> (max)

TABLE 11.2: Wikidata quality indicators (Sicilia et al., 2012) used in the present analysis

taxonomic relations (Sicilia et al., 2012). Inheritance richness (*ir*) is used to understand how knowledge is distributed across the different branches and levels of the ontology. It measures the average number of sub-classes per class. High *ir* values would correspond to a shallower ontology, with classes that tend to represent more general knowledge, whereas lower values typically reflect very specialised, vertical ontologies (Sicilia et al., 2012). Hierarchy depth metrics (*ad*, *md*, and *maxd*) describe the length of (explicit, as opposed to automatically inferred) sub-class relations paths in the taxonomy (Vrandečić, 2010). Deeper ontologies are often seen as more reliable (Fernández et al., 2009), although they may result in being less understandable and usable (Gangemi et al., 2006). Continuously increasing ontology depth and *ir* may point to an ontology becoming too specialised and overly convoluted. We may expect the Wikidata ontology to have unequal depth, with low average depth and high *maxd* values increasing over time. The set from Sicilia et al. (2012) included also the OntoRank metric. We did not include that, as it could not be applied to Wikidata.

Feature	Description	Feature	Description
# edits	Total number of edits in a month.	# Property edits	Total number of edits on Properties in a month.
# ontology edits	Number of edits on classes.	# taxonomy edits	Number of edits on P31 and P279 statements.
# discussion edits	Number of edits on talk pages.	<i>p</i> batch edits	Number of edits done through semi-automated tools.
# modifying edits	Number of revisions on previously existing statements.	Item diversity	Proportion between number of edits and number of items edited.
admin	True if user in an admin user group, false otherwise.	lower admin	True if user in a user group with enhanced user rights, false otherwise.

TABLE 11.3: Features used to cluster users

11.5 Study 2 - User roles

K has been used in several studies to identify emerging user roles (Falconer et al., 2011; Liu and Ram, 2011; Müller-Birn et al., 2015). In our case, it helped us cluster human registered editors according to several features (Table 11.3). The choice of features was informed by prior studies about community dynamics in Wikidata and other platforms. According to what reported by Piscopo et al. (2017c), established users perform more revisions (*# edits*), undertake more often tasks related to the maintenance of the ontology (*# ontology edits*, *# taxonomy edits*, and *# property edits*), patrol the graph to correct errors and uphold quality (*# modifying edits*), and interact more with the community (*# discussion edits*). Because of the varying levels of activity across months, we normalised these variables by their monthly totals, in order to make them comparable across different time frames. Experienced editors are also more likely to carry out more revisions through semi-automated tools, therefore we included the proportion of these types of edits (*p batch edits*). Moreover, the activity of Wikidatians vary also in terms of focusing on larger or smaller ranges of Items (Piscopo et al., 2017b). This is taken into account by the *Item diversity* feature. Finally, formal roles have been observed to be connected to functional roles in other platforms such as Wikipedia (Arazy et al., 2015). We added two binary variables indicating whether a contributor belonged to an administrative user group (e.g. *bureaucrats*, *administrators*, or *stewards*) or to any other user group with extended user rights (e.g. *Property creators* or *rollbackers*) in a determined time frame.

Feature	Value	Feature	Value	Feature	Value
Total Items	38,621,989	<i>nop</i>	3589	<i>rr</i>	0.9
<i>noi</i>	32,858,649	<i>ap</i>	23.6	<i>cr</i>	0.04
<i>noc</i>	1,465,639	<i>mp</i>	0 (0; 0)	<i>ad</i>	46.3
<i>nolc</i>	1,349,963	<i>ir</i>	1.9	<i>md</i>	51 (35;61)
<i>norc</i>	26,265	<i>mir</i>	0 (0; 0)	<i>maxd</i>	96

TABLE 11.4: Ontology metrics figures at 1 October 2017. In brackets, 25th and 75th percentiles.

11.6 Study 3 - Relationship between user roles and ontology quality

We used a lagged multiple regression model to predict changes in an ontology metric between two points in time $metricT_n - metricT_{n-1}$, as described in Chapter 8. The metrics chosen for the dependent variables, computed as $metricT_n - metricT_{n-1}$, represented ontology breadth (*noc*, *norc*, *nolc*), depth (*ad*), and distribution of instances and classes (*ir* and *ap*). Number of edits made by leaders and by contributors were used as independent variables, as they quantified the contribution by each of these editor groups. We controlled for various variables. Besides the initial value of the dependent variable ($metricT_{n-1}$) we added to each model the numbers of bot and anonymous edits, which have been shown to influence the quality of Wikidata (previous chapter and [Piscopo et al. \(2017b\)](#)). Independent and control variables had different scales and were standardised to have mean 0.

11.7 Results

11.7.1 Study 1 - Ontology quality

We implemented the indicators based on the specification from Table 11.2 and computed their values on the data from Section 11.3.1. Summary statistics about the current state of the graph are presented in Table 11.4. The Wikidata ontology has roughly 1.5 million classes, about 2000 times larger than the DBpedia ontology and three times than YAGO ([Paulheim, 2017](#)) (see also Chapter 3). With regard to Properties, Wikidata enables the expression of a vast number of relations, exceeding both that of DBpedia and YAGO (Section 3.4). *noi* is smaller than the total number of Items, meaning that a large number of entities have no defined type information. Concerning other indicators, the differences between mean and median values suggest that instances (*ap* and *mp*) and sub-classes (*ir* and *mir*) are unevenly distributed across the ontology. As regards the evolution of the ontology, all indicator series were tested for significance using a 1-sample *t*-test against the null hypotheses that no trends took place over the

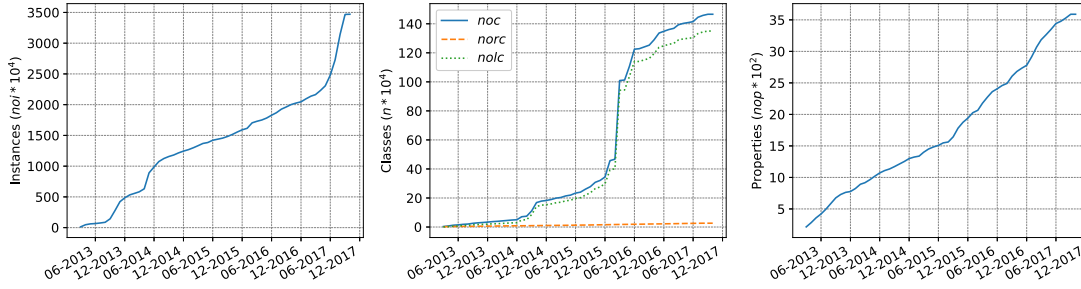


FIGURE 11.1: Evolution of number of entities, classes, and Properties in Wikidata over time

time span observed. All values in Figures 11.1, 11.2, and 11.3 presented a significant trend, except average population (*ap*), max depth (*maxd*), and the medians, whose value was zero. A very large number of classes has no sub-classes at all. This trend is visible since the early months of Wikidata. The totals of classes (*noc*) and leaf classes (*nolc*) increase in a similar fashion, both presenting a spike at the beginning of 2016, and reaching comparable values. By contrast, root classes (*norc*) increase slowly but constantly, reaching a maximum value of ~ 26 thousand (Figure 11.1). Most other indicators tell a similar story. Both population (*ap*) and class richness (*cr*) decrease in the time span considered, suggesting that several classes are either without instances or sub-classes or both (Table 11.2). The median number of sub-classes (*mir*) is stable around zero. If we look at variables related to the size of the taxonomic hierarchy, such as inheritance richness (*ir*) and max depth (*maxd*), it seems that a part of the Wikidata ontology is distributed vertically. *maxd* increases over time, reaching values higher than 80 (Figure 11.3), i.e. hierarchies with more than 80 levels. Finally, Wikidata classes seem to be well defined, with several relations besides P279, which can be seen in the rapid increase of relationship richness (*rr*).

11.7.2 Study 2 - User roles

We collected a total of 783,604 user/timeframes, corresponding to 190,765 unique human registered editors working over 55 months. The average time people contribute to Wikidata is four months (median 1, interquartile range 1,3). 119,943 editors were active for only one month, 28,600 had contributions across more than five months and around 18,000 across more than ten. 143 performed edits throughout the entire time span considered.

We tested values for k between 2 and 8. The maximum value was chosen in order to keep the number of clusters manageable. The gap statistic returned an optimal $k = 2$. An independent t -test across the two clusters showed that the mean of each variable differed significantly, with the exception of *admin* and *lower admin*. We called these roles *contributor* and *leader*. The large majority of user/timeframes (771,044) fell in

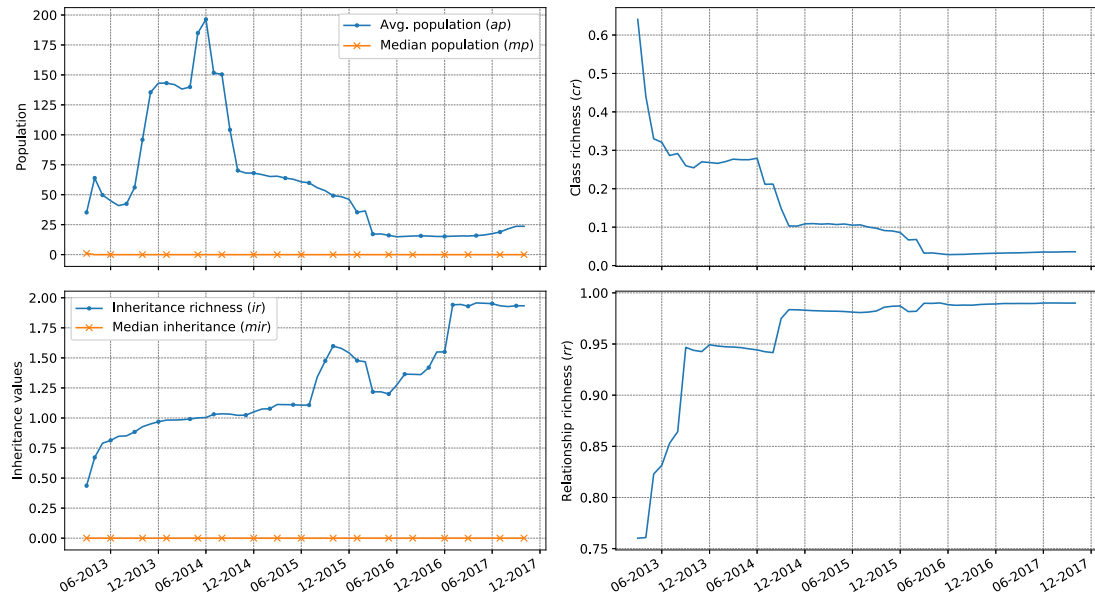


FIGURE 11.2: Wikidata quality assessment

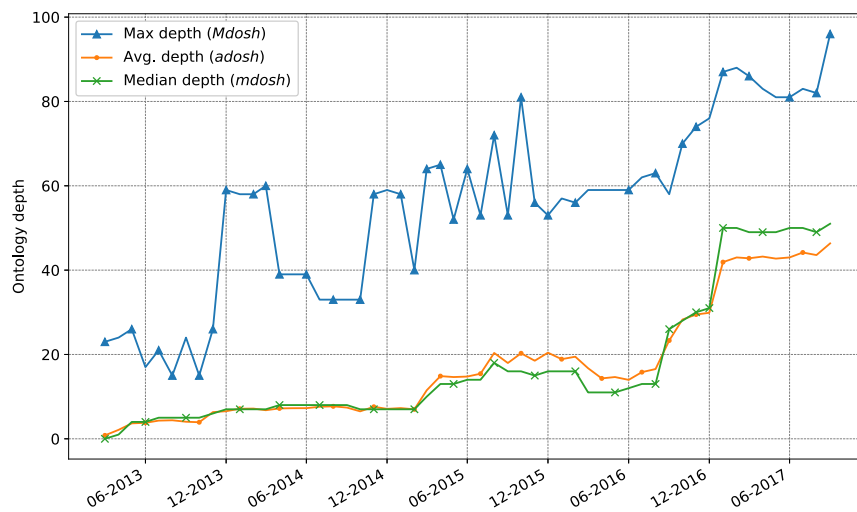


FIGURE 11.3: Ontology depth

the first role, characterised by lower levels of activity for all variables. On average, contributors perform a lower number of edits, preferably without semi-automated editing tools, and participate less in community discussions. Leaders have more sustained editing activity and deeper engagement in community pages. They devote more effort to revising Properties and taxonomic relations (P31 and P279 statements). Besides, they are more active in modifying previously added content and edit more often through semi-automated tools.

