Knowledge Diversity and its Relation to Automation in a Knowledge Base Context: The Case of Wikidata

DISSERTATION

zur Erlangung des Grades eines Doktors der Naturwissenschaften
(Dr. rer. nat.)

am Fachbereich Mathematik und Informatik

der Freien Universität Berlin.

von

Mariam Farda-Sarbas

Berlin, 2024

Erstgutachter: Prof. Dr. Claudia Müller-Birn

Zweitgutachter: Prof. Dr. Bettina Berendt

Tag der Disputation: 08.04.2024

Declaration

I hereby declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those explicitly indicated. I have appropriately marked all statements that are taken verbatim or in content from other writings. This dissertation has not been previously submitted in the same or similar form in any previous doctoral procedure.

I agree to have my thesis examined by a plagiarism examination software.

Berlin, January 2024

Mariam Farda-Sarbas

Abstract

Since its launch in 2012, Wikidata has grown to become the largest open knowledge base (KB), containing more than 100 million data items and over 6 million registered users. Wikidata serves as the structured data backbone of Wikipedia, addressing data inconsistencies, and adhering to the motto of "serving anyone anywhere in the world," a vision realized through the diversity of knowledge. Despite being a collaboratively contributed platform, the Wikidata community heavily relies on bots, automated accounts with batch, and speedy editing rights, for a majority of edits. As Wikidata approaches its first decade, the question arises: How close is Wikidata to achieving its vision of becoming a global KB and how diverse is it in serving the global population? This dissertation investigates the current status of Wikidata's diversity, the role of bot interventions on diversity, and how bots can be leveraged to improve diversity within the context of Wikidata.

The methodologies used in this study are mapping study and content analysis, which led to the development of three datasets: 1) *Wikidata Research Articles Dataset*, covering the literature on Wikidata from its first decade of existence sourced from online databases to inspect its current status; 2) *Wikidata Requests-for-Permissions Dataset*, based on the pages requesting bot rights on the Wikidata website to explore bots from a community perspective; and 3) *Wikidata Revision History Dataset*, compiled from the edit history of Wikidata to investigate bot editing behavior and its impact on diversity, all of which are freely available online.

The insights gained from the mapping study reveal the growing popularity of Wikidata in the research community and its various application areas, indicative of its progress toward the ultimate goal of reaching the global community. However, there is currently no research addressing the topic of diversity in Wikidata, which could shed light on its capacity to serve a diverse global population. To address this gap, this dissertation proposes a diversity measurement concept that defines diversity in a KB context in terms of variety, balance, and disparity and is capable of assessing diversity in a KB from two main angles: user and data. The application of this concept on the domains and classes of the *Wikidata Revision History Dataset* exposes imbalanced content distribution across Wikidata domains, which indicates low data diversity in Wikidata domains.

Further analysis discloses that bots have been active since the inception of Wikidata, and the community embraces their involvement in content editing tasks, often importing data from Wikipedia, which shows a low diversity of sources in bot edits. Bots and human users engage in similar editing tasks but exhibit distinct editing patterns. The findings of this thesis confirm that bots possess the potential to influence diversity within Wikidata by contributing substantial amounts of data to specific classes and domains, leading to an imbalance. However, this potential can also be harnessed to enhance coverage in classes with limited content and restore balance, thus improving diversity. Hence, this study proposes to enhance diversity through automation and demonstrate the practical implementation of the recommendations using a specific use case.

In essence, this research enhances our understanding of diversity in relation to a KB, elucidates the influence of automation on data diversity, and sheds light on diversity improvement within a KB context through the usage of automation.

Zusammenfassung

Seit seiner Einführung im Jahr 2012 hat sich Wikidata zu der größten offenen Wissensdatenbank entwickelt, die mehr als 100 Millionen Datenelemente und über 6 Millionen registrierte Benutzer enthält. Wikidata dient als das strukturierte Rückgrat von Wikipedia, indem es Datenunstimmigkeiten angeht und sich dem Motto verschrieben hat, 'jedem überall auf der Welt zu dienen', eine Vision, die durch die Diversität des Wissens verwirklicht wird. Trotz seiner kooperativen Natur ist die Wikidata-Community in hohem Maße auf Bots, automatisierte Konten mit Batch-Verarbeitung und schnelle Bearbeitungsrechte angewiesen, um die Mehrheit der Bearbeitungen durchzuführen.

Da Wikidata seinem ersten Jahrzehnt entgegengeht, stellt sich die Frage: Wie nahe ist Wikidata daran, seine Vision, eine globale Wissensdatenbank zu werden, zu verwirklichen, und wie ausgeprägt ist seine Dienstleistung für die globale Bevölkerung? Diese Dissertation untersucht den aktuellen Status der Diversität von Wikidata, die Rolle von Bot-Eingriffen in Bezug auf Diversität und wie Bots im Kontext von Wikidata zur Verbesserung der Diversität genutzt werden können.

Die in dieser Studie verwendeten Methoden sind Mapping-Studie und Inhaltsanalyse, die zur Entwicklung von drei Datensätzen geführt haben: 1) *Wikidata Research Articles Dataset*, die die Literatur zu Wikidata aus dem ersten Jahrzehnt aus Online-Datenbanken umfasst, um den aktuellen Stand zu untersuchen; 2) *Request-for-Permission Dataset*, der auf den Seiten zur Beantragung von Bot-Rechten auf der Wikidata-Website basiert, um Bots aus der Perspektive der Gemeinschaft zu untersuchen; und 3) *Wikidata Revision History Dataset*, der aus der Bearbeitungshistorie von Wikidata zusammengestellt wurde, um das Bearbeitungsverhalten von Bots zu untersuchen und dessen Auswirkungen auf die Diversität, die alle online frei verfügbar sind.

Die Erkenntnisse aus der Mapping-Studie zeigen die wachsende Beliebtheit von Wikidata in der Forschungsgemeinschaft und in verschiedenen Anwendungsbereichen, was auf seinen Fortschritt hin zur letztendlichen Zielsetzung hindeutet, die globale Gemeinschaft zu erreichen. Es gibt jedoch derzeit keine Forschung, die sich mit dem Thema der Diversität in Wikidata befasst und Licht auf seine Fähigkeit werfen könnte, eine vielfältige globale Bevölkerung zu bedienen. Um diese Lücke zu schließen, schlägt diese Dissertation ein Konzept zur Messung der Diversität vor, das die Diversität im Kontext einer Wissensbasis anhand von Vielfalt, Balance und Diskrepanz definiert und in der Lage ist, die Diversität aus zwei Hauptperspektiven zu bewerten: Benutzer und Daten.

Die Anwendung dieses Konzepts auf die Bereiche und Klassen des *Wikidata Revision History Dataset* zeigt eine unausgewogene Verteilung des Inhalts über die Bereiche von Wikidata auf, was auf eine geringe Diversität der Daten in den Bereichen von Wikidata hinweist.

Weitere Analysen zeigen, dass Bots seit der Gründung von Wikidata aktiv waren und von der Gemeinschaft inhaltliche Bearbeitungsaufgaben angenommen werden, oft mit Datenimporten aus Wikipedia, was auf eine geringe Diversität der Quellen bei Bot-Bearbeitungen hinweist. Bots und menschliche Benutzer führen ähnliche Bearbeitungsaufgaben aus, zeigen jedoch unterschiedliche Bearbeitungsmuster. Die Ergebnisse dieser Dissertation bestätigen, dass Bots das Potenzial haben, die Diversität in Wikidata zu beeinflussen, indem sie bedeutende Datenmengen zu bestimmten Klassen und Bereichen beitragen, was zu einer Ungleichgewichtung führt. Dieses Potenzial kann jedoch auch genutzt werden, um die Abdeckung in Klassen mit begrenztem Inhalt zu verbessern und das Gleichgewicht wiederherzustellen, um die Diversität zu verbessern. Daher schlägt diese Studie vor, die Diversität durch Automatisierung zu verbessern und die praktische Umsetzung der Empfehlungen anhand eines spezifischen Anwendungsfalls zu demonstrieren.

Kurz gesagt trägt diese Forschung dazu bei, unser Verständnis der Diversität im Kontext einer Wissensbasis zu vertiefen, wirft Licht auf den Einfluss von Automatisierung auf die Diversität von Daten und zeigt die Verbesserung der Diversität im Kontext einer Wissensbasis durch die Verwendung von Automatisierung auf.

Acknowledgement

In the name of Almighty Allah, the Most Compassionate, the Most Merciful, and the source of all knowledge and wisdom. I begin by expressing my deep gratitude and praise for the blessings and guidance that have illuminated my path throughout my life and academic journey.

This journey towards completing my dissertation has gone through the darkest days. In the middle of the pandemic challenges, I lost my land, including my dreams and home. My cultural identity was at risk of elimination, and my family had to migrate because women are no longer considered to have the basic rights of getting an education, working, or even leaving home alone. Although this situation in the 21st century made me lose trust in the term ‘diversity’ and my interest in completing this dissertation faded, I tried to transform this insurmountable challenge into a motivating factor towards empowering myself so that I can work towards defending my cultural identity and enabling myself to help other women in my land. This would not be possible without the continuous support of the following:

I am immensely thankful to my supervisor, Prof. Dr. Claudia Müller-Birn, whose dedication extended beyond academic guidance. Your unwavering support, encouragement, and belief in my potential have been invaluable in shaping both my research and personal growth.

I am deeply grateful to the German Academic Exchange Service (DAAD) for granting me the scholarship that made my PhD studies possible. This scholarship not only supported my academic pursuits but also enabled me to contribute more effectively to the academic community.

My heartfelt thanks go to Freie Universität Berlin and Ernst-Reuter-Gesellschaft for their generous scholarships that provided the essential financial support needed to complete my Ph.D. Your contribution eased my financial burdens and allowed me to focus on my research wholeheartedly.

To my beloved family, I owe a debt of gratitude that words can hardly express. Your unwavering motivation, support, prayers, and love have been my constant inspiration.

I am also grateful to my colleagues and peers who provided constructive feedback, engaged in thought-provoking discussions, and formed a supportive academic community. Your insights have enriched my research and my understanding of the field.

Lastly, I want to express my appreciation to all those whose contributions might not be explicitly mentioned here but have played a role in shaping my academic journey. Your encouragement and support, no matter how small, have made a difference.

In closing, I am deeply indebted to all these individuals and organizations for their unwavering support and encouragement. Their belief in me has been a driving force that has led me to this point of accomplishment.

Mariam Farda-Sarbas

January 2024

Contents

1 INTRODUCTION 1

1.1 Research Motivation 1

1.2 Research Questions 3

1.3 Research Contributions 4

1.4 Thesis Organization 5

2 WIKIDATA 9

2.1 Background 9

2.1.1 Goals and Design Decisions 10

2.1.2 Contributing Community 13

2.1.3 Data Model 15

2.2 Wikidata From a Research Perspective 18

2.2.1 Research Method 19

2.2.1.1 Search Process and Data Sources 19

2.2.1.2 Criteria Exclusion and Inclusion 20

2.2.1.3 Data extraction 22

2.2.1.4 Research Paper Classification 22

2.2.2 Overview of Existing Research on Wikidata 23

2.2.2.1 Frequency of Publication 23

2.2.2.2 Publishers and Publication Types 23

2.2.2.3 Geographical Origins of Research 24

2.2.3 Research Topics of Wikidata 26

2.2.3.1 Community-oriented Research 26

2.2.3.2 Engineering-oriented Research 29

2.2.3.3 Data-oriented Research 32

2.2.3.4 Knowledge Graph Oriented Research 35

2.2.3.5 Application Use Cases 38

2.2.3.6 Implications 43

2.3 Summary 46

3 DIVERSITY & WIKIDATA 47

3.1 The Diversity Notion 48

3.1.1 Defining Diversity 48

3.1.2 Diversity Dimensions 49
3.1.3 Application Domains of Diversity 51
3.1.3.1 Diversity from a Biological Lens 51
3.1.3.2 Diversity from a Sociological and Economical Lens . 52
3.1.3.3 Diversity from a Computational Lens 53
3.1.3.4 Summary of Diversity Application Domains 54
3.1.4 Diversity Measurement 54
3.1.4.1 Types of Diversity Measures 55
3.1.4.2 Implications 60
3.1.5 Summary of Diversity Concept 63
3.2 Concept for Measuring Diversity in Wikidata 64
3.2.1 Diversity in a Knowledge Base Context 64
3.2.1.1 Diversity in Wikipedia 66
3.2.1.2 Diversity in Wikidata 67
3.2.2 The Diversity Conception in Wikidata 69
3.2.2.1 Is Plurality a Synonym for Diversity in Wikidata? . 69
3.2.3 The Proposed Diversity Concept for Wikidata 70
3.2.3.1 Data Diversity 71
3.2.3.2 User Diversity 73
3.3 Summary 76
4 WIKIDATA DIVERSITY STATUS 79
4.1 Overview of Diversity in the Existing Wikidata Literature . 79
4.1.1 User Diversity 80
4.1.1.1 Editor Diversity 80
4.1.1.2 Consumer Diversity 81
4.1.2 Language Diversity 82
4.1.3 Diversity Gaps in Wikidata Research 82
4.1.3.1 Data Diversity in Wikidata 82
4.2 Measuring Data Diversity 83
4.2.1 Diversity in Domain and Class Level 84
4.3 Summary of Existing Wikidata Diversity Status 87
5 RESEARCH DATA & APPROACH 91
5.1 Research Approach 91
5.2 Wikidata-Requests-for-Permissions Dataset 92
5.2.1 Research Methodology 93

5.2.2 Data Collection 94
5.2.3 Classification Process 95
5.3 Wikidata Revision History Dataset 96
5.3.1 Research Method 97
5.3.2 Creating a Sample from the Wikidata Revision History . . . 97
5.3.3 Defining user groups 99
5.3.4 Specifying Edit Types Based on Revisions 100
5.3.5 Analysis of Edit Summaries to determine Edit Types 101
5.4 Summary 102

6 BOTS, DIVERSITY & WIKIDATA 103
6.1 Bots in a Knowledge Base Context 103
6.1.1 Bots in Wikipedia 104
6.1.2 Bots in Wikidata 105
6.2 Bots Social Organization in Wikidata 106
6.2.1 Approving Bot Tasks 106
6.2.1.1 Bot Approval Process in Wikidata 106
6.2.1.2 Task Approval Data 108
6.2.2 Approved Requests 108
6.2.2.1 Activity Type and Focus 109
6.2.2.2 Data Origin 111
6.2.3 Disputed Requests 114
6.2.4 Bots and Wikidata Community 117
6.2.5 Summary of Bots from a Community Perspective 118
6.3 Bot Activities and Contributions in Wikidata 119
6.3.1 Wikidata Edit History 120
6.3.1.1 Overview of Our Sample 120
6.3.2 User edits in Wikidata 122
6.3.2.1 Edit focus Contributions 122
6.3.2.2 Activity Type Usage 122
6.3.2.3 Comparison of Wikidata User Groups 123
6.3.3 Bots in Wikidata Context 127
6.4 Bots and Diversity in Wikidata 129
6.4.1 Diversity of Bots Edits in Wikidata 130
6.4.1.1 Diversity of Edit Domains 130
6.4.1.2 Diversity of Edit Types 134

6.4.2 Impact of Bot Edits on Wikidata Domain Diversity 137

6.5 Summary 139

7 RECOMMENDATIONS ON DIVERSITY IMPROVEMENT 143

7.1 Diversity Improvement through Automation 144

7.1.1 Improving User Diversity 144

7.1.2 Improving Data Diversity 146

7.1.2.1 Domain/ Class Diversity 146

7.1.2.2 Item Diversity 146

7.2 Use Case 148

7.2.1 Herat at the Risk of History Manipulation 148

7.2.2 Proposed Procedure for Preserving the Historical Data 149

7.2.2.1 Establishing a Contributing Community. 150

7.2.2.2 Improving Domain/ Topical Coverage. 152

7.3 Summary 155

8 CONCLUSION 157

8.1 Summary 157

8.2 Future Work 159

A Appendix 161

List of Figures

1.1 Thesis Organization. 7

2.1 An overview of data representation in Wikidata and Wikipedia. 10

2.2 Data model of Wikidata. 16

2.3 Sitelinks and badges on Universe (Q1) Wikidata item. 18

2.4 Article selection process and the number of included search results. (Note: This diagram is inspired by the PRISMA flow diagram [223].) 21

2.5 Frequency of Wikidata research publications per year. Source: *Wikidata Articles Dataset*. 23

2.6 The most popular publishers of Wikidata research. Source: *Wikidata Articles Dataset*. 24

2.7 The most popular conferences on Wikidata research. Source: *Wikidata Articles Dataset*. 24

2.8 Research contributions from countries and continents. Source: *Wikidata Articles Dataset*. (Note: Countries in violet color published research articles before and after 2018 and countries in pink published articles after 2018.) 25

2.9 Most popular research contributing institutions. Source: *Wikidata Articles Dataset*. 25

2.10 Classification of Wikidata research papers (Sum 248) 26

2.11 A comparative view of Wikidata research from 2012 to 2019 and 2019 to 2022. Source: *Wikidata Articles Dataset*. 45

3.1 A visual representation illustrating the qualities of diversity within the context of interdisciplinary analysis, from [301]. 50

3.2 The example which could show contradiction or variety of statements using the *Dmitry Bortniansky* (Q316505) item in Wikidata. 70

3.3 The proposed diversity concept for Wikidata. 71

5.1 An overview of the *Wikidata-Requests-for-Permissions Dataset* preprocessing phase. (Note: This diagram is inspired by the PRISMA flow diagram [223].) 93

5.2 Example of a Wikidata request-for-permissions (RfP) page. 95

5.3 An overview of *Wikidata Edit History Dataset* pre-processing phase. (Note: This diagram is inspired by the PRISMA flow diagram [223].) 98

6.1 Example of a request from the Wikidata community for a bot 107

6.2 Example of a Wikidata requests-for-permissions (RfP) page. 107

6.3 Approved task requests organized into activity type (the upper half part) and edit focus (the lower half part) (n = 794, 262 RfPs are requests with multiple tasks, edit focus *Term* consists of *Alias, Label, and Description*. The edit focus *Miscellaneous* contains the smaller categories, such as *Qualifier*, *Rank*, *Badge*, and *Maintenance related data*). Source: *Wikidata-Requests-for-Permissions Dataset.* 111

6.4 Activity types of approved task requests over time (n = 600, all 81 unsuccessful task requests were not included). Source: *Wikidata-Requests-for-Permissions Dataset.* 111

6.5 Sources of approved task requests over time (n = 600, all 81 unsuccessful task requests were not included). Source: *Wikidata-Requests-for-Permissions Dataset.* 113

6.6 Wikipedia language versions used as sources mentioned on task approval pages (n = 195, all 488 other requests either did not provide a specific Wikipedia language version or did not have Wikipedia as the source). Source: *Wikidata-Requests-for-Permissions Dataset.* 114

6.7 An overview of Edit types per user group (aggregated by edit focus). Source: *Wikidata Revision History Dataset.* 121

6.8 User edits aggregated according to their editing focus. Source: *Wikidata Revision History Dataset.* 123

6.9 Activity types per user group. Source: *Wikidata Revision History Dataset.*123

6.10 Comparison of user groups in the usage of add_alias edit type over time. Source: *Wikidata Revision History Dataset.* 125

6.11 An overview of create_statement edit type per user group over time. Source: *Wikidata Revision History Dataset.* 125

6.12 An overview of create_statement edit type per user groups human and bot over time. Source: *Wikidata Revision History Dataset.* . . . 126

6.13 Heat map of Editor contribution to Wikidata classes. Higher values are represented with lighter turquoise color shades. Data is ordered with the highest values in the top right corner and the lowest values in the left bottom corner of the Figure. In the top right we can see the bot user group followed by humans which had the highest contributions so far, and on the right side are the classes of Media and Geography domains with higher diversity levels, and on the left side classes of Person and Biology domains which have the lowest diversity levels. Source: *Wikidata Revision History Dataset.* 132

6.14 Heatmap of edit types performed by user groups in Wikidata revision sample. Higher values are represented with lighter turquoise color shades. Data is ordered with the highest values in the top right corner and the lowest values in the left bottom corner of the Figure. As can be seen, the most used edit types are the ones performing create, add, and update activities on the right side, while, edit types on the left side are the ones mainly used by human users with low amounts. Source: *Wikidata Revision History Dataset.* 136

List of Tables

2.1 Search results from academic search engines. 20

2.2 Data extraction form 22

3.1 A comparative view of diversity measurement formulas. The notions used in the formulas: N is the category count, p_i the proportion of a category (e.g., class) in an entity (e.g., domain), n_i is the number of elements in the category i, d_{ij} is a disparity between i and j, N_{max} is the total number of elements belonging to the most abundant categories. A is the area between the Lorenz curve (the actual distribution of income or wealth) and the line of perfect equality (a diagonal line representing perfect equality), and B is the area under the line of perfect equality. 56

3.2 An example of Wikidata editor diversity to show the usage of diversity measures. Here we have four Wikidata items with their respective number of editors from different countries. Items A and B are similar except that Item B has a larger sample, Item C has been edited by editors from more countries, and Item D has a more balanced number of editors. (Note: Higher values show better diversity levels, except for HHI where a higher value shows more concentration.) 61

3.3 Wikipedia research papers on diversity with the aspect of diversity being explored and the diversity measurements used. 66

3.4 An overview of Wikidata Diversity Measurement Concept. Each section of user or data diversity angles can be measured using the metrics provided considering the diversity dimensions and the measurement approaches. (Note: In this table V = Variety, B = Balance, D = Disparity, Measure = M, Single-Concept = SC, and Dual-concept = DC. The metrics and diversity dimensions are not limited to the values provided here.) 72

4.1 Descriptive statistics of Wikidata domain diversity levels using the number of classes and items in each domain based on Table A.1. (Note: Higher values mean higher diversity levels and bold values indicate the highest diversity in each column) 85

4.2 Balance of properties across items of the classes and domains in Wikidata based on [65] using the Wikidata Knowledge Imbalance Dashboard tool that runs on Wikidata and limits the number of items to 10,000 (The classes with * are added because they appeared in Wikidata research) 86

5.1 Example of the information extracted from an RfP page. 95

5.2 Overview of the Terminology Used for Classification of Edits in Wikidata. 100
5.3 Definition of activities. 101

6.1 Overview on task requests ($n = 683$, number of all task requests). Source: *Wikidata-Requests-for-Permissions Dataset*. 109
6.2 Task requests categorized into edit focus, activity type, and the request result (approved/denied) ($n = 681$, number of all task requests). Source: *Wikidata-Requests-for-Permissions Dataset*. 110
6.3 All internal, the top 10 external sources most used and unknown sources. Source: *Wikidata-Requests-for-Permissions Dataset*. 112
6.4 Top 5 most edited RfPs closed as unsuccessful. 113
6.5 Main reasons for unsuccessful task requests ($n = 81$, number of unsuccessful task requests). Source: *Wikidata-Requests-for-Permissions Dataset*. 116
6.6 *Number of revisions, users, and items per user group.* The tool users can also appear in the registered and automatic contributions, since users that made semi-automatic contributions only are 201, 854 tool users that were active as humans as well, and 31 tool users also appear as bots. In parenthesis, the relative number is given. Source: *Wikidata Revision History Dataset*. 120
6.7 Domain Diversity based on the number of revisions from *Wikidata Revision History Dataset*. (Note: Higher numbers indicate greater diversity and bold values represent the highest diversity for a domain calculated from the diversity measure in each column.) 131
6.8 Descriptive statistics of user group diversity based on their contributions to Wikidata classes. (Higher values show higher diversity). Source: *Wikidata Revision History Dataset*. 133
6.9 Descriptive statistics of user group diversity based on the usage of edit types. Source: *Wikidata Revision History Dataset*. 135

A.1 Overview of Wikidata Topical Domains and classes showing the distribution of their items and maturity level (20.01.2020) 162
A.2 An overview of *Wikidata Revision History Dataset* aggregated by edit focus. (Note: Bold numbers show highest values in a row.) 163
A.3 Codebook for Mapping Edit Summaries to Edit Types in *Wikidata Revision History Dataset* 164

Definition of Terms

Term	**Definition**
Comment:	Comment-text is a column of the Comment Table, which contains the revision comment on every edit/change made to a page.
District:	A second-level administrative subdivision of a country.
Edit:	A general term used to refer to any activity performed on data within Wikidata.
Entity:	The content of a Wikidata page which is identified with an identifier like an item or property.
Item:	A Wikidata page containing information about a topic or concept which is identified with an identifier starting with the letter *Q* and preceded by numbers. Example: *Q64* represents the item Berlin.
KB:	A Knowledge Base (KB) is a centralized repository of data that stores data in any form, such as in a tabular or graph format.
KG:	A Knowledge Graph (KG) is a knowledge base that stores the data in graph format.
Property:	A Wikidata page identified with an identifier starting with the letter *P* and preceded by numbers. Example: *P1376* is the property for *capital of* which provide the information on *Berlin Q64* being the capital of *Germany Q183* and links both items in a linked data environment.
Revision:	Each Wikidata edit is stored in the database and shows who edited what (see Section 6.3.1).
Wikimedia Community	The Wikimedia community comprises a global network of volunteers from diverse backgrounds who actively contribute to Wikimedia projects. They collaborate to provide free knowledge to all. Community members write, edit, and curate content, add images, translate, and ensure project integrity. Operating on transparency and shared goals, they engage in discussions, make policy decisions, organize events, and uphold core principles like neutrality and verifiability, guided by the ‘Five Pillars of Wikipedia. Their collective efforts drive the growth and improvement of the projects[1].
Wikimedia Foundation	The Wikimedia Foundation is a non-profit organization established in 2003 and offers infrastructure, technical support, and funding to sustain and enhance projects like Wikipedia, Wiktionary, and Wikidata. Staff members work on software development, community outreach, fundraising, legal matters, and public relations. The foundation’s policies ensure project integrity and foster a supportive environment for the global Wikimedia community[2].

[1] https://wikimediafoundation.org/our-work/

[2] https://wikimediafoundation.org/

Diversity Specific Terms

Term	**Definition**
Balance:	In a system with categories, the apportionment of elements into categories shows balance (see Section 3.1.2).
Bias:	a) A personal and sometimes unreasoned judgment. b) Systematic error introduced in sampling or testing by selecting or encouraging one outcome or answer over others [3]
Concentration :	Exclusive attention to one object[4].
Disparity:	A noticeable and usually significant difference or dissimilarity[5]. In a system with categories, the degree of dissimilarity between categories shows disparity (see Section 3.1.2).
Equity:	Freedom from bias or favoritism[6].
Fairness:	Lack of favoritism toward one side or another[7].
Inclusion:	The act or practice of including and accommodating people who have historically been excluded (because of their race, gender, sexuality, or ability)[8].
Plurality:	The state of being numerous[9]. A Wikidata design decision that refers to the coexistence of multiple statements (see Section 3.1.2).

[3] https://www.merriam-webster.com/dictionary/bias

[4] https://www.dictionary.com/browse/concentration

[5] https://www.merriam-webster.com/dictionary/disparity

[6] https://www.merriam-webster.com/dictionary/equity

[7] https://www.merriam-webster.com/dictionary/fairness

[8] https://www.merriam-webster.com/dictionary/inclusion

[9] https://www.merriam-webster.com/dictionary/plurality

Chapter 1

INTRODUCTION

Wikidata, the sister project of Wikipedia, is a freely accessible and structured knowledge base (KB) with the aim of serving the collective knowledge of the world. Reflecting the world knowledge means representing the knowledge in all the existing languages on Earth, including all of the different viewpoints and beliefs. As a project of Western origin, it is pertinent to assess the extent to which Wikidata has succeeded in fulfilling its mission of serving “anyone anywhere in the world.” In other words, how viable is it for Wikidata to store knowledge pertaining to all inhabitants of our planet? In particular, the Wikimedia Foundation[1], which oversees various Wiki projects including Wikidata, has set the goal of achieving “knowledge equity” by 2030 and aims to become the fundamental infrastructure for the ecosystem of free knowledge. Our primary focus is to assess the progress of Wikidata in its endeavor to become a comprehensive repository of global knowledge, especially as it approaches the end of its first decade of existence. This issue becomes more significant when we consider that the majority of edits are carried out by a limited number of automated accounts, overshadowing the contributions of a larger number of human users.

1.1 Research Motivation

In the current digital era, the absence of digitalization can pose challenges to accessing and preserving information, leading to a potential loss of valuable knowledge over time. This limitation can restrict our understanding of the world, as we may only have access to a limited set of facts that are available in digital formats. The use of digital assistants has become a regular part of our daily lives, providing answers to diverse inquiries ranging from weather forecasts and commute updates to historical information. These digital assistants rely on KBs as sources of information, which play a vital role in digitally preserving and providing access to knowledge.

While there are multiple KBs available, Wikidata was specifically developed to serve as a structured repository of knowledge that is accessible not only to humans but also to machines with the slogan of “serving anyone anywhere in the world.” Serving

[1] https://wikimediafoundation.org

the global population means having the capacity to reflect the diverse languages, topics, and opinions/ideas that exist worldwide. Becoming a comprehensive source of world knowledge comes with challenges, particularly in a structured KB where data is stored in a format distinct from plain text and adheres to a defined data model for improved accessibility and utilization. There are numerous disputed issues, and finding a universally accepted statement can be difficult. For example, the status of Kashmir is a subject of dispute between India, China, and Pakistan. Storing such facts requires a structured KB that can accommodate contradictory statements. Wikidata addresses this by allowing the coexistence of multiple statements, known as plurality, which is a key design decision that distinguishes it from other KBs.

Furthermore, in the 21st century, our world is shaped by the ups and downs of human history. All historical events contribute to our collective knowledge and provide insight into our current circumstances. It is crucial for a KB to include relevant historical events in order to preserve our history. Wikidata makes this possible through its pluralism feature, which allows for the storage of multiple values related to specific topics of interest.

So far, Wikidata appears to have the potential to become a global KB. However, as a native Persian speaker with an Eastern background, I am curious to know how well my language and culture are represented in Wikidata. I wonder if I can find my values and topics of interest there, despite its Western origin. Moreover, I am interested in understanding how the plurality feature of Wikidata can help me preserve my history and culture, safeguarding them from being altered by the ongoing crises and political agendas in Afghanistan. In this current period, there is a risk of manipulation and gradual erasure of my language and culture. In August 2021, a terrorist group[2] has taken control of Afghanistan. In addition to numerous other unlawful acts[3], they have initiated the removal of Persian from formal communication areas, such as government letters[4] and signage in government institutions[5]. Moreover, there has been a significant increase in forced displacements of the local population[6], aimed at altering the ethnic fabric of the geography. Attempts to eliminate the Persian language and its speakers in the past [159] have resulted in multiple areas being renamed from Persian to Pashtu, such as Sabzawar renamed to Shindand[7], and Qeratapa renamed Turghundi[8] that were the historical names of these districts [356]. These name changes were intended to showcase the dominance of Pashtun culture for political reasons. Thus, it is essential to preserve our language and culture in order to protect our history from manipulation.

In addition, Wikidata, which serves as a widely utilized source of structured data for both humans and machines, recently celebrated its tenth anniversary. This is an opportune moment to reflect on the extent of Wikidata's success in realizing its objective of providing service globally to individuals across the world. Investigating such a matter necessitates a comprehensive comprehension of the design decisions employed in Wikidata, as they serve as the distinguishing characteristics that set it

[2] United States List of Foreign Terrorist Organizations[Accessed 20.12.2023]

[3] One Year of Taliban Rule in Afghanistan: A Predictable Disaster[Accessed 12.10.2022]

[4] Taliban abolishes the Persian language from Supreme Court bill [Accessed 12.10.2022]

[5] Taliban Group Removes Persian from the Sign Boards at Education Directorate of Herat Province. [Accessed 12.10.2022]

[6] Forced displacements of ethnic groups in Afghanistan[Accessed 13.12.2022]

[7] `https://fa.wikipedia.org/wiki/`

[8] `https://fa.wikipedia.org/wiki/`

apart from other KBs. This understanding can help uncover the primary purpose for which Wikidata was developed. Examining these design decisions, we observe that each of them implicitly relates to the concept of diversity. Therefore, diversity appears to be the overarching objective of Wikidata. One specific design decision, known as plurality, explicitly facilitates support for diversity by allowing the storage of multiple statements together. Despite plurality being one of the key factors that enable Wikidata to serve the diverse human population on Earth, it has received limited attention in existing research on Wikidata. Our knowledge about the concept of plurality and its role as the sole representative of diversity in Wikidata remains limited. Furthermore, existing research on Wikidata highlights an intriguing aspect of the Wikidata community: a small group of contributors perform most of the editing through automation. As contributions to Wikidata are made by a community of volunteers without a defined plan for data input, the data being contributed is more likely to reflect the interests and values of the editors in their respective languages. Consequently, automation can become a dominant factor that overshadows manual edits and presents a challenge to the objective of achieving sufficient diversity to serve the global population.

Hence, while diversity remains a crucial factor in helping Wikidata fulfill its ultimate goal of becoming a global KB, it is essential to acknowledge the long-standing reliance of the Wikidata community on the automation of edits through bots. This reliance has persisted for a decade, posing further considerations for ensuring diversity within Wikidata. Bots, despite being fewer in number (less than 500), have a substantial impact on the volume of edits in Wikidata, surpassing the combined edits of tens of thousands of human users. Given that bots perform the majority of edits, their contributions undoubtedly have a significant influence on the data within Wikidata.

This situation raises the question of how effectively bots contribute to the overarching goal of Wikidata to become a global KB. However, due to the vast scope of this question, it is impractical to address it in a single attempt. Therefore, we break down this question into smaller, more manageable research inquiries.

In the following sections, we present our research questions and outline our approach to addressing them.

1.2 Research Questions

The design principles of Wikidata (cf. Section 2.1.1) emphasize the importance of diversity in achieving its overarching goal. This commitment to diversity aligns with the Movement Strategy Process of the Wikimedia Foundation, which aims to achieve *knowledge equity* by 2030[9].

However, there have been no studies conducted thus far to assess the progress of Wikidata in achieving its overall goal. After a decade, has Wikidata made significant strides toward serving as a repository of global knowledge, or are further measures required to help it fulfill its overarching goal? Notably, research has revealed the dominance of Western languages over other languages in Wikidata, and the majority of edits in this collaboratively edited KB are performed by bots, which are fewer in number (i.e., hundreds) compared to active human users (i.e., 20K+).

[9] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20

Considering our basic understanding of bots, we assume that bot edits are less diverse compared to human edits, as humans tend to automate simple, repetitive, and time-consuming tasks. The substantial volume of bot edits implies the prevalence of repetitive and straightforward changes over more diverse edits. This raises questions such as: How has the extensive use of bots in Wikidata impacted the diversity of topics, perspectives, and languages within the platform? Existing research indicates that Western languages currently dominate Wikidata, while other languages appear to receive less attention. This suggests that Wikidata is primarily utilized in the West and may not be serving everyone everywhere in the world as intended. Therefore, Wikidata has yet to achieve its original goal. Given that bots have played a significant role in editing, it is important to investigate whether they have contributed to the dominance of certain languages over others. Furthermore, we need to explore how the dominance of specific languages affects the diversity objective of Wikidata and identify strategies to foster greater diversity to cater to a more varied audience. The aforementioned issues, among others, may arise when considering this topic. However, our study focuses on providing an overview of the diversity of Wikidata, examining the impact of bot edits on its diversity status, and utilizing bots for diversity improvements. We have defined the following research questions for this study:

RQ1. What is the current status of Wikidata in terms of diversity?

1.1 What does diversity mean in the Wikidata context?

1.2 How to measure the diversity status of Wikidata?

1.3 How diverse is Wikidata at the current point in time?

RQ2. How do bots and their high volume of edits impact diversity in Wikidata?

2.1 How do bots contribute to editing Wikidata and where do they stand in comparison to human users?

2.2 Do bots possess the potential to influence diversity within Wikidata?

2.3 How can bots contribute to enhancing diversity in Wikidata?

By thoroughly investigating these questions and presenting detailed insights, this study aims to contribute to the existing knowledge and understanding of the topic. The following chapters offer a deeper exploration and analysis of the research questions, providing readers with a comprehensive understanding of the research process and the significance of the results obtained.

1.3 Research Contributions

This thesis contributes to the existing body of knowledge in several ways. The main contributions of this thesis are as follows:

- Presenting an overview of the existing literature on Wikidata with a compiled dataset that is freely available online,
- Defining diversity in a KB context and developing a conceptual framework for measuring diversity in a KB,

- Presenting of the current state of diversity in Wikidata,
- In-depth analysis of bots and their impact on diversity in Wikidata,
- Compiling two datasets related to bots that are freely available online with their respective codebooks.

Overall, this thesis makes a significant contribution to the broader understanding of diversity within KBs and the role of automation in shaping it, with a specific emphasis on Wikidata. By examining various dimensions of diversity and investigating the impact of bot edits, the thesis provides valuable insights into the complexities and challenges associated with achieving diversity in structured knowledge repositories.

1.4 Thesis Organization

Figure 1.1 provides an overview of the organization of the chapters of this thesis.

Chapter 2: Wikidata. In this chapter, we provide an in-depth exploration of Wikidata KB. This chapter explores the purpose and objectives behind the creation of Wikidata, shedding light on the design decisions that shape its structure and functionality. Additionally, it examines the contributing community that drives the growth and development of Wikidata.

Furthermore, this chapter offers a comprehensive review of the existing research on Wikidata. By examining previous studies and scholarly work, we aim to provide a state-of-the-art perspective on the advancements and insights gained in the field of Wikidata research. This review serves as the foundation for this study, informing the approach used and contributing to the general understanding of the current state of Wikidata knowledge.

Chapter 3: Diversity & Wikidata. This chapter is dedicated to delving into the topic of diversity within the context of Wikidata. This chapter sets the stage by providing the basics of diversity, starting with the definition of the term "diversity" and exploring how it is interpreted and measured in various other contexts. Based on these insights, the chapter then introduces a unique concept for measuring diversity that is specifically tailored to the Wikidata KB environment.

Chapter 3 serves as a crucial stepping stone in the thesis, offering an in-depth analysis of diversity and laying the groundwork for subsequent chapters. It provides a theoretical framework that informs the research approach undertaken to investigate and improve diversity in Wikidata.

Chapter 4: Wikidata Diversity Status. Using the proposed diversity measurement concept, Chapter 4 offers a comprehensive overview of the current state of diversity within Wikidata. This chapter examines the degree of diversity present in Wikidata and provides a detailed analysis of the findings. Through this analysis, we aim to uncover any existing gaps or imbalances in Wikidata and gain insight into the distribution and representation of information across different categories. The chapter serves as a foundation for further exploration and discussion on enhancing diversity in Wikidata.

Chapter 5: Research Data & Approach. Based on the findings presented in Chapter 4, further investigation is required to understand the underlying reasons for the current diversity status of Wikidata.

Chapter 5 provides a thorough and detailed account of the research approach utilized in this study. It outlines the specific data requirements and describes the procedures employed for data collection. The chapter discusses the essential pre-processing steps undertaken to ensure the data's reliability and suitability for analysis.

Chapter 6: Bots, Diversity & Wikidata. This chapter provides an in-depth exploration of bots in the Wikidata ecosystem, shedding light on their social dynamics, contributions, and their influence on the diversity landscape. Through this examination, we gain a deeper understanding of the intricate relationship between bots and diversity within the context of Wikidata.

Chapter 7: Recommendations on Diversity Improvement. In Chapter 7, we delve into the realm of addressing the diversity gaps identified and the impact of bots on diversity within Wikidata. Building upon our findings, this chapter presents our recommended approach to enhance diversity in Wikidata through the utilization of bots. We outline specific strategies and methods that can be employed to bridge the diversity gaps and promote a more inclusive and balanced representation of knowledge within the platform.

Chapter 8: Conclusion. Following the presentation of our proposed approach, we conclude our study by summarizing the key insights and implications derived from our research. We reflect on the significance of our findings and their implications for the broader field of knowledge curation and representation. Additionally, we offer suggestions for future research directions, highlighting potential avenues for further investigation to advance the understanding and improvement of diversity in knowledge bases such as Wikidata.

With this final chapter, we bring our study to a close, underscoring the importance of diversity and offering a pathway towards its improvement in Wikidata.

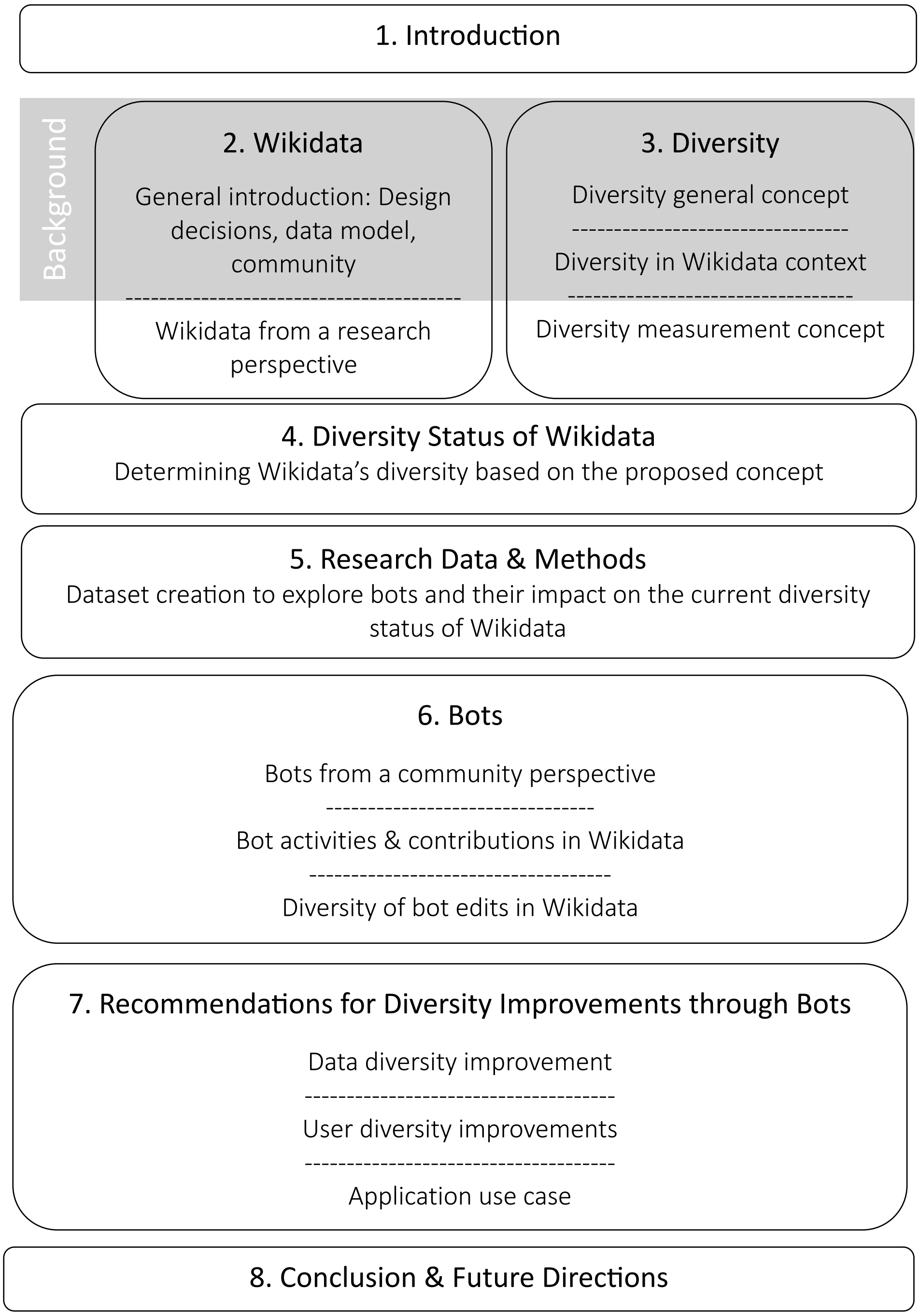

1. Introduction
Background
2. Wikidata
General introduction: Design decisions, data model, community
Wikidata from a research perspective
3. Diversity
Diversity general concept
Diversity in Wikidata context
Diversity measurement concept
4. Diversity Status of Wikidata
Determining Wikidata's diversity based on the proposed concept
5. Research Data & Methods
Dataset creation to explore bots and their impact on the current diversity status of Wikidata
6. Bots
Bots from a community perspective
Bot activities & contributions in Wikidata
Diversity of bot edits in Wikidata
7. Recommendations for Diversity Improvements through Bots
Data diversity improvement
User diversity improvements
Application use case
8. Conclusion & Future Directions

Figure 1.1: Thesis Organization.