It is important to note users can morph into different roles over time, although only a minority of editors ever moved from a contributor to a leader behaviour (Table 11.5).

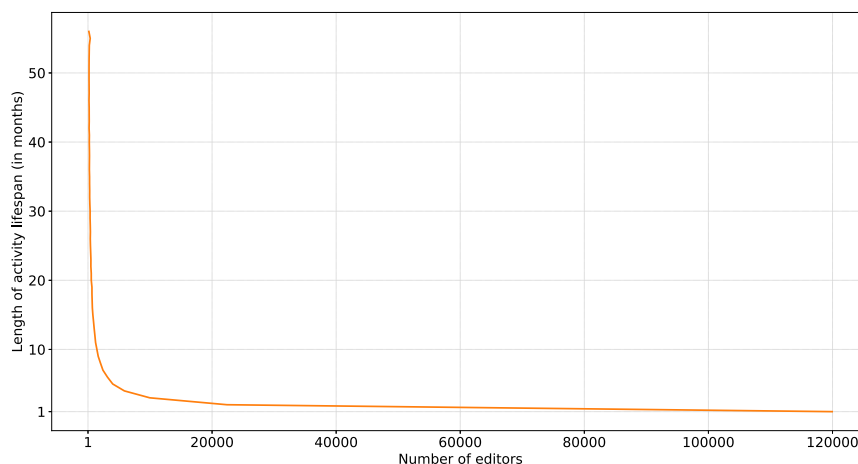


FIGURE 11.4: Number of editors by months of activity on Wikidata.

Only in 0.4% of the timeframes users continued to act as leaders across more than one month.

Editors with at least one month in the leader role had on average a higher activity lifespan on the platform (~ 9 months; interquartile range 1, 11) than editors without (~ 4 months; interquartile range 1, 2). A Mann-Whitney test confirmed that the difference between these two groups is significant ($p < 0.05$). Furthermore, editors' behaviour varies depending on the year they started contributing to Wikidata, as noted in Chapter 6. Early joining users perform a larger number of edits along all the timespans observed. We analysed yearly cohorts also with respect to the user roles studied in this chapter. For all cohorts, the percentage of leaders peaks initially, followed by a drop after around a year. This is less pronounced for 2012–2013 editors (Figure 11.5.d), but it must be considered that our analysis starts on March 2013, leaving out the first months of their activity.

	to	
	contributor	leader
from	contributor	579,786
	leader	5252

TABLE 11.5: Role transition counts

first joined	# users	# leaders	$p(\text{leader} \text{first joined})$
Oct. 2012–Sep. 2013	26,492	1540	0.058
Oct. 2013–Sep. 2014	38,679	1159	0.029
Oct. 2014–Sep. 2015	39,318	1430	0.036
Oct. 2015–Sep. 2016	43,714	1796	0.041
Oct. 2016–Sep. 2017	42,562	2187	0.051
Total	190,765	8112	0.042

TABLE 11.6: Breakdown of users and leaders by yearly cohort. By leader, we refer to anyone who has taken on a leader role in at least one time frame.

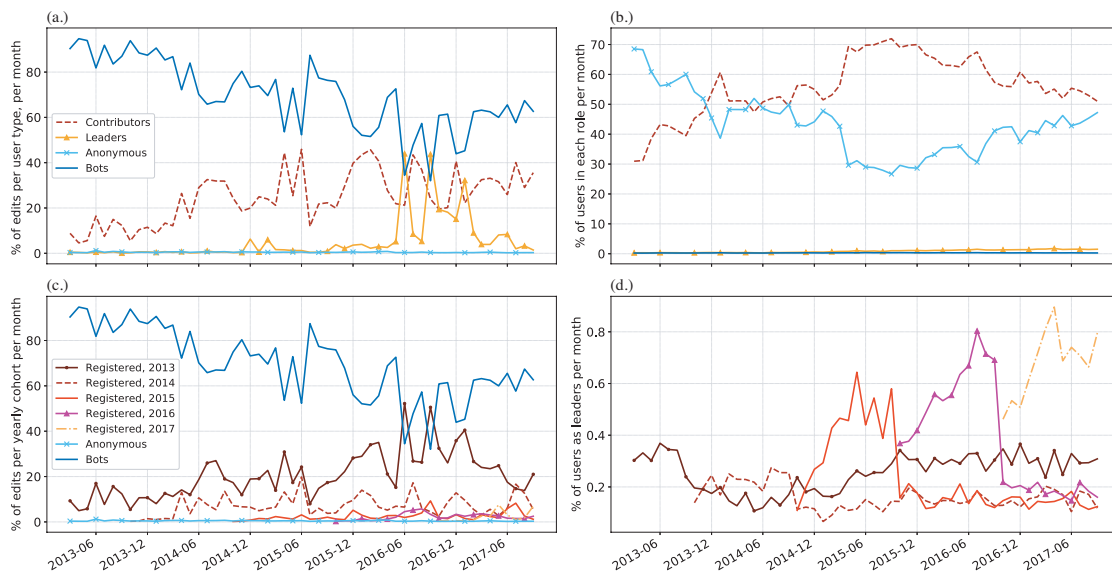


FIGURE 11.5: Proportion of contributions per user type and by yearly cohort over time and percentage of users per type. The count of anonymous users considers unique IP addresses, as these users are only known through them. Nothing prevents editors to connect from different addresses, though. Years in (c.) refer to the period between October of the previous year and September of the following (e.g. 2013 means Oct. 2012–Sep. 2013).

11.7.3 Study 3 - User profiles and ontology quality

Study 3 concerns the investigation of the influence of each Wikidata user role on the quality of the Wikidata ontology. Prior work about Wikidata has highlighted different sets of users that take on quality control tasks concerning the conceptual structure of the graph. Experienced editors report in [Piscopo et al. \(2017c\)](#) to feel as part of their duties the maintenance and cleaning of the ontology, compared to when they were novices on the platform. [Müller-Birn et al. \(2015\)](#) identify two roles that are mainly focused on creating and editing Properties. Similar roles have been identified also in collaborative ontology development projects ([Falconer et al., 2011](#)), where some editors have been shown to work primarily on organisational and hierarchy-cleaning tasks.

In the second study, we have identified two roles that Wikidata editors may take: contributor and leader. Leaders perform more higher-responsibility tasks, including quality control and maintenance of the ontology. Several users change between the two roles over time, although only a minority ever shows a leader activity pattern. Based on the studies cited above, we assume that these users are more familiar with quality issues around the Wikidata ontology and their work on the ontology improves its quality, contrasting harmful behaviour from other users, deleting incorrect P31/P279 statements and working as ‘ontology curators’, e.g. creating new hierarchies and replacing instance of with subclass of relations (and vice versa) when needed. This results in two hypotheses:

Hypothesis 1 Higher levels of leader activity are negatively correlated to number of classes (*noc*), number of root classes (*norc*), and number of leaf classes (*nolc*).

	<i>noc</i>			<i>norc</i>			<i>nolc</i>		
	Coef.	SE	<i>p</i>	Coef.	SE	<i>p</i>	Coef.	SE	<i>p</i>
Intercept	0.002	9927	*	477.5	38.50	**	0.002	9840	
Initial <i>noc—norc—nolc</i>	0.001	0.002		−96.06	98.21		0.001	0.002	
# contributor edits	−6054	0.001		42.69	93.55		−7533	0.001	
# leader edits	−2968	0.001		65.84	49.69		−4592	0.001	
# anonymous edits	0.002	0.001		−0.956	46.44		0.002	0.001	
# bot edits	−0.001	0.001		19.36	41.76		−0.001	0.001	
	<i>ir</i>			<i>ap</i>			<i>ad</i>		
	Coef.	SE	<i>p</i>	Coef.	SE	<i>p</i>	Coef.	SE	<i>p</i>
Intercept	0.035	0.012		−0.430	2.087		0.842	0.260	***
Initial <i>ir—ap—ad</i>	−0.004	0.024		1.826	2.760		0.270	0.552	
# contributor edits	−0.031	0.024		−3.316	2.767		−0.251	0.544	
# leader edits	0.038	0.13	**	−3.316	2.395		1.325	0.301	***
# anonymous edits	−0.021	0.014		−1.034	2.571		−0.400	0.318	
# bot edits	0.039	0.013	**	5.727	2.267	*	0.348	0.293	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

TABLE 11.7: Lagged regression analysis of proportion of activity of each user type on *noc*, *norc*, *nolc*, *ir*, *ap*, and *ad*

Hypothesis 2 Higher levels of leader activity are positively correlated to inheritance richness (*ir*), average population (*ap*), and average depth (*ad*).

The lagged regression analysis did not show any significant relation between *noc*, *norc*, *nolc*, and leader activity levels (Table 11.7). **Hypothesis 1** was thus **not supported**. **Hypothesis 2** was **partially supported**. Leader edits were significantly and positively associated with hierarchy depth (*ad*) and inheritance richness (*ir*). The latter, as well as average population (*ap*), were also positively related to bot edits.

11.8 Discussion

11.8.1 Ontology quality

Different dynamics emerge from our evaluation of the Wikidata ontology. First, it grows over time, in terms of numbers of classes (*noc*), leaf classes (*nolc*), and Properties (*nop*). Second, it also appears to be unevenly maintained and more so as the knowledge graph grows. The combined dynamics of class richness (*cr*) and average population (*ap*), simultaneously dropping in June 2014 and both dwindling afterwards, together with the difference between *ap* and the median number of instances per class *mp* (Figure 11.2) suggest that an increasingly high number of empty classes exists next to a vastly populated core of the ontology, with deep taxonomic hierarchies, as average (*ad*) and max depth (*maxd*) show. Whereas a share of the ontology growth may be attributed to erroneous revisions such as those described in Chapter 6 and discussed in Brasileiro et al.

(2016) and [Piscopo et al. \(2017c\)](#), these are likely to be not the only reason behind the trends observed. An examination of the peaks in [Figures 11.1](#) and [11.2](#) revealed that a single bot (MicrobeBot) added ~ 400 thousand classes in March 2016, causing a sudden rise of *noc* and *nolc*. Two bots (SuccuBot and PLBot) are responsible for June 2014's peak in *ap*, which subsequently falls due to an increase in the number of classes. These mass revisions are not necessarily incorrect. Bots are often programmed to carry out edits within a particular domain, creating new statements by extracting information from a source. MicrobeBot created numerous new classes by adding sub-classes statements to as many protein and gene Items. These edits are not formally incorrect, yet they are questionable under a knowledge engineering point of view, as it is highly unlikely that instances will ever added to these classes. Describing a concept as a class or an instance determines the properties and attributes that can be applied to it and it is an essential part in building an ontology ([Noy et al., 2001](#)). Prior literature has discussed the tension between freedom and standardisation in projects that aim to collaboratively produce structured data ([Hall et al., 2017](#)). The policies of Wikidata put virtually no restrictions on the edits users can make. This liberal approach may potentially lead to imbalances in the conceptual knowledge of the graph such as those identified. Quality metrics such as those from our framework can address this issue, by monitoring the evolution of the ontology.

Quality metrics. The metrics used in Study 1 have been variously related to different characteristics of ontologies. Deeper ontologies—Wikidata's taxonomic tree has an average depth of 40 sub-class relation—are likely to be harder to understand and to be manipulated by users ([Gangemi et al., 2006](#)), although this characteristic has been associated with more reliable semantic content ([Fernández et al., 2009](#)). Others have found connections between ontology accuracy, inheritance (*ir*), and relationship richness (*rr*) ([Lantow, 2016](#)). Nevertheless, our quality framework is limited in expressiveness. While the structural indicators considered work well with the observational data provided by Wikidata and are able to illustrate trends over time, they are hardly comparable across ontologies and do not provide any direct insight into the correctness of the conceptualisation ([Vrandečić, 2010](#)). To deal with the first, we could consider normalised indicators ([Vrandečić, 2010](#)). For the second, we could try to detect inconsistencies, either by inspecting samples of the class hierarchy ([Völker et al., 2005](#)) or by using reasoning software—however, the size of the ontology makes both tasks extremely challenging for state of the art tools. Further on, our framework does not consider cultural aspects of Wikidata. A comparison between Wikipedia articles has found that different language versions overlap only to a very small extent, with regard to the topics they cover ([Hecht and Gergle, 2010](#)). A similar issue was noted in Open Street Map ([Hall et al., 2017](#)), where users often found the knowledge organisations of editors with different cultural backgrounds not meeting their specific needs. Our framework could be extended to include indicators that consider structural changes for different

language spaces as well. To the best of our knowledge, the extent to which Wikidata editors are active beyond their specific countries or cultural areas has not been studied. A sustained activity across cultures might be a proof of the success of the multilingual nature of Wikidata and should be the object of future studies.

To sum up, the metrics computed provide only a partial picture of the quality of the Wikidata ontology. Yet, it is an important part. First, our findings may be a starting point for future studies that want to explore differences in quality between domains in Wikidata conceptual knowledge. Second, the information provided by our metrics may be used to test future design solutions. For example, measures to make the effects of any edit on the hierarchy may be adopted to address the misuse of taxonomic relations, following a suggestion regarding collaborative ontology development contexts in [Vigo et al. \(2015\)](#). An analysis of the metrics selected in this work may be subsequently used to assess the success of such approach.

11.8.2 User roles

We detected two distinct user roles in Wikidata: leaders and contributors. Compared to the latter, the first perform on average more edits on classes, carry out more maintenance work, and get more involved in community discussions. Editors move up and down between these roles. In each time frame of the period examined, leaders are only a small minority ($\sim 1\%$) and that their overall revisions are generally less than those from contributors. When this is not the case (June and August 2016, January 2017, see Figure 11.5.a), the usual proportions are reverted by single human editors who perform revisions at a high rate ($\sim 2\text{M}$ per month) through semi-automated tools. This reinforces the considerations made in Section 11.8.1 concerning the benefits of using our metrics to continuously monitor the quality of the Wikidata ontology. Early editors are more likely to take a leader role (Table 11.6), a behaviour that has been noted also among users of other platforms (e.g. Reddit ([Barbosa et al., 2016](#))) and may be explained by a self-selection bias, i.e. users arriving to Wikidata in its early stage may be on average those who are more inclined to like it ([Barbosa et al., 2016](#)). This hypothesis seems probable, as the Wikidata project had been previously discussed and advertised within the Wikipedia community. In [Piscopo et al. \(2017c\)](#) we have seen that the majority of Wikidatians were already highly active Wikipedia editors, hence highly committed to the aims of the project. Moreover, after a first drop, the percentage of early users taking a leader role remains higher than other cohorts (Figure 11.5.d), which may confirm this hypothesis. Moreover, each cohort shows an initial peak of users working as leaders a few months after joining, followed by a decline. We return on that in the next paragraphs.

Compared to prior studies on roles in Wikidata or in collaborative ontology engineering projects, the Wikidata community seems less structured. A direct comparison may not be appropriate though, since our conceptualisation is built along different dimensions

than earlier ones. Whereas the analysis in Müller-Birn et al. (2015) looks at edit types with more granularity, distinguishing between revisions on the various parts on an Item, we included features related to editors' interaction within the community and use of interfaces, i.e. semi-automated tools. Similar considerations can be done concerning research on collaborative ontology engineering roles, which relies mainly on structural aspects of the concepts edited, e.g. their depth or their function in the ontology hierarchy (Falconer et al., 2011; Walk et al., 2015).

The Reader-to-Leader (R2L) framework (Preece and Shneiderman, 2009) describes people's behaviour in online platforms and has been largely influential in the CSCW community. It is general enough to be applied to a broad range of collaborative platform, and has been widely applied and validated empirically, e.g. on Wikipedia; in addition, it explains transitions between roles. We discuss our findings through the lens of this framework and of previous empirical studies relying on that, focusing mainly on two aspects: (*i.*) the articulation of roles and their connection to administrative responsibilities; (*ii.*) the transitions between roles. Regarding the first point, the R2L framework categorises users into four roles, *reader*, *contributor*, *collaborator*, *leader*, ordered according to their centrality along a core-periphery axis. Readership may prove difficult to detect in Wikidata, as the relevant data can be accessed through APIs and tools that are much less mature than those used in this work. Contributors in R2L move their first steps on the platform, performing small edits and slowly engaging in relationships with the community (Preece and Shneiderman, 2009). On the other hand, collaborators are established members of the community, who cooperate with other participants to achieve more complex outcomes. Wikidata contributors present a mix of the features of both these two roles. Some editors perform a large number of revisions with semi-automated tools, but do not comment on talk pages; others are less active, but communicate more often with the community; other users edit less Items, but stay active for longer times. This heterogeneous behaviour, which characterises also leaders, stands out when compared e.g. to the study of Arazy et al. (2015), whose analysis of Wikipedia's participation dynamics, underpinned by the R2L, brings together organisational and functional features. They show that a specific activity patterns corresponded to each of the nine formal roles found in Wikipedia. On the contrary, formal roles—administrators—were not significant in distinguishing the roles detected in the current study. Wikidata is still an emerging project and its community structure might still be evolving to more defined patterns of activity. However, there is evidence from our analysis that its evolution has so far been distinct from Wikipedia, where the administrators' share of edits increased in the first years (Kittur et al., 2007a). This rise is yet to happen in Wikidata, if it will happen at all. Moreover, the normative burden of rules and policies in Wikidata is still lower than in Wikipedia (Piscopo et al., 2017c). A simpler bureaucracy may require fewer declared roles, which in turn changes how and how much people participate. Our findings may demonstrate that Wikidata's sociotechnical fabric diverges from prior projects in both

the areas of peer-production and collaborative knowledge engineering and may actually represent a new paradigm of collaborative system.

Under a methodological point of view, the clustering algorithm used (k -means) may not perform well with some types of clusters, e.g. non-spherical or unevenly distributed ones. Different algorithms may give different results, although each has its strengths and weaknesses. Future studies should test other approaches to compare their performances.

Regarding role transitions, in R2L they may happen in both directions from periphery to core, although mainly in a sequential fashion (Arazy et al., 2015; Preece and Shneiderman, 2009). The minority of Wikidata editors who ever work as leaders move up and down between roles, seemingly ‘stepping up’ at some point to provide greater support to the project. Several users follow the path to leadership in their first months, but yearly cohorts in Figure 11.5.d show that they often do not beat that path again. This may be a sign of declining participants’ motivation—one challenge initiatives such as Wikidata and Wikipedia may face is the lack of shorter-term, tangible goals and achieving a specific editor status might act as a proxy to them (Kraut and Resnick, 2012). Our analysis, alongside previous studies (Müller-Birn et al., 2015), could inform the definition of these ‘badges’ and help study their uptake and effects.

11.8.3 Ontology quality and user roles

Our results partially support the qualitative findings from Piscopo et al. (2017c). We did not find any significant influence of users in the contributor role on any of the metrics analysed. Leader activity is significantly related to the average number of sub-classes per class (ir) and depth of the ontology (ad), which may be interpreted as a confirmation that more seasoned users concentrate their efforts on vertically extending and consolidating the ontology to ensure that all entities in the KG are optimally represented and organised into classes. Their efforts appear to have limited effect though. Appropriate tools may be designed to help them control Wikidata’s quality. Bot edits are positively related to the average number of sub-classes per class (ir). To a certain extent this can be attributed to the introduction of several Wikidata bots, which e.g. add data from other sources into Wikidata automatically. What is surprising is the lack of any substantial influence on the total number of classes (noc), reinforcing the impression that Wikidata user dynamics differ from those observed in prior collaborative ontology engineering projects.

11.9 Summary

This chapter has addressed RQ3 by evaluating the structural quality of the Wikidata ontology over time and investigating the community dynamics that influence it. We

have articulated our contribution along three studies. First, we have devised a quality framework and applied a set of indicators to assess the Wikidata ontology from its early days until September 2017. Second, we have identified roles for Wikidata editors according to their activity patterns. Finally, we have explored how these roles influence the quality of the ontology.

The Wikidata ontology is large and messy, with numerous underpopulated classes and uneven depth. This confirms prior literature suggesting that several Wikidata contributors fail to use correctly the taxonomic relations P31 (instance of) and P279 (subclass of). On the other hand, we found evidence suggesting that parts of the ontology have higher depth and are likely to be curated by a core of expert users. We identified two activity patterns: *contributors*, i.e. users with lower number of edits and less engaged in community discussions, and *leaders*, who are more active in all of the features considered. Only a minority of users presents a leader activity pattern sometimes during their interaction with the platform. Finally, whereas the activity of leaders seems to influence positively the depth of the ontology, no relation could be proven between any editor category and variables concerning the breadth of the ontology. Future work should explore what variables are at play with regard to that.

Part III — Summary

- The overarching research question posed by this thesis is *how does the socio-technical fabric of Wikidata influence the quality of its data?*.
- **Data** Historic dumps of Wikidata released by the Wikimedia Foundation.
- **RQ1** How do references added by bots and by humans compare with respect to their quality?
Method: Microtask crowdsourcing to evaluate relevant and authoritativeness of references + Machine Learning to enable evaluation on a large scale.
Results: > 60% of references are relevant and authoritative. Only around 40% bot-contributed references are both relevant and authoritative, whereas this percentage increases to over 70% for humans. These findings suggest the need of constant checks on the revisions and activities carried out by bots.
- **RQ2** To what extent does editor group diversity affect Item quality in Wikidata?
Method: Regression analysis to examine effect of proportion of human and bot edits, tenure diversity, and interest diversity on Item quality.
Results: Bot contributions are beneficial for quality, albeit to a lesser extent when they are not balanced with human edits. The work of anonymous users is likely to affect negatively Item quality. Tenure and interest diversity have a significant positive influence on quality.
- **RQ3** What features of editing roles affect the quality of the Wikidata ontology?
Method: (i.) literature survey to devise a suitable ontology evaluation framework for Wikidata; (ii.) clustering algorithm to identify user roles; (iii.) longitudinal regression analysis to examine influence of roles on ontology quality.
Results: The Wikidata ontology has uneven quality, according to our evaluation. A large number of classes has neither super-, nor sub-classes, nor instances, whereas a core of the ontology has deep taxonomic relations and ramified relations. As regards user roles, only two roles emerged from our analysis, *contributors* and *leaders*. The latter perform more edits, are more active within the community, and rely more often on semi-automated tools. The number of leaders' revision has a significant and positive effect on some aspects of the ontology, namely the number of sub-classes per class and its depth.

Chapter 12

Conclusions

In Part III we have addressed the research questions outlined in the Introduction and in Chapter 8. We have first looked at the quality of Wikidata external references and at how these vary according to who has authored them, i.e. bot or human. Then, we have moved to analysing human editors and in particular how the combination of users with different features, i.e. tenure and diversity of interests, affect Item quality. Finally, in Chapter 11 we have investigated the quality of the Wikidata ontology and how this is affected by roles emerging from users activity patterns. We bring together these experiments and their findings in this chapter, drawing conclusions, considering limitations, and proposing future work.

12.1 Contributions and results

In the Introduction we have sketched out the main contribution made by this thesis. We outline them again here:

1. We gauged the data quality of two aspects of Wikidata previously uncovered, i.e. provenance and its ontology;
2. We delved into algorithmic contribution patterns in Wikidata and their role in upholding a constant growth of the knowledge graph and their influence on its quality;
3. We investigated emerging activity patterns of human editors in Wikidata and the effects of these patterns on outcome quality.

We further detail these contribution in the next sections.

12.1.1 Wikidata quality

We evaluated the quality of external references and of the ontology of Wikidata.

Provenance quality Wikidata references—we sampled only from those using P854 and in English (Chapter 9)—appear to be of good quality overall. Around 60% are both relevant and authoritative, according to the requirements set by the Wikidata policy. A larger percentage of sources were authoritative (79.9%) than relevant (68.9). This was largely due to a link used several hundred times that subsequently became broken. This link was automatically added by a bot—we have seen that bots import large number of statements and their related references in large batches. Considering that each reference is edited 1.3 times on average, including the revision in which it was created, it seems like strategies to enforce tighter controls on reference quality are required. We discuss some possible approaches to that in Section 12.3. Finally, internal reference (i.e. those using P248, *stated in*) represent the majority of Wikidata references. Their quality has not been gauged yet and should be the object of future studies.