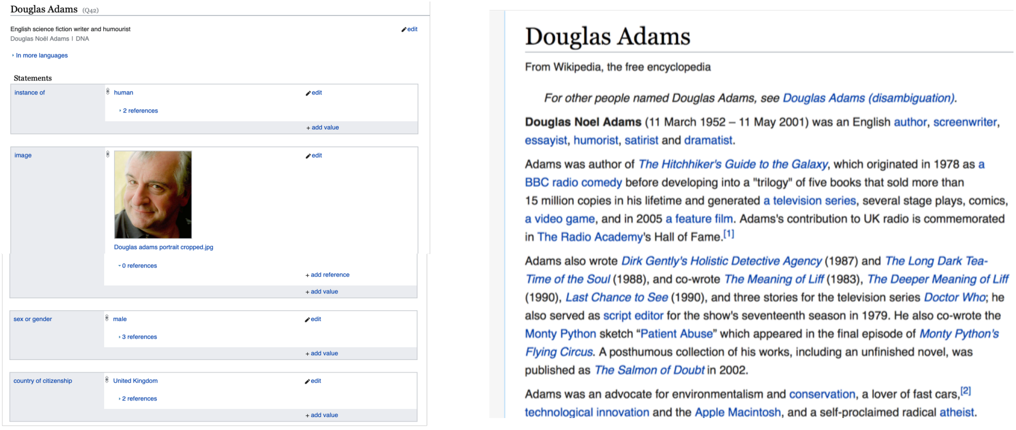
Chapter 2

WIKIDATA

2.1 Background

Wikidata is a structured KB developed by the Wikimedia Deutschland, the German Chapter of the Wikimedia Foundation. It was launched on 29 October 2012, with the primary goal of addressing the data inconsistencies present in Wikipedia [326]. Wikipedia, the most popular collaboratively edited KB by the Wikimedia Foundation, had been launched a decade prior to Wikidata and already contained millions of articles in more than 280 languages. Despite the valuable data stored in Wikipedia, it is in text format, making it difficult to directly access and utilize for purposes such as reuse or analysis. This limitation served as another driving factor behind the development of Wikidata [326]. Hence, Wikidata aims to serve as a centralized storage, providing interwiki links and infobox data that span numerous language editions of Wikipedia. Wikidata works towards enhancing the consistency and quality of Wikipedia articles while making information more accessible in smaller language versions of Wikipedia. It also reduces the maintenance workload of the Wikipedia community volunteer contributors [260]. For example, with Wikidata, there is no longer a need to manually update over 280 language versions of Wikipedia regarding the population of Berlin after a new consensus. This data becomes available to all infoboxes in every language version of Wikipedia upon its entry into Wikidata.

The issue of data inconsistency exists in Wikipedia articles across different language versions because each version of Wikipedia is independent and contains different statements and opinions that reflect the beliefs of its respective community. While Wikipedia allows for a diversity of opinions, accessing or analyzing them together to gain an overview of the most disputed or globally agreed topics is challenging due to the text format of Wikipedia articles (cf. Figure 2.1b on page 10). Furthermore, users outside of that language community typically do not utilize a significant portion of the information in a particular version of the language due to their unfamiliarity with the language. As a result, existing data cannot be efficiently reused or utilized for analysis purposes.



Douglas Adams (Q42)

English science fiction writer and humourist
Douglas Noël Adams | DNA

Statements

instance of: human
2 references

image: Douglas adams portrait cropped.jpg
0 references

sex or gender: male
3 references

country of citizenship: United Kingdom
2 references

(a) Wikidata item in structured format

Douglas Adams

From Wikipedia, the free encyclopedia

For other people named Douglas Adams, see Douglas Adams (disambiguation).

Douglas Noel Adams (11 March 1952 – 11 May 2001) was an English author, screenwriter, essayist, humorist, satirist and dramatist.

Adams was author of *The Hitchhiker's Guide to the Galaxy*, which originated in 1978 as a BBC radio comedy before developing into a "trilogy" of five books that sold more than 15 million copies in his lifetime and generated a television series, several stage plays, comics, a video game, and in 2005 a feature film. Adams's contribution to UK radio is commemorated in The Radio Academy's Hall of Fame.[1]

Adams also wrote *Dirk Gently's Holistic Detective Agency* (1987) and *The Long Dark Tea-Time of the Soul* (1988), and co-wrote *The Meaning of Liff* (1983), *The Deeper Meaning of Liff* (1990), *Last Chance to See* (1990), and three stories for the television series *Doctor Who*; he also served as script editor for the show's seventeenth season in 1979. He also co-wrote the Monty Python sketch "Patient Abuse" which appeared in the final episode of *Monty Python's Flying Circus*. A posthumous collection of his works, including an unfinished novel, was published as *The Salmon of Doubt* in 2002.

Adams was an advocate for environmentalism and conservation, a lover of fast cars,[2] technological innovation and the Apple Macintosh, and a self-proclaimed radical atheist.

(b) Wikipedia article in text format

Figure 2.1: An overview of data representation in Wikidata and Wikipedia.

To address this issue, the concept of a centralized KB was developed, where all data could coexist and be easily accessed and queried. This led to the creation of Wikidata. Despite the existence of structured KBs like DBpedia[1] before the launch of Wikidata, which also utilizes data from Wikipedia [136], Wikidata was developed with distinct design decisions and objectives that set it apart as a unique KB.

This chapter introduces the fundamental concepts related to Wikidata. We begin by discussing the unique features and design decisions that distinguish Wikidata from other existing KBs. An overview of these design decisions not only helps us understand the main goals of Wikidata but also provides insight into the underlying data model, which is further explained in this chapter.

As the data in Wikidata are contributed by a community of volunteers, we also present an introduction to the Wikidata contributing community to shed light on the collaborative editing dynamics within Wikidata. Considering the substantial size of the Wikidata community, we explore the popularity of Wikidata worldwide and its progress toward the ultimate goal of providing free knowledge to all through a research lens.

Furthermore, this chapter delves into the existing literature on Wikidata, extending this introduction from a research perspective. It identifies the primary goals for which Wikidata was developed, examines the progress made thus far, and highlights areas that require further study and exploration.

Thus, this chapter serves as both an introduction to Wikidata and a comprehensive overview of its goals, achievements, and areas for future investigation.

2.1.1 Goals and Design Decisions

To answer the question of why another KB?, while there were many KBs before the launch of Wikidata, we need to find the distinguishing features that Wikidata has by reflecting on the design principles on which Wikidata was developed.

[1] DBpedia is a KB that aims to extract data from various Wikimedia projects in a structured format. It is available at: https://www.dbpedia.org/.

The ultimate goal of Wikidata is to serve individuals globally. This means that Wikidata wants to capture the world's knowledge so that we can serve it to any individual from any part of the world. Considering this goal which distinguishes Wikidata from other existing KBs, a number of design decisions were defined for Wikidata to make it capable of this feature. These design decisions according to Vrandečić and Krötzsch are open editing, community control, plurality, secondary data, multilingual data, easy access, and continuous evolution [326].

Looking into the details of these design decisions gives us the impression that they all refer to the diversity concept in some way or another. In the following, we provide a glance into each of these design decisions and how each one might be related to the diversity concept:

Open editing. Wikidata allows contributions from anyone, with or without a Wikidata user account [326]. This shows that Wikidata is willing to also store the knowledge of individuals who are not part of the Wikidata community. This way it is open to a broader and more diverse range of the world population who don't want to be tied to a community but have something to contribute. Additionally, more contributions from *anyone* can increase the chances of being used by *anyone*.

Community control. The Wikidata community contributions are not limited to data contribution alone, but also to decide the schema of the data [326]. Administrators are periodically elected by the Wikidata community, and all existing issues are discussed, planned, and decided through the Wikidata community portal[2]. The community decides on the properties that relate to each item class. While, adding items is a rather easy task that any user can do, adding new properties needs more detailed discussions in the community. This shows that Wikidata is not always run by a specific group of people, but individuals from any part of the world who show expertise and commitment through their contributions can get the chance to get involved in the administrative issues and decide for Wikidata. Therefore, Wikidata can benefit from the insights and expertise of dedicated contributors with diverse backgrounds from around the world.

Plurality. Wikidata provides a mechanism for storing multiple values for the same data. This is because Wikidata wants to centralize the data and allow all of the existing or possible values to coexist, even if they contradict each other [326]. This allows people with different views to contribute to the same topic and state their part of the fact. Similarly, it allows the world to identify it as a disputed topic without having to judge or decide on the right one [326].

Additionally, this feature provides a more complete picture regarding an entity, e.g., a scientist might have had multiple scientific discoveries that need to be addressed as scientific achievements. Furthermore, it gives the possibility to look at how a topic has evolved over time. For example, when looking at the population of a city or a country in Wikidata, we can see the ups and downs of the population numbers from the last decade or century and see how the population has increased or decreased over time.

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

In Wikidata plurality can exist in the following parts of an item: Multiple labels from different languages, multiple aliases, more than one value for properties, qualifiers, and references, or multiple statements in an item.

Secondary data. Wikidata is not a primary source of knowledge, but a secondary source that stores facts from the primary sources of knowledge like books, articles, or encyclopedias, as some examples. For this reason, any piece of data needs a proper source to be considered reliable in Wikidata [326]. Incorporating data from a range of sources results in a greater diversity of topics and perspectives. This can attract a broader audience, as individuals can discover subjects of personal interest to contribute to or utilize.

Multilingual data. Data in Wikidata are stored in a language-independent form with the help of unique ids for items (Q followed by a number) and properties (P followed by a number). The idea is to have all the data in a centralized form where the values are universal and the labels can be translated into multiple languages [326]. This avoids data inconsistencies that exist in Wikipedia language versions and allows all to have access to the same values in their own languages. This is in contrast to Wikipedia where every language version is independent of other languages and only the language community of a language is capable of editing and consuming that language, while, it remains not understandable and, therefore, inaccessible to the rest of the world outside that specific language community. This feature provides a unified view of the knowledge with the same values in all languages.

Easy access. Wikidata was developed to serve as a structured source of data for Wikipedia [326], however, today it is used in many projects beyond the Wikimedia Foundation. Data are accessible in various formats to reach more audiences, including JavaScript Object Notation (JSON) and Resource Description Framework (RDF) and are also reachable through the Wikidata Query Service (WDQS). [326]. Additionally, data are made available in the form of dump files for researchers dealing with approaches that generate sizable datasets. For standard data access, the user interface and query service assist users in locating the necessary information.

Continuous evolution. Wikidata is developed to evolve with the emerging needs of its community [326]. It is under constant refinement and is being enhanced through community feedback on the existing system and its needed features. Being continuously evolving means being more flexible to counter knowledge from diverse sources with varying formats.

In summary, upon closer examination of these design decisions, it becomes apparent that all of the aforementioned design principles, in various ways, align with the concept of diversity. These design choices collectively contribute to the overarching goal of fostering a diverse and inclusive KB. For instance, open editing allows contributions from both registered and unregistered users, ensuring that diverse voices and perspectives can be represented. Community control ensures that decision-making power is not centralized and that contributors from various backgrounds can shape the direction of Wikidata. Plurality allows for the coexistence of different beliefs and opinions within the KB. The inclusion of data from various sources promotes the integration of diverse information. Multilingual support enables the representation of knowledge in multiple languages, catering to a global audience. Easy access to data ensures that it can be utilized in various ways for different purposes.

Therefore, it can be inferred that diversity plays a crucial role in achieving the ultimate goal of Wikidata, which is to serve “anyone anywhere in the world” by reflecting the diversity of world knowledge. Understanding the extent to which Wikidata has been successful in achieving this goal requires exploring the concept of diversity within the platform.

As mentioned before, one notable aspect of Wikidata is its openness to both registered community members and non-registered users. While anyone can benefit from the Wikidata data, community members have additional responsibilities and decision-making authority that influence the way Wikidata functions. Therefore, exploring the Wikidata community and its dynamics presents an intriguing avenue, especially concerning the topic of diversity, as the content on Wikidata results from the collaborative contributions of its participants. The subsequent section offers a description of the Wikidata community.

2.1.2 Contributing Community

In a collaborative KB like Wikidata where data is considered the focus of the system, users contributing to this data also have an influence on the data and the system. Wikidata’s contributing community, as mentioned earlier, not only contributes data but also decides on the schema of these data. Further, we are interested in understanding how users impact diversity in Wikidata. For this reason, we introduce the existing types of Wikidata contributors.

Existing research on the Wikidata community identifies that the main contributing user groups on Wikidata are human users and bots [206]. Their edits are accordingly categorized as manual and automated edits. Recent studies have also identified tools that exhibit distinguishing features from contributions made by bots and human users. Tools perform a visible number of edits in Wikidata and are called semi-automatic edits [270]. The edits done without being logged in or through registered user accounts are called anonymous edits. As there is not enough information stored regarding these anonymous editors, we cannot determine their automation level.

In Wikidata registered users can be granted any of the user access levels based on the instructions on the Wikidata website[3]. However, since Wikidata is open and allows unregistered edits, not all edits belong to the defined user access levels. Therefore, we follow the definition of the Wikidata user groups based on the differentiating features of their edits in Wikidata, which are revealed in the research and are defined as follows. In the following, we provide an introduction to each of the user groups mentioned which are identified in Wikidata research.

Human Users Human users are the main contributors to any KB, including Wikidata. They are the controllers of Wikidata and, therefore, can get any access level from a normal editor to a bureaucrat or administrator, which are decisive positions. As mentioned earlier, the existing Wikidata research on human users reveals the

[3] Wikidata:User access levels: https://www.wikidata.org/wiki/Wikidata:User_access_levels

role change of these users over time in Wikidata from simple editors to becoming community members and taking more responsibilities than just editing [240, 270].

In Wikidata there are more than 6 million registered users, and among them are around 24K active users[4] (i.e., users who perform at least five edits per month). The majority of these users are human users, other types of users (i.e., bots, tools, and anonymous) each have certain differentiating attributes which we explain in the following. Any users without such differentiating attributes are considered human user accounts.

Bots. According to the Wikidata definition[5], "bots are tools used to make edits without the necessity of human decision-making". They can add any form of data like item, term, statement, and source among other activities they can perform in Wikidata. Although they are user accounts created and operated by humans, their automated and unique high-speed editing style has given bots a differentiating identity. Based on Wikidata policy, bot accounts should contain the word *bot* somewhere in the username of the account for easier distinction from other user groups[6]. The bot policy also mentions that bot accounts need to go through a process called *Approval Process*[7] in order to get permissions to operate as a bot account. Currently, there exists a list of 355 flagged bot accounts on the Wikidata website[8].

In the existing research on Wikidata we could see that despite their high levels of activity and possible impact on Wikidata, bots are yet a rather unexplored user group. A number of studies have only focused on the editing activities of this user group [206, 298, 114], and have shown concern about their possible impact on Wikidata [238]. Although bots are run by human operators, their ability to perform batch edits with their extensive use in Wikidata has given them a distinguishing feature from other human user accounts. For this reason, it makes us curious to know more about bots like what editing patterns they have, how can we define their collaboration with other user groups, and what types of activities or edits are mostly performed through them. In addition, the questions of how they impact the quality and diversity of Wikidata are yet to be explored.

Tools Tools are features introduced in Wikidata to facilitate data access to Wikidata more quickly and comfortably[9], Wikidata Query Service is one of the most popular examples of the tools to search and access Wikidata items [10]. Currently, there exist nearly one thousand Wikimedia related tools[11], of which around 200 tools are Wikidata specific tools[12]. These tools are developed and used for differ-

[4] Wikidata statistics: https://www.wikidata.org/wiki/Special:Statistics[Accessed 20.08.2023]

[5] Wikidata Bots: https://www.wikidata.org/wiki/Wikidata:Bots [Accessed 04.01.2021]

[6] Bot accounts: https://www.wikidata.org/wiki/Wikidata:Bots [Accessed 04.01.2021]

[7] Approval process: https://www.wikidata.org/wiki/Wikidata:Bots [Accessed 04.01.2021]

[8] Wikidata: List of bots: https://www.wikidata.org/wiki/Wikidata:List_of_bots [Accessed 11.05.2023]

[9] Wikidata Tools: https://www.wikidata.org/wiki/Wikidata:Tools [Accessed 05.01.2021]

[10] WDQS: https://query.wikidata.org/ [Accessed 05.01.2021]

[11] Tools Directory: https://hay.toolforge.org/directory/# [Accessed 05.01.2021]

[12] Wikidata Tools: https://hay.toolforge.org/directory/#/search/wikidata [Accessed 05.01.2021]

ent purposes, such as adding and editing items (e.g., QuickStatements[13]), querying Wikidata (e.g., Scholia[14]) and visualizing data (e.g., Reasonator[15]).

The tools are distinct from the bot user group as they are not separate user accounts. Rather, they are features used by Wikidata users to enhance efficiency and reduce the time required for editing. For this reason, edits performed with tools are called semi-automated edits. As the usage of tools also usually speeds up the editing process and has a differentiating effect on the editing behavior of the users, tools are considered a separate user group. Edits performed with tools usually have a trace of the tool name in the comment section of the Wikidata revision history, and therefore, such edits can be identified[16] as done in [270].

Anonymous Earlier we mentioned that Wikidata allows anyone to access and edit Wikidata, even if they are not logged in. Edits performed by unregistered users are called anonymous edits and belong to the anonymous user group. As these users are not logged in, there is no information available about these users except their IP addresses.

In short, in the Wikidata research, we could identify four user groups which are human, bot, tool, and anonymous. Bot and tool user groups have special rights to perform high-speed edits through automation for editing Wikidata, and human and anonymous users perform manual edits. Thus, each user group might have a different editing impact on Wikidata and contribute differently to achieving the overall goal of serving the whole world.

Next, we discuss the most important aspect of a KB, the data model that explains how the data are stored and can be accessed in Wikidata.

2.1.3 Data Model

Wikidata follows the Wikipedia scheme for storing data about a single entity on one page[17]. While this entity in Wikipedia is in human-readable text form represented as an *Article*, in Wikidata such entity pages contain not only human-readable but also machine-readable data called *Item* and *Property* pages[18]. The data model of Wikidata can be seen in Figure 2.2.

Wikidata stores similar pages separated per namespace e.g., all items exist in the *main namespace* (also called *namespace 0*) and all properties are in the *property namespace* also known as *namespace 120.* A complete list of Wikidata namespaces and page types is available on the Wikidata website[19].

The item and property pages are very similar in Wikidata; thus, explaining the item page and its content could give an impression of the property page as well. Wikidata wants to make every portion of data accessible and be used by query services and

[13] QuickStatements: https://quickstatements.toolforge.org/#/[Accessed 05.01.2021]

[14] Scholia: https://scholia.toolforge.org/[Accessed 05.01.2021]

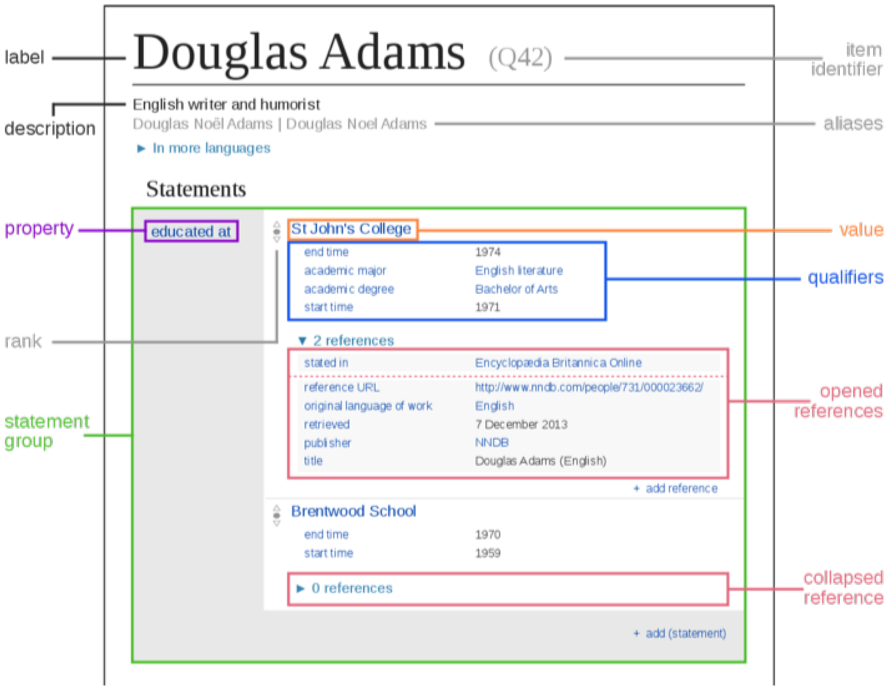
[15] Reasonator: https://reasonator.toolforge.org/ [Accessed 05.01.2021]

[16] Wikidata semi-automated tool edit indicators: https://meta.wikimedia.org/wiki/Research:Understanding_Wikidata%27s_Value/semi-automated_tool_edit_indicators [Accessed 05.01.2021]

[17] Data model: https://www.mediawiki.org/wiki/Wikibase/DataModel

[18] Wikidata Data Model, https://meta.wikimedia.org/wiki/Wikidata/Data_model_update

[19] Namespaces: https://www.wikidata.org/wiki/Help:Namespaces [Accessed: 08.10.2020]



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

► In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

Brentwood School

end time	1970
start time	1959

► 0 references

+ add (statement)

Figure 2.2: Data model of Wikidata.

data analysis tools, so it has defined its data model to have smaller parts in item and property pages. The parts which consist of an item or the contents of an item page in Wikidata are:

Identifier. Each item is stored in Wikidata using a defined item identifier which starts with the letter *Q* and is followed by numbers e.g., *Q42* for *Douglas Adams* in Figure 2.2. Properties also have a property identifier that follows the same pattern as items and starts with the letter *P* and is followed by a number, e.g., *P22* for *has father* or *P25* for *has mother*. The item identifier is used to store the data in a language-independent form and is only readable by machines. To make the data human-readable, multilingual *labels, descriptions* and *aliases* are used. Together, they are called *term*. The term is also part of the item identifier because a label alone cannot provide all the information for an entity. For example, when looking at the Berlin label, one needs to refer to the details provided in the description to make sure that it is the capital of Germany, because Berlin is also the name of a city in the United States[20]. Thus, we need a description and alias to find out which Berlin is talked about here. There can be multiple aliases for one item, and Wikidata can store all as a result of being designed to support plurality.

Statement. The main data about an item is stored in statements. Statements in Wikidata come in the form of property-value pairs to convey information about the item. A statement consists of a claim, a qualifier, a reference, and a rank.

- Claim. The value section of a statement is called a *claim*. This portion of the statement does not consist solely of strings; the value itself could also be an item.

[20] https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer [Accessed 08.10.2020]

For instance, St John's College represents the item (Q691283)[21] and serves as the value denoting the educational institution of Douglas Adams in Figure 2.2. We also observe a second value for this property, once again an item, namely Brentwood School (Q4961791)[22]. Thus, statements link Wikidata items and describe the relationship through the property of the statement such as Douglas Adams being *educated at* Brentwood School.

- Qualifier The statement that Douglas Adams attended St John's College does not convey complete information unless it does not specify the period for this fact. The information adding the time frame when Douglas Adams attended Brentwood School in the value section is called *qualifier* which provides details about the time frame that this claim is true, i.e., 1959 to 1970 (cf. Figure 2.2). Similar to statements, qualifiers are also property-value pairs that provide context to a statement, while statements further elaborate on items.

- Reference. To ensure the reliability of a claim, *references* are added in the statement. In our example, there are two sources that confirm the claim that Douglas Adams attended St John's College, and both are included because Wikidata can store multiple values together, even if they contradict each other.

- Rank. Another part of a statement is *rank* and is used to mark values as preferred, normal, or deprecated to facilitate the selection of value in case of multiple values for a property. It is useful in cases like the population to mark the accurate and up-to-date value as preferred, to indicate the default value, and mark the outdated values as deprecated.

Overall, statements contain two main parts which are claim (value and qualifier) and reference.

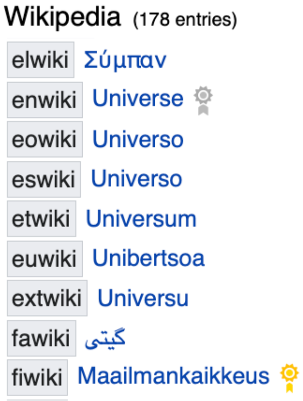
Sitelink. The identification of Wiki pages is done through *sitelinks* also called *interwiki links* or *interlanguage links.* A sitelink is a special link that contains a link and a title and identifies a Wikidata page from an external site on Wikimedia sites like Wikipedia or Wikisource. Sitelinks appear as lists on every item page and are a means of linking Wikidata pages to the Wikimedia site to ensure that every item has at least one corresponding page in the Wikimedia sphere and meets the Wikidata notability criteria[23]. They also aid in finding out which items are unique and which items represent redundant concepts and are subject to merging. Figure 2.3 shows a part of the Wikipedia sitelinks list on the Wikidata item page of Barack Obama (Q76) where the pages in each language version have identifiers like *enwiki* represents English Wikipedia.

- Badge. A *badge* can be attached to a sitelink for marking purposes. For example, the sitelinks on Universe (Q1) item page on Wikidata show two badges which means a featured article on Universe exists in Finnish Wikipedia and a good article is there in English Wikipedia as can be seen in Figure 2.3.

[21] https://www.wikidata.org/wiki/Q691283 [Accessed: 19.09.2022]

[22] https://www.wikidata.org/wiki/Q4961791[Accessed: 19.09.2022].

[23] https://www.wikidata.org/wiki/Wikidata:Notability



Wikipedia (178 entries)

elwiki	Σύμπαν
enwiki	Universe
eowiki	Universo
eswiki	Universo
etwiki	Universum
euwiki	Unibertsoa
extwiki	Universu
fawiki	گیتی
fiwiki	Maailmankaikkeus

Figure 2.3: Sitelinks and badges on Universe (Q1) Wikidata item.

The first section of this chapter has provided a general introduction to Wikidata with a focus on the reasons and design principles for the development of Wikidata, the contributing community, and the data model of this structured KB. Wikidata is moving forward toward the goal of becoming a world KB that can provide service globally to individuals across the world. The design decisions of Wikidata implicitly refer to the concept of diversity and diversity seems to be the factor that enables Wikidata to achieve its overarching goal.

In the next section, we present the one-decade status of Wikidata from a research perspective. We are interested to know where Wikidata stands in the journey of becoming a world KB.

2.2 Wikidata From a Research Perspective

Wikidata aims to reach “anyone anywhere in the world”, to get an impression of how successful Wikidata has been so far in achieving this goal, we narrow our scope to the research community and explore Wikidata from a research perspective to know how popular Wikidata is among the researchers around the globe considering the fact that all of its data is freely accessible.

The research community’s interest in Wikidata has accumulated recently, and this is an indication of its growing popularity. Numerous studies have explored Wikidata from various angles, such as its internal structure, including both, data and community, from a data perspective by looking at its completeness and coverage, from an engineering perspective by looking at the needed tools, and from an application perspective by providing case studies in using Wikidata for projects in medicine, linguistics, or geography as some examples. However, the research on Wikidata seems to be scattered over different research fields and disciplines, and it is challenging to develop a mental map of the existing state of the art of the research. Motivated by this observation, we conducted a mapping study that summarizes and reflects on the insights of existing research and gives an overall overview of what studies have been carried out so far and what topics need to be explored in future research.

In this section, we provide the details of the mapping study method, our defined procedure for data collection and processing, and a statistical overview of this data.

We then, present the classification of the existing Wikidata literature into research topics to identify the most popular areas and bold the research gaps.

2.2.1 Research Method

Our research method is motivated by our goal to provide a general overview of the field by identifying the topics that are well-studied and deriving the open spots in research [70]. Mapping studies are insofar a suitable instrument since they provide the ground and directions for future research as well as educate the members of a community [152]. Our study adopts guidelines for systematic mapping studies which are defined by Petersen et al. in [229].

A *mapping study* provides a "map" of a research area. It helps shape research directions by revealing existing topics that aid in the identification of white spots [70]. A mapping study differs from a systematic literature review insofar as the latter tackles a focused research inquiry [228]. Therefore, a mapping study can be seen as a pre-study of a systematic literature review and is a method for obtaining a comprehensive overview of the research conducted within a specific area of interest. It involves classifying relevant research to gain a deeper understanding of the areas covered and to establish a baseline that supports new research endeavors [153].

In this study, we want to provide an overview of Wikidata from a research perspective. The peer production system Wikipedia, for example, has already drawn research from a myriad of disciplines [220] and the question is, whether we have the same situation in the context of Wikidata. In our mapping study, we summarize what has been researched so far about Wikidata, when, from which origins, and where they were published. We also identify which aspects of Wikidata have received more attention in the research community and which aspects have not yet been given much effort to study, by classifying and categorizing existing research from October 2012 to June 2022. Based on the search results from the academic search engines explained below, we identified 886 search results. All papers were screened, and when necessary, read in more detail to make accurate decisions regarding the inclusion of these papers in the final dataset. Finally, all needed information was extracted from the final set of 248 articles to answer our questions of interest, as listed below.

Q1 What high-quality research has been conducted with Wikidata as a major topic or data source?

Q2 What types of research have been published, when (year) and where (journals or conferences)?

Q3 What are the origins of the research (which countries, and institutions)?

Q4 Which aspects of Wikidata are covered by considered research and which aspects are still to be studied?

In the following, we describe in more detail how and where we searched articles, which papers we included or excluded, respectively, and finally what categories we derived from the articles.

2.2.1.1 Search Process and Data Sources

Data collection is a crucial step in any research since findings are the direct result of the gathered data. We defined the needed keywords which is the first step to

Table 2.1: Search results from academic search engines.

Search Engines	**Search Results**	**First Screening**	**After Selection**
ACM	140	107	75
DBLP	200	112	81
SpringerLink	47	47	47
Google Scholar	499	168	45
Total	886	434	248

search for the literature. As the noun "Wikidata" is only used as the name of the structured data source so far, and has no further meanings, the search string was simply selected as "Wikidata" in order to identify a broad range of related literature. Similarly, since Wikidata was launched in October 2012, the time range was defined from 10/2012 until 05/2022 (some search engines did not support the "month+year" format). The search strategy for this study is an automated search using digital libraries. We obtained Wikidata research from the ACM Digital Library (ACM DL)[24], the Springer Link Digital Library (Springer Link)[25], and the Digital Bibliography & Library Project (DBLP)[26]. ACM DL and DBLP are bibliography search engines specifically for Computer Science. Although Springer Link provides results from a broader range of fields such as social sciences and humanities, we decided to extend the scope of the search in order to achieve a more holistic image of the current state of research on Wikidata from different disciplines. Thus, we also included search results from Google Scholar Search Engine (Google Scholar)[27].

The ACM DL searches for keywords everywhere in the text, and only annual date settings are possible. We received 140 articles. Springer Link was also searched with the same keyword and time range as ACM and returned 47 results. The search interface on DBLP does not provide a time range selection, however, it returned the results from 2012 till now (May 2022), which resulted in 200 papers. The Google Scholar search engine was searched through *Harzing's Publish or Perish*[28] software with the same criteria. The number of Google Scholar search results was 499. This large number is due to the fact that Google Scholar returns technical reports, white papers, and theses as well. The total number of articles in the first stage was 886 (cf. Table 2.1).

2.2.1.2 Criteria Exclusion and Inclusion

We defined inclusion criteria to find the most relevant research papers. The defined criteria for exclusion are duplicates, results in languages other than English, and results that are not published in peer-reviewed journals or conference proceedings, such as websites, reports and data sets, theses, and books.

[24] ACM DL is available at: https://dl.acm.org/.

[25] Springer Link is available at: https://link.springer.com/.

[26] DBPL is available at: https://dblp.uni-trier.de.

[27] Semantic Scholar (https://www.semanticscholar.org) is another source of Wikidata research papers; however, the filtering mechanism of this system was functioning unexpectedly and the results were not reproducible. Although we contacted the SemanticScholar team, the issue could not be solved, and thus, this search engine was not included in the study.

[28] Harzing's "Publish or Perish" provides an interface to use Google Scholar and export all results in a number of formats. In this study, we used the CSV format. The software is available from http://www.harzing.com/pop.htm.

In the first step, we already excluded 117 non-English search results from Google Scholar, second, 230 duplicates (results that were received by more than one search engine), and third, 105 non-papers (presentations, reports, datasets, and books). Therefore, the remaining 434 search results were subject to an inclusion process (cf. Figure 2.4).

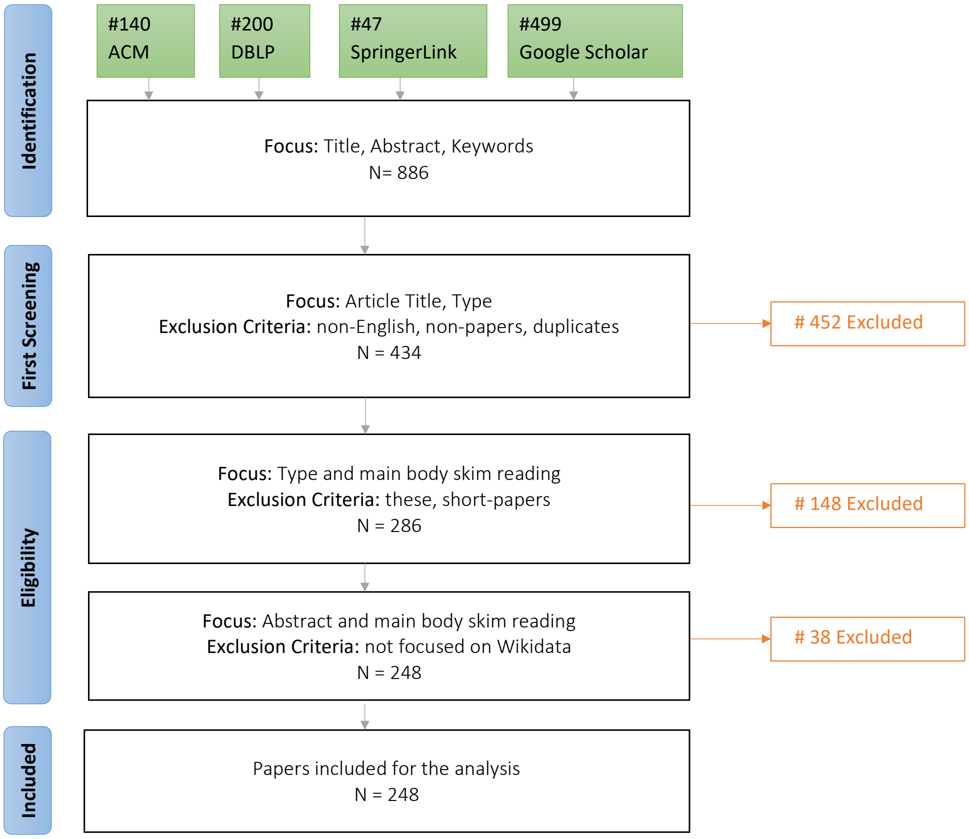

Identification
#140 ACM
#200 DBLP
#47 SpringerLink
#499 Google Scholar
Focus: Title, Abstract, Keywords
N= 886
First Screening
Focus: Article Title, Type
Exclusion Criteria: non-English, non-papers, duplicates
N = 434
# 452 Excluded
Eligibility
Focus: Type and main body skim reading
Exclusion Criteria: these, short-papers
N = 286
# 148 Excluded
Focus: Abstract and main body skim reading
Exclusion Criteria: not focused on Wikidata
N = 248
# 38 Excluded
Included
Papers included for the analysis
N = 248

Figure 2.4: Article selection process and the number of included search results. (Note: This diagram is inspired by the PRISMA flow diagram [223].)

In the second step, we included research papers only if they are published in academic journals or conference proceedings and are full research papers with at least five pages. The latter criterion is based on the reasoning that articles with four or fewer pages are considered short papers and usually are posters, position, or demonstration papers. After applying the aforementioned criteria to the remaining 434 articles, 130 short papers (including papers like position papers or posters) were excluded. In addition, another 18 articles could be identified as (bachelor, master, and doctoral) theses. After reading the abstracts, another 38 papers were excluded because they were not focused on Wikidata.

In total, a majority of the 638 out of 886 articles found were excluded and only 248 papers remained in our sample. The reason for the exclusion of this large number of search results was that we intended to include only the papers that focus solely on Wikidata. Another reason was that Google Scholar returned results that were not only papers.

Table 2.2: Data extraction form

Data Item	**Q relevance**
Title	Q1
Author(s)	Q2
Abstract	Q1, Q4
Date	Q2
Publisher	Q2
Origins of the research (countries and institutions)	Q3

Our further discussion is based on these 248 articles. We have compiled the *Wikidata Research Articles Dataset* [66] from these data that are freely available online[29]. These papers are also listed in the references section and marked with an asterisk.

2.2.1.3 Data extraction

Within the data extraction part of our study, we specified what data we wanted to extract from our data set. Having a uniform data extraction form as in Table 2.2 reduces both, bias and internal validity threats. We developed a data extraction form, to answer the research questions of this study. We extracted the title, author(s), abstract, date, and publisher to answer Q1, Q2, and Q4. However, Q3 required manual extraction of the institutions and countries where the first author of the paper had performed the research. We focused on institutions rather than the first authors themselves, as some authors had published from different institutions. One author, for example, published a research paper from Institution A and later joined Institution B and published a paper there. It would be difficult to select one institution as the origin of that author. Q4 required more insight on each research and therefore we read the article in more detail by focusing on the findings. The tools used for data extraction and analysis are Zotero[30] and Microsoft Excel[31].

2.2.1.4 Research Paper Classification

At this stage, we read the abstract of all articles and if needed the introduction and conclusion parts of the open-access papers to get more insights about each research for categorization. In a number of cases, further sections of the articles had to be read to better understand the scope and topic of the article. After analyzing 248 papers, we manually categorized the papers as shown in Figure 2.10 (on page 26). To define the categories, we started with descriptive labels for each paper. After reading a few papers, we tried to identify categories at a higher level of abstraction. We compared our categories throughout the reading process to make sure that our coding scheme remained consistent.

Since we had performed this mapping study once in 2018 and then did an update of the literature search in 2022, we present the results of papers from 2012 to 2018 (pre-2019) and 2019 to 2022 (post-2018) comparatively. This can give us a view of how the research focus of Wikidata has evolved in its first decade of existence, how

[29] Available at REFUBIUM (FU Berlin Repository): http://dx.doi.org/10.17169/refubium-40231

[30] https://www.zotero.org

[31] https://www.microsoft.com/en-us/microsoft-365/excel

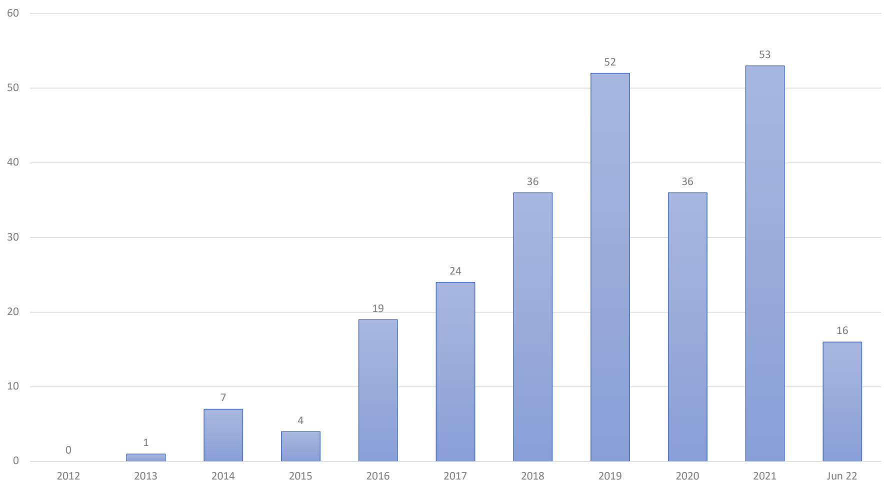


Figure 2.5: Frequency of Wikidata research publications per year. Source: *Wikidata Articles Dataset*.

it was set at first, and how it has evolved after 2018. Our findings of this mapping study are presented in the next sections.

2.2.2 Overview of Existing Research on Wikidata

An overview of Wikidata from the perspective of the research community not only provides us with a glance at the rather large community of more than five million registered Wikidata users[32], but also helps us better understand Wikidata from different angles, such as internal structure, features, or usage domains.

Here, we provide a more detailed description of the resulting dataset from the mapping study, named the *Wikidata Research Articles Dataset*. First, we examine the frequency of publications; second, we identify the publication venues; third, we determine where the research was published. Lastly, we present the geographical origin of the Wikidata research prior to exploring the research themes related to Wikidata.

2.2.2.1 Frequency of Publication

The majority (157, 63%) of the 248 included research papers are recent research from 2019 to 05/2022 (cf. Figure 2.5). This is an indication that Wikidata has gained more awareness in the research community. Since 2013, with the exception of 2015, this count has grown consistently each year. Given this growth and the number of studies until June 2022 (16), it is anticipated that the count of research articles on Wikidata will likely reach 50 to 60 by the end of 2022.

2.2.2.2 Publishers and Publication Types

The most popular publishers for Wikidata research are ACM with 54 articles, Springer with 50, and CEUR with 41 articles. Additionally, Figure 2.6 shows that the most popular journal for publishing Wikidata research articles is *The Semantic Web Journal*. Among the 248 papers in this study, most (192, 78%) are published

[32] Wikidata statistics: https://www.wikidata.org/wiki/Special:Statistics[Accessed 20.11.2022]

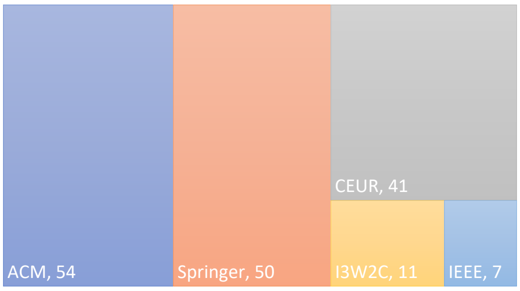


Figure 2.6: The most popular publishers of Wikidata research. Source: *Wikidata Articles Dataset*.