Ontology quality Our evaluation (Chapter 11) has highlighted different levels of quality in the Wikidata ontology. This grows over time, roughly at the same pace of the whole graph, suggesting that a large number of classes may have been defined as such erroneously. Some parts of the ontology present deep subsumption hierarchies and a higher number of instances per class, whereas others are mostly flat, i.e. made by classes with neither super-, nor sub-classes, nor instances. Whereas our analysis did not include any evaluation of the semantics and the consistency of the ontology, the picture we obtain through the structural metrics applied may lead to argue that a core of the ontology is actively curated by the community, although this should be verified in future studies. Moreover, the ontology quality framework we devised is *per se* a valuable contribution, as it may be used in the future to monitor the evolution of Wikidata’s conceptual structure. Due to the editing interface of Wikidata, which allows automated revisions, changes to large swathes of its taxonomic structure may be done with little effort and virtually no time. This means that the quality of the Wikidata ontology may be vulnerable to initiatives of single users or even to possible wars between editors. We do not know whether some of the ontology changes observed in our experiment may have been previously agreed with the community. Projects have been created to coordinate all efforts to model and maintain the Wikidata ontology ¹. Future work might start researching these projects in order to understand community awareness, guidance, and support concerning any action that may strongly affect the conceptual structure of Wikidata.

¹https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology, consulted on 1st May 2019.

12.1.2 Bots

Wikidata is a complex socio-technical system, where technical solutions and collaborative processes are strictly intertwined. Bots are a key element of this socio-technical system. Without their contribution, the steep growth of Wikidata would have been hardly possible. They perform the majority of revisions, importing data from and linking to other sources, creating statements and adding labels in several languages. Furthermore, they periodically scan the graph for various types of quality issues. Besides these control tasks, the influence of bots on quality is manifold. Their work is beneficial for Item quality, although their contributions need to be balanced with those by human editors, as shown in Chapter 10. The Item quality labels compiled by [Yapinus et al. \(2017\)](#) take into account various elements constituting Items, e.g. statements, references, or qualifiers. However, they consider only their completeness, without making any explicit assumption about their correctness. We explored this aspect with respect to references in Chapter 9. Although a large number of sources added by bots is non-relevant—at least, many of them become invalid with time—they are overwhelmingly authoritative. Bots (or, better, the editors who maintain them) seem to be rather well-disciplined in following the norms governing verifiability in Wikidata. This may be explained by Wikidata bot activity policy. Bots must be authorised by the community in order to perform revisions with no restrictions. The approval process requires the bot maintainer to describe in detail the actions that the bot would carry out. When these actions involve importing data from other sources into statements, these sources must be specified and included as references in the newly-created statements. Hence, the approval process of bots determines the differences in editing behaviour we have seen in Chapter 9: when a bot creates a statement, it contextually adds a reference. On the other hand, human-authored statements are often integrated with a reference at a later time. Additionally, the websites used as sources by a bot may not change, even after these have changed and are not valid anymore. Therefore, whereas the authoritativeness of the sources employed by bots is validated in advance through community approval, these may cease to be relevant and the bot continue to add them. This is what was found in Chapter 9, which suggests that simple frequent checks on the links on which bot activity relies upon may prevent the introduction of a large number of errors.

Bots determine also the shape of the Wikidata ontology. The number of their edits is significantly related to the growth of the number of classes. Any initiative to maintain and curate the quality of the Wikidata ontology in the future must take into account ways to address bursts of activity like that seen in Chapter 11. To that end, the framework we developed may be used to monitor changes in the shape and quality of the ontology; moreover, the “rules of engagement” applying to bots may be modified by introducing restrictions upon heavyweight edits—following the definition given in Section 6.1—e.g. revisions affecting Properties such as P31 or P279. Similar rules may also apply to mass-editing through semi-automated tools, albeit this might be a significant change

to Wikidata’s liberal approach to knowledge engineering and its possible effects on and approval within the community should be first investigated.

Finally, bots perform various types of activities, which may affect quality in several ways. Future work should attempt to provide a more fine-grained definition of bots’ work, differentiating these with respects to the types of tasks they carry out, along the lines of what [Clément and Guitton \(2015\)](#) did concerning Wikipedia.

12.2 Wikidatians

Human registered users constitute the backbone of the Wikidata community. Our findings are partly in line with previous research on collaborative platforms. A minority of them does the lion’s share of the work, similarly to what happens e.g. in Wikipedia ([Ortega and González-Barahona, 2007](#)) or in other collaborative ontology development projects ([Strohmaier et al., 2013](#)). Editor can focus either on a few Items, possibly around a topic that interests them, or spread their revisions over a broader range. Moreover, only a few users contribute to Wikidata for longer periods of time in some cases along the whole lifespan of the project. These users are more likely to take a *leader* role, which means to perform a larger number of edits, be more active within the community, and carry out more quality control tasks. Users joining in different times of Wikidata history seem to have different characteristics (Chapter 6 and 11). Early joining users are more likely to be in a leader role and on average have a higher monthly edit count. We explained this phenomenon by a self-selection bias, i.e. people that are most likely to be attracted by Wikidata may be those who actually find it first. This bias has been already observed in other communities, such as Reddit ([Barbosa et al., 2016](#)) or online book review communities ([Li and Hitt, 2008](#)). Wikidata belongs to the Wikimedia ecosystem, which includes several projects, Wikipedia being the largest and best known. Before its launch, Wikidata had a period of gestation, in which users from the various communities of this ecosystem had the chance to discuss the features of the new project. It is likely that enthusiast Wikipedia/other Wikimedia projects editors may have been among the first to know about Wikidata, continuing their support along all its lifespan.

More experienced editors and novices seem to be complementary in building high quality Items (Chapter 10). Notwithstanding, some of the observed differences between users with various levels of tenure may lead to a number of issues in the future. Editors who joined earlier perform a large number of monthly edits on average and are more likely to take a leader role. Further on, the average time people contribute to Wikidata was four months (Chapter 11). High levels of turnover risk to damage Wikidata. Therefore, this topic deserves to be further studied in order to understand whether it may affect the project negatively in later stages of development. Additionally, only a tiny minority

of editors act as a leader. Does this mean that Wikidata has become an oligarchy? This question should be answered by investigating how users become leaders and defining the different forms of involvement in Wikidata. Finally, the extent to which Wikidata editors are active beyond their specific countries or cultural areas has not been studied, to the best of our knowledge. A sustained activity across cultures might be a proof of the success of the multilingual nature of Wikidata and should be the object of future studies.

12.3 Design suggestions

The findings reported in the previous chapters may be used to design tools that can be used to uphold quality in Wikidata. This section provides some ideas for those.

The large-scale reference evaluation approach described in Chapter 9 performed well in detecting not-authoritative and not-relevant sources. Once trained and tested also on non-English sources, it might be deployed to Wikidata to flag possibly problematic references. Furthermore, sources using multiple times may be tested frequently, e.g. either automatically or showing them to human editors, in order to prevent large numbers of references to suddenly become not relevant.

With respect to the Wikidata ontology, warnings may be implemented to advise editors, e.g. administrators, when too many P31 (instance of) or P279 (subclass of) statements are added or modified. Under an organisational point of view, norms can be introduced to require bots special permissions to massively edit statements affecting the taxonomic structure of the graph. Another approach to help editors avoid statements that compromise the structure of the ontology may be to implement a tool that is able to show them the changes that an edit would introduce to a class hierarchy.

Finally, the findings from Chapter 10 could inform the creation of a tool to draw more diverse pools of editors to revise an Item. This could happen in form of recommendations—showing a user an Item in need of edits he/she may be interested in—or as a warning.

12.4 Limitations and future work

We report in this section some limitations of the work presented in this thesis and suggest future work to address these.

Generalisability The experiments reported in this thesis relied on Wikidata datasets updated at different dates. This might represent a treat to the generalisability of our findings, considering also the fast pace at which Wikidata has evolved in recent years.

Future studies may repeat our experiment relying on datasets that are more up to date and refer all to the same timeframe, in order to confirm the validity of our results.

Quality evaluation As mentioned earlier, only English language references have been evaluated in our experiment. However, Wikidata is by design a multilingual platform, which aims to harbour multiple points of view from diverse cultures. Future studies should seek to expand our analysis to sources from other cultural settings and languages. Still concerning the evaluation of references, the approach taken considers as authoritative all sources whose author and publisher type is likely to be it, regardless of the statement they are attached to. However, some sources may be authoritative in combination with a statement, whilst they may be not in supporting another. Future work should devise suitable approaches to address this limitation.

Whereas our ontology framework is able to detect issues affect the structure of the whole of the graph and bursts of activity that may suddenly modify it, it disregards any aspect related to its consistency. As noted in Chapter 11, approaches traditionally used to assess ontologies with respect to this aspect are likely to fail with Wikidata, due to the size of its ontology and to its lacking explicitly declared disjoint classes. Therefore, identifying a suitable approach to evaluate the consistency of Wikidata and applying it it is a matter for future research.

As regards the larger scope of Wikidata quality, this thesis has looked at its provenance and its ontology, disregarding other aspects and dimensions. Whereas some of these, e.g. accuracy, have been overlooked by the literature so far (see Chapter 7), they are key for the success of Wikidata and their relation with its underlying socio-technical processes must be investigated.

Community dynamics In the experiment in Chapter 10 we used as a measure of tenure the time elapsed between a user's first edits and the last day in our dataset. This measure might fail to successfully describe users whose last edits was made long before the end date in the dataset. To measure tenure, future work should use instead the time between the first and the last edit of an editor. Study 2 in Chapter 11 uses k -means as a clustering algorithm and the gap statistic to determine the most suitable k . However, different clustering algorithms may reflect better the underlying features of the data. It would be important to investigate the performance of other algorithms, especially in consideration of the evolving nature of Wikidata. Similarly, an analysis of various heuristics to determine k could be used to unveil different dynamics within the Wikidata community.

12.5 Final remarks and conclusions

Over this thesis, we have investigated the collaborative processes within the Wikidata community and in particular the influence of socio-technical processes on the quality of its outcome. In spite of being still early in its life, Wikidata seems set to become a key resource in the Web of Data. It is one of the first attempts, possibly the first at such a large scale, to empower a community of editors to create and maintain a large multi-purpose structured knowledge resource with little or no constraints in terms of the actions they can perform on the platform. In around six years since its launch, this attempt appears to be successful in terms of breadth, depth, and quality of the resources created. Editors have authored a large ontology without relying on any of the formalised processes normally used in collaborative ontology development projects. This ontology is uneven in quality. Large sections have no hierarchies at all, whereas others are likely to be well-curated and have deep subsumption hierarchies. Bots and other tools to automate revisions have been crucial to achieve all this, although they have sometimes led to a number of quality issues. These may be addressed by the implementation of tools to perform more frequent controls and suggest correct Properties.

Wikidata is still in its infancy, so is the literature about its collaboration processes. This thesis lays the foundations of an area of research that we foresee to grow vastly in the next years.

Appendix A

Properties not requiring a reference

These Properties have been deemed as not requiring a reference, according to the policy in [Wikidata \(2018j\)](#), for the purpose of the analysis conducted in Chapter 9. Properties have been listed with their English label.