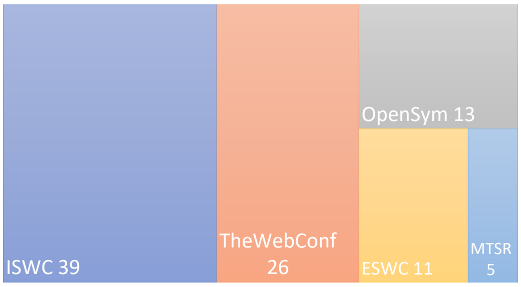


Figure 2.7: The most popular conferences on Wikidata research. Source: *Wikidata Articles Dataset*.

as conference papers, and the rest are journal articles (56, 22%). Thus, conference proceedings are the most popular type of publication in Wikidata.

The most popular conferences where Wikidata research was presented are the *ISWC* (International Semantic Web Conference), *TheWebConf* (The Web Conference)[33], *OpenSym* (The International Symposium on Open Collaboration), *ESWC* (Extended Semantic Web Conference)[34], and *MTSR* (Research Conference on Metadata and Semantics Research) (cf. Figure 2.7).

2.2.2.3 Geographical Origins of Research

We discovered that Europe (61%) is the primary contributor to Wikidata research, with Germany leading among European countries, followed by the United Kingdom in second place and France in third. America (24%) has also made notable contributions to Wikidata research, with the United States of America being the second largest contributor globally, after Germany. (cf. Figure 2.8). We also observe a rising number of research articles originating from various parts of Asia (11%). While Wikidata has become an established research domain in the West, much of the research from Asia is relatively recent, reflecting a growing interest from diverse countries in Wikidata. This trend suggests the potential for increased contributions and an improved diversity status of Wikidata in the future. Figure 2.8 illustrates the increasing interest in Wikidata research, not only in Asia and Africa, but also in Europe and Canada. This trend is promising from a diversity perspective and has the potential to enhance the diversity of content.

[33] Formarly known as International Conference on World Wide Web (WWW)
[34] Formarly known as European Semantic Web Symposium (ESWS)

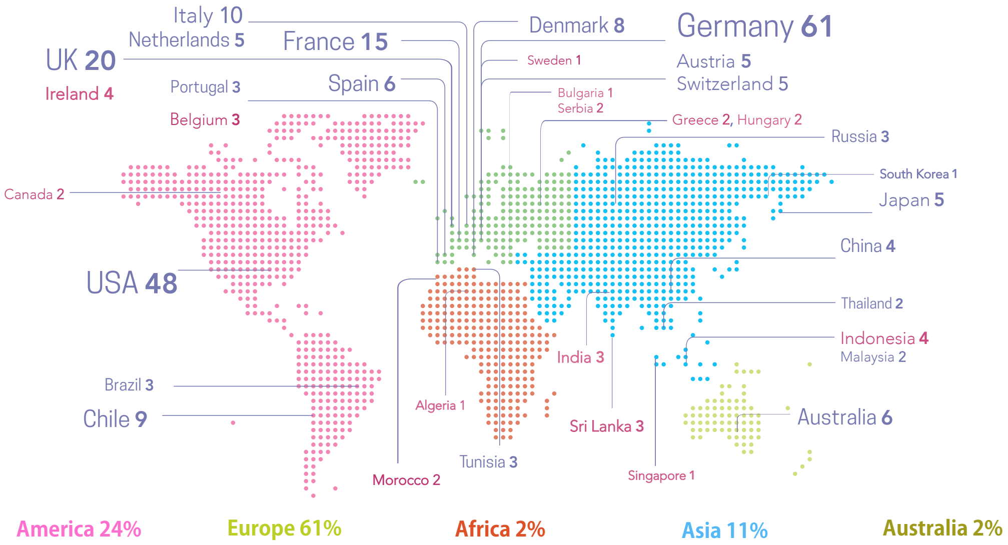


Figure 2.8: Research contributions from countries and continents. Source: *Wikidata Articles Dataset*. (Note: Countries in violet color published research articles before and after 2018 and countries in pink published articles after 2018.)

In terms of the most active institutional contributions (refer to Figure 2.9), the findings reveal that the University of Southampton has made the highest number of contributions (11 articles), followed by the University of Lyon (9 articles), Chile University (8 articles), Technical University of Denmark (7 articles), and the University of Konstanz (6 articles). The remaining top seven contributing institutions include the University of Mannheim and the Technical University of Dresden (each with 5 articles), alongside additional contributions primarily originating from various German universities and research institutions.

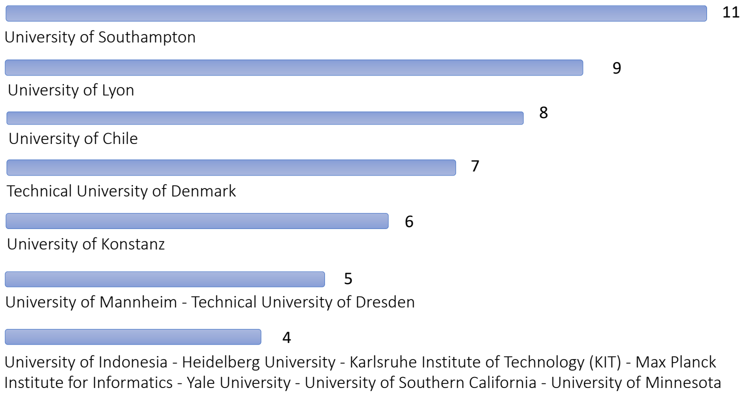


Figure 2.9: Most popular research contributing institutions. Source: *Wikidata Articles Dataset*.

2.2.3 Research Topics of Wikidata

The main research topics of Wikidata which are obtained after classification, are shown in Figure 2.10. Although numerous research topics are already explored within Wikidata, there remains untapped potential for its utilization across a variety of purposes. While conducting a search for literature on Wikidata, it became evident that Wikidata is being widely utilized in numerous studies. Our objective was to identify research papers that specifically delve into Wikidata, elucidating its mechanisms, opportunities for enhancement, or its significant role in addressing prevailing challenges—such as its applications. Consequently, studies in which Wikidata was not the primary focus and played only a partial role within the scope of the study have been excluded.

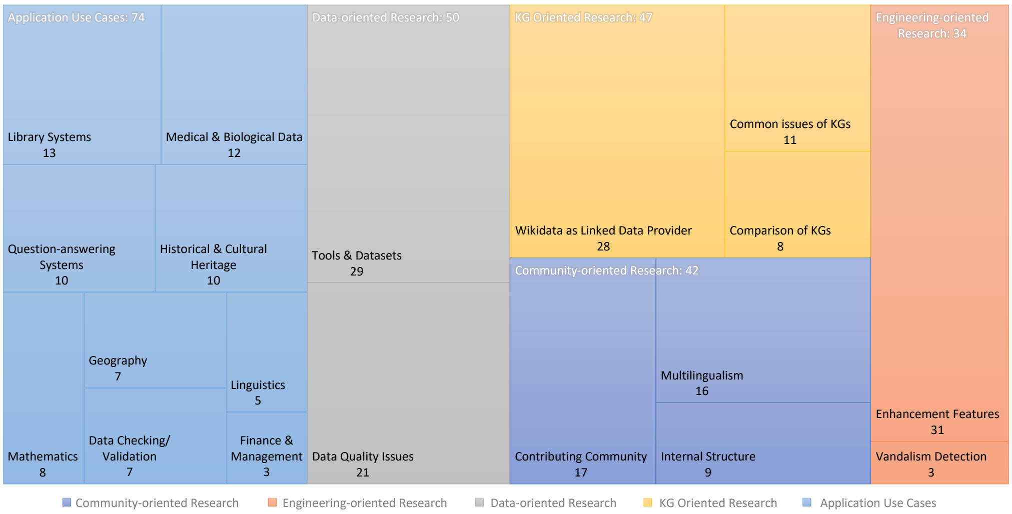


Figure 2.10: Classification of Wikidata research papers (Sum 248)

Here, we reflect on the mapping study of Wikidata in two time frames, namely, from Wikidata's inception in 2012 until the end of 2018, and from the beginning of 2019 to mid-2022. The rationale behind this division is that we initially conducted the mapping study in 2018. Based on our findings, we established our research direction and conducted subsequent research. An updated mapping study in 2022 serves not only to provide insights into the latest state of Wikidata research but also to illustrate the evolution of Wikidata research over this period. We shed light on newly introduced topics, resolved previous issues, identify remaining gaps, and propose potential future directions.

In the following section, we delve into the explanation of each of the five categories depicted in Figure 2.10, along with their respective subcategories and the papers they encompass.

2.2.3.1 Community-oriented Research

Is Wikidata just another peer production system? The research in this category reflects Wikidata's goals, features, and internal structure. Additionally, it contains existing design decisions (esp. multilingualism), and studies on the Wikidata community and their participation patterns.

Internal Structure. The research articles in this category provide mainly an overview of Wikidata and introduce or explain its features and design principles. Most of the papers in this category belong to the early research phase, i.e., from 2012 to 2018.

One of the first articles on Wikidata is by Vrandecic, who motivates the need for integrating existing structured data from the various Wikipedia language versions into one single repository named Wikidata, in order to overcome the existing data inconsistencies of Wikipedia language versions [325]. According to Vrandečić and Krötzsch, the main distinguishing features of Wikidata are being available internationally and support for multilingualism, storing links to facts as a secondary database, and the ability to store contradictory facts to represent knowledge diversity [326]. Malyshev et al. describe Wikidata and its underlying infrastructure as an emerging semantic technology use case [183], while, Yu explains Wikidata and data.gov projects which are structured data sources to provide a deeper understanding of Semantic Web standards at work [345].

Ilievski et al. study the commonsense knowledge coverage in Wikidata and whether it is complementary to other existing commonsense graphs [134]. Kempf gives insight into the mapping process between topical thesaurus concepts and Wikidata items by providing methodological guidance on it [150]. Voß discusses extraction and classification of knowledge organization systems based on Wikidata [324].

Cantallops et al. perform a literature review of Wikidata's existing research and describe Wikidata from a research perspective [33]. Spitz et al. analyze Wikidata from a data consumer's perspective and highlight the existing challenges and possible paths of improvement [296].

Multilingualism. Multilingualism is one of the design principles of Wikidata. Wikidata stores data in a language-independent form and aims to provide data to the whole world population. This section comprises studies that focus primarily on this design principle.

Samuel describes the multilingual collaborative ontology development process in Wikidata by explaining the development process of a new property and its major steps from being proposed to getting approved by the community and finally translated to other languages [264].

Kaffee et al. shed light on the languages covered by Wikidata. Their results suggest that most of the labels and descriptions on Wikidata are only available in a small number of languages like English, Dutch, French, German, Spanish, Italian, and Russian. This stands in contrast to the majority of languages which have close to no coverage [142]. In further studies, Kaffee et al., explore the generation of open domain Wikipedia summaries from Wikidata in "underserved languages" to overcome uneven content distribution [144, 149]. Ta and Anutariya propose a mechanism to enrich Wikidata multilingual content by retrieving "semantic relations based on alignment between info-box properties and Wikidata properties in various languages" [309]. Sáez and Hogan investigate the development of "fully automatic methods" where info-boxes for Wikipedia can be generated from Wikidata descriptions [308]. Kaffee and Simperl study the languages of Wikidata editors and look for any relations between the editor's language and label editing habits of editors [143]. The last studies in the pre-2019 phase are by Samuel, which in one study, develops a tool that visualizes the translation patterns of Wikidata properties [265], and in another, sheds light on the process of multilingual property creation in Wikidata [266].

In the studies post-2019, the research on multilingualism with the property and label focus continues. Samuel develops the WDProp web application to inspect various multilingual aspects of Wikidata properties [268], while Kaffee et al. analyze the hybrid editing of Wikidata labels by humans and bots [145].

Lexems[35] seem to be a recent research focus in the multilingualism area through multiple studies by Nielsen. In his studies, Nielsen presents descriptive statistics of the Wikidata lexemes mainly from the multilingual angle [214] and develops Ordia to display multilingual lexeme data of Wikidata [213]. In his other studies, he mainly focuses on Danish lexemes by demonstrating validation of entity data for Wikidata lexemes through the application of Danish Shape-Expressions [217] and describing Wikidata lexemes and how Danish lexemes are annotated in Wikidata [212].

Further, Sas et al. introduce WikiBank, a resource of partial semantic structures with multilingual capabilities by aligning Wikidata triples with Wikipedia sentences, used to extend the existing resources [271].

Contributing Community. This section reflects the efforts made to understand who are Wikidata's contributors and what participation patterns they follow. The research on the Wikidata community started early and Steiner, for example, developed an application that is capable of monitoring real-time edit activities of all language versions of Wikipedia and Wikidata. The study shows that many Wikipedia language versions, such as English and French contain the most edits by bots. In Wikidata alone 88% of the edits come from bots [298].

Müller-Birn et al. analyze the contribution patterns of the Wikidata community to better understand whether Wikidata community participation patterns follow a peer-production approach like Wikipedia or a collaborative ontology engineering approach. Their findings show that Wikidata stands somewhere between the two mentioned approaches. The study also describes the characteristics of the Wikidata community as, registered users, anonymous (not registered or logged-in users), and bots, where bots perform the highest number of edits [206]. Based on the results of study [206], Cuong and Müller-Birn explore the dynamics of Wikidata community participation process with a focus on human registered users, to know how the participation patterns of the community change over time [42]. Piscopo et al. extend this line of research by studying the participation patterns of the Wikidata human registered community members, from being an editor to becoming a community member, and investigate how these patterns evolve [240]. In another study Piscopo et al. analyze the relationship between group composition of bots, and humans (registered or anonymous) and the item quality in Wikidata [239]. Further, Piscopo and Simperl study the relationship between the quality of Wikidata ontology and different Wikidata user roles [236].

[35] "Lexeme is a lexical element of a language, such as a word, a phrase, or a prefix." https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Glossary

Hall et al. develop a machine classifier to detect bot edits from other user groups using implicit behavior and informal editing characteristics [114]. Bielefeldt et al. analyze the access logs from the SPARQL endpoint and separate the bot-based traffic from the human-based traffic. As expected, the human part is smaller and shows clear trends, e.g., correlated to time of day, in comparison to the bot-based part which is "highly volatile and seems unpredictable even on larger time scales" [25]. The research on Wikidata query logs continues in 2019 and Bonifati et al. analyze Wikidata query logs from different angles like robotic and organic, correctly executed or timed out, and recursive queries [27].

Further studies in this period focus on the analysis of the Wikidata contributors. Kanke, in one article, explores Wikidata editors' participation and editing activities in this collaborative KB [146], and in another article, investigates Wikidata editors' participation and activities through content analysis of discussion pages [147]. Sarasua et al. investigate how editors with different levels of engagement evolve over time and how these different editing behaviors affect the volume of edits and the duration of the contribution of the contributors to Wikidata. The authors focus on the human registered users and distinguish between power and standard editors [270]. Han et al. present a user behavior model to quantify the user/system interaction behavior based on three different datasets including Wikidata editors. According to the study, active editors show more diverse daily behaviors, while, less active editors perform similar daily tasks [116]. The authors Piao and Huáng use multiple classification approaches to predict which Wikidata editors will no longer edit Wikidata [232].

Zhang et al. investigate the motivation of editors in Wikidata, in particular, that a large share of the edits is performed through invisible machines [351]. Farda-Sarbas et al. describe the process which an account needs to go through to get bot privileges in Wikidata [67]. Freire and Isaac evaluate the potential of Wikidata as a resource for bots on the web of data [80].

Despite the fact that bots contribute a significant amount of data, the literature on the Wikidata community pre-2019 was more focused on human users. This trend has continued post-2019, with most studies still prioritizing human users and only a few examining the role of bots.

2.2.3.2 Engineering-oriented Research

This section contains articles that suggest approaches and features for the enhancement of Wikidata's functionality. These features are programmed for two main purposes: first, for the improvement of overall performance like more efficient handling of data, and second, for vandalism detection.

Enhancement Features Wikidata was developed to evolve gradually with the needs of the community. This section contains research that proposes approaches to address existing limitations, facilitate access, or improve the processes of adding data to Wikidata, either manually, through recommender systems, or by using external data sources.

Pellissier Tanon et al. introduces the Primary Sources Tool[36] to facilitate the migration of the content from Freebase to Wikidata [224]. Mousselly Sergieh and Gurevych propose an approach to bridge the missing linguistic information gap of Wikidata by aligning Wikidata with FrameNet[37] lexicon [201]. Zangerle et al. evaluate recommender algorithms, which assist Wikidata contributors in the process of data insertion through property recommendation [348].

In the articles post-2018, recommender systems remain a research interest. Alghamdi et al. present WikidataRec, a hybrid recommender model which suggests editors' items based on their past edits [6], and Gleim et al. introduce a trie-based approach for recommending properties based on the frequentist inference [98]. Further, Chah and Andritsos present a data profiling framework that can be utilized on Wikidata data dumps and can summarize Wikidata's data status through a sequence of granular descriptive statistics. This profiling lays the ground for efficient data access including generating visualizations of data multilingualism across Wikidata domains with the help of machine learning [35]. Klein et al. investigate how to create profiles or obstructive representations of Wikidata entities and make a graph that can be used more efficiently in comparison to the original graph which has certain limitations [156].

Morshed proposes an approach to annotate parts of the Wikidata lexemes, which consist of multiple parts, for the reduction of representation redundancy [200]. Dadalto et al. conduct an empirical analysis of the Wikidata platform to demonstrate the existing problems in modeling types and instances in Wikidata. The study then explains the reason behind these problems and suggests possible solutions [44]. Martín-Chozas et al. propose an approach for the validation of terminological data retrieved from Wikidata [186].

Faiz et al. develop OD2WD tool to enable Wikidata to store tabular data from other Open Data portals. The tool imports tabular data in CSV format into Wikidata after converting them into RDF format [63]. Baskauf and Baskauf present Generating RDF from Tabular Data on the Web (CSV2RDF) which is a W3C Recommendation that provides a mechanism for the mapping of CSV file data to RDF graphs [18]. Hevia et al. propose a system for automated synchronization between RDF data in the control version system and a Wikibase instance [131]. Samuel presents simplification of writing the shape expressions for Wikidata so that entity schema creation can be made easier and more popular [267]. Shanaz and Ragel propose a method for assigning data types for Wikidata entities and create a dataset of the classified Wikidata entities [284]. Arnaout et al. investigates the negative knowledge representation in Wikidata through the Wikinegata platform over Wikidata [14]. Willaert and Roumans perform a micro-level analysis of knowledge representation problems in KBs by looking at the representation of Belgian prime ministers in Wikidata [338]. Ferradji and Benchikha propose an updated version of time-related metrics for a more efficient assessment of linked data in Wikidata [72]. Pellissier Tanon et al. introduce an approach to fix constraint violations through KB edit history [225], and Martin and Patel-Schneider explain constraint handling in their proposed logical framework for Wikidata which is based on multi-attributed relational structures [185].

[36] For more information, please check https://github.com/google/primarysources.

[37] FrameNet is a lexical database of the English language. For more information, please check https://framenet.icsi.berkeley.edu/.

A number of studies have proposed approaches to improve data access through the SPARQL endpoint. Tanon and Suchanek propose a system that can make the edit history of Wikidata accessible through SPARQL end point [310]. Dahir et al. describe how to improve information retrieval results using Wikidata linked data by expanding the queries with attributes' values [45]. Gupta and Berberich propose query operators to optimize Hyper-phrase queries in large document collections and show the efficiency of this proposed approach on Wikidata [106]. Wudage Chekol et al. propose iSPARQL, a lightweight extension of SPARQL to be used as an alternative to SPARQL for querying large temporal Web KGs and solving the query timeout in Wikidata [341]. Sun and Sarwat present Riso-Tree, an indexing framework for spatial entities in graph database management systems which can perform faster execution of graph queries that involve different types of spatial predicates and evaluate it on real datasets including Wikidata [305].

There are also studies that discuss how Wikidata data can be organized in subsets or organized in a way that can improve access to the data in Wikidata. The study by Chalupsky et al. introduces a new query language and processor named KGTK Kypher[38] which allows users to create personalized Wikidata variant on laptop and runs faster than SPARQL [36]. Henselmann and Harth explain an algorithm for creating Wikidata subsets on demand [128]. Similarly, Beghaeiraveri et al. discuss building topical subsets over Wikidata using WDumper for more efficient and flexible usage of data through queries [20]. Hanika et al. discuss how to discover the implicit knowledge hidden in the complex data model of Wikidata using Formal Concept Analysis [117]. Lai et al. improve searching by narrowing down the search space using a new approach based on entity profiling of Wikidata entities [166]. Jozashoori et al. present EABlock, an approach to address the entity alignment issues through capturing knowledge from Wikidata and DBpedia KGs [140].

We could see that the research with a focus on enhancement features of Wikidata before 2019 contained different scattered topics, while, the research after 2019 seems more organized and focused on a number of topics like data handling through recommender systems, enhancing SPARQL query mechanism and making data access easier through providing data in subsets.

Vandalism Detection. Wikidata provides data for Wikipedia and other Wikimedia projects; thus, the integrity and correctness of data are of high importance. Vandalism detection is, therefore, an essential aspect of a knowledge repository and directly influences the data quality and trustworthiness of a KB. In the following, we provide an overview of research that focuses on detecting vandalism and other efforts for making Wikidata more robust.

In their study, Heindorf et al. present a new machine learning-based approach for the automatic detection of vandalism in Wikidata [122]. Sarabadani et al. develop a vandalism detection mechanism for Wikidata by adapting methods from the Wikipedia vandalism detection literature and extending it to Wikidata's structured KB. The mechanism used identifies damaging changes and classifies edits as vandalism in real-time, using a machine classification strategy [269].

The ACM International Conference on Web Search and Data Mining held the competition for developing vandalism detection mechanisms for Wikidata, the WSDM

[38] Kypher documentation: https://kgtk.readthedocs.io/en/latest/transform/query/

Cup 2017.[39] The main goal of this competition was to develop a model for detecting malicious or similarly damaging edits. As a result of participation in WSDM Cup 2017 five submissions presented their own vandalism detection mechanisms[40].

The evaluation of the proposed vandalism detection approaches at the WSDM Cup 2017, is done in [124]. Heindorf et al. evaluate their four baseline approaches[41] along the five submissions [124][42]. The study finds that the best approach is a semi-automatic scenario "where newly arriving revisions are ranked for manual review" is from [41], while, the best approach in a fully automatic detection scenario "where the decision whether or not to revert a given revision is left with the classifier" is the baseline approach by the Wikidata Vandalism Detector (WDVD) system from [122].

The only research on vandalism detection post-2018 is by Heindorf et al. which develop a new vandalism detection system that avoids the bias existing in the previous systems for old editors in comparison to new editors [125].

2.2.3.3 Data-oriented Research

This section contains research that is focused on the topic of data in Wikidata, either as a structured KB or a linked data provider KG. Research articles here address new approaches that could take Wikidata towards data completeness, and discuss new approaches that could utilize the data in Wikidata in a variety of ways. Hence, we organized the papers into two categories: (1) addressing the issues regarding data quality, and (2) the development of new tools on the data from Wikidata and the generation of datasets from Wikidata for various purposes.

Data Quality. The research highlighted in this section is concerned with the data quality aspects of data in Wikidata. A visible number of studies in this category have focused on the completeness angle of data quality. Literature evolving around solving the existing issues that address the existing flaws in data, like gender bias, is also included in this section.

Prasojo et al. discuss "COOL-WD", a tool for supporting the completeness lifecycle of Wikidata and allow to produce and consume completeness data by "data completion tracking, completeness analytics, and query completeness assessment" [245]. Galárraga et al. investigate "different signals to identify the areas where the KB is complete" and perform experiments on Wikidata and YAGO to generate completeness information automatically [86]. Ahmeti et al. and Balaraman et al. in their studies [4] and [16], propose and develop Recoin, a relative completeness tool for evaluating the completeness of entities in Wikidata. Recoin uses information from

[39] For more information, please check https://www.wsdm-cup-2017.org/.

[40] The competition received five submissions: 1) Buffaloberry by Crescenzi et al. [41], 2) Conkerberry by Grigorev [104], 3) Loganberry by Zhu et al. [354], 4) Honeyberry by Yamazaki et al. [342], and 5) Riberry by Yu et al. [346]. We could not include these submissions because [354], [342], and [41] are short papers, [104] and [346] are neither peer-reviewed nor published. Among them [41] reflected on previous work of [122], an automatic data mining approach for vandalism detection in Wikidata, and [104] presented an approach based on a linear classification model, which according to authors, is faster compared to other existing approaches.

[41] The four baseline approaches are: 1) Wikidata Vandalism Detector (WDVD) approach from [122], 2) FILTER, a second baseline which contains trained data from 01.05.2013 to 30.04.2016, 3) ORES, the re-implementation of the approach in [269], and 4) META, a combination of all approaches in [123]. ([123] is a short paper and not included in our study).

[42] [124] is not peer-reviewed and published, so not included in this study.

the class structure of the KG, in order to recommend possible properties for an item on the Wikidata user interface.

Razniewski et al. introduce the problems and limitations of properties in Wikidata and propose entity-specific property ranking for Wikidata [254]. Brasileiro et al. discuss the quality of taxonomic hierarchies in Wikidata to have a consistent data model and representation schema [30]. Piscopo et al. in their studies ([238, 241]) analyze Wikidata quality from the provenance perspective, the relevance and authoritativeness of Wikidata external references.

In the research post-2018, we see Piscopo and Simperl perform a literature survey of Wikidata research regarding data quality [237], while, a number of studies continue to focus on references that enhance data reliability. Amaral et al. inquiry the quality of Wikidata sources based on relevance, ease of access, and authoritativeness of Wikidata references through mixed methods using online crowd-sourcing, descriptive statistics, and machine learning [9]. Shenoy et al. develop a framework that detects low-quality statements and analyzes them through the existing practices of the community [288]. Beghaeiraveri et al. provide a statistical overview of references in Wikidata to help the contributors spot the existing problems and take steps for improving them [19]. Curotto and Hogan discuss the methods which can automate the process of referencing claims in Wikidata through suggesting references from Wikipedia [43].

Completeness also remains a popular area of research in Wikidata. Wisesa et al. develop ProWD, a framework for profiling data completeness in Wikidata. This tool is used to measure the completeness degrees of Wikidata based on Class-Faceted-Attribute [339]. Boschin and Bonald propose an approach to add missing data into Wikidata items from the Wikipedia textual contents [28]. Luggen et al. propose missing properties of Wikidata items through Wikipedia using the Wiki2Prop tool [178]. Freedman et al. propose a method for adding missing data into Wikidata through assignments to undergraduate students [79]. Gómez et al. propose an approach to automatically suggest potential values for Wikidata properties through context matrix [108].

Gender bias is another issue that has got attention in Wikidata research. Zhang and Terveen investigate gender gap in Wikidata content to describe if the reason behind gender content inequality in Wikidata comes from the editors, or it is a reflection of the real world gender bias [350]. Radstok et al. propose an approach to mitigate bias in KGs by balancing the data used for training models, rather than adapting models only. The authors evaluate their approach on Wikidata and DBpedia KGs [248], and Bourli and Pitoura investigate gender occupation bias in Wikidata KG [29].

The literature in the first phase (i.e., 2012 to 2018) is mostly focused on data completeness. In post-2018, we see not only completeness as a research focus but also references and gender bias both of which influence the reliability of data.

Tools & Datasets. This category contains the research that resulted in the development of new methods and tools, which mainly use Wikidata as a backend data source or the datasets which were generated from Wikidata for external services.

Ontodia [340], for example, is an online OWL (Web Ontology Language) and RDF diagramming tool over Wikidata. Scholia [216] is a tool for handling scientific bibliographic information through Wikidata. Lemus-Rojas and Odell use Scholia to generate scholarly profiles for a school at Indiana University which have their data stored in Wikidata [169]. Ferrada et al. present a new web interface for the IMGpedia dataset which can query more than 6 million images of IMGpedia through Wikidata [71]. Sen et al. propose the WikiBrain software framework to access Wikipedia data through Wikidata [281]. Moreno-Vega and Hogan present GraFa, a faceted search and browsing interface over Wikidata [198]. Gatti et al. extend PASS, a football generation system to include multilingual content using Wikidata [87]. Thornton and Seals-Nutt develop the Science Stories web application which creates stories by combining images and structured data from Wikidata [315].

There are also datasets that were developed based on Wikidata for different purposes. Nielsen and Hansen link Wikidata to the pre-trained ImageNet-based deep neural network to augment the model for a multi-modal knowledge representation [215]. Klein et al. develop Wikidata Human Gender Indicators (WHGI), a biographic dataset to monitor gender-related issues in Wikidata [155]. Konieczny and Klein use WHGI along the Wikipedia Gender Indicators (WIGI) to investigate gender inequality based on Wikipedia biographies [157].

Post-2018 we see more tools, new approaches/methods, and datasets built over Wikidata. Taveekarn et al. develop the DATA++ tool for retrieving relevant information from Wikidata as an external open linked data provider [312]. In one article, Metilli et al. developed NBVT, a semi-automated tool, which uses Wikidata to extract information about events and people to construct a visualized narrative [193]. In another publication, the outcomes of an experiment generating the Wikidata Event Graph are discussed by Metilli et al.. This experiment is based on events implicitly stored within Wikidata [194].

Ilievski et al. compare algorithms that compute the similarity of nodes in KBs through their developed user interface which computes the similarity of Qnodes in Wikidata [135]. Delpeuch through the implementation of an standard API[43], provide a reconciliation service for data matching in Wikidata [49]. Graux et al. developed *Wikidata Live*, a real-time dashboard of Wikidata changes through visualizations [103]. Nguyen et al. introduce MTab4Wikidata, an automatic semantic annotation system to match elements of the table (i.e., cell, column, relations between columns) with Wikidata concepts [210].

Rudnik et al. propose a method for producing semantic annotations for news articles using Wikidata KB [262]. Subasic et al. propose an approach to building a KG for domain-specific AI application using the data from Wikidata [304]. Axelsson and Skantze explore how the information presentation session of an interactive agent based on feedback from the audience can be shaped using Wikidata KG [15]. Moya Loustaunau and Hogan propose a method to predict the results of a query in the next version of a dynamic RDF graph using the data from Wikidata and DBpedia KGs [202]. Kume and Kozaki propose a method for domain ontology construction by extracting concepts of a target domain from Wikidata as a validation experiment [164]. Si infers Creative Analogous Relationships from Wikidata [290]. Wasi et al. propose a document classification approach using classification algorithms integrated with Wikidata, where, Wikidata properties are used as features to classify the documents of the same type [329]. Bianchini and Bargioni develop

[43] https://reconciliation-api.github.io/specs/latest/

CCLitBox tool/gadget to classify literally authors and works using faceted classification approach on Wikidata linked open data [23].

Schmelzeisen et al. present a dataset of Wikidata's full revision history [276]. Zhou et al. propose a method for the construction of multimedia entity linking (MEL) datasets and release three MEL datasets including Wikidata-MEL [353].

Hassanzadeh investigates the building of a KB for events and consequences using their defined casual knowledge extraction pipeline on Wikidata [118].

2.2.3.4 Knowledge Graph Oriented Research

Wikidata is maintained by an active community of contributors who create a large amount of structured data. The KB relies on the MediaWiki infrastructure. In the meantime, Wikidata's structured data is stored in RDF and is accessible through SPARQL. Wikidata belongs, therefore, to a group of other general-purpose KGs, such as DBpedia, YAGO, and Cyc.

The research articles in this category look at Wikidata from a KG lens and focus on bringing more strength to Wikidata as a linked data provider, comparing the KGs, or addressing general issues of KGs.

Wikidata as Linked Data Provider. We summarize all articles that propose approaches for storing Wikidata's structured data in RDF, suggest how projects in Wikimedia's ecosystem can use the RDF data, and how to get more benefit from linking Wikidata to other linked data providers.

Erxleben et al. argue that despite being the data platform in the Wikimedia ecosystem, Wikidata provides its data not in RDF, which affects Wikidata's popularity in the Semantic Web community negatively. Thus, the authors propose an RDF encoding for Wikidata and introduce a tool[44] for creating such RDF file exports [62]. From 2015, Wikidata stored its data in RDF based on the Erxleben et al. mapping [62] and provided the data via a SPARQL endpoint, the Wikidata Query Service (WDQS)[45].

Similarly, Hernández et al., in their studies, compare various options for reifying RDF triples from Wikidata [129], and building on that study the efficiency of various database engines for querying Wikidata [130].

Pezoa et al. investigate the feasibility of implementing their formal definition of syntax and semantics as a layer on the top of JSON by demonstrating JSON schema setup and validation for Wikidata [230]. NECKaR [93] is a named entities classifier based on Wikidata, which provides a Wikidata-based named entity data set. Jacobsen et al. propose Wikidata as a hub to connect ontologies in the linked data cloud through resolvable Internationalized Resource Identifiers (IRIs) [137].

Yang et al. uses the data in Wikidata for improving Wikipedia. They discuss that KGs can help machines to analyze plain texts and propose a Relation Linking System for Wikidata (RLSW) which links the Wikidata KG to data in plain text format in Wikipedia [344]. Ismayilov et al. describe the data integration process of Wikidata and DBpedia Data Stack to use Wikidata through DBpedia extractors and describe the structure DBpediaWikidata (DBw) dataset [136].

[44] For more information, please check https://www.mediawiki.org/wiki/Wikidata_Toolkit.

[45] The WDQS is available here https://query.wikidata.org/.

Hachey et al. present a neural network model for mapping structured and unstructured data and investigate the generation of Wikipedia biographic summary sentences from Wikidata [109].

Lubani and Noah use Wikidata for natural language text entities lacking labels to build a vector representation of named entities that facilitate the task of ontology population [176].

The research post-2018 is mostly focused on entity linking and mapping of Wikidata entities to external data sources. Haller et al. investigate how Wikidata is linked to other data sources in the linked data ecosystem [115]. Bhargava et al. build a system to map Wikidata entities to a set of topic-specific predefined concepts [22].

Sakor et al. present Falcon 2.0, a tool used for entity recognition and linking in short text and linking them to Wikidata KG [263]. Filipiak et al. presents a mapping of the ImageNet dataset linked with Wikidata entities which can be used for various computer vision tasks [75]. Cetoli et al. study Named Entity Disambiguation with applied deep learning and neural techniques through comparing the entities in short sentences with Wikidata graphs and develop a new dataset of Wikidata-Disamb [34]. Shanaz and Ragel design an entity linking system to disambiguate persons in news articles [285]. Möller et al. focus on Entity Linking datasets and approaches available on Wikidata [197]. Mulang' et al. investigates how the usage of Wikidata aliases could improve an attentive neural network approach for entity linking on Wikidata [203]. This work by González et al. explains the linking between ESCO (European Skills, Competences, Qualifications, and Occupations ontology) ontology of concrete skills and Wikidata which results in extracting additional knowledge items from Wikidata and being integrated with ESCO [100]. Delpeuch propose a lightweight Named Entity Linking system that can be trained on Wikidata only and stay synchronous with Wikidata in real time [48].

Coladangelo and Ransom Semantically enrich the name authority data related to manuscripts of the pre-modern and present-day scholarly community in the University of Pennsylvania Libraries to Wikidata pages to be queried by SPARQL query language [38]. van Veen Explain how Wikidata could be used as an authority control mechanism with the help of Wikidata identifier of notable entities utilized as a common identifier for connecting resources [322].

Heftberger et al. use Neonion, a semantic annotation tool, to annotate film studies research articles and link them to the data in Wikidata to enable scholars with little or no technical background to work with the structured data required in Linked Data environments without being overwhelmed by the technological concepts [121].

Ostapuk et al. describe their published resource which adds more links between Wikidata and Wikipedia article sections [222]. Ravi et al. propose Cholan, a modular approach for end-to-end entity linking over KGs, which was conducted on Wikidata and Wikipedia KGs [251]. Seidlmayer et al. developed an approach that improves the links between author and publications based on the ORCID database to Wikidata [280]. Shigapov et al. present BBW (boosted by wiki), a semantic annotation tool that matches tabular data to the Wikidata KG [289].

Kovács et al. conducted benchmarking on various graph database implementations using Wikidata [160].

Comparison of KGs. Here, we discuss articles that compare Wikidata with other general domain KGs. Ringler and Paulheim, for example, study DBpedia, Freebase, OpenCyc, Wikidata, and YAGO KGs to find similarities and differences of these KGs [258]. Färber et al., in their research [65], compare the mentioned KGs from a data quality perspective. Razniewski et al. discuss the challenges of asserting completeness in KGs and outline possible solutions. The authors propose a framework for finding the most suitable KG for a given setting [253]. Abián et al. compare Wikidata and DBpedia structured data sources based on the criteria defined in the main data quality frameworks [2]. In a similar study, Thakkar et al. compare DBpedia and Wikidata from a quality assurance perspective and have found that based on the majority of relevant metrics, the quality of Wikidata is higher than DBpedia [313]. The data quality of Wikidata has also been studied from a KG perspective, as in a study from Gad-Elrab et al. which discuss that KGs like DBpedia, Freebase, YAGO, and Wikidata are inevitably incomplete. To address this, the authors analyze the former approach of data correlations and propose a method to overcome the problems with the mentioned approach [85].

There are only two articles post-2018 with a focus on the comparison of KGs. Pillai et al. which explore and compare Wikidata, DBpedia, and YAGO, the most popular cross-domain KGs, from the perspectives of accessibility of the KG, completeness of the relations and timeliness of the data in the KGs [234]. Razniewski and Das analyze the progress of DBpedia and Wikidata coverage in a longitudinal study through question answering and entity summarization [252].

Common Issues of KGs. In their study, Chekol and Stuckenschmidt discuss that KGs, such as YAGO, Wikidata, NELL, and DBpedia, already contain temporal data (facts together with their validity time). The authors propose a "bitemporal" model for KGs, to record the data extraction time from other sources. Currently, only NELL records this time, while, Wikidata only contains the time which is valid about a fact [37]. In another study, Krötzsch discusses the modern knowledge representation technologies and their advantages in information management, such as description logics, and their contribution to KGs, and motivates Wikidata as a use case [163]. González and Hogan propose a data-driven schema for large-scale KGs and evaluate their approach on Wikidata KG for analyzing how versions of this KG have changed over a period of 11 weeks [99]. Hazimeh et al. present an algorithm for adding social links to academic entities as a refinement method to enable the KGs to get data about real-world entities from online social networks and implement it in Wikidata and YAGO KGs [119].

In the post-2018 research, Hellmann et al. discuss how data curation workflow in Wikipedia and Wikidata could be improved where users add a high amount of data from external data sources [127].

Senaratne et al. propose an unsupervised feature-based approach for anomaly detection in KGs and evaluate it on four KGs of YAGO-1, KBpedia, Wikidata, and DSKG [282].

Noullet et al. provide $KORE50^DYWC$ through extending a largely-used gold standard dataset, KORE50, to be used with KGs like DBpedia, YAGO, Wikidata, and Crunchbase for accommodating tasks related to named entity recognition and disambiguation (NERD) [218].

Desouki et al. present two methods to address the ranking problem of KGs which are growing larger day by day and perform a successful test of these methods on Wikidata and LinkedGeoData [52].

Demartini discusses the implicit biased information caused by contributors in KGs and how paid crowd workers can be used to identify such controversial data [50].

Luggen et al. introduce a method for evaluating class size in collaborative KGs with evaluation over Wikidata [177]. Johnson study the transclusion of Wikidata content in Wikipedia language versions with a focus on English Wikipedia [138].

As can be seen, there is not a focused research area and KGs are being studied from various angles.

2.2.3.5 Application Use Cases

From the beginning, Wikidata received much attention from members of various research fields. Many articles described possible use cases for utilizing Wikidata as a central data hub, as we see in the next section.

Medical and Biological Data Medical and biological projects have started using Wikidata as a backend data source, to facilitate data exchange, mapping, and consumption early in 2015. Mitraka et al., for example, propose the usage of Wikidata for addressing the crucial challenges in disseminating and integrating knowledge in life sciences contexts, by linking genes, drugs, and diseases [196]. Pfundner et al. have specified an automated process to integrate data from ONC's[46] high priority DDI[47] list into Wikidata. The authors aim to integrate the data from ONC into Wikidata and then use Wikidata to display the integrated data in articles of different Wikipedia language versions [231]. Burgstaller-Muehlbacher et al. import all human and mouse genes, and all human and mouse proteins into Wikidata to improve the state of biological data, and facilitate data management and data dissemination using the Wikidata Query Service [32]. Putman et al. describe WikiGenomes, a web application based on Wikidata, that facilitates the "consumption and curation of genomic data by the entire biomedical researcher community". WikiGenomes provides access to centralized biomedical data and a simple user interface for non-developer biologists [246].

The research on the utilization and benefits of Wikidata as a resource for medical and biological data continues post-2018. Turki et al. explain the potential of Wikidata as a platform for medical and biological data [320]. Manske et al. show how Wikidata can be beneficial in terms of resources, integrating volunteer and scientific communities, maintenance and enrichment of original data if GeneDB annotations are imported in Wikidata [184]. Dahir et al. suggest an approach to improve information retrieval in the medical domain through Wikidata and DBpedia using query extension [46]. Waagmeester et al. give insight into the breadth and depth of biomedical data stored in Wikidata and the tools they have built which add biomedical knowledge to Wikidata and synchronize it with source databases [327].

[46] The Office of the National Coordinator (ONC) for Health Information Technology is a division of the United States Department of Health and Human Services.

[47] DDI stands for Drug-Drug Interaction, i.e., the effect change of one drug on the body by another drug.

During the event of COVID-19 Pandemic, a visible number of research has been performed to utilize Wikidata for more efficient usage of COVID-19 data. Turki et al. demonstrate the potential of Wikidata for COVID-19 information being stored, analyzed, and visualized for decision support and educational purposes [321]. Waagmeester et al. propose how to use Wikidata as a common ground for linking all COVID-19-related research and studies. Additionally, the study uses Wikidata as a common ground to link disparate resources of medical and biological data and develop a semantic schema for virus strains, genes, and proteins using the Wikidata infrastructure [328]. Lemus-Rojas and Ramirez Rojas share details of collaborations among the employees of campus libraries involved in three Wikidata projects related to the COVID-19 Pandemic [170]. Darari presents COVIWD, a dashboard based on Wikidata which provides information and visualization services covering COVID-19-related topics [47].

Linguistics Wikidata is also used in the linguistics field, either as a dictionary or proposing further approaches for linking lexical datasets or relation extraction.

Turki et al. propose to adopt Wikidata as a dictionary that can be used across multiple dialects of the Arabic language. The authors emphasize that the Arabic language has many dialects and these dialects are not all mutually intelligible, and each one of them has its morphological and phonological, and even semantic and lexical particularities. The study explains how it is possible to convert Wikidata into a multilingual multidialectal dictionary for Arabic dialects and describes how Wikidata (as a multilingual multidialectal dictionary for Arabic dialects) can be used by computational linguistics and computer scientists in the Natural Language Processing of the varieties of the Arabic language [319]. Nielsen describe an ongoing effort for linking ImageNet[48] WordNet[49] synsets to Wikidata [211]. Yu and Qiao present a new approach for meronym relations extraction in Wikidata, which is, building a 13-dimensional feature vector for each hyperlink to be classified with different classification algorithms, based on all 13 different three-node motifs. The high interest of this community might have one driver for the development of the Wikibase Lexeme extension which allows for modeling lexical entities. From 2018, Wikidata includes this new type of data: words, phrases, and sentences [347].