P10 – video,	P361 – part of,
P18 – image,	P373 – Commons category,
P21 – sex or gender,	P396 – SBN author ID,
P50 – author,	P409 – NLA (Australia) ID,
P57 – director,	P443 – pronunciation audio,
P58 – screenwriter,	P480 – FilmAffinity ID,
P161 – cast member,	P496 – ORCID iD,
P162 – producer,	P497 – CBDB ID,
P213 – ISNI,	P502 – HURDAT identifier,
P214 – VIAF ID,	P503 – ISO standard,
P227 – GND ID,	P508 – BNCf Thesaurus ID,
P244 – Library of Congress authority ID,	P625 – coordinate location,
P245 – ULAN ID,	P640 – Léonore ID,
P268 – BnF ID,	P646 – Freebase ID,
P269 – SUDOC authorities ID,	P648 – Open Library ID,
P270 – CALIS ID,	P675 – Google Books ID,
P271 – CiNii author ID (books),	P691 – NKCR AUT ID,
P272 – production company,	P709 – Historic Scotland ID,
P301 – category's main topic,	P718 – Canmore ID,
P344 – director of photography,	P723 – DBNL author ID,
P345 – IMDb ID,	P724 – Internet Archive ID,
P349 – NDL Auth ID,	P727 – Europeana ID,

P745 – Low German Bibliography and Biography ID,	ID,
P760 – DPLA ID,	P1239 – ISFDB publisher ID,
P791 – International Standard Identifier for Libraries,	P1243 – International Standard Recording Code,
P792 – chapter,	P1245 – OmegaWiki Defined Meaning,
P856 – official website,	P1250 – Danish Bibliometric Research Indicator (BFI) SNOCCNO,
P947 – RSL ID (person),	P1251 – ABS ASCL code,
P971 – category combines topics,	P1252 – AUSTLANG code,
P996 – scanned file on Wikimedia Commons,	P1253 – BCU Ecrivainsvd,
P1005 – PTBNP ID,	P1254 – Slovenska biografija ID,
P1006 – National Thesaurus for Author Names ID,	P1255 – HelveticArchives ID,
P1017 – BAV ID,	P1263 – NNDB people ID,
P1043 – IDEO Job ID,	P1270 – Norwegian Register journal ID,
P1051 – PSH ID,	P1271 – Norway Database for Statistics on Higher education publisher ID,
P1052 – Portuguese Job Code CPP-2010,	P1272 – Norway Import Service and Registration Authority periodical code,
P1053 – ResearcherID,	P1273 – CANTIC-ID,
P1054 – NDL bib ID,	P1274 – ISFDB title ID,
P1058 – ERA Journal ID,	P1275 – Norway Import Service and Registration Authority publisher code,
P1187 – Dharma Drum Buddhist College person ID,	P1277 – JUFO ID,
P1188 – Dharma Drum Buddhist College place ID,	P1280 – CONOR ID,
P1208 – ISMN,	P1281 – WOEID,
P1209 – SAPPRFT ID,	P1284 – Munzinger IBA,
P1216 – National Heritage List for England number,	P1285 – Munzinger Sport number,
P1217 – Internet Broadway Database venue ID,	P1286 – Munzinger Pop ID,
P1218 – Internet Broadway Database production ID,	P1287 – KDG Komponisten der Gegenwart,
P1219 – Internet Broadway Database show ID,	P1288 – KLG Kritisches Lexikon der Gegenwartsliteratur,
P1220 – Internet Broadway Database person ID,	P1289 – Critical Dictionary of foreign contemporary literature ID,
P1232 – Linguist list code,	P1291 – Association Authors of Switzerland ID,
P1233 – ISFDB author ID,	P1292 – DNB editions,
P1234 – ISFDB publication ID,	P1293 – Royal Aero Club Aviator's Certificate ID,
P1235 – ISFDB series ID,	P1296 – Gran Enciclopedia Catalana ID,
P1238 – Swedish Football Association player	P1297 – IRS Employer Identification Number,

P1307 – Swiss parliament ID,	P1453 – catholic.ru ID,
P1309 – EGAXA ID,	P1461 – Patientplus ID,
P1310 – statement disputed by,	P1466 – WALS lect code,
P1331 – PACE member ID,	P1533 – family name identical to this given name,
P1343 – described by source,	P1560 – given name version for other gender,
P1375 – NSK ID,	P1601 – Esperantist ID,
P1385 – Enciclopedia Aoriana ID,	P1659 – see also,
P1386 – Japanese High School Code,	P1749 – Parlement & Politiek ID,
P1391 – Index Fungorum ID,	P1793 – format as a regular expression,
P1392 – ComicBookDB ID,	P1803 – Masaryk University person ID,
P1394 – Glottolog code,	P1804 – DNF film ID,
P1395 – National Cancer Institute ID,	P1846 – distribution map,
P1400 – FCC Facility ID,	P1855 – Wikidata property example,
P1415 – Oxford Dictionary of National Biography ID,	P1888 – Dictionary of Medieval Names from European Sources entry,
P1438 – Jewish Encyclopedia ID (Russian),	P2013 – Facebook ID,
P1439 – Norwegian filmography ID,	P2162 – Deutsche Ultramarathon-Vereinigung ID
P1447 – Sports-Reference.com Olympic athlete ID,	

Appendix B

English translations of the epigraphs in Parts I and III

Part I

The invisible cities ([Calvino, 1978](#))

Marco Polo describes a bridge, stone by stone. ‘But which is the stone that supports the bridge?’ Kublai Khan asks.

‘The bridge is not supported by one stone or another,’ Marco answers, ‘but by the line of the arch that they form.’

Kublai Khan remains silent, reflecting. Then he adds: ‘Why do you speak to me of the stones? It is only the arch that matters to me.’

Polo answers: ‘Without stones there is no arch.’

Part III

The library of Babel, in *Collected fictions* ([Borges, 1998](#))

This much is known: For every rational line or forthright statement there are leagues of senseless cacophony, verbal nonsense, and incoherency.

Zinc, in *The periodic table* ([Levi, 2000](#))

In order for the wheel to turn, for life to be lived, impurities are needed, and the impurities of impurities in the soil, too, as is known, if it is to be fertile. Dissension, diversity, the grain of salt and mustard are needed

Bibliography

- Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. Crowdsourcing linked data quality assessment. In *International Semantic Web Conference (2)*, volume 8219 of *Lecture Notes in Computer Science*, pages 260–276. Springer, 2013.
- B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW*, pages 261–270. ACM, 2007.
- Allan Afuah and Christopher L. Tucci. Crowdsourcing as a solution to distant search. *Academy of Management Review*, 37(3):355–375, 2012.
- Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.
- Chintan Amrit and Jos van Hillegersberg. Exploring the impact of socio-technical core-periphery structures in open source software development. *JIT*, 25(2):216–229, 2010.
- Deborah Gladstein Ancona and David F Caldwell. Demography and design: Predictors of new product team performance. *Organization science*, 3(3):321–341, 1992.
- Denise Anthony, Sean W Smith, and Timothy Williamson. Reputation and reliability in collective goods: The case of the online encyclopedia Wikipedia. *Rationality and Society*, 21(3):283–306, 2009.
- Judd Antin and Coye Cheshire. Readers are not free-riders: reading as a form of participation on Wikipedia. In *CSCW*, pages 127–130. ACM, 2010.
- Ofer Arazy, Johannes Daxenberger, Hila Lifshitz-Assaf, Oded Nov, and Iryna Gurevych. Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge coproduction. *Information Systems Research*, 27(4):792–812, 2016.
- Ofer Arazy, Hila Lifshitz-Assaf, Oded Nov, Johannes Daxenberger, Martina Balestra, and Coye Cheshire. On the “how” and “why” of emergent role behaviors in Wikipedia. In *CSCW*, pages 2039–2051. ACM, 2017.
- Ofer Arazy, Oded Nov, Raymond A. Patterson, and M. Lisa Yeo. Information quality in Wikipedia: The effects of group composition and task conflict. *J. of Management Information Systems*, 27(4):71–98, 2011.

- Ofer Arazy, Felipe Ortega, Oded Nov, M. Lisa Yeo, and Adam Balila. Functional roles and career paths in Wikipedia. In *CSCW*, pages 1092–1105. ACM, 2015.
- Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007.
- Georgia Bafoutsou and Gregoris Mentzas. Review and functional classification of collaborative systems. *Int J. Information Management*, 22(4):281–305, 2002.
- Pierre Baldi, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.
- Donald P. Ballou and Harold L. Pazer. Modeling data and process quality in multi-input, multi-output information systems. *Management science*, 31(2):150–162, 1985.
- Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M. Cesar Jr. Averaging gone wrong: Using time-aware analyses to better understand behavior. In *WWW*, pages 829–841. ACM, 2016.
- Jonathan BL Bard and Seung Y Rhee. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*, 5(3):213, 2004.
- Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):16:1–16:52, 2009.
- Gordon D. Baxter and Ian Sommerville. Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, 23(1):4–17, 2011.
- Arthur G Bedeian and Kevin W Mossholder. On the use of the coefficient of variation as a measure of diversity. *Organizational Research Methods*, 3(3):285–297, 2000.
- Ralf Bender and Ulrich Grouven. Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London*, 31(5):546–551, 1997.
- Yochai Benkler, Aaron David Shaw, and Benjamin Mako Hill. Peer production: A modality of collective intelligence. In *Handbook of Collective Intelligence*. MIT Press, 2015.
- Tim Berners-Lee. Linked data. Jul 2006. Online. Retrieved 8 April 2018, from <https://www.w3.org/DesignIssues/LinkedData.html>.

- Tim Berners-Lee, James Hendler, Ora Lassila, et al. The Semantic Web. *Scientific American*, 284(5):28–37, 2001.
- Ivan Beschastnikh, Travis Kriplean, and David W. McDonald. Wikipedian self-governance in action: Motivating the policy lens. In *ICWSM*. The AAAI Press, 2008.
- Christian Bird, Alex Gourley, Premkumar T. Devanbu, Anand Swaminathan, and Greta Hsu. Open borders? immigration in open source projects. In *MSR*, page 6. IEEE Computer Society, 2007.
- Christian Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. PhD thesis, Freie Universität Berlin, Germany, 2007.
- Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009a.
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - A crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165, 2009b.
- Kurt D. Bollacker, Robert P. Cook, and Patrick Tufts. Freebase: A shared database of structured general human knowledge. In *AAAI*, pages 1962–1963. AAAI Press, 2007.
- Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*, pages 1247–1250. ACM, 2008.
- Stephen P. Borgatti and Martin G. Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 2000.
- Jorge Luis Borges. *The library of Babel*. Penguin New York, 1998.
- Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. Network analysis of collaboration structure in Wikipedia. In *WWW*, pages 731–740. ACM, 2009.
- Janez Brank, Marko Grobelnik, and Dunja Mladenic. A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170, 2005.
- Rollin Brant. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, pages 1171–1178, 1990.
- Freddy Brasileiro, João Paulo A. Almeida, Victorio Albani de Carvalho, and Giancarlo Guizzardi. Applying a multi-level modeling theory to assess taxonomic hierarchies in Wikidata. In *WWW (Companion Volume)*, pages 975–980. ACM, 2016.
- John G. Breslin, Alexandre Passant, and Stefan Decker. *The social semantic web*. Springer, 2009.

- D. Brickley and R. V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation, 2004.
- Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP*, pages 1–10. ACM, 2005.
- Brian Butler, Lee Sproull, Sara Kiesler, and Robert Kraut. Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, 1:171–194, 2002.
- Brian S. Butler, Elisabeth Joyce, and Jacqueline Pike. Don’t look now, but we’ve created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *CHI*, pages 1101–1110. ACM, 2008.
- Italo Calvino. *Invisible cities*. Houghton Mifflin Harcourt, 1978.
- Marcelo Cataldo, James D. Herbsleb, and Kathleen M. Carley. Socio-technical congruence: a framework for assessing the impact of technical and work dependencies on software development productivity. In *ESEM*, pages 2–11. ACM, 2008.
- Jilin Chen, Yuqing Ren, and John Riedl. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *CHI*, pages 821–830. ACM, 2010.
- Timothy Chklovski and Yolanda Gil. An analysis of knowledge collected from volunteer contributors. In *AAAI*, pages 564–571. AAAI Press / The MIT Press, 2005.
- Gerda Claeskens, Nils Lid Hjort, et al. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2008.
- Kevin A Clauson, Hyla H Polen, Maged N Kamel Boulos, and Joan H Dzenowagis. Scope, completeness, and accuracy of drug information in Wikipedia. *Annals of Pharmacotherapy*, 42(12):1814–1821, 2008.
- Maxime Clément and Matthieu J. Guitton. Interacting with bots online: Users’ reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior*, 50: 66–75, 2015.
- Kevin Crowston and James Howison. The social structure of free and open source software development. *First Monday*, 10(2), 2005.
- Kevin Crowston, Kangning Wei, Qing Li, and James Howison. Core and periphery in free/libre and open source software team communications. In *HICSS*. IEEE Computer Society, 2006.