There exist only two studies after 2018 on the utilization of Wikidata in the linguistics area. McCrae and Cillessen propose linking Wikidata to WordNet using the techniques such as natural language processing and hapax legomenon links [188]. Thalhath et al. investigate the possibilities of Wikidata being used as a vocabulary resource to boost the use of linkable concepts [314].

Mathematics Recently, a number of studies in Wikidata research have made the effort to get benefit from the structured nature of Wikidata in the context of Mathematical data storage and representation. Scharpf et al. describe how to link identifiers and symbols in Content MathML to Wikidata items in order to utilize the benefit of semantics through annotation of mathematical identifiers or opera-

[48] "ImageNet is an image dataset organized according to the WordNet hierarchy" (http://image-net.org/.

[49] WordNet is a large lexical database of English and contains and groups nouns, verbs, adjectives, and adverbs in the form of sets of cognitive synonyms (synsets). For more information https://wordnet.princeton.edu.

tors [272]. Nguyen and Takeda propose an approach to enhance the performance of semantic labeling for numerical attributes and use Wikidata to for unit conversion and generating more data resources for numerical background KBs [209]. Schubotz et al. in one study, presents the first benchmark dataset that can evaluate the conversion of mathematical formulae between the presentation format and content format and linked to Wikidata entities for improved access through linked data [278]. In another study, Schubotz explains the OpenMath content dictionary which is automatically generated from Wikidata. The study also proposes a Wikidata property to link OpenMath entries with Wikidata Items [277].

In the research articles post 2018 we find continued efforts of Scharpf et al. with a research focus on Mathematics and display a summary of how Wikidata can be improved to reflect the mathematical entities in a proper way through introducing a data model for mathematical statements in Wikidata [274]. Further, Scharpf et al. presents an approach to speed up mathematical entity linking in Wikidata [273]. Furthermore, Scharpf et al. developed MathQA, a Question Answering system to answer mathematical questions based on the data from Wikidata, Wikipedia, and arXiv preprint repository [275]. Another study by Elizarov et al. proposes methods for building a digital mathematical library and uses Wikidata to supplement metadata [61].

Geography. Wikidata's semantic nature can provide benefits in the geographical data context as can be seen from the research below. Almeida et al. introduce a tool that harmonizes street names from OpenStreetMap[50] and the entities they refer to are accessed through Wikidata [8]. Leyh and Filho discuss the opportunities and challenges of Wikidata as a central integration facility by interlinking it with OpenStreetMap [173]. Spitz et al. present an approach for constructing a network of locations from Wikipedia by computing the similarity of locations based on their distances and linking it to Wikidata as a knowledge source [294]. In another study, Spitz et al. introduce ranking methods for the extraction of complex location relations from Wikipedia articles and Wikidata KB that are not hierarchical [295].

Wikidata remains a research focus in the geography context after 2018. Shanaz and Ragel develop a location entity linking system used to disambiguate named entities related to locations mentioned in English news articles utilizing the semantic data of Wikidata KG [286]. Gurtovoy and Gottschalk present StreetToPerson, a new approach for linking street names in OpenStreetMap to persons in Wikidata [107]. Yang et al. propose a hybrid search application that takes the text and geospatial data and extracts knowledge from Wikidata and HydroSHEDS dataset regarding the rivers worldwide [343].

Question-Answering Systems. Question-answering systems have more recently utilized the structured data of Wikidata. Diefenbach et al. present and discuss WDAqua-core, a new Questions Answering component, which uses DBpedia and Wikidata [54, 56]. Tanon et al. explain the development of Platypus, a multilingual Quesion-answering system on Wikidata [311]. Striewe describe how to dynamically generate assessment items from Wikidata [303].

[50] For more information please check: https://www.openstreetmap.org/

A visible number of research in this area comes after 2018. Ma and Ma propose a complete framework that can automatically generate quiz questions based on video subtitles of MOOC (Massive Open Online Courses) through Wikidata [179]. Korablinov and Braslavski present RuBQ, a Russian Question Answering KB dataset built over Wikidata [158]. Ploumis et al. present an approach for question-answering systems which analyses the question and looks for the answer in the Wikidata through SPARQL queries [242]. Dubey et al. explain the creation of the Large-scale Complex Question Answering Dataset (LC-QuAD 2.0) with compatible SPARQL queries of Wikidata and DBpedia [60]. In one study, Perevalov et al. present the extended version of the KG question-answering dataset (QALD-9) by adding translation of the existing questions in eight languages and transferring the queries from DBpedia to Wikidata [227]. In another study, Perevalov et al. use the semantic potential of Wikidata along machine translation approaches over extended QALD-9 dataset to provide questions and answers in multilingual form [226].

Historical and Cultural Heritage. Archived data can take advantage of semantic technologies to become accessible and be used more efficiently. Wikidata provides this facility and has got popular in providing visibility to such archived data very recently. The research focused on the usage of Wikidata as a linked data provider for historical and cultural heritage data has emerged recently and so far we see all contributions in this research focus post-2018 except for one.

Veen et al. use Wikidata to improve access to the collection of Dutch historical newspapers [323]. Kapsalis looks at how Wikidata could be utilized by cultural heritage organizations to provide cultural heritage data more visibility and help them become accessible knowledge resources [148]. Putra et al. present their approach for extracting cultural heritage data from multiple formats and constructing a KB using an RDF data model where the entities of this KB are imported and linked into Wikidata for greater interoperability of cultural heritage information [247]. Heberlein reports on the project of modeling numismatic descriptive metadata using Functional Requirements for Bibliographic Records(FRBR-oo)[51] ontology and Wikidata [120]. Faraj and Micsik automate the process of detecting and extending links between the entities of COURAGE[52] project and Wikidata cultural heritage data [64]. Cooey discusses the usage of Wikidata to expand and enhance the authority records for Holocaust-era camps and ghettos in European Holocaust Research Infrastructure (EHRI) portal [40]. Denis discusses the usage of Wikidata as a means of getting advantage from linked open data for the archived cartographic heritage resources [51].

Origlia et al. present a methodology to create a graph database from Wikidata, Wikipedia, and Flickr that supports cultural heritage data [221]. Freire and Proença experiment several approaches to find a solution for reasoning problems in large ontologies. The authors designed a method for this purpose and evaluated it on Schema.org and Wikidata Cultural Heritage data [81]. Colla et al. present a new feature for an ontology-driven annotation system that generates suggestions on historical entities based on their extracted information from Wikidata [39].

[51] "The FRBRoo is a formal ontology intended to capture and represent the underlying semantics of bibliographic information." https://www.cidoc-crm.org/frbroo

[52] COURAGE (Cultural Opposition – Understanding the CultuRal HeritAGE of Dissent in the Former Socialist Countries) is a three-year international research project funded by Horizon 2020, the EU Framework Programme for Research and Innovation. http://cultural-opposition.eu

Library Systems. The concept of using linked data in Library systems has newly emerged through the usage of Wikidata in the context of library systems and digital preservation. The research in this area has begun in 2017 by Thornton et al. with a study that explores the potential of Wikidata to serve as a technical metadata repository and how it provides distinct advantages for usage in the domain of digital preservation [316]. Allison-Cassin and Scott describe how Wikidata is a low-barrier option for creating and using linked open data in libraries [7]. Thornton et al. prepare datasets related to digital preservation from Wikidata for their developed software portal of Wikidata for Digital Preservation [317].

The main body of research in this area comes post-2018. Seals-Nutt and Thornton present Wikidp which is a digital preservation portal and allows people to access digital preservation-related data from the Wikidata KB [279]. Lemus-Rojas and Lee share the experience of pilot projects from three university libraries to prepare data and link it to Wikidata properties with the aim of broadening the representation and enhancing the visibility of women in STEM [168].

Snyder et al. explain the usage of linked open data in libraries to enhance the information displayed in library discovery systems through the linkage between Library of Congress Subject Heading (LCSH) and Wikidata [292]. Pohl explains the process of creating spatial classification entries from Wikidata for the North Rhine-Westphalian Bibliography (NWBib) [243]. Nešić et al. take a number of use cases as examples to describe the integration of Wikidata with digital libraries and external systems and how the process of data preparation and import could speed up in Wikidata [208]. Stanković and Davidović use Infotheca, the journal of digital humanities, to explain the integration of Wikidata with digital libraries and external systems [297]. Spencer et al. create the Siegfried/Wikidata integration tool which aims to make Yale University Library's data in Wikidata consumable through the Siegfried[53] utility for the file identification purposes [293]. Alexiev et al. introduce BIDL[54] which is the virtual encyclopedia of Bulgarian Icons and how it can be exported into Wikidata [5]. Fukuda examines the advantage of utilizing Wikidata to catalog video games at the Center for Game Studies, Ritsumeikan University (RCGS) in order to construct an authority of works for video games [84]. Bianchini and Sardo investigate how the semantic nature of Wikidata could serve as a new perspective towards universal bibliographic control [24].

Finance and Management. The very recent research on Wikidata reveals the potential of Wikidata in providing service to management and financial systems. Wikidata could be used as a multipurpose KB, and this study by Krabina and Polleres examines the use of Wikidata as a platform for processing public finances [161].

Ang and Lim propose the Knowledge-Enriched Company Embedding (KECE) model which combines multimodal information of companies from KBs and takes advantage of KG relationships in Wikidata for generating company entity embeddings. These entities are used to improve the performance of downstream investment management tasks [10]. Portisch et al. present a hypernym detection system for the financial services domain named FinMatcher which leverages the semantic benefits of Wikidata, WordNet, and WebIsALOD [244].

[53] Siegfried is a file format identification tool.

[54] Bulgarian Iconographic Digital Library http://bidl.cc.bas.bg/

Data Checking/Validation. Besides, all of the above-mentioned application areas of Wikidata, recent research shows the ability of Wikidata to provide fact-checking and validation. Khandelwal and Kumar is focused on a new method for fact-checking using unstructured data from Wikipedia and structured data from Wikidata for determining the validity of facts [151]. Zubiaga and Jiang introduce a method to detect hoaxes in social media through Wikidata KB in a semi-automated manner [355].

Abián et al. propose the contemporary constraint concept for information consistency for KBs like Wikidata [1]. Goodrich et al. propose a model-based metric to indicate the factual accuracy of generated text based on their introduced dataset from Wikipedia and Wikidata [102]. Frey et al. define their proposed process for creating rich language-specific datasets from DBpedia and evaluate them against Wikidata and DBpedia KBs [82].

Lim et al. perform a preliminary study of topic modeling on Twitter by incorporating spatial and temporal data from Wikidata [174]. Dooley and Bozic examine the correlation between Twitter trending hashtags and Wikidata revisions page titles in a specific time frame [57].

2.2.3.6 Implications

Our exploration of Wikidata research in the first decade of Wikidata's existence display where the Wikidata community and researchers stand from the research perspective, where they show more interest, and where we still see research gaps.

The articles have a prevalence of computer science articles which we expected from the chosen databases which are mainly Computer Science related (ACM, Springer Link, DBLP). However, by including Google Scholar, we expected to identify more research from disciplines such as sociology or communication science. Unfortunately, our results suggest that this approach was less successful.

Like other peer production communities, Wikidata provides a valuable opportunity to deepen our understanding of existing community practice. It might be interesting, for example, to explore existing differences with Wikipedia. There is still resistance to the use of Wikidata within various Wikipedia language versions. Further research is needed, to better understand existing reservations. Another interesting less studied aspect in Wikidata is the existing human-bot-collaboration [206]. Wikidata might be, besides Wikipedia, an interesting use case to better understand the social-technical infrastructure of a peer production community.

The results suggest that research on Wikidata seems to be entirely concentrated on specific institutions, such as the University of Southampton or the University of Lyon, or countries, for example, Germany and USA. It might be the origin of Wikidata as a European project initiated by members of the Semantic Web community which causes the research on Wikidata to be more popular in Europe. We wonder, how this Western perspective on knowledge representation might exclude other understandings of knowledge. For example, the indigenous peoples give their knowledge orally from generation to generation. Research, which deals with the question of how this knowledge or the potential occurrence of such knowledge can be represented, would undoubtedly be useful to achieve the aim of becoming a global universal KB, which can be used by anyone for any purpose [326].

While there have been studies on the multilingualism aspect of Wikidata, the data is still not present in every language. Current findings show that there are some dominant languages (e.g., English, French, German, Spanish), while, many other languages are 'underserved'. This indicates that, although there have been some efforts in addressing the issue of uneven language distribution, further studies are needed to overcome the language gap in Wikidata. Furthermore, these studies focus on the descriptions and labels of an item. It might be interesting to understand better when Wikidata's data model fails because a one-to-one relationship between two words from different languages is not possible.

Continuous evolution is one of the design decisions of Wikidata, which means Wikidata grows with its community and tasks, and new features are deployed incrementally [326]. While we see numerous articles focusing on the enhancement of existing Wikidata features, the findings suggest only a little research on improving the usability of the user interface. User studies concerning aspects such as learnability or explainability are still rare on Wikidata. From our own experiences in conducting Wikidata workshops, it can be said, that people struggle with understanding Wikidata's central concepts, for example, the difference between a class and an instance. It seems that Wikidata has still untapped potential in becoming accessible for non-technical experts.

Many efforts are made to sustain and improve the quality and completeness of data in Wikidata. One issue in this context is, for example, the handling of vandalism and data integrity. In the context of data quality, we call for more research on the effects of plurality, i.e., the co-existence of contradictory information, in order to enhance the trustworthiness of Wikidata content. However, if anyone can add contradictory information, further research is needed to provide such mechanisms in the user interface as well as in the WDQS for providing this information in a possible format.

As opposed to Wikidata, Wikipedia is studied from a variety of disciplines, such as humanities (e.g, history, literature, philosophy), logic and mathematics, natural sciences (biology, chemistry), social sciences (e.g., communications, education, economics, law, journalism) and interdisciplinary (anthropology, computer science, health, industrial ecology, and information science) [220]. While Wikidata has the competence to be used in different disciplines, investigations are needed to find out whether Wikidata can be beneficial in the same areas where Wikipedia was used. Even though our study reveals the usage of Wikidata in various contexts, the use cases come from a limited number of application areas such as the biomedical domain, linguistics, cultural heritage, or library systems. Although we see a recent increase in the areas where Wikidata is utilized like geography or mathematics, it might be valuable to see more use cases from other disciplines, such as social sciences or humanities. Leveraging Wikidata in educational environments, for instance, could offer significant value.

Earlier, as we reflected on the design decisions, we recognized that diversity serves as a means to achieve the overarching goal of Wikidata. Given the significance of this concept within the Wikidata context, we observe a research gap that pertains to diversity. Thus, one of the recommended directions of research on Wikidata is to study the concept of diversity in the Wikidata context. This would help to better understand how close Wikidata is to achieving its main goal of being diverse enough to provide service for anyone on the planet. Since the data in Wikidata comes from the community of contributors, understanding the contributing community of Wikidata is one of the important steps in drawing the diversity picture of Wikidata.

Figure 2.11 shows the number of published papers in the time frame 2012 to 2018 in comparison to the time frame of 2019 to 2022. We can observe that the focus of research in the era pre-2019 was the development of tools and datasets over Wikidata, enabling Wikidata to serve as a linked-data provider and the usage of Wikidata in different application use cases. In addition, in this time period, a visible number of research also exists on the topics of multilingualism, data quality, and the contributing community of Wikidata. Overall, in the pre-2019 time frame, we see a rather balanced contribution to the Wikidata research categories with more focus on Application Use Cases categories.

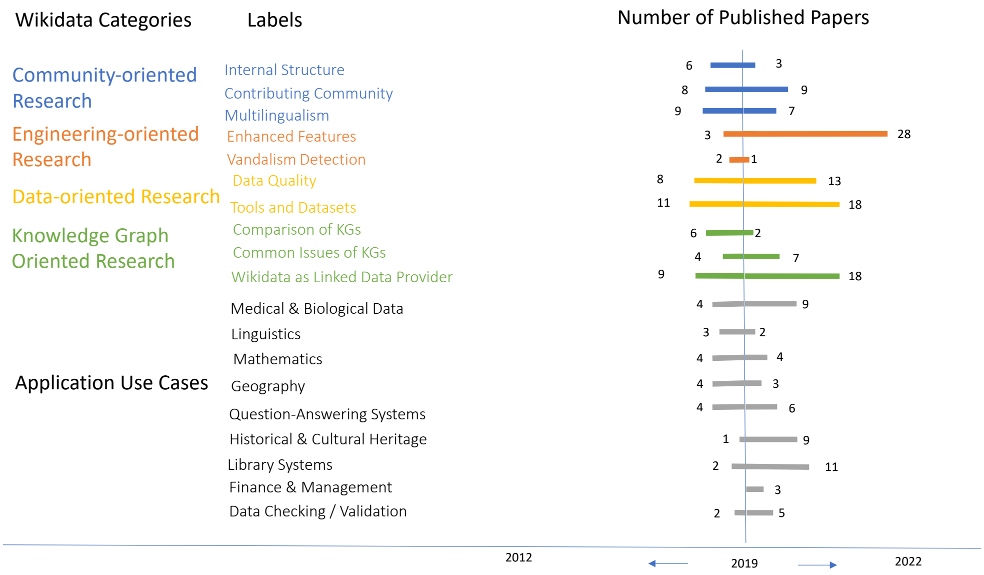


Figure 2.11: A comparative view of Wikidata research from 2012 to 2019 and 2019 to 2022. Source: *Wikidata Articles Dataset*.

In the time frame 2019 onward, research article distribution across Wikidata categories seems less balanced. In contrast to the pre-2019 period, we see a visible increase in the Engineering-oriented Research category, similarly, the Application Use Cases category is on the rise more than before. Having a high number of papers focused on enhancing Wikidata features is expected because Wikidata was designed to continuously evolve, addressing the needs of the community. Moreover, the more Wikidata is used by more people for a higher variety of purposes, the more existing shortcomings are discovered, and the need for new functionalities is revealed. Development of new tools, building datasets over the data of Wikidata for various purposes, and strengthening the ability of Wikidata as a linked data provider remain popular research topics in the Wikidata research community. It seems that the focus of the research community is currently on Engineering-oriented research and the Data-oriented research categories with a visible increase in the areas of data quality and providing means to utilize the data in Wikidata through new tools and datasets. Further, more efforts towards bringing more power to Wikidata as a linked data provider and usage of Wikidata in various application areas indicate that the research focus is mainly on the data aspect of Wikidata.

Overall, we are left with the impression that Wikidata is consistently evolving to address the evolving needs of its community. Furthermore, Wikidata is progressively being integrated and applied across various application domains. This indicates Wikidata's success in furnishing data to diverse systems for multiple purposes, ultimately serving individuals worldwide. While this progress is promising, numerous challenges remain to be explored and examined in order to ascertain the true status and advancement of Wikidata toward its ultimate goal.

2.3 Summary

Wikidata is an open, collaborative KB launched with the goal of reflecting world knowledge and serving "anyone, anywhere in the world." To achieve this objective, certain design decisions were considered in the development of Wikidata that would enable it to reflect world knowledge and distinguish it from other KBs. These design decisions include open editing, secondary data, multilingualism, plurality, community control, and continuous evolution - all of which refer to the concept of diversity. Therefore, we get the impression that being diverse enough to serve the diverse people in the world is the ultimate goal of Wikidata.

Diversity is not only a distinguishing feature of Wikidata but also the 2030 goal of the entire Wikimedia Foundation. The Foundation initiated a Movement Strategy Process and one of the resulting two goals set for the 2030 horizon is to reach "knowledge equity." This includes the objective to "counteract structural inequalities to ensure a just representation of knowledge and people in the Wikimedia movement." However, the concept of diversity in the context of Wikidata is not yet clearly defined and explored. It is only discussed from the angle of plurality, i.e., the coexistence of contradictory statements. Whether plurality is a synonym for diversity in the Wikidata context is yet to be explored.

Existing research has shown that the Wikidata community consists of both humans and bots, with bots being the most active contributors to Wikidata. Furthermore, very few studies exist on bots, and they are a rather unexplored user group. Therefore, it remains to be studied how bots might impact Wikidata, especially its diversity aspect.

In the next sections, we will tackle both issues, the diversity concept in Wikidata and bots in the Wikidata context. The results of both sections will help us understand what diversity means in the Wikidata context and how bots might contribute to or impact it.

Chapter 3

DIVERSITY & WIKIDATA

In the Wikidata chapter, we learned that Wikidata aims to become a universal KB, and to achieve this goal, it needs to serve diverse enough data in terms of providing data in a variety of languages, covering different topical domains, and reflecting diverse opinions. Hence, diversity is a means to enable Wikidata to achieve its ultimate goal of serving world knowledge.

Despite the significance of diversity in the context of Wikidata and the presence of built-in support for it through the plurality design decision, diversity has not been given sufficient attention as a research focus. Further, Wikidata provides the possibility to be edited even by unregistered users, thus, it is open and welcomes participation from anyone willing to contribute in any language. However, having the ability to support diversity does not guarantee a high level of diversity in the data. This is because Wikidata research has shown that providing support for multilingualism alone is not enough to ensure that all data are available in a multilingual form, as there exist many languages that are overlooked [142].

In addition, there does not appear to be a common definition of diversity in the context of Wikidata. Some studies focus on gender inequality [283, 248, 29, 350] and assess gender bias, highlighting that not all genders receive equal attention in Wikidata. However, it is not directly established how these studies relate to the concept of plurality in Wikidata. Thus, the exploration of a common definition of diversity and the understanding of whether plurality is a true synonym for diversity are areas that remain to be explored.

The concept of diversity is not only of interest in the context of Wikidata but also aligns with the Wikimedia 2030 goal[1]. The strategic direction of the Wikimedia movement toward diversity in 2030 focuses on knowledge equity and knowledge as a service, highlighting the importance of diversity for the entire Wikimedia community. As Wikidata is a project under the umbrella of the Wikimedia Foundation, which was designed with diversity in mind, examining the current status of diversity in Wikidata after a decade of its existence holds relevance not only for the Wikidata

[1] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20

community but also for the broader Wikimedia community and any system aiming to provide free knowledge for all.

Furthermore, previous research has indicated that a significant portion of Wikidata edits are conducted by bots, leading to the potential influence of bots on diversity. Bots are known for automating repetitive tasks, often resulting in more uniform edits. There are concerns that bot edits in Wikidata may be less diverse compared to human edits [238]. This raises the question of how the involvement of bots within the Wikidata community could influence the objective of promoting diversity to serve to a worldwide audience.

In this chapter, our objective is to provide an overview of the diversity concept and explore its application within a KB context with a focus on Wikidata. Given our interest in assessing the existing diversity status of Wikidata, we propose a concept for measuring diversity in Wikidata and present our suggested approach for measuring the current diversity status. This measurement will later enable us to examine the impact of bot edits on diversity in Wikidata.

We commence by offering a general introduction to diversity and delving into the various dimensions and interpretations of this term.

3.1 The Diversity Notion

Diversity is a rather general concept that is present in numerous contexts. One of the earliest definitions of diversity dates back to 1949 [291] in the field of economics. This definition introduced the concept of measuring diversity using the diversity index, which assesses the level of concentration within a group of individuals when they are categorized into various classifications. Since then, diversity has been a topic of attention in many fields like biology, ecology, sociology, and archaeology, to name a few. Due to its extensive usage in different fields for various purposes, it is a challenge to provide a universal definition for the term diversity that could satisfy every context where diversity is used [59]. Since we are focusing on the concept of diversity within the context of a KB (i.e., Wikidata), it is important to first gain a clear understanding of what diversity entails and how it is interpreted in other contexts. Therefore, we begin our exploration of the diversity concept by exploring the definition of the term “diversity” and then delve into its interpretations and applications in various fields. This process allows us to establish a foundational understanding of diversity in general, which will serve as a basis for discussing the concept of diversity within the realm of Wikidata.

3.1.1 Defining Diversity

According to the Merriam-Webster Online Dictionary[2] the term diversity is defined as “the condition of having or being composed of differing elements” which is very close to the term variety “the quality or state of having different forms or types.” Thus, variety is commonly misunderstood to be an alternative to diversity. While diversity is a means to measure variety, the diversity concept is not limited to representing variety only [171] and has further dimensions explained later.

[2] https://www.merriam-webster.com/dictionary/diversity [Accessed 03-08-2020]

Leonard and Jones define diversity as “the nature or degree of apportionment of a quantity to a set of well-defined categories” [171]. This definition which has emerged from the field of archaeology shows that diversity can be a topic of interest in any field which consists of categories and, thus, could be used in a variety of fields. For instance, social sciences deal with people from different backgrounds, economics is interested in the income levels of individuals in a society, and biology where species types are the focus.

There are also terms that show specific types of diversity, such as knowledge diversity and information diversity. According to Helberger et al. the idea of *information diversity* is that in a “democratic society informed citizens collect information about the world from a diverse mix of sources with different viewpoints so that they can make balanced and well-considered decisions” [126]. Giunchiglia et al. defines diversity in the context of media content analysis and knowledge diversity as “the co-existence of contradictory opinions and/or statements (some typically non-factual or referring to opposing beliefs/opinions)” [96].

In the above-mentioned definitions, we can see the terms variety, balance (i.e., degree of apportionment), and difference/ disparity (i.e., contradictory) to define diversity. This confirms the fact that there exists no common definition of diversity that could fit every situation, and the existing definitions from one field or context are not accurately applicable to all other contexts. For example, assuming diversity as an alternative to variety is popular in biodiversity, but is not applicable in the field of economics where diversity is understood in terms of inequality and balance [94]. Or, diversity may not always show contradictory information as defined by [96], for example, the diversity of languages shows the variety of languages that do not contradict each other.

Nevertheless, we could learn that the diversity concept can be a topic of interest in any system where elements can be differentiated and grouped together based on their common factors to form categories. In other words, diversity is the property of any system which consists of categories of elements. However, a thorough understanding of the diversity concept in a context requires a deeper insight into the diversity properties and how they are applied in different contexts in which diversity is used.

3.1.2 Diversity Dimensions

The main question that still remains to be answered is: What do we truly mean when we refer to the diversity of a system? In other words, is it feasible to compare diversity across two different systems? And if so, what are some shared attributes that can be used to measure diversity, irrespective of the context in which a system exists? As an example, we previously saw that in the economic context, diversity refers to the distribution of wealth among citizens of a country. However, diversity in a biodiversity context means the richness of the species in a defined area. Even though, both are examples of diversity, it is a challenge to find a common ground between them and look at them from a single lens due to different or context-specific interpretations of diversity in each field.

Although diversity is mostly studied in the context of specific disciplines, attempts have been made to answer such questions using diversity in an interdisciplinary manner [300]. Despite all these different interpretations of diversity in diverse fields, Stirling was able to spot the most general attributes or properties of diversity and

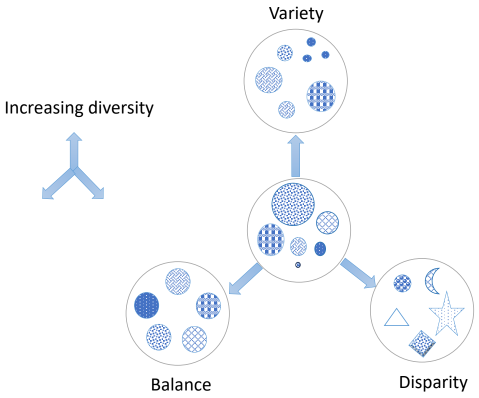


Figure 3.1: A visual representation illustrating the qualities of diversity within the context of interdisciplinary analysis, from [301].

define a common ground for the diversity concept applicable to all contexts. After exploring numerous fields, Stirling has proposed a general framework for analyzing diversity in science, technology, and society [301]. The author defines three properties of diversity, which are variety, balance, and disparity, as follows:

- "Variety is the number of categories into which system elements are apportioned." Variety answers the question of "how many types of things do we have." The relation of variety to diversity is based on: "All else being equal, the greater the variety, the greater the diversity."
- "Balance is a function of the pattern of apportionment of elements across categories." Balance answers the question of "how much of each type of thing do we have." The relation of balance to diversity is based on: "All else being equal, the more even is the balance, the greater the diversity."
- "Disparity refers to the manner and degree in which the elements may be distinguished." Disparity answers the question of "how different from each other are the types of things that we have." The relation of balance to diversity is based on: "All else being equal, the more disparate are the represented elements, the greater the diversity."

According to Stirling each system has some combination of these properties, and every property on its own does not have the potential to be a sole representative of diversity as can be seen in Figure 3.1. Variety and balance, for example, can show how many different species exist and how they are distributed across the species categories. In other words, with variety alone, we could only count the number of different species but would not be able to compare two systems with the same variety but different balance levels. Nevertheless, we first need disparity to distinguish different groups or categories and then measure the variety and balance. Hence, while each of these properties individually offers only a limited understanding of the system's diversity, they are most valuable when utilized together to create a comprehensive depiction.

To further elucidate the concepts of variety, balance, and disparity and their potential impact on diversity, we examine the instance of editor diversity within a Wikidata item. Diversity increases with the rise in the number of countries to which the editors belong. For instance, an item edited by contributors from five distinct countries is expected to exhibit more diverse content than an item edited by contributors from only three countries. Similarly, achieving a balanced distribution of editors from various countries indicates a more diverse content composition compared to a scenario where the majority of content originates from a single dominant race or country. A concentration of editors from one particular country or race can result in an imbalanced representation of viewpoints and a reduction in content diversity within the entry. Likewise, having editors hailing from diverse countries across the globe, characterized by differences in language, culture, race, and geographical location, contributes to greater diversity than an item edited by individuals solely from a similar number of European countries.

Hence, the general concept of diversity finds widespread use across various fields; however, its necessity and application vary in each context. In certain domains, diversity represents a straightforward notion, referring primarily to variety — as the mere count of categories in a distribution Jong and Bates. On the contrary, diversity is presented as a two-dimensional concept, encompassing not only the count of categories (referred to as richness or variety) but also the distribution of elements within those categories (known as evenness or balance) [189]. Adding a recent facet to this concept, disparity complements the widely acknowledged dimensions of the diversity concept focused on variety and balance, as proposed by Stirling.

Next, we explore some application domains of diversity and see how each of the above-mentioned properties or dimensions are mapped into each context.

3.1.3 Application Domains of Diversity

As mentioned before, diversity is an established concept in a variety of fields, like, archaeology, economics, social sciences, biology, ecology, and computer science. To better understand how diversity is defined and what the usage of diversity is like in these contexts, we take a deeper look into some of these fields where diversity is established and used.

3.1.3.1 Diversity from a Biological Lens

Diversity is a central aspect of biological studies and there exists a specific field focused on biological and ecological diversity, named biodiversity. According to the definition of the United Nations Environmental Program (UNEP), biodiversity is a study of variability between living organisms of all types in an ecosystem with a focus on both, within-species and between-species diversity [132]. In another definition, biodiversity is “synonymous with species richness and relative species abundance in space and time” [133].

Biodiversity studies focus on living organisms in a defined area to find the variety, rarity, and abundance of the species in that geographical boundary. The results of these studies are used to compare the diversity levels of two areas or compare the results of the same area in different time slots to better understand ecological changes and to find out if certain species are in danger of distinction.

To measure diversity levels in the context of biodiversity, numerous approaches are used. The oldest measure of biological diversity is called *species richness* which refers to the number of species in a defined area of study [190]. Despite its apparent simplicity, measuring richness demands more effort when applied to samples or populations, as species are not uniformly distributed. Consequently, various approaches exist to quantify the richness of a given area, as exemplified by studies such as [181], [219], and [192]. Entropy is another extensively used diversity measure in this context, also called the Shannon Index (cf. Section 3.1.4).

While numerous methods are at our disposal for measuring species diversity, these measurements do not always yield absolute accuracy. This is due to the fact that, on one hand, not all species are fully identified and categorized, and on the other hand, species do not uniformly exist in equal proportions. Consequently, further inquiries are necessary within this context, and diversity remains an ongoing focal point within biological research.

3.1.3.2 Diversity from a Sociological and Economical Lens

Social sciences represent another domain where diversity stands as a prevalent subject of interest. The ultimate goal of humanity and any truly free society is to respect each individual as they are, embracing their unique differences. Hence, in a sociological context, the essence of diversity revolves around the presence of individuals with diverse backgrounds. A synonymous term pertinent to this domain is *fairness*. Fairness, defined as the “lack of favoritism toward one side or another”[3], contrasts the concept of bias. Fairness is used to support the diversity of a community by looking at the community members without “any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics” [191] such as the origin, gender, culture, language, and beliefs (e.g., political, religious). Fairness encapsulates the concept of an ideal society where each individual is treated equitably, without bias or judgment.

Another vastly used synonym of diversity in the area of social sciences is called *inclusion*. Communities are suggested to include more people from a variety of backgrounds, to not only reduce discrimination but also, bring more experience, innovation, and knowledge together [83].

One example of the sociological diversity research area is organizational/ workplace diversity which categorizes diversity attributes into the following four categories[4]: a) Internal diversity which contains the attributes of a person by birth like race, culture, gender, skin color, b) External diversity that consists of the attributes of a person influenced by his/her surroundings like education, religion, citizenship, c) Organizational diversity that shows characteristics that differentiate one employee from the other like job function, rank, department, and d) Worldview diversity which is the attribute developed by the combination of the above three attributes, like political beliefs.

As can be seen, here the interest lies in having people from a variety of backgrounds and their abundance in a community to demonstrate higher fairness or inclusion levels which are representations of diversity.

[3] Merriam Webster Dictionary: https://www.merriam-webster.com/thesaurus/fairness

[4] What Are the 4 Types of Diversity? https://www.alliant.edu/blog/what-are-4-types-diversity [Accessed: 22.12.2020]

Another perspective to gain a deeper understanding of diversity within a society is through the lens of economics. The economy is a collective asset of society, and assessing the economic well-being of its citizens is a key facet of exploring diversity.

Diversity in economics predominantly centers around matters such as income, wages, and consumption. In this context, diversity is framed through the concepts of 'inequality' and 'concentration.' Researchers are keen on uncovering disparities in individuals' economic statuses, focusing on variations in income levels within a society (i.e., inequality) and the equitable distribution of wealth among its members (e.g., within a country, referred to as concentration).

The most commonly used metric for measuring diversity in this field is the Gini coefficient (cf. Section 3.1.4) [94].

3.1.3.3 Diversity from a Computational Lens

Diversity within a technical context can be exceptionally impactful due to the extensive integration of technology into various facets of daily life. Algorithms and decision-making systems not only simplify complex calculations and classifications but also tend to outperform human capabilities with greater accuracy. Nonetheless, if these systems aren't trained on sufficiently diverse data, they may exhibit bias towards dominant data elements or categories, inadvertently disregarding less represented ones. This imbalance can lead to an overemphasis on and magnification of the dominant elements or categories. Consequently, such biases have far-reaching consequences on diversity within the specific contexts where these systems are employed. For example, consider organizations that, in today's landscape, rely on algorithms to filter applicant curriculum vitae (CVs) for interviews[5]. If the training data lacks information from a diverse array of backgrounds, there's a higher probability that CVs from individuals with different backgrounds might not make it onto the interview shortlist. Instead, only candidates hailing from backgrounds that are more prevalent could end up being selected. Another example can be found within the judicial system, where automated decision-making systems generate a risk assessment score. This score is used in courtrooms to guide determinations about individuals who are less likely to commit future criminal acts and can thus be granted release[6]. However, it's been observed that these algorithms can exhibit bias toward individuals from specific backgrounds due to the composition of the training data. Mass media is another domain significantly impacted by technology. News outlets are progressively transitioning to online platforms, often utilizing recommender systems. Although news should empower citizens to remain well-informed and take informed actions by offering a diverse range of topics, sources, and balanced perspectives, algorithms can inadvertently contribute to selective information exposure and lead to partial information blindness [110].

The impact of diversity in big data is a relatively recent research focus, with numerous studies exploring various aspects. Areas like *Information Retrieval* delve into methods for enriching result sets by considering diverse facets of a query. Such diversification aims to address different information needs that underlie a query [306]. In

[5] Come 2021, AI-based tools will shortlist your resume https://indianexpress.com/article/jobs/come-2021-ai-based-tools-will-shortlist-your-resume-7119609/[Accessed: 27.07.2022]

[6] Source: 'Machine Bias: Risk Assessments in Criminal Sentencing,' ProPublica, Accessed: 16.09.2020

Recommender Systems, diversification aims to counter overfitting and overspecialization issues associated with recommendations solely based on user preferences. The goal is to provide users with a broader array of recommendations [165]. The realm of *Algorithms & Machine Learning* investigates how deploying diverse algorithms to tackle the same problem can lead to enhanced outcomes [207]. Moreover, the topics of fairness and bias in machine learning have received considerable attention [191].

In a comprehensive view, the technical perspective on diversity appears to encompass the integration of diverse data types and algorithms, aiming for more efficient and realistic results.

3.1.3.4 Summary of Diversity Application Domains

In the preceding section, we explored the notion of diversity and observed different interpretations and understandings of diversity in the contexts mentioned above, encompassing the realms of biology/ecology, economics, and various fields within the social and computer sciences. Due to the diversity of these contexts, it is evident that a universal definition of diversity cannot be uniformly applied. Furthermore, the concept of diversity may carry different connotations depending on the specific field and context under consideration. For instance, in a workplace environment, diversity often refers to the presence of individuals from diverse backgrounds. On the other hand, in the field of economics, diversity may pertain to a more balanced distribution of wealth, with individuals sharing similar living conditions and income levels, thereby minimizing extremes of wealth disparity.

In an effort to establish general characteristics that can be universally applicable and foster a shared understanding of diversity across all domains, researchers have presented common or overarching properties of diversity in their studies [300]. These properties serve as a framework for assessing and comparing the levels of diversity across different systems. These shared properties or dimensions are applicable to all contexts and can be employed to measure the levels of diversity within a given system, regardless of the specific field. In the subsequent section, we elaborate on popular and widely used approaches for measuring diversity, providing insight into the methodologies employed for this purpose.

3.1.4 Diversity Measurement

As noted above, the concept of diversity is pervasive and relevant in almost every field. However, each field brings its own unique interpretation of diversity and focuses on specific aspects of interest. While the three established dimensions of diversity remain fundamental characteristics regardless of the context in which the concept is applied, different fields may emphasize one dimension over the others.

Consequently, a multitude of diversity measurements has emerged from various fields, tailored to address the specific requirements of those domains. Interestingly, some of these measurements have extended beyond their initial domains of origin, gaining recognition as widely used diversity measures. Studies such as [189] and [301] have already undertaken comparisons of existing diversity measures and explored their efficacy across different contexts.

In this context, we introduce the most prevalent diversity measures highlighted in these studies. We elaborate on their intended purposes and the contexts in which they are applied. This exploration aims to facilitate the selection of the most suitable diversity measures for assessing diversity within the Wikidata framework.

3.1.4.1 Types of Diversity Measures

It has been decades since diversity came into focus and used in different contexts. Like every other concept, it has gone through refinements and today we have numerous diversity measures which range from simple ones to more complex ones. As mentioned earlier, in the early days due to the closeness of the terms diversity and variety, diversity was considered to be a simple concept and a synonym to the term 'variety' also called *richness*, especially in the fields of biology and ecology [190]. In the same manner, in fields like economics, it simply focused on balance to measure diversity [95]. Later on, the concept was extended beyond just one aspect and included evenness or balance besides variety. This is called *heterogeneity* [101] and is represented through a single value called *diversity index*. Heterogeneity and evenness measures are typically categorized into two groups: parametric and non-parametric measures [182]. Parametric measures are applied to data that adhere to specific distribution assumptions. However, these measures are less commonly used, as diversity measurements often deal with samples where making assumptions about the data is not feasible. Consequently, the measures we discuss below fall into the category of non-parametric measures.

Considering the myriad existing approaches to measuring diversity across various fields, an exhaustive review of all these measures is time-consuming and falls beyond the scope of this study. More importantly, previous attempts to explore existing diversity measures and examine their behavior in certain cases have concluded that all of these diversity measures give similar results with very little difference [301, 189]. For this reason, in order to provide a glance into the existing diversity measures, we here narrow our scope and introduce the measures of diversity from the studies [189] and [301]) which are not limited to the scope of one field but are generally applicable in various contexts.

Here, these measurements are organized into categories of single-concept and dual-concept diversity measures considering the number of dimensions they take into account while measuring diversity. Table 3.1 presents the formulas associated with the aforementioned diversity measurement approaches, offering a quick comparative overview of these measures.

To provide a more comprehensive understanding of these measures and how they operate, later in this section, we furnish an example using a small sample that represents the diversity of editors across four Wikidata items (as shown in Table 3.2).

I. Single-concept Diversity Measures: Following are the measures that are called single-concept diversity measures because they only focus on one dimension/ property of the diversity, i.e., variety, balance, or disparity. As mentioned earlier, our understanding of the three main properties of diversity is as follows:

- Variety is the count of the number of categories into which the elements of the system are apportioned.

Table 3.1: A comparative view of diversity measurement formulas. The notions used in the formulas: N is the category count, p_i the proportion of a category (e.g., class) in an entity (e.g., domain), n_i is the number of elements in the category i, d_{ij} is a disparity between i and j, N_{max} is the total number of elements belonging to the most abundant categories. A is the area between the Lorenz curve (the actual distribution of income or wealth) and the line of perfect equality (a diagonal line representing perfect equality), and B is the area under the line of perfect equality.

Diversity Measure	**Formula**	**Source**
Single-concept		
Richness	$\sum_i (p_i^0)$	MacArthur 1965
Disparity	$\sum_{ij} d_{ij}/N$	Stirling 1998
Gini-coefficient	$A/(A+B)$	Gini 1936
Dual-concept		
Berger-Parker Index	N_{max}/N	Berger and Parker 1970
Shannon Entropy	$-\sum_i (p_i log p_i)$	Shannon 1948
Brillouin	$(lnN! - \sum lnn_i!)/N$	Brillouin 1956
Simpson's Index	$\sum_i n_i(n_i-1)/N(N-1)$	Simpson 1949
HHI	$\sum_i (p_i^2)$	Rhoades 1993
Gini-Simpson Index	$1 - \sum_i (p_i^2)$	Gini 1912
McIntosh	$(N - \sqrt{\sum_i n_i^2})/(N - \sqrt{N})$	McIntosh 1967
Rao-Stirling	$\sum_{ij} (d_{ij})^{\alpha} (p_i p_j)^{\beta}$	Stirling 2007

- Evenness is counting the number of the elements of each category in relation to each other and measures the balance of the elements' distribution across categories.

- Disparity is the measurement of the degree of similarity or difference of the system's elements

Here, we introduce the most popular measures that use one single dimension of diversity in consideration.

- Richness/ Variety. Richness is the simplest form of measuring diversity, also called variety. This is a popular measure in the field of biodiversity. Biodiversity aims to determine the number of different species present in a particular area, and the greater the number of species in that specific area, the higher the level of diversity is considered. Since richness is sometimes considered an alternative to diversity in the fields of biology and ecology (i.e., biodiversity), there exist different ways to measure species richness in this field due to the fact that not all species are known or possible to identify in all occasions in addition to dealing with samples. For instance, if we would like to know the richness of a species S in an area where considering the total number of individuals, we could find out the richness either by simply counting the species number (i.e., Richness = S) [180], or through other defined approaches explained in [181], [219], or [192]. Since other approaches are used in different situations to tackle specific issues in the context of species and their ecosystem and our focus is not on the species, we only consider richness or variety as S which presents the total number of elements.

- Disparity. Another important dimension of diversity is disparity or dissimilarity between categories or elements, although, it has been often neglected or relatively less used than variety and balance in the measurement of diversity [300]. Disparity focus is how different or dissimilar the categories of a system are. The disparity is closely related to the approaches measuring similarity and distance. Euclidean Distance is one of the commonly used distance metrics and measures the straight-line distance between two points in a multi-dimensional space [250].

- Balance: Gini-coefficient. This measure was originally developed in the field of economics to quantify income or wealth inequality within a population [94]. The coefficient is calculated by comparing the cumulative distribution of income or wealth to a hypothetical perfectly equal distribution. In other words, it is calculated and graphically represented using the Lorenz curve graph [58] using the $A/(A+B)$ formula where A represents the area between the Lorenz curve[7] and the line of perfect equality (an ideal diagonal line symbolizing complete equality), and B is the area under the line of perfect equality. The Gini coefficient is then expressed as a decimal value between 0 and 1 where 0 represents perfect equality (every individual has an equal share) and 1 indicates maximum inequality (one individual possesses all the income or wealth)[8]. The Gini index is a perfect indicator of balance and doesn't depend on variety [261].

Overall, single-concept diversity measures are used when only one property or dimension of diversity is concerned. Next, we look at the dual-concept measures.

II. Dual-concept (Heterogeneity) Diversity Measures: These measures take more than one dimension of diversity into account, e.g., a combination of variety and balance, hence they are called heterogeneity measures. Although single-concept diversity measures exist, their usage is rather bound to specific contexts where only one dimension of diversity is the focus. However, as discussed earlier, the diversity of a system is generally calculated by considering a mixture of the above dimensions.

Here, we observe the combination of variety and balance as the prevailing diversity measurement approach, often referred to as dual-concept measures. While these measures incorporate both properties into their diversity calculations, they may yield distinct results depending on their inclination towards one of these properties. The example in Table 3.2 illustrates these measures and their behavior in various scenarios. The subsequent section explains the heterogeneity measures in detail.

- Berger-Parker Index. This index is one of the most easily calculated measures of diversity through the proportional abundance of the most abundant category [21]. In other words, the Berger-Parker Index counts the number of individuals for each species in the community and determines the species with the highest number of individuals N_{max}. Then, sums up the total number of individuals across all species N and finally, divides the number of individuals in the most abundant species by the

[7] Lorenz curve exhibits the actual distribution of data (e.g., income) that is achieved after the following steps: 1) Arranging data from lowest to highest, 2) calculating the cumulative percentage of the population and the cumulative percentage of total income at each data point, 3) plotting the cumulative share of the population on the x-axis and the cumulative share of income on the y-axis to create the Lorenz curve.

[8] For example, the map of wealth distribution among countries in 2019 is available at: https://en.wikipedia.org/wiki/List_of_countries_by_wealth_equality [Accessed 23.09.2020]

total number of individuals in the community (cf. Table 3.1). The Berger-Parker Index ranges from 0 to 1, where 0 indicates perfect evenness (no dominance), and 1 represents complete dominance by a single species. The higher the Berger-Parker Index, the greater the dominance of the most abundant species in the community.

It is considered a dominance measure and gives importance to the most dominant category. It behaves differently when dealing with data with a higher and smaller number of categories [182].

- Shannon Entropy. This is one of the old measures that dates back to 1948 and is also known as Shannon's diversity index and Shannon–Wiener index. It is also popular as entropy and has been the source of inspiration for many entropy-based indices [257]. This measure originated from information theory by Claude Shannon for the purpose of quantifying the degree of uncertainty of the information content [287]. It is representing the uncertainty level about determining the category of an element, thus it is also used to measure diversity. It rates a system as highly diverse if the elements of the system categories are evenly distributed and an unknown element has an equal chance of belonging to any of the categories, hence leading to high prediction uncertainty. Conversely, a system will be considered less diverse when certain categories dominate other categories of the system and lower uncertainty levels of predicting the category of an unknown element [287].

The formula for Shannon entropy includes the sum of the probability of each possible outcome (p_i) multiplied by the log (base 2) of that probability:

$$-\sum_i (p_i log p_i)$$

This logarithmic transformation allows for a more intuitive understanding of entropy, as it emphasizes the relative uncertainty or surprise associated with each possible outcome [287].

It is called a heterogeneity measure since it considers both the number of categories present and their proportional representation. It allows us to compare different datasets or subsets based on the distribution of categories and provides a standardized measure to quantify diversity across various contexts.

This is one of the robust and most commonly used diversity measurements available. It provides reliable outcomes when dealing with two extreme scenarios: highly evenly distributed and highly unevenly distributed systems. However, its accuracy is compromised when values lie somewhere in between these extremes, resulting in a skewed representation [233].

- Brillouin Index. The Brillouin diversity index often provides identical estimates of diversity like the Shannon index, thus, both indices are very similar. The difference between the two indices comes from the fact that the Brillouin index deals with collections (or a non-random sample) and the Shannon index is used with random samples [181]. The Brillouin index is measured using the category count (N) and the number of elements in each category (n_i) based on the natural logarithm (ln) [31]:

$$(lnN! - \sum lnn_i!)/N$$

As can be seen, Brillouin also is based on logarithms like Shannon Entropy but the focus of each approach generates more accurate results in their defined scopes.

- Simpsons' Index. This index is also called Simpson's Diversity Index. It is also one of the earliest indexes of diversity (i.e., from the year 1949) which was introduced by Simpson in the field of ecology as a means to measure the probability of two random individuals from a community belonging to the same category, or concentration of individuals classified into groups [291].

It measures both balance and variety, however, it has less sensitivity towards variety/ richness and gives more weight to the most abundant categories. Thus, it is useful in fields like ecology, sociology, or economics to identify if certain categories are more prevalent than others. This index is calculated by considering the number of elements in a category (n_i) and the category count *(N)* in the calculation of diversity as:

$$\sum_i n_i(n_i - 1)/N(N - 1)$$

Hence, it's called the count-based formula. The result shows higher diversity with lower values, and conversely, lower diversity with higher values.

Other than the count-based formula explained above, there is also a probability-based formula of Simpson's Index that calculates the index by summing the squared proportions (p_i) of individuals in each category:

$$(\sum_i (p_i^2))$$

In other words, it shows the sum of the probabilities, for each *i*, that two randomly selected items will both be categorized as belonging to category *i*, assuming an equal probability of selecting any individual within the community [291].

Additionally, the *Herfindahl-Hirschman Index (HHI)*, which is a diversity index used in the field of economics to measure market concentration [256] is also calculated using $(\sum_i (p_i^2))$. Hence, here higher values indicate higher concentration.

Thus, Simpson's Index is not only one of 'the most meaningful and robust diversity measures' available [182], but is also the core of many other diversity measures in different disciplines, as explained below.

- Gini-Simpson Index. This index is based on Simpson's Index. It was proposed to compute the probability of two random elements taken from a sample [95]. In other words, it shows how probable it is that two random elements taken from a sample belong to the same category. Since this index is a measure of balance and balance is inverse of concentration, it is calculated by subtracting the concentration index from 1:

$$1 - (\sum_i (p_i^2)$$

The resulting value ranges from 0 to 1, where a value of 0 indicates complete dominance by a single element, while a value of 1 represents maximum diversity where all elements have equal abundances.

This concept in the context of social sciences is called *Blau Index* and is used to quantify the degree to which individuals are concentrated in certain occupations based on their demographic characteristics, such as gender, race, or ethnicity [26].

- McIntosh Diversity Index. This is one of the less-used diversity indices that is also based on Simpson's Index. If we consider a community of species as a point in an S-dimensional hypervolume, the Euclidean distance between the community and its origin is calculated using this diversity measure. The formula measuring the McIntosh index applies square root on the category count (N) and the number of elements in each category (n_i) and is:

$$(N - \sqrt{\sum_i n_i^2})/(N - \sqrt{N})$$

Rao-Stirling's Index. This index includes all of the three properties/dimensions of diversity, i.e., variety, balance, and disparity, and is inspired by the 'triple concept' diversity measure of [141]. It originated in the field of economics as the sum of pairwise disparities, adjusted based on the individual system elements' contributions (D) [300] and is illustrated as:

$$D = \sum_{ij(i \neq j)} d_{ij} p_i p_j$$

This index was originally called the Rao Index and was, later on, updated and called the Rao-Stirling Index after providing the option of assigning weights α and β considering the importance of disparity or balance. Rao-Stirling Index was introduced as a result of a general framework for analyzing diversity in science, technology, and society [301]. Hence, having p_i and p_j as the proportions of a category (e.g., class) in an entity (e.g., domain) and d_{ij} as the disparity between categories i and j, the formula calculating the Rao-Stirling Index is:

$$\sum_{ij(i \neq j)} (d_{ij})^{\alpha} (p_i p_j)^{\beta}$$

When all of the properties of a system are equally important ($\alpha = \beta = 1$), then Rao Index and Rao-Stirling Indexes are equivalent. Rao-Stirling is preferred when some properties are given more importance than others.

3.1.4.2 Implications for Diversity Measures in the Wikidata Context

Diversity is a general term that is interpreted according to the context of its usage and can be measured using the three general properties which are variety, balance, and disparity. In the Wikidata context, we are interested in the variety of topics, languages, sources, and viewpoints that come from all corners of the world and are represented in a balanced manner. In other words, the data in Wikidata deals with classes, items, and statements and we can consider diversity in Wikidata as variety, disparity, and balance of data in Wikidata.

Multiple diversity measures exist that have emerged from different fields to answer context-related issues and focus on different dimensions of diversity. Diversity measures are categorized as single-concept and dual-concept based on the number of dimensions or diversity properties they take into account when measuring the diversity levels of a system.

Single-concept diversity measures can be used in the Wikidata context if only one dimension of diversity is concerned. For example, when comparing the number

Table 3.2: An example of Wikidata editor diversity to show the usage of diversity measures. Here we have four Wikidata items with their respective number of editors from different countries. Items A and B are similar except that Item B has a larger sample, Item C has been edited by editors from more countries, and Item D has a more balanced number of editors. (Note: Higher values show better diversity levels, except for HHI where a higher value shows more concentration.)

Editors Origin	**Item A**	**Item B**	**Item C**	**Item D**
Brazil	0	0	3	0
France	20	40	20	12
Germany	24	48	24	13
India	3	6	3	14
Italy	7	14	7	15
Tunisia	0	0	1	0
Richness/ Variety:	4	4	6	4
Disparity:	186.510	186.510	482.494	186.510
Gini-coefficient:	0.35	0.35	0.49	0.05
Berger-Parker Index:	2.250	2.250	2.417	3.60
Shannon Entropy:	1.15	1.15	1.36	1.38
Brillouin Index:	1.05	1.09	1.23	1.27
Simpson's Index:	0.658	0.651	0.702	0.762
HHI:	0.355	0.355	0.310	0.252
Gini-Simpson Index:	0.645	0.645	0.690	0.748
Rao-Stirling:	7.738	7.738	9.662	11.61

of classes or items across various Wikidata domains, or examining the diversity of properties utilized in Wikidata items, focusing solely on variety might suffice. Similarly, if the objective centers on analyzing the distribution of items within classes or the distribution of properties/statements across Wikidata items—regardless of the variety of items or properties (refer to Table 4.2)—a balance measure can prove useful. On the other hand, heterogeneity or dual-concept measures cover more than one dimension. The dual-concept diversity measures focusing on the dimensions of variety and balance are widely used and recently a third dimension (i.e., disparity) has also been added to the measurement of a diversity index.

As mentioned before, many of the dual-concept diversity measures have shared origins and were extended to tackle specific issues in the contexts they are used. Among them, Simpson's Index and Shannon Entropy are the original measures widely used. Studies exist that have compared these measures and concluded that all of the diversity measures produce identical results. For instance, a study by Mcdonald and Dimmick has compared twelve dual-concept diversity measures which take the dimensions of variety and balance. The authors have applied these diversity measures to a dataset containing a time series of thirty years of prime-time network radio programs classified by program type. The authors suggest that nearly all of these measures are good indicators of diversity and provide very similar results, thus, using one measure does not make a huge difference over using any other measure [189]. Another more recent study is by Morris et al. that compared six diversity indexes on a dataset in the field of biodiversity. They confirm that there is no single diversity index that is superior to others, but each measure can provide more accurate results when used in its defined criteria. The study also suggests using more than one measure for a better understanding and more considerable results [199]. Hence, we provide a comparative example of diversity measures in Table 3.2 to be able to apply these measures to an example in the Wikidata context and find the appropriate measures. So, our aim for a detailed description of these measures is not limited to providing a basic understanding and description of these diversity measures and how they work, but also a means to explore and pick the more suitable ones in the Wikidata context.

The mentioned definitions and formulas don't seem sufficient for a comparative view of these measures, thus, we apply these measures to four different datasets and use their results in the comparison as well. We apply these measures to an example of four Wikidata items (A, B, C, and D) that are edited by different numbers of editors from different backgrounds (countries of origin) and in different amounts. In this example, datasets for *Item A* and *B* exhibit similar levels of variety (in terms of countries of origin) and balance (regarding the distribution of edits across these countries of origin). The primary distinction between these two datasets lies in their sizes, with *Item B* being twice the size of *Item A* and boasting the largest number of editors among all. Moving to *Item C*, we note the inclusion of editors from two additional countries, resulting in a higher variety compared to all other datasets. Examining *Item D*, we observe a more equitable distribution of editors across countries of origin in contrast to the other three datasets, where editors from France and Germany dominate in number.

Our findings show that *Item C* has a higher variety of employee origins, and thus is considered more diverse from the richness angle as richness measurement is reliant on variety. Looking at disparity[9], we see that *Item C* is more different from items *A, B* and *D*. The findings indicate that disparity has shown more sensitivity towards variety and changes in balance have not affected the disparity level as can be seen in Items *A, B* and *D*. Additionally, we can see that Gini-coefficient has the highest value in *Item D* but the lowest in *Item C* despite *Item C* having the highest variety of editors, while, based on variety and disparity measures *Item C* is considered the most diverse.

Our example could also reveal that only Brillouin and Simpson's Indexes have shown a slight change to the altered sample size, while, all other measures treated both sample sizes as equal. This confirms that the Brillion Index was developed to deal with collections, and thus is sensitive to the size of the data. Additionally, our example gives an overview of these measures comparatively and shows that most of these measures are more sensitive toward dominant categories rather than rare ones. In other words, diversity levels will change only if a visible number of editors from different origins are included. Thus, *Item D* which has a higher balance is considered more diverse than *Item C* which has a higher variety.

We observe that the outcomes generated by these measures are identical. According to most of these metrics, *Item D* emerges as the most diverse entity, while in terms of richness and disparity measures, *Item C* showcases higher diversity than the other two. Conversely, all these metrics indicate that items *A* and *B* are the least diverse. This comparison of diversity measures confirms the fact that these metrics yield similar outcomes, with the choice of one over the other not significantly impacting the results. Thus, if we had employed the Brillion index, Berger-Parker

[9] Disparity is measured using Euclidean distance.

index, Simpson's index, or any other heterogeneity measures, we would still identify *Item D* as the most diverse entry. Consequently, for general-purpose contexts such as measuring diversity within Wikidata, any of the aforementioned measures can be utilized. Among them, Simpson's index and Shannon entropy are particularly versatile diversity metrics [189]. Additionally, the relatively recent Rao-Stirling approach considers disparity when assessing diversity. These three metrics—Simpson's index, Shannon entropy, and Rao-Stirling—can be deemed comprehensive diversity measures applicable across various disciplines.

Moreover, several of the aforementioned diversity measures are derived from Simpson's Index, which is esteemed as a robust and widely accepted metric. In addition, Shannon Entropy serves as a versatile and widely recognized diversity measure across various disciplines. It distinguishes itself from other measures by employing logarithms in the diversity assessment. As recommended for a more comprehensive understanding of results [199], we opt to include Shannon Entropy alongside Simpson's Index as a candidate for measuring diversity in Wikidata. This choice allows us to approach diversity quantification from a distinct perspective. We prefer Shannon Entropy over the Brillouin Index due to the random sample nature of our data. Furthermore, Stirling introduced a novel index that encompasses all three dimensions of diversity. We include this index in our list of selected diversity measures to capture the full spectrum of diversity attributes within Wikidata. However, within the realm of single-concept diversity measures, the Gini-coefficient stands out as a widely used measure for assessing balance in Wikipedia research on diversity (refer to Table 3.3 on page 66). Given its applicability in a context highly similar to Wikidata, we consider the Gini coefficient a valuable choice in Wikidata when focusing solely on balance.

3.1.5 Summary of Diversity Concept

Diversity is a pervasive concept that holds significance across various domains and situations. In the case of Wikidata, diversity is considered an important means of achieving its overarching goal, yet there is a lack of research assessing the extent to which Wikidata has achieved this goal in its first decade. Although designed with diversity in mind, no formal definition or concept of diversity specific to Wikidata has been established. Therefore, this section delves into the fundamental aspects of the diversity concept, aiming to enhance our comprehension and establish the basis for our proposed diversity concept within the Wikidata context.

While diversity is a widely recognized concept, a universally applicable definition remains elusive. However, there are common properties of diversity, such as variety, balance, and disparity, that can be employed to assess diversity levels in any system, irrespective of its specific context. These properties offer a shared understanding of diversity that cuts across different fields and contexts.

Diversity in a system is determined by the presence of various categories and elements. When a system includes a greater number of elements from a wide range of categories in a balanced manner, its diversity levels are considered higher. Numerous measurement approaches have been developed to quantify the diversity of a system, originating from different fields of study. Some approaches focus on a single dimension, i.e., variety, balance, or disparity, and are known as single-concept measures. Others take into account multiple dimensions and are referred to as dual-concept or heterogeneity diversity measures. Most heterogeneity measures consider the dimensions of variety and balance. Interestingly, a comparative analysis of these measures confirms that they often yield similar results probably due to the fact that many of these approaches share a common foundation and are variations of the same fundamental formula, adapted to address specific scenarios.

Next, we delve into the concept of diversity within the context of Wikidata, aiming to identify the specific aspects that need to be measured and the approaches to be employed. Building upon the knowledge gained in this chapter, we will explore the characteristics of diversity that are relevant to Wikidata, enabling us to determine the key elements to measure in order to assess its diversity levels. By aligning our understanding with the unique requirements and structure of Wikidata, we can develop a comprehensive framework for measuring diversity within this KB.

3.2 Concept for Measuring Diversity in Wikidata

As mentioned before, diversity is a pervasive concept that holds significance across various domains and situations. In the case of Wikidata, diversity is considered a means to achieve its overarching goal. However, there is a lack of research assessing the extent to which Wikidata has achieved this goal in its first decade. Although designed with diversity in mind, no formal definition or concept of diversity specific to Wikidata has been established. With a solid grasp of the diversity concept, its key dimensions, and established measurement approaches, our focus now shifts towards applying this knowledge to determine which aspects to measure in Wikidata, serving as indicators of its diversity levels. Currently, the only reference to diversity in Wikidata is in the design principle of plurality. However, the challenge lies in establishing a method for quantifying plurality, which requires further investigation into how plurality is defined within the concept of diversity specifically in the context of Wikidata.

Hence, before applying any measures, it is crucial to have a clear understanding of what needs to be measured in a KB. Given the existing research gap on the topic of diversity in Wikidata, we recognize the need to develop a diversity measurement concept specifically tailored for Wikidata. Therefore, in the upcoming section, we start by defining diversity within the context of a KB and identifying the features that can be measured using these measures. This will enable us to provide an accurate assessment of the diversity landscape in Wikidata.

Therefore, we will now direct our attention to examining the concept of diversity within the context of a KB, drawing insights from diversity in the Wikipedia context and placing particular emphasis on Wikidata. We will explore approaches to gain a comprehensive understanding of the current diversity status within Wikidata, aiming to uncover valuable insights and findings.

3.2.1 Diversity in a Knowledge Base Context

In this setting, our focus narrows to diversity within the realm of knowledge bases (KBs), primarily because Wikidata operates as a KB. In a KB, information is typically structured into discrete data units, such as articles in Wikipedia or items in Wikidata. These units can be organized into topical domains, enabling the capture of diversity [97]. Therefore, in a KB context, diversity is generally assessed by examining the variety of topic domains it encompasses, as well as the balance or concentration of data units within these categories.

Furthermore, KBs host user communities that often hail from diverse backgrounds, utilizing the KB for purposes ranging from data editing to data consumption. This presents an opportunity to evaluate the diversity of editors or the diverse ways in which data is consumed.

In essence, in a KB like Wikidata, diversity extends beyond the data itself to encompass the individuals who contribute this data or knowledge—a concept referred to as *knowledge diversity*. Fundamentally, data within a KB follows a cyclic trajectory: editors contribute data, data finds a home within the KB, and consumers utilize this data. Bearing this cycle in mind, diversity within a KB context can be defined. Knowledge diversity encompasses both data diversity and user diversity, as illustrated in Figure 3.3.

Despite diversity being a known topic for decades, it is not given much attention in the context of KBs. Giunchiglia et al. introduce the concept of a diversity-aware KB to overcome one of the barriers towards the use and success of semantics which is lack of background knowledge [97]. The study proposes the creation of an extensible diversity-aware KB aimed at capturing the diverse range of background knowledge. This approach is based on the faceted methodology of library science, which revolves around domain and facet concepts. Domains offer a comprehensive perspective of the entire field of knowledge, while facets provide a detailed analysis of each component within a domain [97]. Hence, the data in this diversity-aware KB needs to be stored in domains that can be divided into multiple facets or classes, with each facet/ class encompassing a distinct aspect of the domain [97]. The study also compares the existing KBs, YAGO[10], CYC[11], OpenCyc[12], SUMO[13], MILO[14], DBpedia[15], and Freebase[16] with their proposed diversity-aware KB in terms of their support for diversity. The authors find that none of the mentioned KBs have the required support for diversity better than their proposed one. This study does not include Wikidata because it was conducted before the launch of Wikidata.

Similar to other KBs, diversity as a research focus has not been given enough attention in Wikidata. However, there exist a number of studies on diversity in Wikipedia. Wikipedia was launched years before Wikidata and is considered the sister project of Wikidata. Due to the existence of similarities between both projects, we use the research on Wikipedia as a starting point for understanding diversity in Wikidata.

[10] "Yet Another Great Ontology (YAGO)is a large KB with general knowledge about people, cities, countries, movies, and organizations." https://yago-knowledge.org

[11] Cyc is an artificial intelligence project that aims to compile an extensive ontology and KB of common sense knowledge, aiming to empower AI applications with the ability to engage in human-like reasoning. https://cyc.com

[12] A subset and open source version of Cyc [187].

[13] SUMO (Suggested Upper Merged Ontology) is a formal public ontology owned by IEEE. https://www.ontologyportal.org

[14] MILO (MId-Level Ontology) is the extension of SUMO. https://github.com/ontologyportal/sumo/blob/master/Mid-level-ontology.kif

[15] "DBpedia is a crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects." https://www.dbpedia.org/

[16] Freebase is a graph database that has been collaboratively constructed to organize and structure human knowledge. [224]

Table 3.3: Wikipedia research papers on diversity with the aspect of diversity being explored and the diversity measurements used.

Paper	**Diversity focus**	**Measure Used**
Arazy and Nov [12]	Measure the impact of coordination and contributor inequality on content quality considering the inequality at a local level (i.e., articles) and global (i.e., overall Wikipedia).	Gini-coefficient
Tsikerdekis [318]	Experience diversity and implicit coordination with their effect on content quality improvement.	Gini-coefficient
Zhang et al. [352]	Measure the impact of editors tenure diversity on article quality	Gini-coefficient
Ren and Yan [255]	Contribution diversity of editors in Wikipedia articles and its' effect on performance and article quality.	Coefficient of variation
Robert and Romero [259]	Measure the effects of group size and group diversity on crowd performance.	Gini-coefficient
Kittur and Kraut [154]	Measure inequality in group structure in Wikipedia	Gini-coefficient
Sydow et al. [307]	The effect of editor and team diversity on the quality of virtual cooperative work in Wikipedia	Shannon entropy
Flöck and Rodchenko [76]	Calculate word concentration and inequality authorship	Gini-coefficient
Halavais and Lackaff [111]	Explore the topical coverage diversity of Wikipedia	-
Flöck et al. [77]	Discuss the effect of diversity on Wikipedia content quality	-

3.2.1.1 Diversity in Wikipedia

Wikipedia is an online encyclopedia where topics from numerous domains are contributed by a community of contributors with diverse backgrounds. There are several studies on Wikipedia from a diversity perspective that we explore in the following sections and provide an overview of in Table 3.3. Diversity within Wikipedia primarily revolves around the community and the impact of editors' diversity on article quality with the Gini coefficient being the most commonly employed diversity measure in this area.

Arazy and Nov in their study investigate the contribution and coordination inequality (i.e., the balance of user roles) in Wikipedia and confirm the effects of inequality on article quality [12]. Tsikerdekis studies experience diversity and implicit coordination with their effect on content quality improvement in Wikipedia [318]. Zhang et al. measure the impact of tenure diversity on article quality in Wikipedia [352]. Ren and Yan investigate crowd diversity (balance of user contribution) in Wikipedia articles and diversity effect on performance and article quality [255].Robert and Romero measure the effects of group size and group diversity on crowd performance [259]. Kittur and Kraut explore how communication occurs in online platforms and use diversity to measure inequality in group structure in Wikipedia [154]. Sydow et al. look into the effect of editor and team diversity on the quality of virtual cooperative work in Wikipedia [307].

Few studies exist that use diversity from the data or the article perspective in Wikipedia. Flöck and Rodchenko calculate word concentration and use Gini-coefficient as an inequality measure of authorship [76]. Halavais and Lackaff explore the topical diversity of Wikipedia in their study "An Analysis of Topical Coverage of Wikipedia". The authors compare a sample of 500 English Wikipedia articles with the bibliographic database of Bowker's Books In Print and three field-specific encyclopedias to find out the degree of Wikipedia's content diversity [111]. Flöck et al. discuss and analyze the effect of diversity on Wikipedia content quality through the survey of existing research and identifies future directions for research and development in this area. The authors also present a diversity-minded content management approach within Wikipedia which would give the community matrices and indicators to identify bias and knowledge imbalance across Wikipedia articles. Their proposed approach is implemented in the form of Render. Render is a tool, as part of the Render project[17], developed to measure diversity in Wikipedia. The tool is based on diversity aspects of thematic coverage, timeliness, and neutrality in Wikipedia [11]. Wikipedia Diversity Observatory is another project to overcome the diversity gap of Wikipedia language editions [195]. This project displays the concepts that are not present or shared across languages through dashboards with visualizations and tools.

Our journey of the Wikipedia research on diversity shows that the focus of Wikipedia diversity research evolves around the diversity of community and data which confirms the impact of community diversity on data in a collaboratively developed KB like Wikipedia and Wikidata. The approaches used to measure Wikipedia community diversity are the Gini coefficient and Shannon entropy, which are explained in Section 3.1.4. Since Wikipedia's research on diversity is more focused on the community and is concerned about the balance or inequality of edits on articles, the Gini coefficient has been vastly used. However, we see no common definition of diversity in these papers and they focus on balance, concentration/ inequality.

Next, we look at the existing research on Wikidata from a diversity perspective.

3.2.1.2 Diversity in Wikidata

Here, we present the existing Wikidata literature with any indication or traces of research on diversity that could help us build upon it. We have earlier seen that diversity in Wikipedia mostly means diversity of editors, however, diversity in Wikidata, which is more focused on data, is not much investigated. To the best of our knowledge, no research has studied diversity as a main topic in Wikidata, nevertheless, there exist studies on gender bias in Wikidata that refer to diversity gaps in the gender data in Wikidata. Shaik et al. examine the race and citizenship bias in Wikidata with a focus on people with STEM (Science, Technology, Engineering, and Mathematics) backgrounds and computer scientists. The study finds an over-representation of white Western individuals from Europe and North America in comparison to all others in the globe [283]. Zhang and Terveen investigated the gender content gap in Wikidata to find out if the lower representation of women in Wikidata is due to the editors' bias or a reflection of the real-world data. The authors find that the most popular professions in Wikidata are male-dominant professions

[17] Render- Reflecting Knowledge Diversity available at: http://render-project.eu

(e.g., American football). They also concluded that Wikidata's representation of women is very close to their real-world representation [350].

Further, there are papers where diversity is not the focus of the study, however, they have used diversity measures to investigate a part of their research. Sarasua et al. study the editing behavior of human editors in Wikidata to predict power and standard users. The study measures the diversity of edit types using Shannon entropy [270]. Cuong and Müller-Birn, in their research on the applicability of sequence analysis methods on human user participation patterns in Wikidata, measure diversity of states or user roles using entropy [42]. As part of their study, Piscopo et al. in the investigation of Wikidata provenance analysis show that bots add less diverse references to Wikidata than humans [238].

We can observe that the existing literature does not prioritize the topic of diversity. One reason for the research gap in the area of diversity could be the assumption by researchers that the findings from diversity research in Wikipedia can be directly applied to Wikidata due to the close relationship between the two projects. For instance, research has revealed that Wikidata mirrors Wikipedia in displaying Western, male-oriented biases, with nearly four times as many statements about Western artists compared to non-Western artists, and an even more pronounced ratio of nine times as many statements about Western masterpieces as non-Western ones [3]. Nevertheless, despite the similarities between the two projects, the significant volume of edits made by bots differentiates the Wikidata community from Wikipedia. Consequently, further research is required to gain insights into how the introduction of bots influences the diversity and quality of edits.

In addition, despite gender diversity being an important issue, there exist many other issues in the context of a KB that also need to be addressed like the diversity of language, topic, or diversity of opinions that need to be addressed. Another reason for the diversity gap in Wikidata research, or in general in the KB context, could be that we don't see a proper definition or understanding of diversity from a research perspective. As mentioned before, diversity in a KB is a broad concept and could be looked at from either, user or data perspectives. Hence, we see a need first to provide a general concept of diversity in the Wikidata context and then search for any existing literature based on our defined diversity angles. In Wikipedia diversity is mostly looked up from a social science perspective, i.e., it focuses on the community of editors.

Additionally, we observe a close relationship between diversity and data quality topics in Wikipedia research. While diversity is not explicitly defined in the data quality frameworks [349], many studies on diversity in Wikipedia have suggested the influence of editor diversity on article quality (cf. Table 3.3). This underscores the significance of diversity within data in a KB context. In Wikidata, diversity is one of the design decisions, hence, diversity should have a higher impact on Wikidata's data and quality.

As mentioned above, Wikidata receives the majority of its edits from automated assistants, i.e. bots, and this distinguishes the Wikidata community from Wikipedia. Another issue that may affect diversity differently between the two communities is the structured nature of the data in Wikidata versus the text-based data of Wikipedia, especially since Wikidata implements diversity in the form of plurality. This raises the question of whether diversity in Wikidata should also be more focused on the community of editors, or whether it should have other aspects beyond that, which we want to address next.

3.2.2 The Diversity Conception in Wikidata

Despite being one of the essential aspects of Wikidata, there is currently no diversity model in place that provides an overview of the existing state and identifies gaps. Furthermore, the definition of diversity within the context of Wikidata still needs to be established, utilizing the knowledge diversity concept within a KB framework as mentioned earlier, which revolves around the dual perspectives of users and data. So far, our understanding of diversity in Wikidata has been based on the *plurality* design decision, which allows contradictory statements to coexist [326]. In this section, we delve into the concept of plurality and its potential relationship with the knowledge diversity concept in Wikidata. We aim to answer the question of whether plurality is indeed a true synonym for diversity in Wikidata. Following this discussion, we will explore diversity within the Wikidata context and proceed to assess the current state of diversity in Wikidata.

3.2.2.1 Is Plurality a Synonym for Diversity in Wikidata?

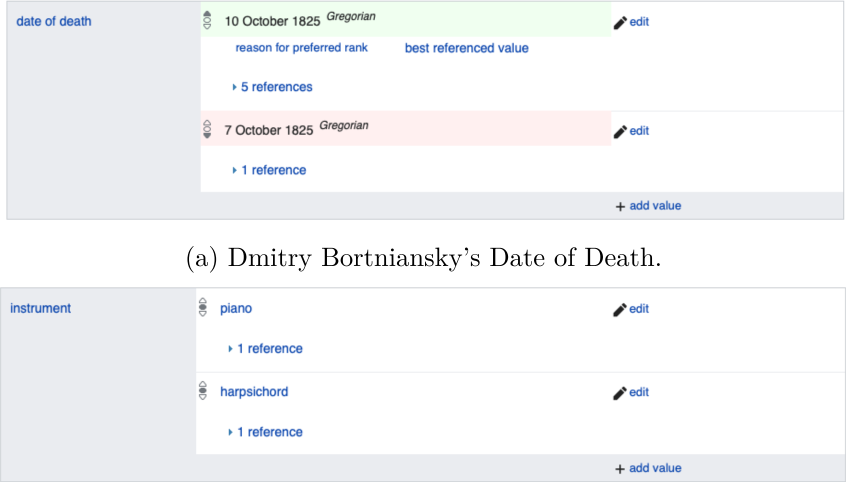
Wikidata supports diversity by nature by implementing it in the form of plurality. Plurality is used to reflect world data where not all data is globally agreed and there exist uncertain and disputed data [326]. This shows that Wikidata already has a mechanism to store and reflect this diversity from the data perspective. However, it becomes apparent that plurality is centered solely on data, whereas our definition of knowledge diversity in a KB context encompasses both user and data diversity. This distinction suggests that plurality does not serve as a genuine alternative to the diversity concept in Wikidata. Thus, further exploration of plurality is necessary to gain a comprehensive understanding of its implications within the diversity framework of Wikidata. Earlier we mentioned that plurality is a design principle that enables Wikidata to store multiple statements together, even if they are contradictory [326]. Nevertheless, we see a research gap in this area; there is no overview of the number of diverse statements in Wikidata, the items to which they belong, and, whether, they all show contradiction.

An example of a contradictory statement could be having more than one statement as the date of death. For instance, Dmytro Bortniansky[18] has two dates of death according to different sources, as can be seen in Figure 3.2a. Similarly, we observe two statements as the musical instruments he used, piano and harpsichord. However, these statements do not indicate a contradiction, but rather the variety of musical instruments this Russian composer could play (cf. Figure 3.2b). Further, the statements showing variety can be explained with an example of the item Q43 which belongs to the country Turkey[19]. In the continent's property, Turkey has two values of Asia and Europe which are not contradictory but show the fact that the country is a transcontinental country and is located in West Asia and East Europe. While most countries have only one value for their continent property, having more than one value here gives a more complete picture of this country's location which is

[18] Dmytro Bortniansky: https://www.wikidata.org/wiki/Q316505 [Accessed: 14.04.2021]

[19] Turkey: https://www.wikidata.org/wiki/Q43

not contradictory. These examples show that multiple values for the same property are not necessarily always contradictory, but could also show variety.



date of death

10 October 1825 Gregorian — edit

reason for preferred rank: best referenced value

5 references

7 October 1825 Gregorian — edit

1 reference

+ add value

(a) Dmitry Bortniansky's Date of Death.

instrument

piano — edit

1 reference

harpsichord — edit

1 reference

+ add value

(b) Music instruments Dmitry Bortniansky played.

Figure 3.2: The example which could show contradiction or variety of statements using the *Dmitry Bortniansky* (Q316505) item in Wikidata.

Keeping in mind that diversity is not only difference but also variety and balance, we could refer to plurality as part of diversity which is capable of showing disparity (i.e., contradictions) and variety at the item level. In addition, the plurality also provides the possibility of storing labels in different languages, various descriptions, and aliases; plurality here mainly shows variety. Plurality also allows multiple sources, however, sources are stored in references which are part of Wikidata statements. Thus, we could consider plurality in Wikidata as the coexistence of multiple statements, whether showing contradiction or variety. Further, earlier we saw that diversity is the property of any system that consists of categories (cf. 3.1.1), and we also know that Wikidata is a KB made of data in the form of categories of classes and items along the side of a contributor community. For this reason, we need to define diversity in the Wikidata context beyond the limits of the item and statements level (i.e., plurality) and consider further aspects of knowledge diversity which are domain/ class and user aspects. Hence, next present our proposed diversity concept for Wikidata with a broader view that can capture the topical coverage of Wikidata, a better understanding of the globally agreed or disputed data, and a glance at the contributors of this cumulative knowledge.

3.2.3 The Proposed Diversity Concept for Wikidata

We develop a concept for the measurement of diversity in Wikidata so that we can get an overview of the existing diversity status and spot the gaps. Only then it is possible to take steps towards bridging the gaps and take Wikidata towards becoming a world knowledge.

As mentioned earlier, Wikidata is edited collaboratively by a volunteer community. This means that the diversity of knowledge in Wikidata is directly dependent on the editors' knowledge and background diversity. The more editors from diverse backgrounds, the more diverse topics, statements, and sources are expected. In other words, knowledge diversity in the Wikidata context does not relate to data only, rather, this knowledge flows in a cycle; editors contribute data, data is stored in Wikidata, and consumers use this data (cf. Figure 3.3). Gaining an overview of diversity in Wikidata depends on several factors: a) exploring the diversity of those who contribute to this data (i.e., editors), b) assessing the variety of topic domains and the balance of data within those domains, and c) understanding how this data is used by various individuals (i.e., consumers) for different purposes. Following the knowledge diversity concept we define the Wikidata diversity measurement concept from the two main angles of user and data. *User diversity* contains editor and consumer diversity categories, while, *data diversity* consists of the data diversity in domain and item levels, see Figure 3.3.

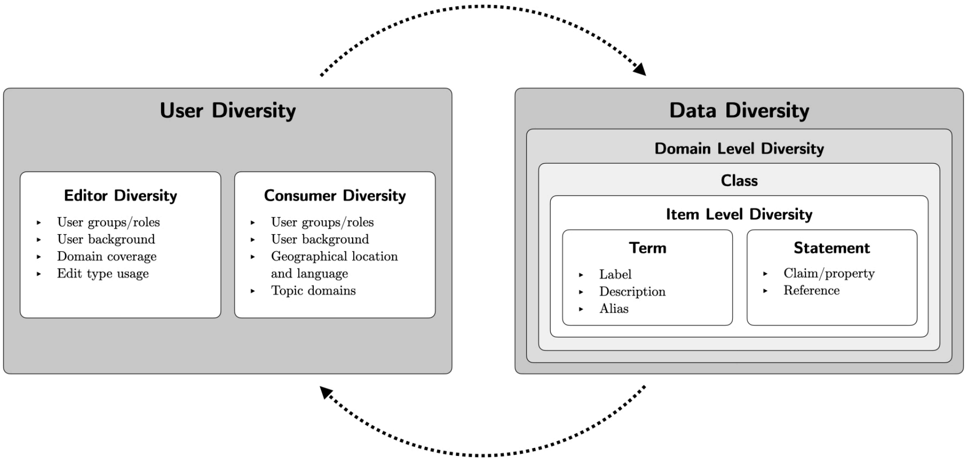


Figure 3.3: The proposed diversity concept for Wikidata.

Below, we provide a detailed explanation of our proposed concept and its components. Table 3.4 provides a summary of the proposed concept for the measurement of diversity in Wikidata. This concept serves as a guideline in exploring how close Wikidata is to the goal of serving the diverse population in the world.

3.2.3.1 Data Diversity

Data in Wikidata is already available in categories in the form of classes. Items are associated with their respective classes and each item has its own collection of statements. For example, the class Human(Q5)[20] in Wikidata has many instances/ items, one of which is Douglas Adams represented as Q42[21]. This item in turn has many multilingual labels and numerous statements in the form of property-value pairs which follow the data model defined for Wikidata items (cf. Section 2.1.3). Hence, Wikidata items are not single entities like Wikipedia articles, they consist of further parts, i.e., *term* (label, description, alias), *statement* (claim (property-value, qualifier, rank), reference), and *sitelink*. Although diversity is the property of a system as a whole, in Wikidata we could look at data diversity from different

[20] https://www.wikidata.org/wiki/Q5
[21] https://www.wikidata.org/wiki/Q42

Table 3.4: An overview of Wikidata Diversity Measurement Concept. Each section of user or data diversity angles can be measured using the metrics provided considering the diversity dimensions and the measurement approaches. (Note: In this table V = Variety, B = Balance, D = Disparity, Measure = M, Single-Concept = SC, and Dual-concept = DC. The metrics and diversity dimensions are not limited to the values provided here.)

	Section	**Metric**	**Dimension(s)**	**M**
Data Diversity	Domain/ Class	#Classes per Domain	V of topics	SC
		#Items per Class	B of topics	SC
	Item/ Statement	#Language Term	V, B & D of languages	DC
		#Properties per Item	B of items	SC
		#References per Item	V, B & D of sources	DC
		#References per Statement	V, B & D of sources	DC
		#Claims per Statement	V & D of data	DC
User Diversity	Editor	Background	V, B & D of background	DC
		User groups/ roles	V & B of groups / roles	DC
		Edit type usage	V & B of edit types	DC
		Domain contribution	V, B & D of domains	DC
	Consumer	Background	V, B & D of background	DC
		User groups/ roles	V & B of groups / roles	DC
		Consumption purpose	V & B of purpose	DC
		Queries usage	V & B of queries	DC
		Domain usage	V & B of domain	DC

angles. The advantage of this approach would be a clearer understanding of the concept of diversity, easier measurement, identification of gaps, and more efficient efforts to fill these gaps.

We follow the approach proposed for diversity-aware knowledge bases (KBs) [97], which is built on the concepts of domains and facets. Domains offer a comprehensive perspective on the entire field of knowledge, while facets provide a detailed analysis of each component within a domain.

At the domain level, diversity can be assessed by considering the topical coverage of each category or domain. This approach provides an overview of the variety of topical domains a KB encompasses by counting the number of domains it covers. By counting the number of classes within each domain and similarly the number of items in those classes, we gain insight into the diversity within these topical domains. This analysis allows us to determine whether these domains and their classes maintain balanced levels of data or if certain domains and classes receive more attention while others are overlooked.

In the same manner, we could look at the item level to measure the variety of language labels, the balance of property usage among items of the same class, and the variety of references used as source data in Wikidata. While measuring the diversity of labels and properties in items is straightforward by just counting the number of language labels and the number of properties per item of the same class, measuring the diversity of statements needs a defined mechanism of measurement. This is because Wikidata supports plurality and having multiple statements could show either variety or disparity. To achieve one of the benefits of diversity which is to bold disputed data over globally agreed data, we need to define a mechanism to differentiate between variety and disparity of statements. One possibility could be the identification of properties where multiple statements show variety and the properties that could show contradiction. For instance, the property child (P40)[22] shows variety if contains more than one value as in Barack Obama (Q76)[23]. Similarly, the property population (P1082)[24] could show multiple numbers in different time intervals which are not contradictory, as in item(Q64), i.e., Berlin[25]. On the other hand, multiple values for properties like date of birth(P569)[26] or place of birth (P19)[27] would show contradictions, because it is possible to have only one date of birth or place of birth.

At the item level, we could also look at how diverse data types are in the items of the classes. For example, we could verify how many of the items contain not only text but also other varieties of data types like image, audio, link, or geo-coordinates data. These different forms of data are stored in properties or claim levels. At the moment, the presence of images and audio is not common in all items. One reason could be the issue of the free licensing condition in Wikidata based on which the data in Wikidata is free to be used by anyone for any purpose. Since not all images and audio are available under this license, these data types are not present in every item. Further, not all items need data types like geo-coordinates. For example, items in the classes related to Protein, Fish, or Planet are not very related to this property, so they lack this data type. For this reason, the absence of this data type in the items of certain classes does not show low diversity, so further research is needed to define a proper approach regarding this issue.