- Armando Cruz, António Correia, Hugo Paredes, Benjamim Fonseca, Leonel Morgado, and Paulo Martins. Towards an overarching classification model of CSCW and groupware: A socio-technical perspective. In *CRIWG*, volume 7493 of *Lecture Notes in Computer Science*, pages 41–56. Springer, 2012.
- Linus Dahlander and Siobhan O’Mahony. Progressing to the center: Coordinating project work. *Organization Science*, 22(4):961–979, 2011.
- Sherae L. Daniel, Ritu Agarwal, and Katherine J. Stewart. The effects of diversity in global, distributed collectives: A study of open source project success. *Information Systems Research*, 24(2):312–333, 2013.
- Fariz Darari, Simon Razniewski, Radityo Eko Prasajo, and Werner Nutt. Enabling fine-grained RDF data completeness assessment. In *ICWE*, volume 9671 of *Lecture Notes in Computer Science*, pages 170–187. Springer, 2016.
- Paul B De Laat. Governance of open source software: state of the art. *Journal of Management & Governance*, 11(2):165–177, 2007.
- Gerardine De Sanctis and R. Brent Gallupe. A foundation for the study of group decision support systems. *Management Science*, 33(5):589 – 609, 1987. ISSN 00251909.
- Jeremy Debattista, Sören Auer, and Christoph Lange. Luzzu - A framework for linked data quality assessment. In *ICSC*, pages 124–131. IEEE Computer Society, 2016.
- Vanessa Paz Dennen. Becoming a blogger: Trajectories, norms, and activities in a community of practice. *Computers in Human Behavior*, 36:350–358, 2014.
- Lara Devgan, Neil Powe, Brittony Blakey, and Martin Makary. Wiki-surgery? internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons*, 205(3):S76–S77, 2007.
- Mariangiola Dezani-Ciancaglini, Ross Horne, and Vladimiro Sassone. Tracing where and who provenance in linked data: A calculus. *Theor. Comput. Sci.*, 464:113–129, 2012.
- Martin Doerr. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75, 2003.
- Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmman, Shaohua Sun, and Wei Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *KDD*, pages 601–610. ACM, 2014.
- Maisa C. Duarte and Estevam R. Hruschka Jr. How to read the web in portuguese using the never-ending language learner’s principles. In *ISDA*, pages 162–167. IEEE, 2014.
- Maísa Cristina Duarte and Pierre Maret. Vers une instance française de NELL : chaîne TLN multilingue et modélisation d’ontologie. In *EGC*, volume E-33 of *RNTI*, pages 469–472. Éditions RNTI, 2017.

- Nicolas Ducheneaut. Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work*, 14(4):323–368, 2005.
- Paul Duguid. Limits of self-organization: Peer production and ”laws of quality”. *First Monday*, 11(10), 2006.
- Carsten Eickhoff and Arjen P. de Vries. How crowdsourcable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the 4th ACM international conference on web search and data mining (WSDM)*, pages 11–14, 2011.
- Carsten Eickhoff and Arjen P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Inf. Retr.*, 16(2):121–137, 2013.
- Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. T-Rex: A large scale alignment of natural language with knowledge base triples. In *LREC*. European Language Resources Association (ELRA), 2018.
- Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. Introducing Wikidata to the linked data web. In *Semantic Web Conference (1)*, volume 8796 of *Lecture Notes in Computer Science*, pages 50–65. Springer, 2014.
- Sean M. Falconer, Tania Tudorache, and Natalya Fridman Noy. An analysis of collaborative patterns in large-scale ontology development projects. In *K-CAP*, pages 25–32. ACM, 2011.
- Samer Faraj, Sirkka L. Jarvenpaa, and Ann Majchrzak. Knowledge collaboration in online communities. *Organization Science*, 22(5):1224–1239, 2011.
- Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1):77–129, 2018.
- Miriam Fernández, Chwhynny Overbeeke, Marta Sabou, and Enrico Motta. What makes a good ontology? A case-study in fine-grained knowledge reuse. In *ASWC*, volume 5926 of *Lecture Notes in Computer Science*, pages 61–75. Springer, 2009.
- Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. Finding news citations for wikipedia. In *CIKM*, pages 337–346. ACM, 2016.
- Danyel Fisher, Marc A. Smith, and Howard T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. In *HICSS*. IEEE Computer Society, 2006.
- Andrew J. Flanagin and Miriam J. Metzger. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, 9(2):319–342, 2007.

- Fabian Flöck, Denny Vrandečić, and Elena Simperl. Towards a diversity-minded Wikipedia. In *WebSci*, pages 5:1–5:8. ACM, 2011.
- Heather Ford, Shilad Sen, David R. Musicant, and Nathaniel Miller. Getting to the source: where does Wikipedia get its information from? In *Proceedings of the 9th International Symposium on Open Collaboration, Hong Kong, China, August 05 - 07, 2013*, pages 9:1–9:10, 2013.
- George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explorations*, 12(1):49–57, 2010.
- Andrea Forte and Amy Bruckman. Scaling consensus: Increasing decentralization in Wikipedia governance. In *HICSS*, page 157. IEEE Computer Society, 2008.
- Andrea Forte, Niki Kittur, Vanessa Larco, Haiyi Zhu, Amy Bruckman, and Robert E. Kraut. Coordination and beyond: social functions of groups in open content production. In *CSCW*, pages 417–426. ACM, 2012.
- Philipp Frischmuth, Michael Martin, Sebastian Tramp, Thomas Riechert, and Sören Auer. Ontowiki - an authoring, publication and visualization interface for the data web. *Semantic Web*, 6(3):215–240, 2015.
- Christian Fürber and Martin Hepp. Swiqa - a semantic web information quality assessment framework. In *ECIS*, page 76, 2011.
- Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jos Lehmann. Modelling ontology evaluation and validation. In *ESWC*, volume 4011 of *Lecture Notes in Computer Science*, pages 140–154. Springer, 2006.
- Rich Gazan. Redesign as an act of violence: disrupted interaction patterns and the fragmenting of a social q&a community. In *CHI*, pages 2847–2856. ACM, 2011.
- R. Stuart Geiger and Aaron Halfaker. When the levee breaks: without bots, what happens to Wikipedia’s quality control processes? In *OpenSym*, pages 6:1–6:6. ACM, 2013.
- Yolanda Gil and Varun Ratnakar. Knowledge capture in the wild: a perspective from semantic wiki communities. In *K-CAP*, pages 49–56. ACM, 2013.
- Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.
- Corrado Gini. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208:73–79, 1936.
- Boris Glavic and Klaus R. Dittrich. Data provenance: A categorization of existing approaches. In *BTW*, volume 103 of *LNI*, pages 227–241. GI, 2007.

- Asunción Gómez-Pérez, Mariano Fernández-López, and Óscar Corcho. *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. Advanced Information and Knowledge Processing. Springer, 2004.
- Bette Gray. Informal learning in an online community of practice. *Journal of Distance Education*, 19(1):20–35, 2004.
- Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.
- Jonathan Grudin. Computer-supported cooperative work: History and focus. *IEEE Computer*, 27(5):19–26, 1994.
- Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. Making peripheral participation legitimate: reader engagement experiments in Wikipedia. In *CSCW*, pages 849–860. ACM, 2013.
- Andrew Hall, Sarah McRoberts, Jacob Thebault-Spieker, Yilun Lin, Shilad Sen, Brent J. Hecht, and Loren G. Terveen. Freedom versus standardization: Structured data generation in a peer production community. In *CHI*, pages 6352–6362. ACM, 2017.
- Julia Handl, Joshua Knowles, and Douglas B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.
- Steve Harris, Andy Seaborne, and Eric Prud’hommeaux (eds.). SPARQL 1.1 Query Language. W3C Recommendation, W3C, October 2013. Online. Retrieved 9 May 2018, from <https://www.w3.org/TR/sparql11-query/>.
- Olaf Hartig. Provenance information in the web of data. In *LDOW*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.
- Olaf Hartig and Jun Zhao. Using web data provenance for quality assessment. In *SWPM*, volume 526 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.
- Caroline Haythornthwaite. Crowds and communities: Light and heavyweight models of peer production. In *HICSS*, pages 1–10. IEEE Computer Society, 2009.
- Caroline Haythornthwaite and Barry Wellman. Work, friendship, and media use for information exchange in a networked organization. *JASIS*, 49(12):1101–1114, 1998.
- Brent J. Hecht and Darren Gergle. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *CHI*, pages 291–300. ACM, 2010.
- Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph (eds.). OWL 2 Web Ontology Language: Primer. W3C Recommendation, W3C, October 2009. Online. Retrieved 9 May 2018, from <http://www.w3.org/TR/owl2-primer/>.

- Hlomani Hlomani and Deborah Stacey. Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal*, 1(5):1–11, 2014.
- Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.
- Lucy Holman Rector. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review*, 36(1):7–22, 2008.
- Corey Jergensen, Anita Sarma, and Patrick Wagstrom. The onion patch: migration in open source ecosystems. In *SIGSOFT FSE*, pages 70–80. ACM, 2011.
- Nicolas Jullien. What we know about Wikipedia: A review of the literature analyzing the project(s). Technical report, 2012. Online. Retrieved 10 February 2019, from <https://hal.archives-ouvertes.fr/hal-00857208>.
- Joseph M. Juran. *Quality control handbook*. McGraw-Hill, 1962.
- Lucie-Aimée Kaffee and Elena Simperl. The human face of the web of data: A cross-sectional study of labels. In *SEMANTICS*, volume 137 of *Procedia Computer Science*, pages 66–77. Elsevier, 2018.
- Michal Kakol, Michal Jankowski-Lorek, Katarzyna Abramczuk, Adam Wierzbicki, and Michele Catasta. On the subjectivity and bias of web content credibility evaluations. In *WWW (Companion Volume)*, pages 1131–1136. International World Wide Web Conferences Steering Committee / ACM, 2013.
- Dimitris Karampinas and Peter Triantafillou. Crowdsourcing taxonomies. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, volume 7295 of *Lecture Notes in Computer Science*, pages 545–559. Springer, 2012.
- Kensaku Kawamoto, Yasuhiko Kitamura, and Yuri A. Tijerino. Kawawiki: A semantic wiki based on RDF templates. In *IAT Workshops*, pages 425–432. IEEE Computer Society, 2006.
- Brian Keegan and Casey Fiesler. The evolution and consequences of peer producing Wikipedia’s rules. In *ICWSM*, pages 112–121. AAAI Press, 2017.
- Andrew Keen. *The Cult of the Amateur: How blogs, MySpace, YouTube and the rest of today’s user-generated media are killing our culture and economy*. Hachette UK, 2011.
- Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web*, 1(2):19, 2007a.

- Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *CHI*, pages 453–456. ACM, 2008.
- Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *CSCW*, pages 37–46. ACM, 2008.
- Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in Wikipedia. In *CHI*, pages 453–462. ACM, 2007b.
- Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5): 604–632, 1999.
- Peter Kollock. The economics of online cooperation: Gifts and public goods in cyberspace. In M. Smith and P. Kollock, editors, *Communities in Cyberspace*, pages 220–239. Routledge, New York, 1998.
- Piotr Konieczny. Governance, organization, and democracy on the internet: The iron law and the evolution of Wikipedia. In *Sociological Forum*, volume 24, pages 162–192. Wiley Online Library, 2009.
- Robert E. Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.
- Markus Krötzsch, Sebastian Schaffert, and Denny Vrandečić. Reasoning in semantic wikis. In *Reasoning Web*, pages 310–329. Springer, 2007.
- Markus Krötzsch, Denny Vrandečić, and Max Völkel. Semantic mediawiki. In *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 935–942. Springer, 2006.
- Kari Kuutti. Activity theory as a potential framework for human-computer interaction research. *Context and consciousness: Activity theory and human-computer interaction*, 17, 1996.
- Shyong K. Lam, Jawed Karim, and John Riedl. The effects of group composition on decision quality in a social production community. In *GROUP*, pages 55–64. ACM, 2010.
- Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. Motivations to participate in online communities. In *CHI*, pages 1927–1936. ACM, 2010.
- Birger Lantow. Ontometrics: Application of on-line ontology metric calculation. In *BIR Workshops*, volume 1684 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.
- Jens Lehmann, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. Defacto - deep fact validation. In *International Semantic Web Conference (1)*, volume 7649 of *Lecture Notes in Computer Science*, pages 312–327. Springer, 2012.

- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015.
- Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):32–38, 1995.
- Primo Levi. *The Periodic Table*. Penguin Modern Classics, 2000.
- John M. Levine and Richard L. Moreland. Progress in small group research. *Annual review of psychology*, 41(1):585–634, 1990.
- Xinxin Li and Lorin M. Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008.
- Hsiu-Fen Lin and Gwo-Guang Lee. Determinants of success for online communities: an empirical study. *Behaviour & IT*, 25(6):479–488, 2006.
- Jun Liu and Sudha Ram. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. *ACM Trans. Management Inf. Syst.*, 2(2):11:1–11:23, 2011.
- Yuan Long and Keng Siau. Social network structures in open source software development teams. *J. Database Manag.*, 18(2):25–40, 2007.
- Teun Lucassen and Jan Maarten Schraagen. Trust in Wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th ACM Workshop on Information Credibility on the Web, WICOW 2010, Raleigh, North Carolina, USA, April 27, 2010*, pages 19–26. ACM, 2010.
- Roman Lukyanenko, Jeffrey Parsons, and Yolanda F. Wiersma. The IQ of the crowd: Understanding and improving information quality in structured user-generated content. *Information Systems Research*, 25(4):669–689, 2014.
- Meng Ma and Ritu Agarwal. Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Information Systems Research*, 18(1):42–67, 2007.
- Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. YAGO3: A knowledge base from multilingual Wikipedias. In *CIDR*. www.cidrdb.org, 2015.
- Sanna Malinen. Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in Human Behavior*, 46:228–238, 2015.

- Diane Maloney-Krichmar and Jennifer J. Preece. A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Trans. Comput.-Hum. Interact.*, 12(2):201–232, 2005.
- Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. Getting the most out of Wikidata: Semantic technology usage in Wikipedia’s knowledge graph. In *International Semantic Web Conference (2)*, volume 11137 of *Lecture Notes in Computer Science*, pages 376–394. Springer, 2018.
- Sergio L. Toral Marín, M. Rocío Martínez-Torres, and Federico Barrero. Analysis of virtual communities supporting OSS projects using social network analysis. *Information & Software Technology*, 52(3):296–303, 2010.
- Owen S. Martin. A wikipedia literature review. *CoRR*, abs/1110.5863, 2011.
- Joseph E. McGrath. Time, interaction, and performance (tip) a theory of groups. *Small group research*, 22(2):147–174, 1991.
- Pamela J. McKenzie, Jacquelyn A. Burkell, Lola Wong, Caroline Whippley, Samuel E. Trosow, and Michael B. McNally. User-generated online content 1: overview, current state and context. *First Monday*, 17(6), 2012.
- Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: linked data quality assessment and fusion. In *EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.
- Jean Mercer. Wikipedia and open source mental health information. *Scientific Review of Mental Health Practice*, 5(1):88–92, 2007.
- Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. “the sum of all human knowledge”: A systematic review of scholarly research on the content of Wikipedia. *JASIST*, 66(2):219–245, 2015.
- Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. Social and heuristic approaches to credibility evaluation online. *Journal of communication*, 60(3):413–439, 2010.
- Frances J. Milliken and Luis L. Martins. Searching for common threads: Understanding the multiple effects of diversity in organizational groups. *The Academy of Management Review*, 21(2):402–433, 1996. ISSN 03637425.
- Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha Pratim Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. Never-ending learning. In *AAAI*, pages 2302–2310. AAAI Press, 2015.

- Audris Mockus, Roy T. Fielding, and James D. Herbsleb. A case study of open source software development: the apache server. In *ICSE*, pages 263–272. ACM, 2000.
- Audris Mockus, Roy T. Fielding, and James D. Herbsleb. Two case studies of open source software development: Apache and mozilla. *ACM Trans. Softw. Eng. Methodol.*, 11(3):309–346, 2002.
- Luc Moreau, Juliana Freire, Joe Futrelle, Robert E. McGrath, Jim Myers, and Patrick R. Paulson. The open provenance model: An overview. In *Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers*, volume 5272 of *Lecture Notes in Computer Science*, pages 323–326. Springer, 2008.
- Richard L. Moreland and John M. Levine. Socialization in organizations and work groups. *Groups at Work: Theory and Research*, page 69, 2014.
- Lev Muchnik, Sen Pei, Lucas C. Parra, Saulo D. S. Reis, José S. Andrade Jr., Shlomo Havlin, and Hernán A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3:1783, 2013.
- Claudia Müller-Birn, Leonhard Dobusch, and James D. Herbsleb. Work-to-rule: the emergence of algorithmic governance in Wikipedia. In *C&T*, pages 80–89. ACM, 2013.
- Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. Peer-production system or collaborative ontology engineering effort: what is Wiki-data? In *OpenSym*, pages 20:1–20:10. ACM, 2015.
- Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. Scalable knowledge harvesting with high precision and high recall. In *WSDM*, pages 227–236. ACM, 2011.
- Grzegorz J. Nalepa. Collective knowledge engineering with semantic wikis. *J. UCS*, 16(7):1006–1023, 2010.
- Alessandro Narduzzo and Alessandro Rossi. The role of modularity in free/open source software development. In *Free/Open source software development*, pages 84–102. Igi Global, 2005.
- Felix Naumann. *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes in Computer Science*. Springer, 2002.
- Juan Francisco García Navarro, Francisco José García-Peñalvo, and Roberto Therón. A survey on ontology metrics. In *WSKS (1)*, volume 111 of *Communications in Computer and Information Science*, pages 22–27. Springer, 2010.
- Sabine Niederer and José van Dijck. Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society*, 12(8):1368–1387, 2010.

- Blair Nonnecke and Jennifer J. Preece. Lurker demographics: counting the silent. In *CHI*, pages 73–80. ACM, 2000.
- Oded Nov. What motivates Wikipedians? *Commun. ACM*, 50(11):60–64, 2007.
- Natalya F. Noy, Deborah L. McGuinness, et al. Ontology development 101: A guide to creating your first ontology. Technical report, Stanford knowledge systems laboratory technical report KSL-01-05 and , 2001.
- Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *JASIST*, 65(12):2381–2403, 2014.
- Siobhán O’Mahony and Fabrizio Ferraro. The emergence of governance in an open source community. *Academy of Management Journal*, 50(5):1079–1106, 2007.
- Eyal Oren. Semperwiki: a semantic personal wiki. In *Semantic Desktop Workshop*, volume 175 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2005.
- Anthony M. Orme, Haining Yao, and Letha H. Etzkorn. Coupling metrics for ontology-based systems. *IEEE Software*, 23(2):102–108, Mar 2006.
- Felipe Ortega and Jesús M. González-Barahona. Quantitative analysis of the Wikipedia community of users. In *Int. Sym. Wikis*, pages 75–86. ACM, 2007.
- Felipe Ortega, Jesús M. González-Barahona, and Gregorio Robles. On the inequality of contributions to Wikipedia. In *HICSS*, page 304. IEEE Computer Society, 2008.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
- Katherine A. Panciera, Reid Priedhorsky, Thomas Erickson, and Loren G. Terveen. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. In *CHI*, pages 1917–1926. ACM, 2010.
- Jarutas Pattanaphanchai, Kieron O’Hara, and Wendy Hall. Trustworthiness criteria for supporting users to assess the credibility of web information. In *WWW (Companion Volume)*, pages 1123–1130. International World Wide Web Conferences Steering Committee / ACM, 2013.
- Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508, 2017.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

- Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *ICDE*, pages 215–224. IEEE Computer Society, 2001.
- Lisa Hope Pelled, Kathleen M. Eisenhardt, and Katherine R. Xin. Exploring the black box: An analysis of work group diversity, conflict, and performance. *Administrative Science Quarterly*, 44(1):1–28, 1999.
- Michael P. Pender, Kaye Lasserre, Lisa Kruesi, Christopher Del Mar, and Satyamurthy Anuradha. Putting Wikipedia to the test: a case study. In *The Special Libraries Association Annual Conference*, pages 1–16, 2008.
- Barbara Pernici and Monica Scannapieco. Data quality in web information systems. In *Journal on Data Semantics I*, pages 48–68. Springer, 2003.
- Leo Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, 2002.
- Alessandro Piscopo, Lucie-Aimée Kaffee, Christopher Phethean, and Elena Simperl. Provenance information in a collaborative knowledge graph: An evaluation of Wikidata external references. In *International Semantic Web Conference (1)*, volume 10587 of *Lecture Notes in Computer Science*, pages 542–558. Springer, 2017a.
- Alessandro Piscopo, Christopher Phethean, and Elena Simperl. What makes a good collaborative knowledge graph: Group composition and quality in Wikidata. In *SocInfo (1)*, volume 10539 of *Lecture Notes in Computer Science*, pages 305–322. Springer, 2017b.
- Alessandro Piscopo, Christopher Phethean, and Elena Simperl. Wikidatians are born: Paths to full participation in a collaborative structured knowledge base. In *HICSS. AIS Electronic Library (AISeL)*, 2017c.
- Alessandro Piscopo and Elena Simperl. Who models the world?: Collaborative ontology creation and user roles in Wikidata. *PACMHCI*, 2(CSCW):141:1–141:18, 2018.
- Alessandro Piscopo and Elena Simperl. What we talk about when we talk about Wikidata quality: a literature survey. In *OpenSym*, pages 17:1–17:11. ACM, 2019.
- Alessandro Piscopo, Pavlos Vougiouklis, Lucie-Aimée Kaffee, Christopher Phethean, Jonathon S. Hare, and Elena Simperl. What do Wikidata and Wikipedia have in common? An analysis of their use of external references. In *OpenSym*, pages 1:1–1:10. ACM, 2017d.
- Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in Wikipedia. In *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 663–668. Springer, 2008.

- Radityo Eko Prasajo, Fariz Darari, Simon Razniewski, and Werner Nutt. Managing and consuming completeness information for Wikidata using COOL-WD. In *COLD@ISWC*, volume 1666 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.
- Jennifer J. Preece. Designing usability, supporting sociability: Questions participants ask about online communities. In *INTERACT*, pages 3–12. IOS Press, 2001a.
- Jennifer J. Preece. Sociability and usability in online communities: determining and measuring success. *Behaviour & IT*, 20(5):347–356, 2001b.
- Jennifer J. Preece, Blair Nonnecke, and Dorine Andrews. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2):201–223, 2004.
- Jennifer J. Preece and Ben Shneiderman. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, 1(1):13–32, 2009.
- Reid Priedhorsky, Jilin Chen, Shyong K. Lam, Katherine A. Panciera, Loren G. Terveen, and John Riedl. Creating, destroying, and restoring value in Wikipedia. In *GROUP*, pages 259–268. ACM, 2007.
- Malolan S Rajagopalan, Vineet Khanna, M Stott, Y Leiter, TN Showalter, A Dicker, and YR Lawrence. Accuracy of cancer information on the internet: A comparison of a wiki with a professionally maintained database. *Journal of Clinical Oncology*, 28(15-suppl):6058–6058, 2010.
- Eric S. Raymond. *The cathedral and the bazaar - musings on Linux and open source by an accidental revolutionary (rev. ed.)*. O’Reilly, 2001.
- Thomas C Redman. *Data quality: the field guide*. Digital press, 2001.
- Ruqin Ren and Bei Yan. Crowd diversity and performance in Wikipedia: The mediating effects of task conflict and communication. In *CHI*, pages 6342–6351. ACM, 2017.
- Yuqing Ren, Jilin Chen, and John Riedl. The impact and evolution of group diversity in online open collaboration. *Management Science*, 62(6):1668–1686, 2016.
- Howard Rheingold. A slice of my life in my virtual community. *High noon on the electronic frontier: Conceptual issues in cyberspace*, pages 413–36, 1996.
- Ignacio Traverso Ribón, Maria-Esther Vidal, Benedikt Kämpgen, and York Sure-Vetter. GADES: A graph-based semantic similarity measure. In *SEMANTICS*, pages 101–104. ACM, 2016.
- Daniel Ringler and Heiko Paulheim. One knowledge graph to rule them all? analyzing the differences between DBpedia, YAGO, Wikidata & co. In *KI*, volume 10505 of *Lecture Notes in Computer Science*, pages 366–372. Springer, 2017.