In summary, to measure knowledge diversity from the data perspective, in other words, data diversity, we could perform this measurement on two levels. First, we could use the combination of variety (i.e., to show how many different domains, classes and items exist), balance (i.e., to show if all domains contain the same number of classes and items, or some domains are dominant and some are overlooked) and disparity (i.e., to show if the domains really cover different topics reflecting the world knowledge or they cover rather similar topics) to measure the diversity of classes and their items. Second, we could measure item-level diversity (i.e., variety, balance, and disparity) of language labels, sources, and statements in items which can also refer to plurality.

3.2.3.2 User Diversity

We have previously emphasized the significance of data as a dimension of diversity within the Wikidata model. However, it's important to acknowledge that this data is contributed by users in a collaborative manner. Given that editors are unlikely to adhere to a specific structured approach in their contributions and instead voluntarily contribute data driven by personal interests, we can anticipate that they will concentrate on their preferred topics, languages, viewpoints, and values. Conse-

[22] https://www.wikidata.org/wiki/Property:P40
[23] https://www.wikidata.org/wiki/Q76
[24] https://www.wikidata.org/wiki/Property:P1082
[25] https://www.wikidata.org/wiki/Q64
[26] https://www.wikidata.org/wiki/Property:P569
[27] https://www.wikidata.org/wiki/Property:P19

quently, a higher degree of user diversity will lead to a broader range of contributed topics and opinions, enhancing the overall diversity of the content.

Users in Wikidata could refer to both, editors and data consumers. According to the research, the contributions to Wikidata come from humans and bots. Similarly, this data is used by different consumers for different purposes, like Wikipedia the online encyclopedia, Amazon's voice assistant Alexa, and Euro-wings in-flight app, as some examples. We differentiate between users who edit Wikidata and those who utilize or query these data simply because they access Wikidata for different purposes. Additionally, editor diversity could impact consumer diversity and vice versa. For example, data contributions in diverse languages could allow more diverse consumers to use the data. Likewise, demand for a certain topic or language by consumers could encourage editors to contribute accordingly. As a result, user diversity could cause more diverse data. Thus, measuring user diversity from the two mentioned angles could help in the more precise identification of diversity gaps and deal with gaps in a lower level of complexity.

Looking at diversity from a user's perspective could have further two levels. The first one would be to shed light on how diverse Wikidata contributors are. User background information like age, country of origin, language, and many other personal attributes could be indicators of user diversity. The second one would be the activities these users perform in Wikidata. Measuring the diversity of user backgrounds along with their contribution/ consumption could show us how diverse the Wikidata community is. Nevertheless, in Wikidata, getting user background information is not straightforward and would need defined approaches for collecting this information[28]. In a Wikidata community survey by Wikimedia Foundation [53] for example, the participants were asked to fill out a questionnaire regarding their basic background information. In a study of analyzing editor's languages in Wikidata user languages were extracted from user pages with Babelbox[29] which store user spoken languages [143]. However, not all Wikimedia users enable Babelbox as in this study only 4,120 users had it enabled among the 19,333 active users of the 2,930,072 total registered users. In a Wikipedia research, article geo-coordinates were used to show the coverage of topics or articles on the world map [11]. Hence, user background information might need different approaches to be captured.

Having more users from a variety of backgrounds could be an indicator of more diverse contributions, however, only being diverse doesn't ensure these diverse users will all have a balanced participation in Wikidata. In addition, automation capabilities have introduced new editing patterns and have changed the contribution balance. Thus, for getting an insight into how diverse Wikidata is, we need to also look from the angle of how diverse data contribution/ consumption they make. This angle could look at the diversity of languages, edit types/ queries, and topical domains Wikidata users edit or access. In the following, we describe editor and consumer diversity in Wikidata.

[28] In MediaWiki every editor can enable a user page which can show personal information like languages spoken, user privileges, and roles, however, it is not mandatory.https://www.wikidata.org/wiki/Wikidata:Userboxes [Accessed: 29.04.22]

[29] A template for showing user languages. https://en.wikipedia.org/wiki/Wikipedia:Babel[Accessed: 29.04.22]

- Editor Diversity. Research has shown that Wikidata edits come from three levels of automation, a) manual edits through humans, b) semi-automated edits through tools, and c) automated edits through bots. Anonymous is another group of users who edit Wikidata unregistered and are identified by IP address only, hence, no further information on their level of automation is available. Since automation can visibly alter the editing speed and volume, Wikidata users are grouped based on the level of their automation into groups of bots, tools, humans, and anonymous.

In Wikipedia, where most of the edits come from humans, the diversity of editors is claimed to have an effect on the quality of the articles (cf. 3.2.1.1). In Wikidata, where most of the edits come from bots, the results of Wikipedia might not be fully applicable. Further investigation is needed to better understand editors' diversity in Wikidata, in particular diversity of bot edits and their impact on Wikidata quality.

In Wikidata editor diversity could be measured using: user background (e.g., country, language, age, religion), participation patterns (i.e., user experience, roles, ...), contributions in topic domains (are they active in many domains or only focused on some), the volume of edits (balance of edits among domains) and the types of edits they perform, like adding, removing or updating and in which data level, like item, term or statement.

Understanding the diversity of editors contributes to a more comprehensive assessment of the diversity within the Wikidata data. That is, greater diversity among editors may indicate greater coverage of diverse topics and opinions, and vice versa.

- Consumer Diversity. Wikidata data is accessed primarily through the provided SPARQL[30] end point of Wikidata Query Service (WDQS)[31]. The data are also available through the MediaWiki API[32] and in the form of dumps, like XML (eXtensible Markup Language), RDF (Resource Description Framework) and JSON (JavaScript Object Notation) file formats[33].

Wikidata has the potential to be used in a variety of areas. Existing research shows the usage of Wikidata in linguistics, medical, and biological fields, while, the further usages of Wikidata remain to be studied. Studies exist on the usage of Wikidata for knowledge dissemination and integration of genes, drugs, and diseases [196], human and mouse genes and proteins from different sources [32]. Similarly, Wikidata is used as a multi-lingual multi-dialectal dictionary for Arabic dialects [319].

There are also a number of tools developed on top of Wikidata and consume Wikidata for different purposes. Some examples of these tools are: WDAqua-core which is a question-answering component [55], Ontodia is an online OWL and RDF diagramming tool [340] and Scholia is used for handling scientific bibliographic information [216]. Furthermore, Wikidata is used to obtain details of street names using OpenStreetMap [8].

[30] A Resource Description Framework (RDF) query language. https://www.w3.org/TR/sparql11-overview/ [Accessed: 14.04.2020]

[31] WDQS is a software package to query Wikidata dataset https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual [Accessed: 14.04.2020]

[32] `https://www.wikidata.org/w/api.php`[Accessed: 01.11.2020]

[33] Wikidata database dumps: `https://www.wikidata.org/wiki/Wikidata:Database_download`[Accessed: 01.11.2020]

Consumer diversity could be measured using consumer backgrounds (to find out how different countries and languages use the data from Wikidata and where are white spots), consumption of topic domains (to know how different domains are of more interest and which are not and why), consumption purposes, data access approaches (e.g., WDQS or dumps) or automation level of data access.

In essence, consumer diversity offers information on the demographics of those who access Wikidata from “anyone, anywhere in the world” revealing potential usage gaps. This could lay the foundation for identifying factors that contribute to limited access in specific regions or languages and facilitate efforts to address these issues. In addition, despite diversity being the property of an overall system, in Wikidata diversity could be measured from two main angles, i.e., data, and user. Measuring the diversity of the overall Wikidata is a complex task, and defining different angles for looking at diversity in Wikidata makes it more feasible to measure and easier to understand. In addition, it aids in exploring the impact of diversity increase/ decrease of one part on other parts. For example, an increase in multilingual labels could lead to an increase in reference diversity. This is because users who can find and read the information in their language can add the existing data sources of their language to Wikidata. Thus, we could encourage contributors to add more diverse sources through the improvement of label diversity in certain languages. Or, demand for data in certain classes or domains could increase the number of edits in those classes or domains, and vice versa. With this in mind, new approaches and various mechanisms could be developed to improve diversity in a way that can affect more than one angle of diversity.

3.3 Summary

Diversity is an important factor in achieving the overarching goal of Wikidata, as Wikidata can “serve anyone anywhere in the world” if it is diverse enough to address the needs of diverse people around the world. Additionally, all of the design decisions of Wikidata also refer to the concept of diversity, and therefore, make diversity a central point of focus in Wikidata. Despite its importance in Wikidata, there exists no research that has studied diversity in the Wikidata context. Plurality is one of the design decisions in Wikidata that explicitly represents diversity by allowing the coexistence of multiple statements. The fact that plurality is a real alternative to diversity in the Wikidata context is yet to be explored, so we begin by understanding the term diversity and how it is applied and interpreted in other contexts.

Diversity is a widely used concept in numerous fields and is measured using the three general properties or dimensions of diversity, which are variety, balance, and disparity. Based on our understanding of the basics of diversity in the existing fields, we propose our concept of diversity for Wikidata with a focus on the measurement of Wikidata diversity. Our proposed approach looks at diversity from two different angles of data and user because Wikidata is a KB where data is contributed and used by users. So, to measure diversity it is not important to only measure the data, but a glance at who contributes data and who uses it can also provide details on how diverse the community is and how diverse data we can expect. In Wikipedia research, the focus of diversity is on editors’ diversity and how it might impact the quality of data.

To measure diversity in Wikidata we proposed further angles for each of the two main aspects of user diversity and data diversity. User diversity can be measured from both the editor's and the consumer's perspectives. Similarly, data diversity can be looked up from a broader angle of domain/ class or a detailed angle of items, terms, and statements.

In the next chapter, we utilized our proposed approach to measure the current diversity status of Wikidata.

Chapter 4

WIKIDATA DIVERSITY STATUS

In this section, we employ our proposed concept for measuring diversity in Wikidata to obtain an overview of the current diversity status within the knowledge base. The results of our mapping study revealed a limited coverage of diversity research in relation to Wikidata. While diversity is a broad term encompassing multiple factors as defined in our proposed concept, our initial search specifically targeting the term ‘diversity’ may have resulted in overlooking certain indicators that are now included in our proposed diversity concept. To address this, we conducted a thorough review of the existing Wikidata literature, this time considering the diversity angles defined within our proposed concept for Wikidata (refer to Figure 3.3).

This section begins by providing an overview of the existing knowledge on diversity in the context of Wikidata, while also highlighting any gaps or areas where research on diversity is lacking. As data forms the core of Wikidata, we examine the available evidence pertaining to data diversity within the literature. However, given the limited explicit references to data diversity, we proceeded to collect the necessary information directly from Wikidata in order to perform a diversity measure at the domain/class level, using our proposed concept as a framework. We ultimately present a glimpse into the diversity landscape of Wikidata based on our proposed concept. By applying our concept and utilizing the collected information, we are able to provide a preliminary assessment of the diversity status within Wikidata. This serves as an initial step towards understanding and evaluating the extent to which Wikidata has achieved its diversity goal to become a world KB.

4.1 Overview of Diversity in the Existing Wikidata Literature

Before starting to measure diversity in Wikidata, we want to provide an overview of what is already known about diversity in Wikidata research. Hence, we used our proposed concept for diversity in Wikidata to re-explore the existing Wikidata literature. Since our previous search of the word diversity resulted in very little information (cf. Section 3.2.1.2), this time we looked for every element mentioned in our diversity measurement concept (cf. Figure 3.3) to capture the available information about the indicators of diversity and build the existing diversity status of Wikidata. Our results once again showed that research on diversity is not yet an established topic in Wikidata. Nevertheless, we could get an overview of the diversity status based on the parts mentioned in our diversity measurement concept for Wikidata which we explain below.

Commencing with user diversity, we examine both editor and consumer diversity to explore existing factors that can aid our comprehension of the current state of diversity within Wikidata.

4.1.1 User Diversity

The diversity of users is a topic of interest in any collaboratively developed KB like Wikidata. Volunteers contribute knowledge or data based on their own knowledge level which is deeply tied to their backgrounds. So, studying user diversity is another way of obtaining insight into data diversity to find out the reason for the existing data diversity status of Wikidata.

Users can either edit Wikidata or utilize the data for some purpose. In Wikidata, users are of importance because the data comes from user contributions, and in the same way they make use of this data which is the reason Wikidata was developed to serve as a structured data source. Wikimedia Foundation in a survey of the Wikidata community has found that most of the users in the survey which represent the Wikidata community, are young males who joined Wikidata early (i.e., 2012/2013) and live in the global north [53]. Another study by Shaik et al. compared Wikidata queries to real-world datasets to examine the race and country of citizenship bias in general with a focus on STEM and computer scientists in Wikidata. The authors find an over-representation of white Western individuals from Europe and North America, while, the rest of the world is underrepresented [283].

Following we explain diversity from the editor and consumer perspectives based on the existing research.

4.1.1.1 Editor Diversity

Users who edit are of importance in the concept of diversity in Wikidata because it is the editors who contribute their knowledge to this KB and make it accessible to all. In collaborative KBs like Wikidata where the content is dependent on the community, editors are the determiners of what would get into the KB. The more diverse backgrounds they have, the more diverse topics, language coverage, and ideas to expect. In the existing literature on Wikidata, a number of studies have studied multilingualism in Wikidata.

The study by Kaffee et al. has analyzed multilingual label editing in Wikidata with a focus on three user groups, i.e., registered editors, bots, and anonymous editors. The authors found that registered editors (i.e., humans) tend to edit labels in more languages than bots. On the other hand, bots add a high number of labels, but, only in specific languages [145]. Thus, bots' edits are less diverse than humans from the multilingualism angle.

Further, Piscopo et al. in their study of Wikidata external references with a focus on sources in English, show that despite similar numbers of references added by humans and bots, humans added more diverse web domains than bots. They also show that a high percentage of references added by bots are not authoritative[1] and not relevant. Although there are similarly invalid references added by humans, they are much fewer than the ones added by bots [238]. Thus, bots could have an impact on the diversity of sources in Wikidata due to the high volume of data they import from a small number of sources. Farda-Sarbas et al. have shown that bots tend to import most of the data from Wikipedia, especially the Western languages like the English language version. This tendency of bot edits and their imports from Western languages can be blamed for the change in the language balance in Wikidata that made Western languages the dominating ones over other languages.

In general, there exists an editing imbalance among the defined user groups of Wikidata, where, bots perform the lion's share of edits among all.

4.1.1.2 Consumer Diversity

Wikidata data is mainly accessed through the provided SPARQL[2] end point of Wikidata Query Service (WDQS)[3]. In their study of the Wikidata SPARQL query logs from 2017, Bielefeldt et al. have classified the Wikidata queries as organic (i.e., by humans) and robotic (i.e., program), where organic queries make only 0.31% of the whole queries. This shows the dominance of robotic queries in Wikidata. In other words, bots are not only performing the majority of edits in Wikidata but are also the most prominent consumers. Organic queries retrieve data to fulfill an immediate information requirement of a human user, whereas robotic queries retrieve data autonomously for subsequent automated processing [25]. Hence, organic queries are more diverse, and robotic queries are more uniform. They also found that fewer users from Asia access Wikidata through SPARQL-based services and again Western languages, i.e., English and other European languages are on the top of the data querying list. This shows that Wikidata is not only developed and mainly edited by Westerners, but also used and consumed by them as well. Additionally, the research in Wikidata is mostly focused on data editing, and the data consumption aspect is rather less explored [296]. The Wikidata properties and data hierarchies are some examples of the challenges that limit data extraction and access to the data in Wikidata by consumers. Therefore, further work is needed to make Wikidata more diverse for serving globally to individuals across the world.

In summary, the existing research shows that Wikidata is mainly edited and consumed through automation. Since this automation is performed and controlled by computer experts, we could speculate that Wikidata knowledge is largely contributed by white and Western computer scientists, mostly living in the global north.

[1] 'Authoritative sources refer to sources of information that are deemed trustworthy, up-to-date, and free of bias for supporting a particular statement on Wikidata.' https://www.wikidata.org/wiki/Wikidata:Verifiability [Accessed 10.05.2020]

[2] A Resource Description Framework (RDF) query language. https://www.w3.org/TR/sparql11-overview/ [Accessed: 14.04.2020]

[3] WDQS is a software package to query Wikidata dataset https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual [Accessed: 14.04.2020]

4.1.2 Language Diversity

Wikidata contains data in more than 300 languages all in one place. The presence of Wikidata terms (i.e., label, description, and alias which make Wikidata items human-readable) in a high variety of languages shows diversity when considering single-concept diversity, i.e., variety of languages. However, here it is important to proceed with dual-concept diversity, i.e., to find out if all of the languages are equally present or not. The term diversity in Wikidata is an alternative to the multilinguality of terms.

As mentioned earlier, Kaffee et al. in their study of multilingualism in Wikidata demonstrate the uneven distribution of labels and descriptions across languages in Wikidata. They find that most of the terms exist in a small number of languages which are mainly Western languages. On the other hand, the majority of languages in Wikidata have close to no coverage [142]. This shows that despite having a high variety of languages in Wikidata, their representation through terms is heavily imbalanced and, hence, Wikidata language diversity is low. The existence of a higher number of bot requests for Western languages [67] could be the reason for this language imbalance in Wikidata.

4.1.3 Diversity Gaps in Wikidata Research

Despite the fact that Wikidata was designed with the ability to serve data that could represent the diversity of the world, diversity remains a research gap in Wikidata literature. While diversity is measured for a system as a whole, we defined aspects in our proposed concept for diversity measurement that allow us to look at diversity from different angles. Using these aspects we re-explored the Wikidata literature and could see that there exists some basic information about user diversity in Wikidata from both angles of editor and consumer. From the data angle, we could only find some papers on the multilingualism of labels in Wikidata, however, data diversity as a whole remains a white point and we have very little knowledge of the existing diversity of topical domains, sources, and statements in Wikidata.

4.1.3.1 Data Diversity in Wikidata

Data is the core of Wikidata and based on our proposed concept can be measured from angles that are domain/class, and item.

Items are the core representation of data in Wikidata. As mentioned earlier, they consist of further parts which are term (label, description, and alias), statement (claim and reference), and sitelink (cf. Section 2.1.3). Diversity in item level looks deeper into these parts of the item.

Term. Considering the description and alias parts of the term, there exists no study that could provide diversity status information in these areas. While, labels and descriptions can be multilingual, for the same item there could exist multiple aliases in each language, and further research is needed to provide insights into these parts of the term in Wikidata items.

Statement. In Wikidata, the main body of data about an item is represented by the relevant properties of that item, also called a claim. Wikidata highly suggests providing a source (i.e., reference) with every claim to ensure data reliability. The combination of claim and reference makes a statement in Wikidata. So far, there is no estimate of statement diversity in Wikidata with a focus on the claim, however, there is a clue of reference diversity by Piscopo et al. which show that Wikidata seems to be less focused on Anglo-American sources than Wikipedia and, thus, contains knowledge from more diverse sources in comparison to Wikipedia [241].

To the best of our knowledge, there is no existing research that can shed light on the diversity at the statement level, also referred to as plurality. Performing diversity measurement at the statement level is not as straightforward as measuring the diversity of terms, references, or domain-level diversity. The reason, as earlier mentioned, is the challenge in the identification of statements showing variety from the statements that show contradiction. Since this classification of statements needs further investigation (cf. Section 3.2.3.1), our knowledge of diversity at a statement level remains a white spot.

Additionally, Wikidata items are associated with their respective classes. Nevertheless, we found no studies that could provide an overview of Wikidata class coverage or diversity status in the existing literature. Hence, domain/class diversity also remains a gap in the Wikidata research.

Overall, we witnessed a big gap in Wikidata's existing literature regarding research on data diversity. Data diversity is yet to be explored from both, domain/class and item angles. At the item level, there exist studies on multilingualism of Wikidata labels which also refer to the language imbalance. Nevertheless, other parts of the item like description, alias, reference, and claim are not yet studied from the diversity angle. Statement level diversity which is also referred to as plurality and aims to bold globally agreed data over contradictory ones is also a big white spot in the Wikidata research.

For this reason, in the next section, we apply our proposed concept for measuring diversity in Wikidata explaining the measurement procedure and presenting an overview of the existing data diversity in Wikidata.

4.2 Measuring Data Diversity

We perform our measurement on the domain/ class level to provide an eagle-eye overview of the data in Wikidata including the item coverage of properties/ statements. An overall picture of the topical domains and their class coverage can provide a glance into how diverse Wikidata domains are and where there might be diversity gaps in those domains. In addition, beginning with measuring the domain/ class diversity has the benefit that it can highlight which domains/ classes are more diverse and which are less diverse. This way we can look into the item and statement diversity of these domains to compare them and find out the reason.

Wikidata items are already linked to their corresponding classes; however, these classes are not organized into distinct domains. To address this, we adopt the domains and classes classification defined by Färber et al., who delineated five domains: people, media, organizations, geography, and biology[4]. These are the domains where

[4] The list of these domains along with their respective classes is available at: http://km.aifb.kit.edu/sites/knowledge-graph-comparison/[Accessed: 06.12.2019]

Wikidata classes can be further categorized [65]. As mentioned before, Färber et al. have used these domains to compare Wikidata with other KGs, i.e., YAGO, DBpedia, Freebase, and OpenCyc to measure the quality of data in these KGs. We use this available domain categorization approach to shed light on the domain coverage and distribution of data on Wikidata, the only KB that has a built-in diversity mechanism for its data.

Regarding the usage of diversity measures to assess the domain level diversity, we use *Simpson's Index*, *Shannon Entropy*, and *Rao-Stirling Index* as heterogeneity measures when considering more than one diversity dimension and *Gini-coefficient* when focusing on balance only, as explained in 3.1.4.2. All of the mentioned measures result in a number or index, where higher numbers show higher diversity levels and vice versa. We calculate this number for each domain and then compare these domains to find out if all domains are on the same level of diversity or not. We rank the domains with higher numbers as the more diverse domains and the domains with lower indexes as the less diverse domains in Wikidata.

We present the details of the above-mentioned measurements in the following.

4.2.1 Diversity in Domain and Class Level

Here, our aim is to address the question of how diverse domains in Wikidata are. To make it feasible to measure, we break this question into the following two questions:

a) How balanced are the classes in Wikidata domains?

b) How balanced are the items of those classes?

As mentioned before, our focus here is on the distribution of items across Wikidata domains/ classes and the distribution of properties/statements across Wikidata items. In other words, we are interested in finding out if data is distributed evenly in Wikidata domains/ classes or if we have data concentration in some domains.

In the first step, we measure the diversity of Wikidata domains based on the number of classes in each domain and the number of items in those classes. To measure this, we take the list of domains and classes from [65]. Using Wikidata query service (WDQS) we retrieved the number of items per class considering both properties, i.e., *instance-of (P31)* and *subclass-of (P279)*. Table A.1 (in the Appendix on page 162) contains the list of these Wikidata domains, classes, and the number of items per class. We used the columns *Domain*, *Subclass*, and *#unique items* for this measurement using heterogeneity measures mentioned in Section 3.1.4.2 available in the diverse package of R programming language in RStudio software. Table 4.1 displays the results of this diversity measurement in each domain in regard to the number of classes and items each domain has.

As can be seen, the results show uneven diversity numbers across Wikidata domains. This could indicate that not all Wikidata domains and classes are given equal attention. Here we can see that using two dimensions (i.e., variety and balance), the domain *Media* is the most diverse and the domain *Geography* comes in second. However, using three dimensions (i.e., variety, balance, and disparity) [5]), the most diverse domain is *Geography*, and here the domain *Media* comes in second

[5] Here the diversity measurement is done using the diverse package of RStudio software, which uses the Euclidean distance method as the default option for measuring disparity.

Table 4.1: Descriptive statistics of Wikidata domain diversity levels using the number of classes and items in each domain based on Table A.1. (Note: Higher values mean higher diversity levels and bold values indicate the highest diversity in each column)

	Entropy	Simpson Index	Rao Stirling Index
Biology	0.773	0.508	931.737
Geography	1.199	0.670	**64188.895**
Media	**1.425**	**0.725**	45155.799
Organization	1.034	0.521	15752.442
Person	0.141	0.048	1400.811

place. Similarly, the domains *Person* and *Biology* are shown to be the least diverse domains. While we were able to determine which domains have higher diversity levels and which domains are less diverse in comparison, we want to look into the details to know why some domains are more diverse than others and vice versa.

For this reason, in the second step, we get an overall overview of whether all items are on the same level of completeness and depth of information in the classes of these domains. In other words, do all items have a balanced coverage of properties/statements? or do some classes have items with a rather higher content coverage and some have more items with basic information or are empty, and therefore, we have imbalanced domains? Thus, we focus on each class in these domains and provide a more detailed view of the item coverage of these classes using the Wikidata Knowledge Imbalance Dashboard[6] tool to measure the balance of classes in the previously mentioned five domains. This tool extracts items based on the *instance of (P31)* property and quantifies knowledge imbalances on Wikidata using the Gini coefficient. As discussed in Section 3.1.4, the Gini-coefficient is a measure of balance that is widely used in Wikipedia (cf. Table 3.3) and is applicable in the Wikidata context. Although the result of this tool might not show the actual class-item relations because it only focuses on *instance of* relation, we can still get an overview of the classes and their items property coverage in Table 4.2.

In Table 4.2 we can see that most of the classes are imbalanced. In the Organization domain, all classes are imbalanced, while, *Biology* and *Person* domains also have classes that are heavily imbalanced, i.e., *Politician* and *Mammal.* For the class *Writer* there are no direct instances; thus, the tool could not generate any results. For the class *Grass* there is only one instance, thus, no comparison of balance is possible. The only balanced classes are *Film, Book, Album, Mountain* and *Country* which belong to the *Media* and *Geography* domains. One possible reason for the balanced and higher data coverage of *Geography* domain could be that OpenStreetMap[7] extracts geographic data from Wikidata [173] and this demand for geographic information has attracted more attention from editors in this domain.

Overall, the imbalance in Wikidata domains is not only in the number of items per class but the property coverage of Wikidata items is also imbalanced. As can be seen, the domains of *Person* and *Biology* have a lower number of items per class than the remaining three domains. The only exception is with the class *Gene* which

[6] Wikidata Knowledge Imbalance Dashboard available at: `https://prowd.netlify.app` [Accessed: 14.04.2020]

[7] https://www.openstreetmap.org/node/597204161#map=18/58.38974/13.85165

Table 4.2: Balance of properties across items of the classes and domains in Wikidata based on [65] using the Wikidata Knowledge Imbalance Dashboard tool that runs on Wikidata and limits the number of items to 10,000 (The classes with * are added because they appeared in Wikidata research)

Domain	Class	#Items	Gini-Coef.	Diversity Level
Person	Human*	10,000+	0.235	🟧 Imbalanced
	Musician	41	0.339	🟧 Imbalanced
	Athlete	33	0.312	🟧 Imbalanced
	Writer	-	-	-
	Politician	9	0.41	🟥 Heavily Imbalanced
Media	Film	10,000+	0.19	🟩 Balanced
	TV Series	10,000+	0.304	🟧 Imbalanced
	Book	6,906	0.142	🟩 Balanced
	Magazine	10,000+	0.347	🟧 Imbalanced
	Album	10,000+	0.187	🟩 Balanced
Organization	Bank	2,181	0.32	🟧 Imbalanced
	Airlines	4,600	0.287	🟧 Imbalanced
	University	10,000+	0.275	🟧 Imbalanced
	Sports club	10,000+	0.236	🟧 Imbalanced
	Political Party	10,000+	0.366	🟧 Imbalanced
Geography	Lake	10,000+	0.263	🟧 Imbalanced
	River	10,000+	0.212	🟧 Imbalanced
	Mountain	10,000+	0.17	🟩 Balanced
	Country	180	0.137	🟩 Balanced
	City	8,969	0.388	🟧 Imbalanced
	Road*	10,000+	0.213	🟧 Imbalanced
Biology	Mammal	9	0.543	🟥 Heavily Imbalanced
	Bird	27	0.287	🟧 Imbalanced
	Fish	9	0.307	🟧 Imbalanced
	Tree	533	0.256	🟧 Imbalanced
	Grass	1	0	-
	Gene*	10,000+	0.117	🟩 Balanced

has more than 10,000 items and is balanced. Again, this could be due to the usage of Wikidata as a genomic data source by biologists [32, 246].

Measuring domain and class level diversity in Wikidata is not as easy as it looks because the Wikidata class hierarchy is not well-ordered and not all classes are instantiated in the same way. For example, within the *Musician* class, some items are instantiated using the *subclass of* property, while others are instantiated using the *instance of* property. As a result, using the Wikidata Imbalance Dashboard that extracts based solely on the *instance of* property, we obtained 41 items (cf. Table 4.2). However, when querying Wikidata through WDQS and considering both *instance of* and *subclass of* properties, we obtained 617 items (cf. Table A.1 on page 162 in the Appendix).

Furthermore, the Wikidata class hierarchy is structured as a system of subclasses, where each class represents a more specific type compared to the class above it, and can itself be subdivided into even more specialized subclasses. As a result, the current method of calculating Wikidata diversity based on domains and classes, using the count of items per class alone, may not yield entirely accurate results. Nonetheless, it can still offer insights and an overview of the prevailing diversity status within Wikidata. Achieving greater precision would necessitate additional efforts, but it would only be feasible once the issue of class hierarchy in Wikidata is addressed. Since resolving the class hierarchy issue exceeds the scope of this study, we rely on the available results as indicators of diversity gaps within Wikidata domains and classes. These results are consistent with those presented in Table 4.1, which assess domain-level diversity by considering items associated with classes through both the *instance of* and *subclass of* properties.

4.3 Summary of Existing Wikidata Diversity Status

The application of our proposed model has helped to gain a general overview of where Wikidata is on the path to achieving the overall goal of serving "anyone anywhere in the world." Through our analysis, we have observed a notable absence of comprehensive research on the topic of diversity within the existing body of Wikidata literature. Only a limited number of references to diversity could be found, indicating the need for further exploration and development of the diversity concept within the Wikidata context.

In this study, we have taken an initial step toward investigating diversity in Wikidata by introducing a concept for measuring diversity. Given the vast scope and complexity of the topic, we have focused on specific angles to examine Wikidata's diversity landscape. This approach has facilitated a more practical approach to assessing diversity and addressing any identified gaps.

At the data level, our analysis has shed light on the issue of domain coverage imbalance, revealing areas where diversity is lacking in Wikidata. Additionally, in existing research, we have observed a significant language imbalance in terms of language labels, with a dominance of Western languages and limited representation of other languages. However, the diversity of descriptions and aliases in Wikidata remains unexplored and requires further investigation. Furthermore, the concept of plurality, which is a design principle of Wikidata at the item level, remains a significant gap in our understanding of diversity within the platform. Currently, there is a lack of information on diversity in this particular area, and more research is needed to explore and evaluate the diversity aspects related to plurality.

When examining user diversity in Wikidata, we observe that users are typically categorized into groups such as humans, bots, tools, and anonymous, based on their level of automation. However, there is a lack of in-depth studies on the editing patterns of these user groups, which would provide valuable insights into the diversity of user contributions in Wikidata and how this diversity may have influenced the data. Of particular interest is the unique nature of the Wikidata community, where a significant portion of edits is contributed in large volumes by a small number of automated accounts. Consequently, within this community, bots emerge as the most active contributors, potentially exerting a greater influence on Wikidata's data compared to other user groups. Existing research suggests that bot edits tend to be less diverse compared to human edits, and bot queries are often more uniform and less diverse than human queries. Therefore, it is important to investigate the editing patterns of different user groups, with a specific focus on bots, in order to understand how bot contributions may have influenced data diversity in Wikidata, particularly in relation to the existing domain-level imbalances. Moreover, gaining deeper insights into the diversity of editors allows us to recognize the potential of bot participation and discuss potential mechanisms to enhance the diversity levels of data in Wikidata. By examining and addressing the impact of user diversity, efforts can be directed toward improving the overall diversity and inclusiveness of the knowledge representation in Wikidata.

The current state of data diversity in Wikidata reveals a diversity gap among its domains. Some domains receive more attention from editors, while others are overlooked. This issue arises due to the collaborative nature of data contribution, where a small number of bots are the most active editing users. Diversity gaps also exist in multilingual labels, with languages predominantly edited by bots occupying dominant positions. Additionally, the most active contributors in Wikidata predominantly come from Europe and North America, which could explain the overrepresentation of data from Western regions. In general, Wikidata has not yet fully achieved its goal of serving 'anyone anywhere.' For instance, currently, Wikidata contains nearly four times as many statements concerning Western artists as it does for non-Western artists and approximately nine times as many statements about Western masterpieces as it does for their non-Western counterparts [3]. Based on the aforementioned findings, there is a concentration of Western data in Wikidata, highlighting the need for further work to achieve a more balanced representation that provides equal services for users worldwide. By addressing these diversity gaps, Wikidata can enhance its inclusivity and ensure equitable access and services for users from all parts of the globe.

Given that the data in Wikidata is contributed voluntarily by the community, gaining insight into the contributing community can provide insights into the factors contributing to the low diversity status and data imbalance in Wikidata. Existing research suggests that in the Wikidata community, a small number of bot operators employ bot accounts, leveraging their ability to execute high-speed and batch edits to import substantial volumes of data aligned with their interests. This practice can subsequently impact the overall data composition in Wikidata. Therefore, it is likely that bot edits play a role in the diversity gaps observed in Wikidata. However, our knowledge about bots and their editing behavior in the context of Wikidata is limited. To investigate whether bots contribute to the data imbalance in Wikidata's domains and classes, a comprehensive study and understanding of bots in the Wikidata context is required.

Considering the limited availability of comprehensive information regarding bot editing behavior in existing research, we curated our own datasets to acquire insights into bot activities. Initially, we gathered online bot request forms from the Wikidata website to comprehend the motives and methodologies behind bots' involvement in editing activities. This initial effort formed the foundation of our first dataset, named the *Wikidata-Requests-for-Permissions Dataset.* Following this, we extracted real bot edits from the Wikidata database, constructing the *Wikidata Revision History Dataset.* This dataset allowed us to compare bot editing patterns with those of other users within the Wikidata community. Our objective was to determine whether or not bots have indeed played a role in the imbalances that have been observed in the various domains of Wikidata.

In the next section, we will present our research approach for creating the aforementioned datasets that we used to study bots. We will then analyze bot edits using the aforementioned datasets to examine their impact on diversity within Wikidata.

Chapter 5

RESEARCH DATA & APPROACH

In our quest to understand the current diversity status of Wikidata, we discovered a lack of research dedicated to diversity in this context. Utilizing our proposed concept for measuring diversity, examining the available literature on Wikidata, and gathering data from the platform itself, we were able to provide an overview of diversity within Wikidata. Our analysis revealed an imbalance in the coverage of data across various Wikidata domains.

To delve deeper into the underlying reasons for this imbalance and to address the issue, we turned our attention to the contributors who have contributed to this data. In particular, we recognized the need for a comprehensive investigation into bots, as they are responsible for the majority of contributions in Wikidata. Bots play an important role in shaping the data landscape and understanding their impact is crucial to addressing diversity imbalance.

In this chapter, we present the research data we collected and the methodology employed in the subsequent chapters, which will shed further light on the diversity landscape of Wikidata.

5.1 Research Approach

Finding answers to research questions in a scientific way is only possible through defined scientific methods and approaches. Research methods could be general and applicable in any field like literature review, or could have a rather narrowed area of usage like design science methodology. In general, there are three main approaches to follow, which are qualitative research, quantitative research, and mixed method, which refers to a combination of both. In our study, in the area of human-computer interaction (HCI), we use a combination of qualitative and quantitative research. For addressing our defined research questions in Section 1.2 (on page 3), we have mainly used a mapping study to answer the first research question, and content analysis to answer the second research question. The details of the mapping study along with its results are already discussed in Section 2.2 (on page 18). The content analysis approach used to answer the remaining research questions is described in detail in the following sections.

In order to understand the current status of Wikidata from the diversity perspective, the focus of the first research question, we need to have a detailed view of the Wikidata literature and a clear understanding of the diversity concept in the Wikidata context. We began with a mapping study, a general form of a literature survey, of Wikidata and shed light on the angles from which it has been studied so far to find if there exist any clues on diversity in the Wikidata literature. We explain the mapping study methodology in Section 2.2.1 (on page 19) and reflect on the results of this study in Section 2.2.3 (on page 26). We also performed a literature survey to explore and understand the diversity concept in general and what it means in a KB context. The result of this literature survey is given in Section 3 in order to form a basis for understanding the diversity concept itself and its interpretations from different application areas. Using the implications of these results we define and propose a concept for measuring diversity in Wikidata.

The second research question involves dealing with the data stored in Wikidata. Since the data of our focus are from a) Requests-for-Permissions pages in Wikidata which are available in unstructured (i.e., plain text) form, and b) revisions from the Wikidata edit history which are in semi-structured form, hence, we used the content analysis method.

We used two datasets here, the first dataset, *Wikidata-Requests-for-Permissions Dataset* [69] in Section 5.2, contains requests for bot accounts called RfP (Requests-for-Permissions) pages. We use RfPs to dig deeper into what bots are, what they intend to do on Wikidata, and how the Wikidata community deals with bots. This dataset and its results are presented in Section 6.2 (on page 106). Only after getting answers to the above-mentioned questions about bots can we get a better understanding of the Wikidata community and compare the role of bots with human users. To shed light on the potential of bots regarding the diversity status of Wikidata, we analyze the second dataset, *Wikidata Revision History Dataset* [68] in Section 5.3. This dataset is created from the Wikidata revision history and contains every edit in Wikidata by any user, registered or unregistered, in a semi-structured form. The pre-processing of these data was a rather long process, which is explained in Section 6.3 (on page 119). With the answers to these questions, we could recommend ways to improve the diversity status of Wikidata through bots.

In summary, the approach used to address the first research question on measuring the current diversity status of Wikidata is a literature survey, and to tackle the second research question on the role of bots on diversity in Wikidata, the content analysis method is used.

Next, we explain each of the mentioned datasets in detail and give an overview of the methods used for the development of these datasets.

5.2 Wikidata-Requests-for-Permissions Dataset

As mentioned before, despite the high activity levels that bots show in Wikidata, they have remained a rather less investigated user group. In order to find the answer to our second research question on the role of bots on diversity in Wikidata, we begin by studying what exactly bots are and to know this, we have targeted the request pages for bot permission on Wikidata website which act as the very first step in a bot's life to get into the Wikidata community as a contributor. These pages are called Requests-for-Permissions (RfP) pages and contain details such as bot name, intended tasks, data sources, further information during community inspection, and discussions along with the final decision made on the request.

Figure 5.1 displays a glance into the data pre-processing phase. The details of our selected approach to data collection and coding come next.

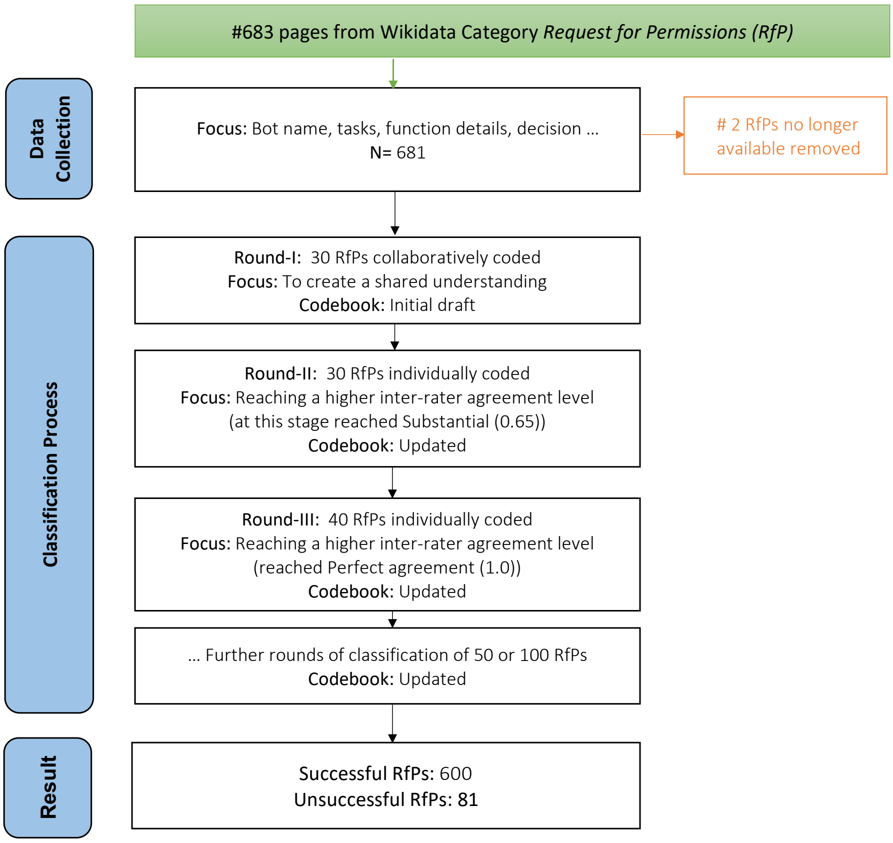

#683 pages from Wikidata Category *Request for Permissions (RfP)*
Data Collection
Focus: Bot name, tasks, function details, decision …
N= 681
# 2 RfPs no longer available removed
Classification Process
Round-I: 30 RfPs collaboratively coded
Focus: To create a shared understanding
Codebook: Initial draft
Round-II: 30 RfPs individually coded
Focus: Reaching a higher inter-rater agreement level
(at this stage reached Substantial (0.65))
Codebook: Updated
Round-III: 40 RfPs individually coded
Focus: Reaching a higher inter-rater agreement level
(reached Perfect agreement (1.0))
Codebook: Updated
… Further rounds of classification of 50 or 100 RfPs
Codebook: Updated
Result
Successful RfPs: 600
Unsuccessful RfPs: 81

Figure 5.1: An overview of the *Wikidata-Requests-for-Permissions Dataset* pre-processing phase. (Note: This diagram is inspired by the PRISMA flow diagram [223].)

5.2.1 Research Methodology

As mentioned above, RfP pages contain the details regarding a bot account. These details are stored as sentences which we call text format or unstructured format and use the content analysis approach to process them.

Next, we explain the content analysis method that is used to answer RQ2.

Content analysis is a form of qualitative analysis that deals with unstructured content or data and processes the data to develop a representative description of the unstructured text [167]. It defines systematic and reproducible techniques for compressing longer words through coding schemes into shorter word categories [299]. This approach originated from communication studies in the 1950s and was used to analyze newspaper data by coding text into categories and bringing it into a quantitative form to be analyzed by statistical tools [162]. For this reason, content analysis is usually considered a more quantitative approach.

Nevertheless, content analysis can also have a qualitative approach that can be distinguished from the quantitative one. The distinction between quantitative and qualitative approaches comes from the way coding is performed.

According to Lazar et al., coding is the process of categorizing unstructured content into defined categories by assigning descriptors to the unstructured content. He emphasizes that coding is not simply "paraphrasing the text and counting the number of keywords in the text" but doing much more like comparing the data, deriving concepts from the data, or additional details like properties and dimensions for data.

Lazar et al. also describe the two existing approaches to coding. *A priori* coding refers to quantitative analysis where the codes are already defined and present, either from an earlier study or from the previous investigations of the same topic. The *emergent* coding, on the contrary, is a qualitative approach used when there is no established method to guide the coding, and thus, codes emerge during the coding process by noting interesting concepts in the data and continuously refining the code book. Emergent coding is beneficial when working on a new topic where finding established theories or sufficient literature to develop the coding categories in advance is a challenge [167].

5.2.2 Data Collection

We collected all bot requests that were in the final approval stage in July 2018[1], i.e., we ignored all tasks without a final decision, from the Wikidata's archive for requests[2]. We collected our data based on web scraping, i.e., web data extraction programs implemented in Python. Bot requests that were listed several times in different archives were only parsed once[3]. This resulted in 683 task approval pages.

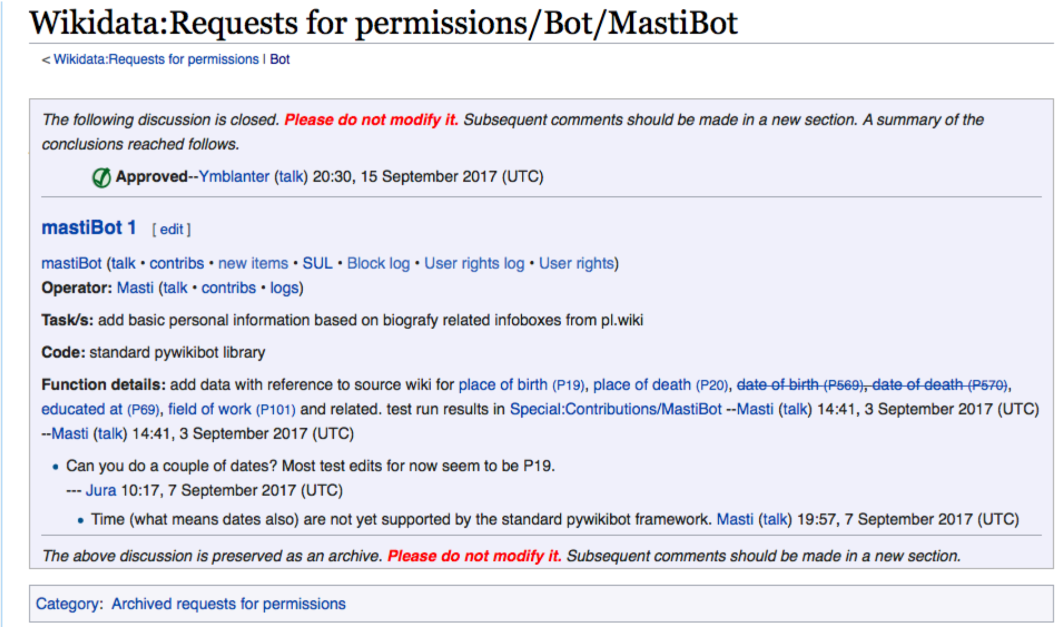
We extracted the following information from each page (cf. Figure 5.2): URL, bot and operator name, decision, decision date, tasks, code, and function details. Additionally, we collected the date of the first and last page edits[4], the number of page edits, and the number of distinct editors who contributed to the request for

[1] The first request in our data set was opened on October 31, 2012, and the last one was on June 29, 2018.

[2] www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Archive#Requests_for_bot_flags.

[3] For example, www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/VIAFbot is listed in www.wikidata.org/wiki/Wikidata:Requests_for_permissions/RfBot/March_2013 and in www.wikidata.org/wiki/Wikidata:Requests_for_permissions/RfBot/April_2013 and only the latter one was used.

[4] The first edit can be interpreted as the request opening and last edit as the request closing or date of archiving.



Wikidata:Requests for permissions/Bot/MastiBot

< Wikidata:Requests for permissions | Bot

*The following discussion is closed. **Please do not modify it.** Subsequent comments should be made in a new section. A summary of the conclusions reached follows.*

✓ **Approved**--Ymblanter (talk) 20:30, 15 September 2017 (UTC)

mastiBot 1 [edit]

mastiBot (talk • contribs • new items • SUL • Block log • User rights log • User rights)
Operator: Masti (talk • contribs • logs)

Task/s: add basic personal information based on biografy related infoboxes from pl.wiki

Code: standard pywikibot library

Function details: add data with reference to source wiki for place of birth (P19), place of death (P20), ~~date of birth (P569), date of death (P570)~~, educated at (P69), field of work (P101) and related. test run results in Special:Contributions/MastiBot --Masti (talk) 14:41, 3 September 2017 (UTC) --Masti (talk) 14:41, 3 September 2017 (UTC)

- Can you do a couple of dates? Most test edits for now seem to be P19. --- Jura 10:17, 7 September 2017 (UTC)
  - Time (what means dates also) are not yet supported by the standard pywikibot framework. Masti (talk) 19:57, 7 September 2017 (UTC)

*The above discussion is preserved as an archive. **Please do not modify it.** Subsequent comments should be made in a new section.*

Category: Archived requests for permissions

Figure 5.2: Example of a Wikidata request-for-permissions (RfP) page.

each page using the MediaWiki API[5]. This extracted information (cf. Table 5.1) was processed and stored in a relational database[6]

Table 5.1: Example of the information extracted from an RfP page.

URL	www.wikidata.org/wiki/[...]/Bot//MastiBot
Bot Name	mastiBot
Operator Name	Masti
Task and Function	add basic personal information based on biography-related infoboxes from pl.wiki
Decision	Approved
Decision Date	15/09/2017
First Edit	03/09/2017
Last Edit	21/02/2018
No. of Edits	8
No. of Editors	5

5.2.3 Classification Process

We classified the data collected manually to gain a deeper understanding of the different tasks that bot operators carry out on Wikidata. In the following, we describe this process in detail.

We employed the qualitative method of content analysis that is used to analyze unstructured content [167] which in our case is the text in the RfPs. We classified the data manually using the open coding approach. The data we were dealing with were applicants' own sentences (in vivo codes); thus, we developed an emergent

[5] www.mediawiki.org/wiki/API:Main_page.

[6] The dataset is released under a public license on REFUBIUM: http://dx.doi.org/10.17169/refubium-40234.

coding scheme and categorized the textual information. Two of the authors[7] coded the data in a three-phase process by ensuring the consistency of the codes and reducing possible bias during coding.

We read a sample of the data collected and discussed the various requests in detail. We noticed that some common categories could be found within all requests based on some particular perspectives, such as the potential actions of the requesting bots and the main sources of the data to be imported from bots. We focused, thus, on task and function details of the RfP page and extracted the intended use (task) and the data source of the bot edits which were approved or closed as successful.

In the first phase, we collaboratively categorized 30 randomly chosen RfPs to develop a shared understanding. Based on the first categories, we carried out a second round in which another set of randomly chosen 30 RfPs was categorized individually. We then compared the results and checked the agreement level[8]. After discussing diverging cases and cross-validating our category set, we continued with another round of categorizing the data (40 RfPs) individually, and we had 39 agreements out of the 40 cases. We continued with further rounds of coding individually; we met frequently on a regular basis and discussed new or unclear cases and cross-validated our category sets to ensure the consistency of our data classification.

We developed a codebook[9] as the primary reference of the classification process starting from the first phase and continually updating it during the classification period. Codes are structured in verb-noun pairs denoting content entities and operations, which are inspired by the work of Müller-Birn [205] and were originally drafted after the first 30 RfP classifications. On the basis of this, we further developed it by structuring all newly presented information into verb-noun pairs and adding them to the codebook.

The overview of the *Wikidata-Requests-for-Permissions Dataset* is present in Section 6.2.1.2 (on page 108) followed by the results and discussions.

Next, we describe the *Wikidata Revision History Dataset* that contains actual bot edits to confirm bot activities from RfPs and explore the bot editing patterns compared to humans. The analysis of Dataset-II provides insights into what bots are and what they claim to be doing with bot rights being granted. This only answers one-half of the research question on understanding the potential of bot edits on diversity in Wikidata. For this reason, we continue with our research on bots, and this time take a look at the Wikidata edit history to trace bot edits among other user groups, in particular human users.

5.3 Wikidata Revision History Dataset

We use the Wikidata edit history to get an idea of how each editor edits Wikidata. Data in Wikidata are accessible mainly through the user interface and regular dumps

[7] The first and second authors of the paper: Approving automation: Analyzing requests for bot permissions in Wikidata.

[8] We measured the inter-rater reliability by using Cohen's kappa. At this stage, we had a substantial agreement level, i.e., 0.65.

[9] The codebook along with the data is provided on GitHub: https://github.com/FUB-HCC/wikidata-bot-request-analysis.

in different formats on a weekly basis[10]. How this data should be accessed can be determined by the purpose behind data usage. For example, to look for known items in Wikidata, one can use the user interface utilizing the Special Search[11], when looking for known items in formats other than just HTML, Linked Data Interface (URI)[12] can be used, and in cases other than the known items, it is recommended to use Wikidata Query Service[13] or Wikidata database dumps[14] when dealing with large portions of data.

Additionally, Wikimedia provides a complete history of all edits carried out at *Wikimedia Toolforge*[15]. Toolforge provides access to replica databases of all Wikimedia projects; thus, we used the revision data from the Wikidata database that contains details on who edited what and when.

In the following, we explain our approach to creating this data set as detailed as possible to make the process clear and reproducible. Figure 5.3 presents an outline of the pre-processing phase.

5.3.1 Research Method

In the revision history of Wikidata, the details of a performed edit are called a comment and are in a semi-structured form. The comments consist of two parts, the structured part, and the unstructured part, like $/*wbsetdescription-set:1|fr*/$ *Ville de Suisse et chef-lieu du canton de Bâle-Ville.* To process these data we have used the content analysis method explained in Section 5.2.1. Below, we explain our steps for the development of this dataset to answer the RQ2 of this study on understanding bot editing behavior in Wikidata.

5.3.2 Creating a Sample from the Wikidata Revision History

We collected data from seven tables of the database (`user`, `user group`, `user former group`, `actor`, `revision`, `userindex` and `page`) for a stratified sample by considering two sampling dimensions: the topic of an item and its maturity level in terms of a number of revisions for the following reasons:

Firstly, previous research has shown that Wikidata's coverage depends on the topical domain of the items [65]. We assume that the different coverage might reflect the varying interests of the Wikidata community and that the topic of the item might influence the editing behavior. We wanted to capture these possible differences and

[10] The different ways Wikidata could be accessed are: https://www.wikidata.org/wiki/Wikidata:Data_access

[11] https://www.wikidata.org/wiki/Special:Search

[12] https://www.wikidata.org/wiki/Special:EntityData where the format of the entity data can be specified by appending .json, .rdf, .ttl, .nt or .jsonld extensions to the data URL.

[13] Wikidata Query Service (WDQS) is a public service that provides access to Wikidata's KG via an SPARQL endpoint since September 2015 [25]. Originally, Wikidata contents are not stored in RDF format. For this reason, data is mapped from its internal representation to RDF format and exported as RDF. These RDF data are stored in a graph database named BlazeGraph to be queried by the SPARQL query service, which is an RDF query language.

[14] The copies of Wikidata content is available for download in various formats such as JSON, RDF, or XML formats on the Wikidata website. Using dumps is best when expecting a large set of results that could affect query performance or cause timeouts.

[15] Toolforge is a hosting environment for Wikimedia-related software. Further information is available at: https://tools.wmflabs.org [Accessed: 06.12.2019]

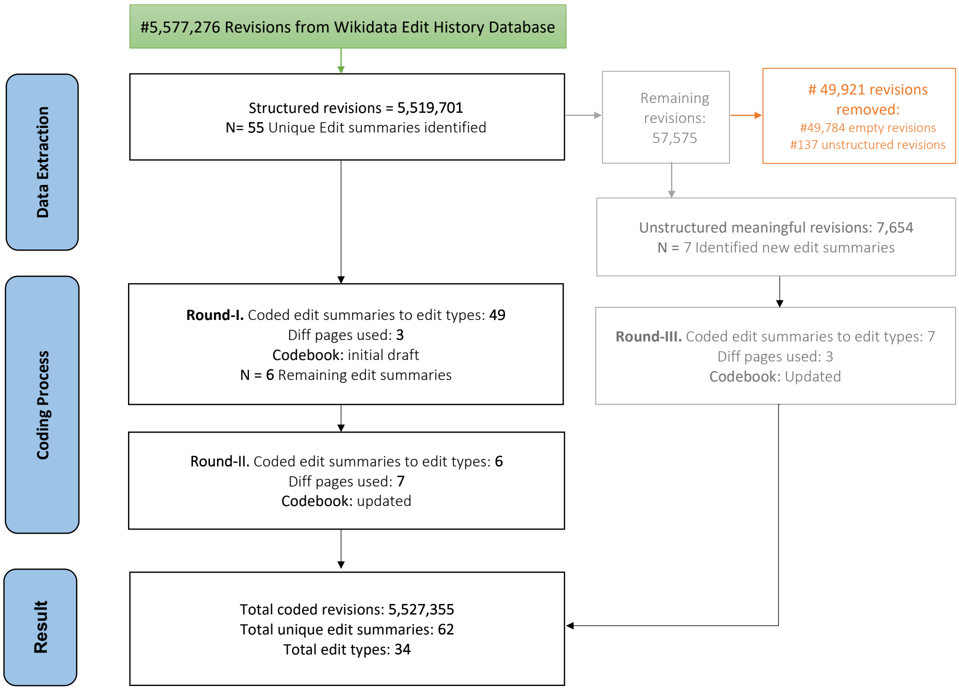

#5,577,276 Revisions from Wikidata Edit History Database
Data Extraction
Structured revisions = 5,519,701
N= 55 Unique Edit summaries identified
Remaining revisions: 57,575
# 49,921 revisions removed:
#49,784 empty revisions
#137 unstructured revisions
Unstructured meaningful revisions: 7,654
N = 7 Identified new edit summaries
Coding Process
Round-I. Coded edit summaries to edit types: 49
Diff pages used: 3
Codebook: initial draft
N = 6 Remaining edit summaries
Round-II. Coded edit summaries to edit types: 6
Diff pages used: 7
Codebook: updated
Round-III. Coded edit summaries to edit types: 7
Diff pages used: 3
Codebook: Updated
Result
Total coded revisions: 5,527,355
Total unique edit summaries: 62
Total edit types: 34

Figure 5.3: An overview of *Wikidata Edit History Dataset* pre-processing phase. (Note: This diagram is inspired by the PRISMA flow diagram [223].)

used the topical domains – people, media, organizations, geography, and biology – defined by Färber et al. [65] for selecting items on Wikidata.[16]

These domains represent the terminological knowledge of Wikidata and refer to ten classes which are further divided into 24 subclasses. By using the Wikidata Query Service[17], we queried for each domain, i.e., class or subclass, the belonging items with their unique identifier (Q-id). Based on that, we used Toolforge[18] to collect all revisions of these items. This resulted in 2,243,390 distinct items, of which the topical breadth is shown in the first fifth columns of Table A.1. The highest number of items (526,898) belongs to the class *Mountain* of domain *Geography*, while the class *Grass* of domain *Biology* contributes with the least number of items (6).

Secondly, we account for the different maturity levels of the items by adopting the collection strategy of Arazy et al. [13]. We used four maturity strata which represent the development stages of items in Wikidata: *inception* with $[1; 10]$ revisions, *creation* with $[11; 100]$ revisions, *growth* with $[101; 1,000]$ revisions, and $[1,001; \infty]$ revisions refer to the *maturity* stage. An overview of these maturity levels is given in the last four columns of Table A.1. The majority of items (1,818,170) reside in the creation stage, and the least number of items (291) exist in the maturity stage.

[16] The authors list the topics in the form of classes in one of the files ("m_cPop.xlsx") and provide this information on a website http://km.aifb.kit.edu/sites/knowledge-graph-comparison/ [Accessed: 06.12.2019].

[17] https://www.mediawiki.org/wiki/Wikidata_Query_Service

[18] https://tools.wmflabs.org

Based on the 24 subclasses and the four maturity levels, we randomly selected from each of the resulting 96 cells (cp. Table A.1), 15 percent of the items (at least one, if available). This resulted in a sample data set with 224,343 items, with 5,577,276 revisions that are created by 37,455 unique contributors.

5.3.3 Defining user groups

Of the 37,455 unique users who contributed to Wikidata in our sample, 12,728 (0.33%) edited Wikidata anonymously. We identified from the non-anonymous contributions three different user groups that consider the algorithmic support as described next. The user groups, the number of people represented, the number of revisions, and the number of edited items in each group are shown in Table 6.6.

At first, we identified all registered users in the data set. For this, we again used the Wikidata database from Toolforge, as described in the previous section. Regular contributions represent the largest user group in our sample with the largest number of revisions. However, the contributions of these registered contributors seem to be primarily focused and concern 30% of the items only in the dataset.

In the second step, we identified all semi-automated contributors, i.e., contributors that use tools for supporting their editing activities. Many of these tools leave traces in the comment section of the revisions. The Quickstatement-Tool[19], for example, which is being used for batch edits on Wikidata, adds automatically the tag `#quickstatements` to each revision. We used a list provided by Sarasua et al. [270] that contains various of these tags and queried the `change_tag` table on Toolforge to identify such tool usages. Although a small number of people have used tools (3%), their contributions affected 34% of all the items in the sample.

In the third step, we identified all algorithmic users, i.e., the bots. For this identification step, we used four sources. First, we labeled all user accounts as bots if they are in the user group ”bot”[20]. In a second step, we identified all bots that have a related requests-for-permissions (RfP) page[21]. For this, we used a data set provided by Farda-Sarbas et al. [67]. In the third step, we identified all contributors in our data set, where the user name contains the word *bot*. In the last step, we checked all available bot lists on Wikidata, to identify further bots: Bots with a botflag[22], bots without a botflag[23], extension bots[24] and list of bots[25]. We matched our user data with all these sources to ensure that we identify all bots. In the data set, bots are the smallest user group, however, they are responsible for a quarter of all revisions and their edits affect, as semi-automatic contributions, 34% of all items.

[19] Further information is given at https://www.wikidata.org/wiki/Help:QuickStatements.

[20] We checked the users account against both `user_groups` which contains current bots and `user_former_groups` which contains the user accounts which were earlier in group bot but are no more in that group.

[21] The formal way of requesting a bot flag is described on the page `https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot` [Accessed: 25.11.2019].

[22] https://www.wikidata.org/wiki/Category:Bots_with_botflag [Accessed: 25.11.2019].

[23] https://www.wikidata.org/wiki/Category:Bots_without_botflag [Accessed: 25.11.2019].

[24] Information on Extension Bots: https://www.wikidata.org/wiki/Category:Extension_bots [Accessed: 25.11.2019].

[25] List of bots: https://www.wikidata.org/wikiWikidata:List_of_bots [Accessed: 25.11.2019].

5.3.4 Specifying Edit Types Based on Revisions

Each revision has an edit comment. We can use these comments to contextualize the nature of a revision. By adopting the trace ethnography methodology suggested by Geiger et al. [92], we used these comments to explore the contributions and their interdependencies in more detail.

Manual inspection showed that the majority of comments are semi-structured; for example, *"/*wbsetsitelink-set:1|dewiki*/Japan"*, consisting of a structured and an unstructured part. The structured part is between the /∗ and : symbols; we call this part the *edit summary*. By parsing all revisions, we identified 55 distinct edit summaries.[26] An *edit summary* gives a first indication of the actual type of an edit (cf. Table 5.3.5). However, in some cases the edit summary is ambiguous, for example, comments *wbeditentity-update* or *wblinktitles-connect* are not self-explanatory.

Thus, we decided to classify these edit summaries manually within a three-round process. As a result of this process, we assigned to each edit summary an *edit type*. We adopted an a priori coding scheme based on a Müller-Birn et al. [206] study, which investigates the contextual information of the edit comments in detail.

Table 5.2: Overview of the Terminology Used for Classification of Edits in Wikidata.

Terminology	**Definition**	**Example**
Edit summary	Structured part retrieved from the comment between */** and *:* mostly with a *wb* prefix.	wbsetsitelink-set
Activity type	Defines the type of change carried on the edit focus, can come at the end of the edit summary after a hyphen (e.g., wbsetdescription-add) or between the *wb* prefix and edit focus.	set
Edit focus	Defines the focus or target of the change that is being carried out on a page (item, property).	sitelink
Edit type	Actual activity that is carried out on a page (item, property).	set sitelink

In the first round, four of the authors individually assigned the *edit types* to these 55 edit summaries based on three random diff pages[27] for each edit summary. The coding process consisted of additional dimensions, as described in the next section on analyzing edit summaries by edit type. At the end of the first round, we compared the individually coded results and found several cases that we coded differently. We discussed these ambiguous cases to reach a common understanding. This discussion resulted in 49 edit summaries and we agreed on the edit types. At this point, we had 49 edit summaries mapped to 33 edit types. Table A.3 provides more details on the difference between the number of edit summaries and edit types. However, for example, the *wbsetentity* and *wbeditentity* edit summaries were still ambiguous, so in a second round we increased the number of exemplary revisions, i.e., diff pages. Based on these examples, we could assign an edit type to the remaining six edit

[26] As part of this manual inspection, we reviewed the documentation of the MediaWiki software (https://www.wikidata.org/w/api.php?action=help&modules=main [Accessed: 2020-01-07]) to get ourselves familiar with the used wording in the edit summaries. We also used diff pages for understanding the difference caused by edits).

[27] diff pages show the difference between two revisions of an article (https://meta.wikimedia.org/wiki/Help:Diff [Accessed: 2020-01-01]

summaries. Based on these 33 edit types, we classified 5,519,701 revisions out of a total of 5,577,276, so 57,575 revisions (0.01%) were defined as unstructured.

We were curious about these unstructured comments and did a further manual inspection. It turned out that 49,784 revisions had empty comments[28]. Looking at the remaining 7,791, we realized that we could identify existing edit summaries in these revisions and could even identify new edit summaries[29]. We found, for example, ”restore” and ”undo” in the comments, which are already present in our codebook. In other cases, for example, ”revert” (also ”reverting”) or ”clean” (also ”cleaning”, ”clean up”), we defined new edit summaries and one new edit type, i.e., *protect item*. As a result of this process, 7,654 additional revisions, most of which are reverts, could be classified based on our codebook[30], which finally consisted of 62 edit summaries from which we specified 34 edit types. Table A.3 provides details on the mapping of edit summaries to their respective edit types.

In the end, 137 revisions were finally classified as unstructured. We removed these 137 and the 49,784 empty revisions from our data set. After this step, the sample comprises 5,527,355 revisions, which we use for our further analysis.

This data is openly available[31] in the form of the *Wikidata Revision History Dataset*[68]. Tables A.2 and A.1 in the Appendix provide a statistical overview of the dataset.

5.3.5 Analysis of Edit Summaries to determine Edit Types

Following the approach by Müller-Birn et al. [206], we defined edit types as verb-noun pairs (e.g., *update item*) which show the activity (i.e.,*update*) performed on data model (i.e., *item*) that we call edit focus. We first, identified activity type and edit focus from each edit summary, and then combined them together to obtain the edit type. The only exception is with *revert* which does not have an edit focus. Table 5.3 contains our definition of the activities in this study. In Table A.2 we provide an overview of the edit types aggregated by edit focus. Statement, claim, qualifier, rank, and reference, are aggregated as edit focus *Statement*, for instance.

Table 5.3: Definition of activities.

Activity	**Definition**
Add	Add something new to an already existing item
Create	Creating new item
Merge	Merging items
Protect	To keep a page or item from further changes, locking to avoid edits
Remove	Remove something from an already existing item
Revert	Undo an edit
Set	Update or Add new to an already existing item
Update	Add and Remove something from an already existing item

There are many edit summaries that refer to the same types of edits, thus, more than one edit summary could be assigned to one edit type. For instance, both

[28] 12 of these empty comments were deleted comments, and the remaining 49,772 were just empty.

[29] We extracted the first words of the comments as edit summaries.

[30] The codebook is available in Table A.3 in the Appendix.

[31] *Wikidata Revision History Dataset* is available on FU Berlin Primo at http://dx.doi.org/10.17169/refubium-40243

clientsitelink-remove and *wbsetsitelink-remove* is assigned to edit type *remove sitelink*. Table A.3 shows the mapping of the edit summaries to edit types.

5.4 Summary

In this chapter, we have outlined the data sources and methods used to address our research questions. Specifically, we employed content analysis as our method to investigate the impact of bots on diversity in Wikidata. To obtain the necessary data for analysis, we created two datasets: the *Wikidata-Requests-for-Permissions Dataset* and *Wikidata Revision History Dataset*. These datasets were carefully constructed from different sources within Wikidata, enabling us to gain insights into the role of bots and their potential influence on diversity in Wikidata.

Given the limited research available on bots in the context of Wikidata, we embarked on a comprehensive exploration of the topic. This involved thoroughly understanding what bots are and their activities within Wikidata, allowing us to trace any potential impact they may have on the diversity status of the platform.

The next chapter will focus on the analysis of these datasets, where we will delve into the details of each data source, explain why they were selected, and present the results of the data analysis. The purpose of this chapter, however, was to provide a clear overview of the methodology employed in developing these datasets, ensuring the reproducibility of our work.

Chapter 6

BOTS, DIVERSITY & WIKIDATA

As we have observed in the previous chapters, the contributing community of Wikidata consists of both bots and humans. Bots, due to their capability to execute high-speed edits, are responsible for a significant portion of the activities within Wikidata. However, this user group, despite its substantial contribution, remains relatively understudied. There is still much to learn about the editing behavior of bots and the impact of their edits on Wikidata, particularly in relation to diversity, which is an important goal of Wikidata. The imbalanced data coverage in Wikidata domains further emphasizes the need to investigate the role of bots and their potential influence on diversity. By gaining a deeper understanding of bot behavior and their contributions, we can better comprehend their impact on the overall diversity of Wikidata.

6.1 Bots in a Knowledge Base Context

The term bot is considered an alternative word to software script, software agent, and robot all of which primarily automate and speed up tasks. Bots are defined as software programs that automate tasks, usually repetitive or routine tasks that humans consider time-consuming and tedious (e.g., [88, 89, 206, 302]). They are operated and controlled by humans.

A visible area where bots are active is the collaboratively developed systems. Very different bots populate shared contribution communities: These bots collect information, execute functions independently, create content, or mimic humans. The effects of bots on collaborative content creation systems and our society are increasingly being discussed, for example, when influencing voting behaviour [73] or imitating human behaviour [172]. Bots have been used in Wikipedia from early on[1]. Wikidata's community has profited from these experiences when handling their bots. We

[1] https://en.wikipedia.org/wiki/Wikipedia:History_of_Wikipedia_bots#"rambot"_and_other_small-town_bots.

review, therefore, existing insights into the bot community on Wikipedia and building on that, highlight research on Wikidata that considers bot activity. Looking at bots from a social lens reveals the reasons why bots are created in Wikidata. Additionally, we can see what type of edits the community shares with bots, what type of data and from which sources they import through bots, and whether the community is open and allows anyone to operate bots or apply hard conditions so that only a small number of operators can make to run their bot in Wikidata. Furthermore, investigating the editing history of Wikidata reveals whether bot editing patterns differ noticeably from those of human users, indicating their potential impact on diversity within Wikidata. This analysis helps us ascertain if bots are responsible for the observed low diversity status of Wikidata domains and classes.

In this chapter[2], we focus on bots in detail, from how and why they come into being to what they actually do. We reflect on bots' social organization to understand them from the community perspective. We then compare them with human users to find out bot editing behavior and if both groups edit in a similar manner and, then, investigate the potential of the bot user group to impact diversity in Wikidata.

We begin with a glance into the usage of bots in Wikimedia projects, i.e., Wikipedia and Wikidata.

6.1.1 Bots in Wikipedia

Wikipedia has developed a stable and increasingly active bot community over the years, although, bots were not widely accepted and trusted in the beginning [175]. Halfaker and Riedl distinguishes four types of bots in Wikipedia [112]: (1) Bots that transfer data from public databases into articles, (2) bots that monitor and curate articles, (3) bots that extend the existing software functionality of the underlying Wikipedia MediaWiki software and (4) bots that protect against vandalism. Similarly, based on a study of the German bot community, Müller-Birn et al. differentiate bot responsibilities into content maintenance (e.g., updating templates, creating archive pages, detecting spam), bot coordinating (e.g., sending notifications or creating task lists), and community support (e.g., welcoming new users and counting votes) [204].

The majority of research focuses on bots that protect against vandalism. Geiger and Ribes for example, investigated the process of fighting vandalism in Wikipedia by using trace ethnography [92]. They show how human editors and bots work together to fight vandalism in Wikipedia. They conjecture that such distribution of concerns to human and algorithmic editors might change the moral order in Wikipedia. Halfaker et al. show how a distributed cognitive network of social and algorithmic actors works efficiently together to detect and revert vandalism on Wikipedia [113]. In another study, Geiger and Halfaker investigated the impact of a counter-vandalism bot downtime on the quality control network of Wikipedia and found that during this downtime, the quality control network performed slower but was still effective [90].

Another piece of research focuses on how tools like these alter the dynamics of editing and user engagement. Geiger shows how a subtle yet existing social norm was transformed into a technological participant in a contentious manner [88]. He refers to the example of the HagermanBot, which has been implemented to sign

[2] Parts of this chapter are already published in [67].

unsigned discussion entries in Wikipedia. Halfaker and Riedl show that bots are not only responsible for the enforcement of existing guidelines on a larger scale but also that their activities can have unexpected effects. The number of reverts of newcomers' edits, for example, has elevated, while (surprisingly) the quality of those edits has stayed almost constant. The authors show that editors increasingly apply algorithmic tools for monitoring the edits of newcomers. In 2010, 40 percent of rejections of newcomers' contributions were based on this algorithmic tool [112]. This contradicts attempts of the community to engage more new editors. Moreover, Geiger and Halfaker defined bots as "assemblages of code and a human develop" and show that the bot activity is well aligned with Wikipedia's policy environment [91].

The research suggests that bots are more critical to the success of the Wikipedia project than expected previously, despite the reluctance of the Wikipedia community to allow bots at the beginning [175]. Bots have a significant role in maintaining this text-based KB, especially in fighting vandalism. As bots in Wikidata have their roots in Wikipedia, we expect to see similarities between bots in both peer production systems - Wikipedia and Wikidata. Before we look closer to see if the same areas of use of bot activities emerge from Wikidata, we give an overview of the existing insights into the bot community in Wikidata.

6.1.2 Bots in Wikidata

Wikidata inherited bots from its sister project Wikipedia and bots started editing Wikidata with its launch by linking Wikidata item pages to their respective Wikipedia language pages. The current research on Wikidata bots shows that bots perform most of the edits in Wikidata [298, 206]. Steiner, in his research, aims to understand the editing distribution of editors on Wikidata and Wikipedia. He provides a web application to observe real-time edit activity on Wikidata for bots and logged-in and anonymous users. The study shows that the number of bots vs. the number of edits has grown in a linear form and most of Wikidata's edits, i.e., 88%, account for bot edits [298].

Müller-Birn et al. can confirm these insights in a later study by studying the community editing patterns of Wikidata through a cluster analysis of contributors' editing activities. They determine six editing patterns of the participating community (i.e., reference editor, item creator, item editor, item expert, property editor, and property engineer) and show that bots are responsible for simpler editing patterns, such as creating items, editing items, statements, or sitelinks [206].

Further studies focus on how bot edits contribute to data quality in Wikidata. In one study on Wikidata's external references, Piscopo et al. find that the diversity of external sources in bot edits is lower than in human edits [238]. In another study, Piscopo et al. explore the influence of bots and human (registered and anonymous) contributions on the quality of data in Wikidata. The research shows that equal contributions of humans and bots have a positive impact on data quality, while more anonymous edits lead to a lower quality [241].

Hall et al. analyzed these anonymous edits on Wikidata to detect bots in this group that had not previously been identified in Wikidata. The study shows that two to three percent of the contributions (more than 1 million edits), considered as human contributions previously came from unidentified bots. They emphasize that it might be a concerning issue for Wikidata and all projects relying on these data. Even if vandalism causes a small portion of these edits, the damaging effect on Wikidata might be high [114].

This danger is reflected by Piscopo and Simperl, who highlights significant challenges for Wikidata that might endanger its sustainability: Namely, a lack of quality control because of the large amount of data added by bots, a lack of diversity because of the usage of a few sources only and existing threats to user participation because of bot usage [236].

Existing research focuses primarily on the activity levels of bots in Wikidata based on their edits. Some work (e.g., [236]) conveys the impression that bots are independent of humans. However, this ignores the fact that humans operate bots. The community officially grants most bot activities in a well-defined process.

It is visible from the literature that bots are, so far, studied mainly from their activity angle both in Wikipedia and Wikidata. In Wikipedia, bots are used primarily for quality assurance tasks, i.e., vandalism detection and maintenance tasks, for example, removing/replacing templates on articles, while Wikidata's bots are performing tasks mostly related to content editing. A possible reason could be the structured nature of Wikidata content, which is less challenging to deal with than the unstructured data in Wikipedia. It is intriguing to delve into the content editing activities that bots have requested the most in Wikidata and which have been approved by the community. While the increased content added through bots seems to improve not just data diversity but also data quality in terms of completeness, it raises the question of whether this data is sourced and contributes to data quality in terms of trustworthiness. Additionally, it piques interest in understanding how these aspects relate to data diversity. However, these inquiries fall beyond the scope of this study. We delve deeper into the Request for Permissions (RfP) process for bots, outlining which tasks are deemed useful by Wikidata's community for bots, and in which cases the community does not support a bot request.

6.2 Bots Social Organization in Wikidata

In the Wikidata context, bots are separate user accounts with special privileges that are run by human users, as described in Section 2.1.2. Bots need to go through a defined process to gain bot privileges which we explain in detail in the next section.

6.2.1 Approving Bot Tasks

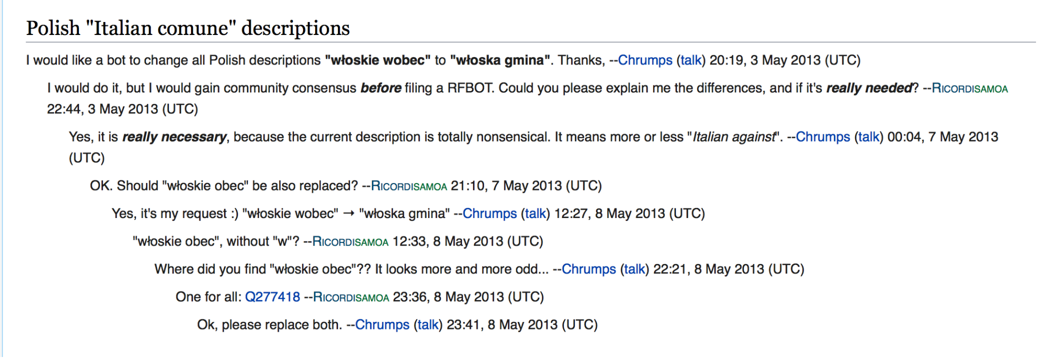
In this section, we describe the bot approval process on Wikidata. The explanation of our data collection and the details of the classification process are present in Section 5.2. Here, we give an overview of the resulting dataset, the findings of this study, and our implications.

6.2.1.1 Bot Approval Process in Wikidata

Requests for permissions (RfP) pages are a formal way of requesting bot rights for an account in Wikidata, where the decisions are based on community consensus.

Building on the experiences of the Wikipedia community, Wikidata has had a well-defined policy system for bots almost from the beginning (November 2012)[3]. Except for cases like editing Wikidata's sandbox[4] or their own user pages, every (semi-) automatic task carried out by a bot needs to be approved by the community. It means that before operators can run their bots on Wikidata, they have to open an RfP for their bot[5]. The RfP is, thus, the formal way of requesting bot rights for an account where the decisions on RfPs are based on the community consensus.

An RfP is caused by either an editor's need or by a community request as can be seen in Figure 6.1.



Polish "Italian comune" descriptions

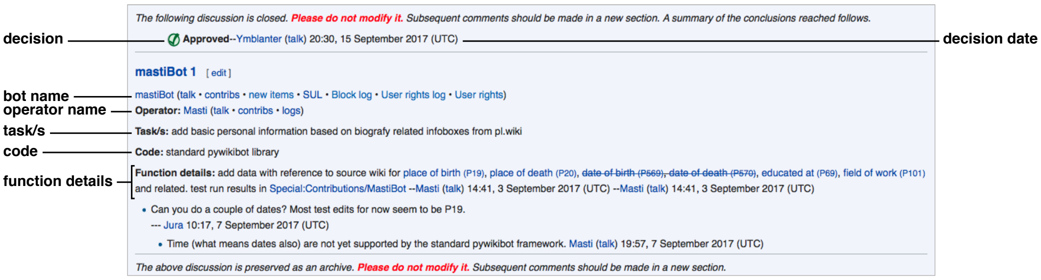
I would like a bot to change all Polish descriptions **"włoskie wobec"** to **"włoska gmina"**. Thanks, --Chrumps (talk) 20:19, 3 May 2013 (UTC)

> I would do it, but I would gain community consensus ***before*** filing a RFBOT. Could you please explain me the differences, and if it's ***really needed***? --Ricordisamoa 22:44, 3 May 2013 (UTC)
>
> > Yes, it is ***really necessary***, because the current description is totally nonsensical. It means more or less "*Italian against*". --Chrumps (talk) 00:04, 7 May 2013 (UTC)
> >
> > > OK. Should "włoskie obec" be also replaced? --Ricordisamoa 21:10, 7 May 2013 (UTC)
> > >
> > > > Yes, it's my request :) "włoskie wobec" → "włoska gmina" --Chrumps (talk) 12:27, 8 May 2013 (UTC)
> > > >
> > > > > "włoskie obec", without "w"? --Ricordisamoa 12:33, 8 May 2013 (UTC)
> > > > >
> > > > > > Where did you find "włoskie obec"?? It looks more and more odd... --Chrumps (talk) 22:21, 8 May 2013 (UTC)
> > > > > >
> > > > > > > One for all: Q277418 --Ricordisamoa 23:36, 8 May 2013 (UTC)
> > > > > > >
> > > > > > > > Ok, please replace both. --Chrumps (talk) 23:41, 8 May 2013 (UTC)

Figure 6.1: Example of a request from the Wikidata community for a bot

Such a bot request is well documented and available to all interested community members on Wikidata.

Figure 6.2 shows a typical RfP page. It consists of a bot name[6], an operator name (bot owner), tasks (a brief description of what the bot intends to do), a link to the source code, function details (a detailed description of bot tasks and sources of imported data), and the final decision (RfP approved or not approved with a reason).



*The following discussion is closed. **Please do not modify it.** Subsequent comments should be made in a new section. A summary of the conclusions reached follows.*

✓ **Approved**--Ymblanter (talk) 20:30, 15 September 2017 (UTC)

mastiBot 1 [edit]

mastiBot (talk • contribs • new items • SUL • Block log • User rights log • User rights)

Operator: Masti (talk • contribs • logs)

Task/s: add basic personal information based on biografy related infoboxes from pl.wiki

Code: standard pywikibot library

Function details: add data with reference to source wiki for place of birth (P19), place of death (P20), ~~date of birth (P569), date of death (P570)~~, educated at (P69), field of work (P101) and related. test run results in Special:Contributions/MastiBot --Masti (talk) 14:41, 3 September 2017 (UTC) --Masti (talk) 14:41, 3 September 2017 (UTC)

- Can you do a couple of dates? Most test edits for now seem to be P19.
  --- Jura 10:17, 7 September 2017 (UTC)
  - Time (what means dates also) are not yet supported by the standard pywikibot framework. Masti (talk) 19:57, 7 September 2017 (UTC)

*The above discussion is preserved as an archive. **Please do not modify it.** Subsequent comments should be made in a new section.*

Figure 6.2: Example of a Wikidata requests-for-permissions (RfP) page.

[3] https://www.wikidata.org/w/index.php?title=Wikidata:Bots&oldid=549166.

[4] Wikidata's sandbox can be used for test edits: https://www.wikidata.org/wiki/Wikidata:Sandbox.

[5] All open requests are available at www.wikidata.org/wiki/Wikidata:Request_for_permissions/Bot.

[6] According to the bot policy, bot accounts are recommended to have the word bot in their names.

Bot owners have to use a template for providing all information for the decision-making process and need to clarify questions regarding their requests during the decision-making process. They also often have to provide a number of test edits (50 to 250 edits). The community handles RfPs in accordance with the Bot Approval Process[7]. The community discusses whether the task requested is needed and whether the implementation of the task functions properly. After the decision has been made, an administrator or a bureaucrat closes the request, if approved, a bureaucrat will flag the account, and if not, the request is closed stating the reason for the unsuccessful request.

After a successful request, each bot operator has to list the task the bot performs with a link to the RfP on its user page. However, the bot operator is required to open a new RfP if there is a substantial change in the tasks the bot performs. Consequently, bots can have multiple requests which are shown on the user page. Furthermore, in special cases, a bot is also allowed to carry out administrative tasks, such as blocking users or deleting or protecting pages. In this case, the operator needs to apply for both, the RfP and the administrator status[8].

To learn more about bots, we analyze these RfP pages by developing a dataset containing these RfPs. The details of the methodology for data collection and coding is present in Section 5.2. Next, we provide an overview of our developed dataset from the bot requests i.e., *Wikidata-Requests-for-Permissions Dataset*.

6.2.1.2 Task Approval Data

We collected 683 distinct requests from Wikidata as explained in Section 5.2.2; however, two of them are no longer available and, so, we excluded them. Of the resulting 681 requests, 600 requests (88%) were successful, and 81 (12%) were unsuccessful.

An average of five people participated in an approval request, for example, by taking part in the discussion or stating the final decision. These people accounted for an average of slightly above ten edits for each request.

Based on the requests collected and processing them through the classification scheme in Section 5.2.3, we identified 391 distinct bots and 366 distinct bot operators on Wikidata (cf. Table 6.1). Some operators applied for more than one task for their bot in one request (e.g., update label, update description, update alias), with five being the highest number. Furthermore, bot owners can operate more than one bot. The majority of operators (319 editors) have only one bot, while three editors were running three bots and there was even one operator managing seven bots simultaneously. Similarly, one bot can also be operated by more than one user, for example, three editors were managing the ProteinBoxBot together.

6.2.2 Approved Requests

Exploring RfPs is a way to understand the intention of the Wikidata community behind the usage of bots. Since bots are special accounts run by operators, we would like to see how easy or difficult it is to create a bot, what types of tasks are allowed/disallowed to be automated, and which data sources are the most popular among

[7] A detailed description is available at www.wikidata.org/wiki/Wikidata:Bots.

[8] Further information can be found at https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Administrator.

Table 6.1: Overview on task requests ($n = 683$, number of all task requests). Source: *Wikidata-Requests-for-Permissions Dataset.*

Decision Result	Requests	No. of Bots	Operators
Successful	600	323	299
Unsuccessful	81	68	67
Both	681	391	366

bots in Wikidata. This already provides a glance into the diversity of bot edits, i.e., how diverse tasks bots have requested and from how diverse sources they intend to import data.

In this part, we show the types of activities which were approved, the sources they used for editing, and the relationship between bots and their operators. We focus on three aspects of the tasks: The activity focus, the activity type, and the origin of the data used by the bots.

6.2.2.1 Activity Type and Focus

The activity focus considers what the bots have planned to edit in Wikidata. In other words, are bot requests concentrated on certain activities or community allows automation of all diverse activities human users perform in Wikidata?