- Richard Rogers. *Digital methods*. MIT press, 2013.
- Roy Rosenzweig. Can history be open source? Wikipedia and the future of the past. *The journal of American history*, 93(1):117–146, 2006.
- Camille Roth, Dario Taraborelli, and Nigel Gilbert. Measuring wiki viability: an empirical assessment of the social dynamics of a large sample of wikis. In *Int. Sym. Wikis*. ACM, 2008.
- Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. Building automated vandalism detection tools for Wikidata. In *WWW (Companion Volume)*, pages 1647–1654. ACM, 2017.
- Cristina Sarasua, Alessandro Checco, Gianluca Demartini, Djellel Eddine Difallah, Michael Feldman, and Lydia Pintscher. The evolution of power and standard Wikidata editors: Comparing editing behavior over time to predict lifespan and volume of edits. *Computer Supported Cooperative Work*, 28(5):843–882, 2019.
- Walt Scacchi. Free/open source software development: recent research results and emerging opportunities. In *ESEC/SIGSOFT FSE (Companion)*, pages 459–468. ACM, 2007.
- Sebastian Schaffert. Ikewiki: A semantic wiki for collaborative knowledge management. In *WETICE*, pages 388–396. IEEE Computer Society, 2006.
- Sebastian Schaffert, François Bry, Joachim Baumeister, and Malte Kiesel. Semantic wikis. *IEEE Software*, 25(4):8–11, 2008.
- Guus Schreiber, Manola Raimond, Yves Frank, Eric Miller, and Brian McBride (eds.). RDF 1.1 Primer. W3C Working Group Note, W3C, June 2014. Online. Retrieved 9 May 2018, from <https://www.w3.org/TR/rdf11-primer/>.
- Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.
- Nigel R. Shadbolt, Daniel A. Smith, Elena Simperl, Max Van Kleek, Yang Yang, and Wendy Hall. Towards a classification framework for social machines. In *WWW (Companion Volume)*, pages 905–912. International World Wide Web Conferences Steering Committee / ACM, 2013.
- Aaron D. Shaw and Benjamin Mako Hill. Laboratories of oligarchy? how the iron law extends to peer production. *CoRR*, abs/1407.0323, 2014.
- Miguel-Ángel Sicilia, Daniel Rodríguez, Elena García Barriocanal, and Salvador Sánchez Alonso. Empirical findings on ontology metrics. *Expert Syst. Appl.*, 39(8):6706–6711, 2012.
- Elena Simperl. How to use crowdsourcing effectively: Guidelines and examples. *Liber Quarterly*, 25(1):18–39, 2015.

- Elena Simperl and Markus Luczak-Rösch. Collaborative ontology engineering: a survey. *Knowledge Eng. Review*, 29(1):101–131, 2014.
- Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *CoopIS/DOA/ODBASE*, volume 2519 of *Lecture Notes in Computer Science*, pages 1223–1237. Springer, 2002.
- Amit Singhal. Introducing the knowledge graph: things, not strings. In *Google Blog*. May 2012. Online. Retrieved 1 February 2019, from <https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>.
- Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*, pages 254–263. ACL, 2008.
- Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451. AAAI Press, 2017.
- Thomas Steiner. Bots vs. Wikipedians, Anons vs. Logged-ins (redux): A global study of edit activity on Wikipedia and Wikidata. In *OpenSym*, pages 25:1–25:7. ACM, 2014.
- Markus Strohmaier, Simon Walk, Jan Pöschko, Daniel Lamprecht, Tania Tudorache, Csongor Nyulas, Mark A. Musen, and Natalya Fridman Noy. How ontologies are made: Studying the hidden social dynamics behind collaborative ontology engineering projects. *J. Web Sem.*, 20:18–34, 2013.
- Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. Information quality work organization in Wikipedia. *JASIST*, 59(6):983–1001, 2008.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: a core of semantic knowledge. In *WWW*, pages 697–706. ACM, 2007.
- Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. The singularity is not near: slowing growth of Wikipedia. In *Int. Sym. Wikis*. ACM, 2009.
- James Surowiecki. *The wisdom of crowds*. Anchor, 2005.
- Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. From Freebase to Wikidata: The great migration. In *WWW*, pages 1419–1428. ACM, 2016.

- Samir Tartir and Ismailcem Budak Arpinar. Ontology evaluation and ranking using ontoqa. In *ICSC*, pages 185–192. IEEE Computer Society, 2007.
- Adolfo Lozano Tello and Asunción Gómez-Pérez. ONTOMETRIC: A method to choose the appropriate ontology. *J. Database Manag.*, 15(2):1–18, 2004.
- Christoph Tempich, Elena Simperl, Markus Luczak, Rudi Studer, and H Sofia Pinto. Argumentation-based ontology engineering. *IEEE Intelligent Systems*, (6):52–59, 2007.
- Harsh Thakkar, Kemele M. Endris, José M. Giménez-García, Jeremy Debattista, Christoph Lange, and Sören Auer. Are linked datasets fit for open-domain question answering? A quality assessment. In *WIMS*, pages 19:1–19:12. ACM, 2016.
- Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.
- Ramine Tinati, Markus Luczak-Rösch, Nigel Shadbolt, and Wendy Hall. Using wikiprojects to measure the health of Wikipedia. In *WWW (Companion Volume)*, pages 369–370. ACM, 2015.
- Daan Van Knippenberg, Carsten KW De Dreu, and Astrid C Homan. Work group diversity and group performance: an integrative model and research agenda. *Journal of applied psychology*, 89(6):1008, 2004.
- Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with *history flow* visualizations. In *CHI*, pages 575–582. ACM, 2004.
- Markel Vigo, Caroline Jay, and Robert Stevens. Constructing conceptual knowledge artefacts: Activity patterns in the ontology authoring process. In *CHI*, pages 3385–3394. ACM, 2015.
- Nicholas Vincent, Isaac L. Johnson, and Brent J. Hecht. Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia’s relationships with other large-scale online communities. In *CHI*, page 566. ACM, 2018.
- Johanna Völker, Denny Vrandečić, and York Sure. Automatic evaluation of ontologies (AEON). In *International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 716–731. Springer, 2005.
- Denny Vrandečić. Restricting the world. In *Wikimedia Blog*. February 2013. Online. Retrieved 8 April 2018, from <http://blog.wikimedia.de/2013/02/22/restricting-the-world/>.
- Denny Vrandečić. *Ontology evaluation*. PhD thesis, Karlsruhe Institute of Technology, 2010.

- Denny Vrandečić. The rise of Wikidata. *IEEE Intelligent Systems*, 28(4):90–95, 2013.
- Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledge base. *Commun. ACM*, 57(10):78–85, 2014.
- Clifford H. Wagner. Simpson’s paradox in real life. *The American Statistician*, 36(1):46–48, 1982.
- Simon Walk, Philipp Singer, Lisette Espin Noboa, Tania Tudorache, Mark A. Musen, and Markus Strohmaier. Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In *International Semantic Web Conference (1)*, volume 9366 of *Lecture Notes in Computer Science*, pages 551–568. Springer, 2015.
- Hao Wang, Tania Tudorache, Dejing Dou, Natalya Fridman Noy, and Mark A. Musen. Analysis and prediction of user editing patterns in ontology development projects. *J. Data Semantics*, 4(2):117–132, 2015.
- Richard Y. Wang, Henry B. Kon, and Stuart E. Madnick. Data quality requirements analysis and modeling. In *ICDE*, pages 670–677. IEEE Computer Society, 1993.
- Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12(4):5–33, 1996.
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph and text jointly embedding. In *EMNLP*, pages 1591–1601. ACL, 2014.
- Morten Warncke-Wang, Vladislav R. Ayukaev, Brent J. Hecht, and Loren G. Terveen. The success and failure of quality improvement projects in peer production communities. In *CSCW*, pages 743–756. ACM, 2015.
- C. Nadine Wathen and Jacquelyn A. Burkell. Believe it or not: Factors influencing credibility on the web. *JASIST*, 53(2):134–144, 2002.
- Steven Weber. *The Success of Open Source*. Harvard University Press, Cambridge, MA, USA, 2004. ISBN 0674012925.
- Etienne Wenger and Jean Lave. *Situated Learning: Legitimate Peripheral Participation (Learning in Doing: Social, Cognitive and Computational Perspectives)*. Cambridge University Press, Cambridge, UK, 1991.
- Steve Whittaker, Loren G. Terveen, William C. Hill, and Lynn Cherny. The dynamics of mass interaction. In *CSCW*, pages 257–264. ACM, 1998.
- Brian Whitworth. The social requirements of technical systems. In *Handbook of Research on Socio-Technical Design and Social Networking Systems*, pages 2–22. IGI Global, 2009.

- Wikidata. Bots. In *Wikidata, The Free Knowledge Base*. 2018a. Online. Retrieved 8 April 2018, from <https://www.wikidata.org/w/index.php?title=Wikidata:Bots&oldid=611067206>.
- Wikidata. Help:Aliases. In *Wikidata, The Free Knowledge Base*. 2018b. Online. Retrieved 8 April 2018, from <https://www.wikidata.org/w/index.php?title=Help:Aliases&oldid=635297952>.
- Wikidata. Help:Description. In *Wikidata, The Free Knowledge Base*. 2018c. Online. Retrieved 8 April 2018, from <https://www.wikidata.org/w/index.php?title=Help:Description&oldid=847950716>.
- Wikidata. Help:Label. In *Wikidata, The Free Knowledge Base*. 2018d. Online. Retrieved 8 April 2018, from <https://www.wikidata.org/w/index.php?title=Help:Label&oldid=657767014>.
- Wikidata. Help:Qualifiers. In *Wikidata, The Free Knowledge Base*. 2018e. Online. Retrieved 8 April 2018, from <https://www.wikidata.org/w/index.php?title=Help:Qualifiers&oldid=616291280>.
- Wikidata. Help:Sources. In *Wikidata, The Free Knowledge Base*. 2018f. Online. Retrieved 9 May 2018, from <https://www.wikidata.org/w/index.php?title=Help:Sources&oldid=612214941>.
- Wikidata. Help:Sources/Items not needing sources. In *Wikidata, The Free Knowledge Base*. 2018g. Online. Retrieved 9 May 2018, from https://www.wikidata.org/w/index.php?title=Help:Sources/Items_not_needing_sources&oldid=565986976.
- Wikidata. Showcase Items. In *Wikidata, The Free Knowledge Base*. 2018h. Online. Retrieved 31 August 2018, from https://www.wikidata.org/w/index.php?title=Wikidata:Showcase_items&oldid=718790084.
- Wikidata. User access levels. In *Wikidata, The Free Knowledge Base*. 2018i. Online. Retrieved 8 April 2018, from https://www.wikidata.org/w/index.php?title=Wikidata:User_access_levels&oldid=637336949.
- Wikidata. Verifiability. In *Wikidata, The Free Knowledge Base*. 2018j. Online. Retrieved 9 May 2018, from <https://www.wikidata.org/w/index.php?title=Wikidata:Verifiability&oldid=552124443>.
- Wikipedia. Verifiability. In *Wikipedia, The Free Encyclopedia*. 2018a. Online. Retrieved 9 May 2018, from <https://en.wikipedia.org/w/index.php?title=Wikipedia:Verifiability&oldid=838827946>.
- Wikipedia. Wikipedia, the Free Encyclopedia. In *Wikipedia, The Free Encyclopedia*. 2018b. Online. Retrieved 9 May 2018, from <https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=839464552>.

- Heather Woltman, Andrea Feldstain, Christine J. MacKay, and Meredith Rocchi. An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, 8(1):52–69, 2012.
- Heng-Li Yang and Cheng-Yu Lai. Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377–1383, 2010.
- Haining Yao, Anthony Mark Orme, and Letha Etzkorn. Cohesion metrics for ontology design and application. *Journal of Computer science*, 1(1):107–113, 2005.
- Glorian Yapinus, Amir Sarabadani, and Aaron Halfaker. Wikidata Item quality labels, 5 2017. Dataset. Retrieved 9 May 2018, from https://figshare.com/articles/Wikidata_item_quality_labels/5035796.
- Yunwen Ye and Kouichi Kishida. Toward an understanding of the motivation of open source software developers. In *ICSE*, pages 419–429. IEEE Computer Society, 2003.
- Jonathan Yu, James A. Thom, and Audrey M. Tam. Ontology evaluation using Wikipedia categories for browsing. In *CIKM*, pages 223–232. ACM, 2007.
- Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.
- Yang Zhe, Dalu Zhang, and Ye Chuan. Evaluation metrics for ontology complexity and evolution analysis. In *e-Business Engineering, 2006. ICEBE'06. IEEE International Conference on*, pages 162–170. IEEE, 2006.
- Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. The impact of membership overlap on the survival of online communities. In *CHI*, pages 281–290. ACM, 2014.
- Pujan Ziaie. A model for context in the design of open production communities. *ACM Comput. Surv.*, 47(2):29:1–29:29, 2014.