As it is already known, within the Wikidata data model, items are composed of various components on which we can engage in an activity referred to as edit focus. Table 6.2 is created from the task requests and provides a view of these bot requests aggregated by edit focus. The different categories in Table 6.2 represent the Wikidata item including the different components of an item based on Wikidata data model 2.1.3. Additionally, we have a category that contains requested tasks that are beyond the item scope.

Table 6.2 shows that a high number of task requests aim to edit items in Wikidata. The number of approved task requests reveals that the majority of bots seem to focus on adding data to statements in general and claims more specifically.

Figure 6.3 shows the distribution of approved tasks into activity types and edit focuses. Again, the majority of task requests deal with the addition of data.

Furthermore, there are 30 requests dealing with community concerns regarding maintenance activities, such as archiving pages/ sections, moving pages, reverting edits, removing duplicates, and generating statistical reports as some examples.

In addition to this, there are 24 requests concluded as unknown tasks, half of which are written extremely briefly[9] or in an unclear way[10], so that it was difficult for us to identify their potential tasks. Some other cases can only be classified to a specific task based on particular assumptions, for instance, ShBot[11] uses a Bot-Framework QuickStatements[12] to batch edits, which shows a great possibility of importing data to statements. However, this tool could also be used to remove statements or import

[9] e.g., www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Glavkos_bot

[10] e.g., www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/BotMultichillT

[11] www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/ShBot.

[12] Further information are available at meta.wikimedia.org/wiki/QuickStatements.

Table 6.2: Task requests categorized into edit focus, activity type, and the request result (approved/denied) ($n = 681$, number of all task requests). Source: *Wikidata-Requests-for-Permissions Dataset.*

Edit focus	**Activity**	**Approved**	**Denied**
Item	add	85	12
	update	40	3
	merge	6	1
	mark-deleted	1	0
Term	add	22	5
	update	11	0
Label	add	31	3
	update	15	1
	delete	1	0
Description	add	33	5
	update	10	1
Alias	add	4	0
	update	1	0
	delete	1	1
Statement	add	141	19
	update	15	0
	delete	1	1
Claim	add	161	18
	update	33	7
	delete	13	3
Qualifier	add	7	0
Reference	add	9	3
	update	4	0
	delete	1	0
Rank	update	1	0
Sitelink	add	52	3
	update	23	4
	delete	1	0
	merge	1	0
	revert	1	0
Badge	add	4	0
Page	add	0	1
	update	8	1
	delete	1	0
Community	maintain	25	5
Misc.	−−	26	13

terms, so, there is still a challenge to identify the requests' focus without analyzing the edits. We categorized all tasks abiding strictly by our codebook, thus, all the RfPs which needed assumptions are therefore classified as unknown.

We assumed that the data addition to Wikidata is related primarily to Wikidata's inception, i.e., that the majority of task approval would be more at the beginning of Wikidata's lifetime. However, Figure 6.4 shows that the most often requested activity types in the first six months were ”add” and ”update”. At the end of 2017, these two activity types were often requested again. We found the peak in 2017 for requesting tasks such as adding and updating items unusual and investigated it

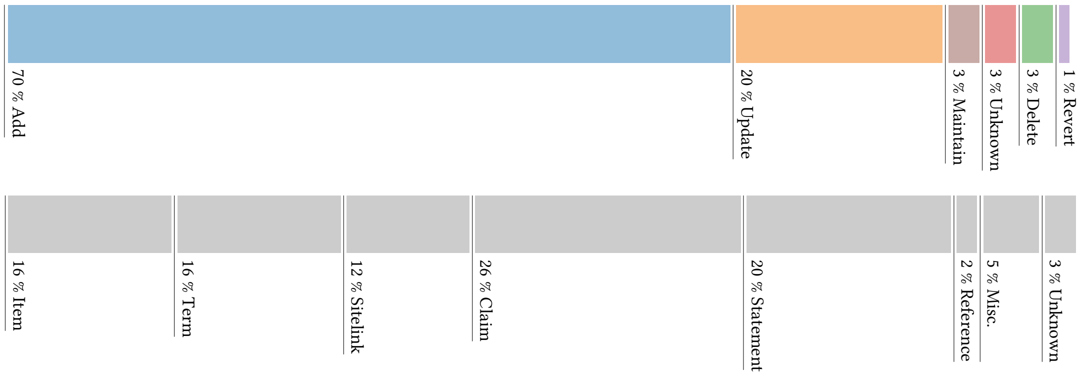


Figure 6.3: Approved task requests organized into activity type (the upper half part) and edit focus (the lower half part)
(n = 794, 262 RfPs are requests with multiple tasks, edit focus *Term* consists of *Alias, Label, and Description.* The edit focus *Miscellaneous* contains the smaller categories, such as *Qualifier*, *Rank*, *Badge*, and *Maintenance related data*). Source: *Wikidata-Requests-for-Permissions Dataset.*

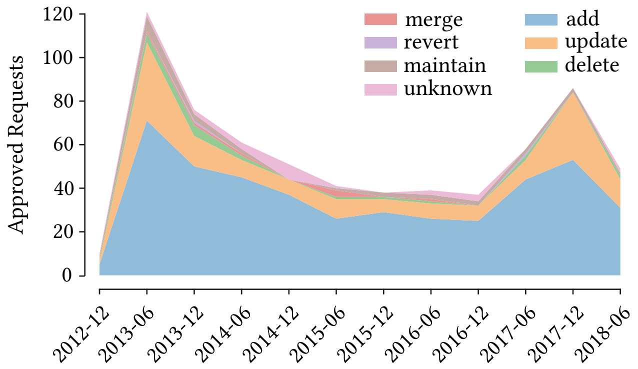


Figure 6.4: Activity types of approved task requests over time ($n = 600$, all 81 unsuccessful task requests were not included). Source: *Wikidata-Requests-for-Permissions Dataset.*

further. Two operators[13] carried out 31 requests in 2017 for importing items from different languages. Since all these requests were permitted, it explains the sudden acceleration of importing and updating tasks shown in the graph.

In the following step, we looked more closely at the sources of the data that bots are supposed to add to Wikidata.

6.2.2.2 Data Origin

In addition to the types of tasks in RfPs, the sources from which data are imported also exist in the RfP. Analysis of data sources can reveal if bots are used to import data from multiple and diverse sources, or if the community allows data import from certain sources only. We tried to identify all sources independently of the focus or activity type in the classification of all task approval pages. We differentiate internal, external, and unknown sources:

[13] Both editors are employed by Wikimedia Sverige and were engaged in a GLAM initiative.

Table 6.3: All internal, the top 10 external sources most used and unknown sources. Source: *Wikidata-Requests-for-Permissions Dataset.*

Origin	Source	Task Requests
Internal (455)	Wikipedia	263
	Wikidata	110
	Wiki-loves-monuments	30
	Wikicommons	13
	Wikisource	5
	Wikivoyage	4
	Wikinews	2
	Wikiquote	2
	WikiPathways	2
	Wikimedia Project	24
External (38)	MusicBrainz	8
	VIAF	7
	OpenStreetMap	4
	DBpedia	4
	Freebase	4
	GRID	3
	GND	2
	GitHub	2
	YouTube	2
	Disease Ontology Project	2
Unknown	Unknown	107

1. Internal: including all sources from the Wikimedia ecosystem,
2. External: All sources outside Wikimedia's ecosystem, and
3. Unknown: All cases where the source of data was a local file/database or not clearly stated.

Table 6.3 shows the number of tasks approved per data source. The majority of approved task requests (454) deal with data from the Wikimedia ecosystem, with Wikipedia providing most of the data. In addition to Wikipedia, the Wiki Loves Monuments (WLM) initiative, is organized worldwide by Wikipedia community members, and Wikimedia projects. The category Wikimedia project refers to requests where the operator provided information on the source being restricted to the Wikimedia projects' scope, however, a specific attribution to one of the projects was not possible.

Another important source refers to Wikidata itself. We found that only a small number of these data are used in maintenance tasks[14]; the majority of requests are for tasks, such as retrieving information from Wikidata, adding descriptions, and updating or adding new labels by translating the existing labels in other languages.

There are 128 task approval requests related to the task of importing external data into Wikidata. The most frequently mentioned external sources are MusicBrainz[15]

[14] Among 104 requests with a data source of Wikidata, there are 16 requests holding tasks of maintenance, such as archive discussions or update database reports.

[15] MusicBrainz is an open and publicly available music encyclopedia that collects music metadata, available at www.musicbrainz.org.

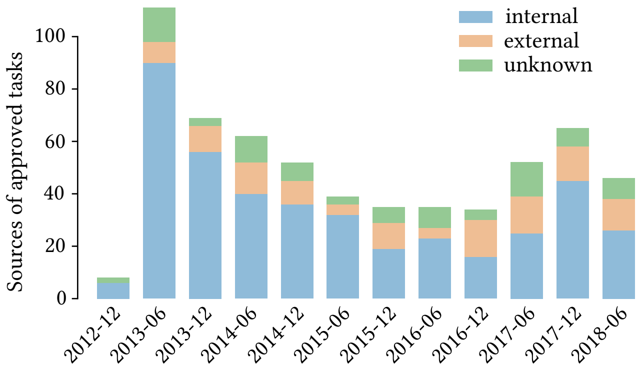


Figure 6.5: Sources of approved task requests over time ($n = 600$, all 81 unsuccessful task requests were not included). Source: *Wikidata-Requests-for-Permissions Dataset*.

and VIAF[16] (cf. Table 6.3). Both sources are mainly used to add identifiers to Wikidata items. Other external sources that are used to import data into Wikidata are OpenStreetMap[17], DBpedia[18], and Freebase[19].

As shown in Figure 6.5, the internal data sources have remained those most often mentioned in the task requests. External sources and unknown sources are on the same level and external sources show a slight increase over time.

Most data from Wikipedia comes from 43 different language versions (cf. Figure 6.6) with English, French, German, Russian, and Persian being the top five languages. The results show that bots' contribution to certain languages had made these languages more prominent than others in Wikidata.

There is a total of 109 RfPs in total with unknown sources; 85 of them were approved; some of these requests mentioned a local file or database as a data source. Other RfPs are not adding data to Wikidata, but are, for example, moving pages.

Table 6.4: Top 5 most edited RfPs closed as unsuccessful.

Reasons	#Edits	#Editors	Created	Ref.
Data source, introduced errors	88	10	2018-03	[335]
Bot name, bot edits, bot performance	43	6	2015-12	[333]
Automatic adding of implicit statements	41	21	2013-04	[332]
Support from less active users, duplicate task	32	16	2018-03	[336]
Bot name, conflict with users, running without flag	32	11	2014-10	[334]

[16] VIAF is an international authority file that stands for Virtual International Authority File, available at www.viaf.org.

[17] OpenStreetMap is an open-license map of the world, available at www.openstreetmap.org.

[18] DBpedia is also a Wikipedia-based KG wiki.dbpedia.org.

[19] Freebase was a collaborative KB launched in 2007 and closed in 2014. All data from Freebase was subsequently imported into Wikidata.

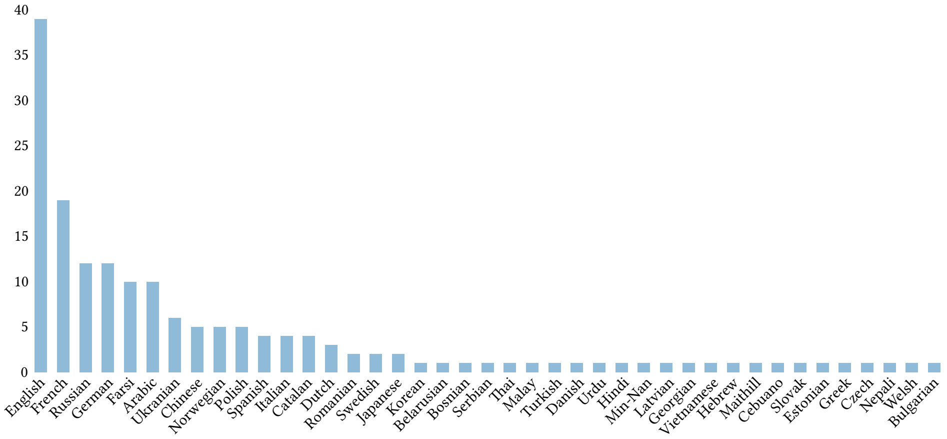


Figure 6.6: Wikipedia language versions used as sources mentioned on task approval pages ($n = 195$, all 488 other requests either did not provide a specific Wikipedia language version or did not have Wikipedia as the source). Source: *Wikidata-Requests-for-Permissions Dataset*.

6.2.3 Disputed Requests

There are 81 RfPs in our dataset which were closed unsuccessfully. We wanted to understand better why the community decided to decline these requests, thus, we investigated the main reasons for all unsuccessful RfPs. Exploring the unsuccessful cases can reveal the main reasons for rejecting a bot request and the openness of the community in dealing with operators. We explored whether there are certain types of tasks that bots are not allowed to perform or whether anyone can operate a bot and can create any kind of bot they want.

Our data show that operators themselves are most commonly responsible for their RfPs being rejected. Most of the time, operators were not responding to questions in a timely manner or did not provide enough information when required. There are only a few cases where RfPs were not approved by the community, i.e., no community consensus was reached or the bot violated the bot policy. In one case[20], for instance, an editor was asking for bot rights for its own user account instead of the one for the bot. This violates bot policy as operators are required to create a new account for their bots and include the word ”bot” in the account name.

Table 6.5 shows the main reasons why tasks were rejected: RfPs which had already been implemented by other bots (duplicate), or RfPs requesting tasks that were not needed, such as one bot which wanted to remove obsolete claims related to the property (P107)[21] but there were no more items associated with P107.

Furthermore, RfPs were closed as unsuccessful because the community could not trust the operators' behaviors. One bot, for example, has requested many RfPs which were not accepted. The community was questioning its editing behavior and required this user to gain community trust first. It can be seen that unsuccessful closing is not only related to the task types that users had requested. MechQuesterBot [337] is another bot that wanted to add descriptions to villages of China in

[20] www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Florentyna.

[21] P107 was a property representing GND type.

Chinese and English. The RfP was not approved after a number of questions were asked by the community, and the community blocked the owner for sock-puppetry finally. BotMultichill [330], on the contrary, was approved after many support and opposing opinions of the community. The operator had claimed to be one of the maintainers of Pywikipedia and had experience operating about a dozen bots and several million edits. The RfP was not for asking for any specific task but instead for playing around with Wikidata. While there were opposing opinions against unconditional blanket approval, there were many supporting opinions by community members who knew the operator, and finally, the bot was approved.

In the following, we describe six RfPs which were edited and discussed most of all (cf. Table 6.4). We defined the number of edits on an RfP page as a proxy for ”challenging” cases.

Phenobot [335] was asking for correcting species names based on the UniProt Taxonomy database. The community doubted the data source (UniProt Taxonomy database) and insisted on more reliable information. The bot has introduced some error reports, therefore, the community had to ask for edit rollbacks. Since then the operator has been inactive for two years which, finally, led to this request being closed unsuccessfully.

WikiLovesESBot [336], which wanted to add statements related to Spanish Municipalities from Spanish Wikipedia, gained continuous support from nine users until a user asked for more active Wikidata users to comment. After this, one active user checked the bot contributions and pointed out that this bot was blocked. Moreover, another active user figured out that this request was asking for duplicate tasks since there was already another user also working on Spanish municipalities. The RfP remained open for about half a year without any operator activities and came to a procedural close.

Another remarkable case is VlsergeyBot [334], which intended to update local dictionaries, transfer data from Infobox to Wikidata, and update properties and reports in Wikidata. At first, the account was named “Secretary” which was against the bot policy. After a community suggestion, it was renamed to “VlsergeyBot.” After that, a user opposed the request and said “Vlsergey‘s activity generates a number of large conflicts with different users in ruwiki” and then another user stated that “maybe you have a personal conflict with Vlsergey” and discussions looked like a personal conflict. The community, then, considered whether this account needed a bot flag, and found that the bot was already running without a bot flag for hours and even broke the speed limit for bots with a flag. The operator promised to update his code by adding restrictions before the next run. Finally, the operator withdrew his request stating that he ”does not need a bot flag that much”.

ImplicatorBot [332] was designed to automatically add implicit statements to Wikidata. Implicit statements are relationships between items that can be inferred from existing relationships. For example, a property of an article that says that a famous mother has a daughter implies that the daughter has a mother, even though it is not yet represented. The task sounds fairly straightforward, but the community was reluctant to approve it because of the possibility of vandalism. Also, changes to properties are still common, and such a bot could cause a tremendous amount of additional work in the case of those changes. Even after the operator shared his code, the discussion continued with the highest number of editors discussing a

Table 6.5: Main reasons for unsuccessful task requests ($n = 81$, number of unsuccessful task requests). Source: *Wikidata-Requests-for-Permissions Dataset.*

Reason	**Description**	**Freq.**
No activity from the operator	Operators are not active or do not respond to the questions regarding their bots.	35
Withdrawn	Operators wanted to close the requested task.	14
Duplicate	Requested tasks are redundant and are already done through other bots or tools.	9
No community consensus	Community opposed the requested tasks.	6
User not trusted	Operator has questionable editing behavior and ability to fix any bugs introduced.	6
Task not needed	Tasks do not need a bot or are not a significant problem for the time being.	5
Don't need a bot flag	The requested tasks can be performed without a bot flag.	3
No info provided	The fields in the RfP page are left empty.	2
Against bot policy	Task is not in line with bot policy.	1

request (i.e., 21), and the maintainer eventually stated that he was ”tired of the bickering. Despite the community's insistence that it needed the task, the operator withdrew it.

Similar to ImplicatorBot, Structor [333] was withdrawn by the operator because the operator was ”tired of bureaucracy.” The bot wanted to add claim, label, and description, particularly to provide structured information about species items. There were many questions raised by a small number of community members (i.e., a total of six different users participated in the discussion), including the bot name, the source of edits, duplicate entries, the edit summary, and edit test performances. The operator was actively responding to all these questions. However, the whole process was so surprisingly complex (the discussions continued for over seven months from the start, that the operator was inactive for a long time, then withdrew his request.

Requests denied because the user is not trusted, are also interesting. ElphiBot[22] [331], for example, which is managed together by three operators, sent a request along with the source code for updating the interwiki links in Wikidata after the category in Arabic Wikipedia has been moved. The community preferred, however, someone who is more responsible and more familiar with the technical aspects of Wikipedia and Wikidata. What concerned the discussants most was that the operators were not capable enough to fix the bugs introduced by their bot. The community did not feel comfortable approving this request because the operators' treatment for fixing bugs was simply to reverse the mistakes manually. The requirement to run a bot should not only be to know how to code but surely also to respond in a timely manner and notify those who can fix the issue.

[22] This RfP has a total of 29 edits and 8 editors contributed in the discussions.

6.2.4 Bots and Wikidata Community

Bots' performance in Wikidata has been very bold, hence, we explored bots from a social lens to know what kind of tasks the community allows to be automated or shared with bots.

Looking at the type and focus of the tasks that were mentioned most often in RfPs, we found that the dominant request type over time is "adding data." This observation suggests that Wikidata's community uses bots to increase the coverage of the provided data which aligns with the vision of the Wikimania Foundation[23]. Updating data is the second highest requested task which shows bots also take part in correcting mistakes or refreshing data according to changes (e.g., updating the sitelinks of moved pages) and contribute to data completeness, which is defined as one data quality dimension [65].

However, we were surprised that there were fewer requests regarding adding or updating references that could support the trustworthiness of data imported by bots. This result supports existing analyses on the edit activities of bots [238]. However, it would be interesting to bring these two lines of research together - the task approval and edit perspective - to understand more closely the development over time. Thus, we can infer that bots are allowed to assist the Wikidata community in increasing data coverage, however, this is less visible from the source coverage angle. In comparison to Wikipedia, where bots perform primarily maintenance tasks, bot requests in Wikidata concentrate mainly on the content, i.e., data perspective.

The majority of data sources mentioned in RfPs come from inside Wikimedia projects, mainly Wikipedia (cf. Section 6.2.2.2). This observation implies that Wikidata is on its way to serving as the structured data source for Wikimedia projects, one of the main goals for the development of Wikidata. Among the five most commonly used language versions of Wikipedia, namely English, French, Russian, German, and Persian (cf. Figure 6.6), the first three languages exhibit the prevalence of Western knowledge in bot imports. This observation suggests that bots might play a role in the imbalanced language coverage in Wikidata, aligning with earlier findings that highlighted the dominance of Western languages in Wikidata [142]. Consequently, we can infer that the better coverage of these languages compared to others could be attributed to the presence of bots, as language coverage is not necessarily linked to the number of speakers [142]. This further supports our assumption that bots possess the potential to influence data balance within Wikidata.

We can see from the external sources that bots use these sources mostly for importing identifiers (e.g., VIAF) or for importing data from other databases (e.g., DBpedia, Freebase). This insight supports Piscopo's [236] argument that bot edits need to be more diverse. We suggest that further efforts should be made to import or link data from different sources, for example from research institutions and libraries. With the increased usage of the Wikibase software[24], we assume that more data might be linked or imported to Wikidata. Our classification revealed that bots import data from local files or databases already. However, such data imports often rely on the community trusting the bot operators and do not seem large-scale.

[23] Information on Wikimedia strategy can be found on: https://wikimediafoundation.org/about/vision/.

[24] Wikibase is based on MediaWiki software with the Wikibase extension https://www.mediawiki.org/wiki/Wikibase.

The requested maintenance tasks within the RfP show similar patterns to those in the Wikipedia community. Bots are being used to overcome the limitations of the MediaWiki software [112]. An administrative task, such as deletion, is not often requested in RfPs. However, bots on Wikidata are allowed to delete sitelinks only. Similar to Wikipedia, the Wikidata community comes closer to a common understanding of what bots are supposed to do and what not.

The issue of unknown tasks, which was not clearly defined in RfPs, shows the trust the community has in single operators, probably due to their previous participation history. There is a noticeable number of approved RfPs which we coded as unknown due to their vague task description, while there are also cases where task descriptions were clear, but the community did not trust the operators and, thus, they were not approved. These two observations indicate that trust is also given importance by the community in addition to their defined policy and procedures.

The success rate of RfPs is relatively high since only a small number of RfPs are closed as unsuccessful. Our findings show that among the 81 unsuccessful RfPs, only some of the RfPs were rejected by the community directly; the majority of them were unsuccessful due to the reason that the operators were not responding or had withdrawn the request. Therefore, the operator's inactivity is a higher reason for failure than community refusal. In some cases (i.e., the RfPs discussed most) we can see that the community considers every detail of the tasks and then comes to a decision, however, in some cases, they approve the request without a detailed discussion, as can be seen in the cases of unknown tasks and unknown sources. This result could indicate that in addition to the defined procedure for RfPs, the community applies a more flexible approach when deciding on RfPs considering the context (e.g., editor experience) of the application into account.

In essence, the high approval rate of RfPs indicates that the Wikidata community holds a positive and felxible stance towards bot operations within Wikidata, showing a willingness to harness task automation through bots. While the community permits a diverse range of content editing tasks to be undertaken by bots, the primary role of bots has been that of importing data from various Wikipedia language versions. The Wikidata community has effectively capitalized on the experiences of the Wikipedia community, successfully establishing efficient human-bot processes in a relatively short span of time.

6.2.5 Summary of Bots from a Community Perspective

Bots are the most actively editing group in Wikidata and, despite this, we know very little about them. As a first step to understanding this user group, we studied the formal process of requesting bot rights in Wikidata to find out why bots are created and what kind of task types are allowed to be automated by bots. Our study provides a detailed description of the RfP process. We retrieved the closed RfPs from the Wikidata archive up to mid-2018. We defined a scheme, in the form of a codebook, to classify the RfPs and developed our dataset. The RfPs were studied mainly from two perspectives: 1) What information is provided during the time the bot rights are requested, and 2) how the community handles these requests.

We found that bots are created on community demand or the operator's will. There is a defined mechanism for asking for bot rights for an account. The main tasks requested are adding claims, statements, terms, and sitelinks into Wikidata which means that bots in Wikidata are capable of performing content editing tasks like human users and they are created to import content into Wikidata. The main sources of bot edits having their roots in Wikipedia show that so far bot operators tend to focus on importing from Wikipedia and are not creating bots that would import from more diverse data sources. This could be an indicator of concern about the diversity of sources in bot edits. Overall, this contrasts with Wikipedia where bots are performing mostly maintenance tasks. Our findings also show that most of the RfPs were approved and a small number of them were unsuccessful mainly because operators had withdrawn or there was no activity from the operators. Hence, the Wikidata community is open to automation and anyone who can display the ability and provide a convincing reason, can apply for bot rights and operate bot accounts. In other words, the community is open to all but wants to ensure anyone asking for bot rights has the potential to fix problems caused by their bots.

Next, this research will focus on analyzing the actual activities of bots and comparing them to the results of this study to determine whether bots are effectively performing the tasks assigned to them. However, our primary goal is to examine the editing behavior of bots in comparison to that of human editors, and specifically to investigate whether bot edits have a distinct impact on the diversity of data in Wikidata. As mentioned earlier, we have observed data imbalance in Wikidata domains, and it is known that bots request similar tasks as human users. Our current objective is to explore whether bots execute edits akin to human users, or if distinctions in bot edits contribute to the data imbalance observed in various domains within Wikidata.

Furthermore, this analysis will provide insights into the evolution of bots over time and shed light on whether the community effectively controls bot activities based on their permitted tasks, or if these bot accounts are capable of making edits beyond their designated permissions.

6.3 Bot Activities and Contributions in Wikidata

Earlier, in the findings of Section 6.1.2 we saw that operators (i.e., human users) ask bot privileges to add or update data at a higher speed. However, the fact that all of these bots are still active and perform their requested tasks is yet to be explored. Especially, it is unknown which of these approved bots remained active and which slept after performing the requested tasks. For this reason, here we look at the actual Wikidata revision history to find out what bots actually do in Wikidata in comparison to human users. This will serve as the basis for understanding the diversity of bot edits in Wikidata, i.e., how diverse activities they perform, how diverse topic domains they edit, and how their contributions might impact the overall diversity of Wikidata.

In the previous section, we looked at bot-requested tasks or edits from the two angles of *activity type* and *edit focus* to know which activities are most popular to be shared with bots and which parts of the Wikidata data model is considered to be edited mainly through automation. Following this pattern, we want to look at actual edits in Wikidata and compare bot and human edits from activity type and edit focus angles which together make an edit type. This way we can explore the diversity of edits based on the different edit focuses or the variety of activity types used.

Table 6.6: *Number of revisions, users, and items per user group.* The tool users can also appear in the registered and automatic contributions, since users that made semi-automatic contributions only are 201, 854 tool users that were active as humans as well, and 31 tool users also appear as bots. In parenthesis, the relative number is given. Source: *Wikidata Revision History Dataset.*

	# Users	**# Revisions**	**# Items**
# Anonymous contributions	12,728 (0.33)	40,296 (0.01)	11,555 (0.02)
# Registered contributions	24,252 (0.63)	1,705,674 (0.44)	187,803 (0.30)
# Semi-automatic, reg. cont. (tools)	1,086 (0.03)	1,399,897 (0.30)	215,543 (0.34)
# Automatic, reg. cont. (bots)	274 (0.01)	2,431,409 (0.25)	212,551 (0.34)
	38,340	**5,577,276**	**627,452**

Here, we will explain the Wikidata edit history and our reasons and approach to preparing our dataset out of this data. Further, we will present the results with a focus on a comparison of human and bot users' behaviors. We will sum up this section with what we learned and what next steps to take.

6.3.1 Wikidata Edit History

Wikidata stores every edit made by every contributor at any time in its edit history as revisions[25]. Each edit is stored along its metadata and is available for access and utilization in the form of dumps[26] or through *Wikimedia Toolforge*[27]. We explore this edit history to inspect the Wikidata contributing community through the lens of their edits, with a focus on bot contributions. A detailed description of the *Wikidata Revision History Dataset* creation process is provided in Section 5.3. Below, we present our findings and discussions starting with providing a statistical overview of this dataset.

6.3.1.1 Overview of Our Sample

Developing a dataset out of the revision/ edit history of Wikidata was a long and rather effort-taking process which is already explained in Section 5.3.

After the data pre-processing step we identified four user groups based on their level of automation which are humans, tools, bots, and anonymous. Table 6.6 shows a summary of the identified user groups, the number of users in each user group, their edits in terms of the number of revisions each user group performed, and the number of items each user group contributed.

In Table 6.6 we can see that users identified as bots are only 1% of the total users (i.e., *274*) in comparison to other user groups in our sample, while, humans contributors make 63% of the whole users. Nevertheless, we can observe that bot edits, in terms of the number of revisions and items, are nearly double that of human users.

[25] The Revision table of the MediaWiki database stores metadata related to each edit on Wiki pages. https://www.mediawiki.org/wiki/Manual:Revision_table[Accessed 01.03.2023]

[26] Wikidata database dumps: https://www.wikidata.org/wiki/Wikidata:Database_download[Accessed: 01.11.2020]

[27] Toolforge is a hosting environment for Wikimedia-related software. Further information is available at: https://tools.wmflabs.org[Accessed:06.12.2019]

The tool is the second user group which, despite fewer users, has a high contribution rate. On the opposite side, is the anonymous user group which has users much more in number than the bot and tool user groups, but a much lower contribution level than both of the mentioned users. In this regard, it can be inferred that human and anonymous user groups, while having a larger number of users, primarily engage in manual edits, leading to lower contribution levels. Conversely, the bot and tool user groups, comprising fewer users, exhibit the highest contribution levels due to their automation capabilities, allowing for automated and semi-automated edits. In total, our dataset consists of 38,340 users from the mentioned four user groups, 5,577,276 revisions of 627,452 items.

From these revisions, we extracted edit types which consist of two parts, i.e., activity type and edit focus (cf. Table 5.2 on page 100). Therefore, we could aggregate the edits using activity type and edit focus to have a glance at the user groups from their edits angle and explore how diverse edits they perform.

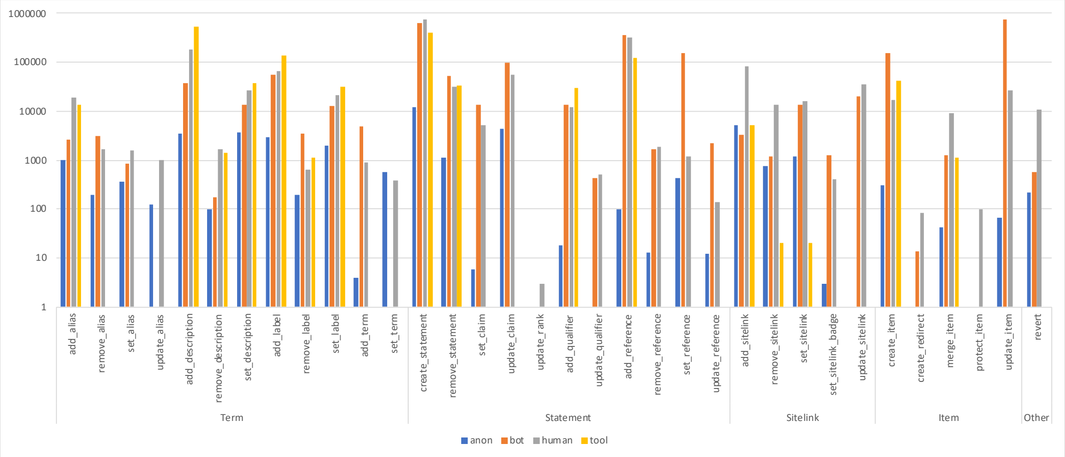


Figure 6.7: An overview of Edit types per user group (aggregated by edit focus). Source: *Wikidata Revision History Dataset.*

Figure 6.7 provides an overview of the edit types per user group, aggregated by four main edit focuses, and demonstrates that almost all edit focuses are edited by all user groups, yet one user group stands out with the majority of revisions (cf. Table 6.6). Among the 34 edit types depicted in Figure 6.7, only four edit types (i.e., update_alias, set_term, update_rank, and protect_item) are not performed by bots, in contrast to humans who engage in all of these mentioned edit types. This suggests a similarity in the usage of edit types between humans and bots, albeit with variations in the volume of edits. Nevertheless, delving deeper into the data is necessary to ascertain the true nature of the similarities and differences between these user groups.

In the upcoming sections, our focus shifts towards a comparative analysis of edit types employed by bots in contrast to other user groups, particularly humans. Moreover, we conduct a comparison of the editing patterns of these user groups, exploring both the volume of edits and the utilization of different edit types, aiming to gain insights into the task similarities and differences between humans and bots.

6.3.2 User edits in Wikidata

In this section, we delve into the findings in a more detailed manner to gain a comprehensive understanding of how bots contribute to Wikidata and how their editing behavior compares to that of other users.

Figures 6.8 and 6.9 illustrate the distribution of edit types among user groups from the perspectives of activity type and edit focus. This visual representation highlights the varying contribution levels of different user groups across different edit focuses. For instance, we see that *statement* and *item* received higher editing volumes than *alias* and *sitelink*. Similarly, there is a noticeable concentration of effort on certain activities, such as *add*, *create*, and *update* which appear to be the most prevalent. This suggests that the Wikidata community has primarily directed its efforts towards adding content to the platform. Overall, we observe that all user groups engage in nearly all of the activities within the specified edit focuses. However, differences in edit volumes among these user groups persist, making it challenging to definitively ascertain similarities or differences based solely on these visualizations. Given our interest in understanding the editing behavior of these user groups, with a particular focus on bots, our next step involves comparing these user groups in more detail.

6.3.2.1 Edit focus Contributions

The edit focus refers to specific parts of an item, as outlined in the Wikidata data model (refer to Section 2.1.3 on page 15), where various activities can be executed. For example, the activity *add* can be carried out on the edit focus *alias*.

Figure 6.8 provides a summarized representation of edits per edit focus conducted by each user group. While the figure illustrates that all user groups contribute to edits in each edit focus, it is evident that bots are most active in the areas of items, references, claims, and terms. Conversely, humans predominantly perform edits within the alias, statement, and sitelink edit focuses. Tools have shown to be actively editing the description, label, and qualifier focuses, while, anonymous users had no dominance in any of the mentioned edit focuses. Once again, it is not possible to pinpoint any commonalities or disparities between these user groups from these visualizations at this point.

6.3.2.2 Activity Type Usage

Activity types refer to the specific edits or actions performed by a user, such as creating or removing (cf. Table 5.3 on page 101 for more details). Figure 6.9 is a visual representation of who performed which activities in our dataset. According to this figure, bots have carried out most of the activities related to creating, removing, setting, and updating activity types on Wikidata data. Tools exhibit similar behavior to bots when it comes to the “add” activity. Humans, apart from being visibly active in performing various activities, have demonstrated the highest level of engagement in maintenance activities such as merging, protecting, and reverting. On the other hand, anonymous users do not exhibit a dominant activity type and have shown the least participation in each activity category.

Overall, it is evident that all user groups engage in nearly all activity types and edit focuses. However, maintenance-related tasks, though fewer in number, are exclusively carried out by humans. However, these results alone do not provide insight

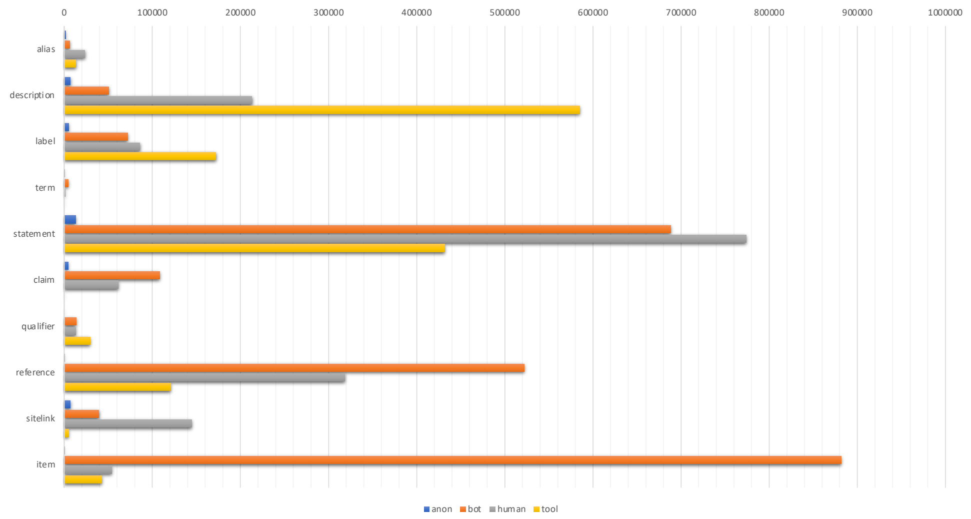


Figure 6.8: User edits aggregated according to their editing focus. Source: *Wikidata Revision History Dataset.*

into how these user groups compare to each other. Therefore, a more elaborate mechanism is needed for this purpose, as elaborated in the following section.

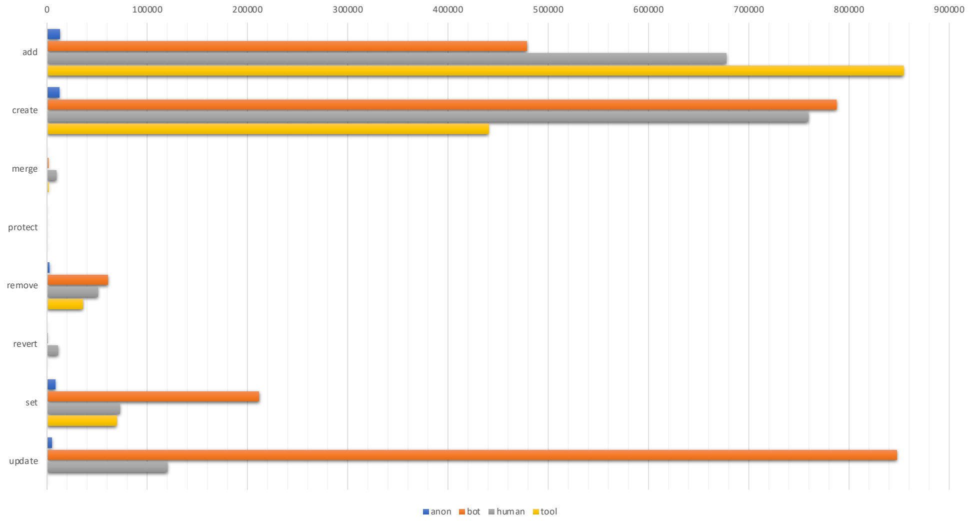


Figure 6.9: Activity types per user group. Source: *Wikidata Revision History Dataset.*

6.3.2.3 Comparison of Wikidata User Groups

In this section, we statistically assess the similarity or difference of edits among Wikidata user groups. This assessment is based on their edits that are available as revisions in *Wikidata Revision History Dataset.* In this dataset, we have identified 34 different edit types (cf. Table A.3 on page 164) which are used in different amounts by the user groups Anonymous, Bot, Human, and Tool. Since the overview of our dataset shows the usage of almost all edit types by all user groups, our hypothesis is that all user groups perform similar edits in Wikidata and are not different from each other. There are many methods that could be used to assess this hypothesis and compare these user groups, however, due to the criteria of our dataset, explained below, we used the Chi-squared test.

Our data is categorical and we have two variables which are user group (Anonymous, Bot, Human, and Tool) and edit type (the 34 mentioned edit types in Table A.3 on page 164). Both of these variables consist of categories and do not follow any intrinsic ordering. A method that can find differences between groups with categorical variables and frequency values is the Chi-Squared Test [74]. We inspect if the usage of edit types is dependent on the user groups or not. As our study evolves around the editing patterns of humans and bots, here we focus on these two user groups. In other words, do Human and Bot user groups differ from each other based on the usage of edit types or not? For this reason, we use the Chi-square independent test and define the null hypothesis H_0 stating no relation between the user group and edit type, and the alternative hypothesis H_1 that states dependence between the user group and edit type. The result of the Chi-square independent test confirms one hypothesis and rejects the other.

We examined the relationship between user groups (Human, Bot) and edit types (31 of the 34 edit types[28] in Table A.3) using the Chi-square independent test after verifying it's assumptions (the observations/ edits were drawn independent from the user groups and the values in each cell were larger than 5). A p-value less than 0.05 is an indicator of statistical significance.

The result of the Chi-square independent test rejects the null hypothesis and shows a significant association between the edit types and user groups in Wikidata, i.e.:

$$X^2(30) = 1101760, p < .001$$

This means that the editing behavior (i.e., edit type usage and edit volume) is significantly different between humans and bots considering the one in a thousand chance of results being random. This behavior can also be seen if we look in detail at edit types over time for each user group.

This difference is not visible in a general overview of the edit types per user group (cf. Figure 6.7). A detailed look into a single edit type may reveal the difference between user groups in using each type and shed light on how each user group contributes to this edit type, showing the editing behavior of each user group. Hence, we can see if the amount of edits is the only differentiating factor between the user groups or we can spot further factors as well. To visualize this difference we take *add_alias*, the first edit type in Figure 6.7, and compare each user group editing patterns per month from 2012 till the end of 2019.

As can be seen from Figure 6.10, user groups Bot and Tool that perform automated edits have similar editing patterns over time. They have highly active months and at the same time many inactive months. On the other hand, Human and Anonymous user groups that perform manual edits also have similar editing patterns per month. The later user groups perform more steady editing volume over time and have no

[28] We excluded three of the edit types (set_term, update_alias, update_rank) which had zero values, before performing the Chi-square independent test to meet the assumptions of this test.

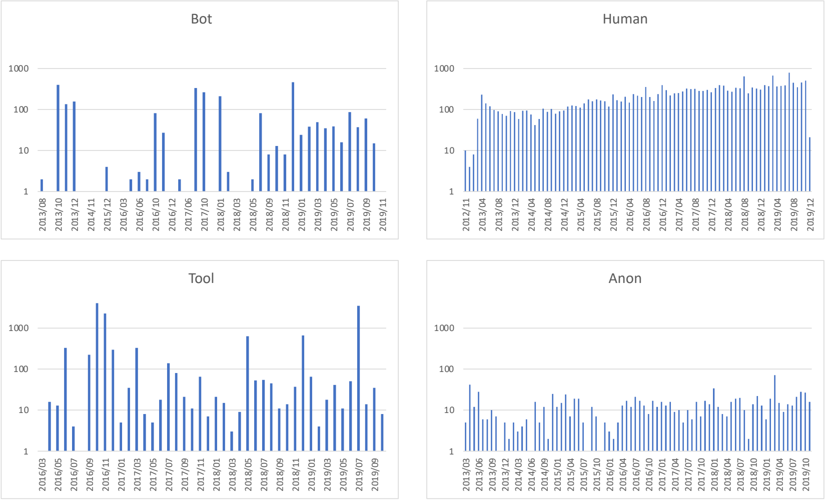


Figure 6.10: Comparison of user groups in the usage of add_alias edit type over time. Source: *Wikidata Revision History Dataset.*

inactive months. Nevertheless, this edit type (i.e., add_alias) is one of the few edit types where the user group Human has the highest edit volume, therefore, it might not represent the majority of edit types where bots are more active than other user groups. For this reason, we look at Figure 6.7 and select *create_statement* edit type for a detailed editing pattern of user groups over time. *Create_statement* seems the best candidate because all of the user groups have visible participation levels and, in particular, the user groups Human and Bot have rather similar levels of participation.

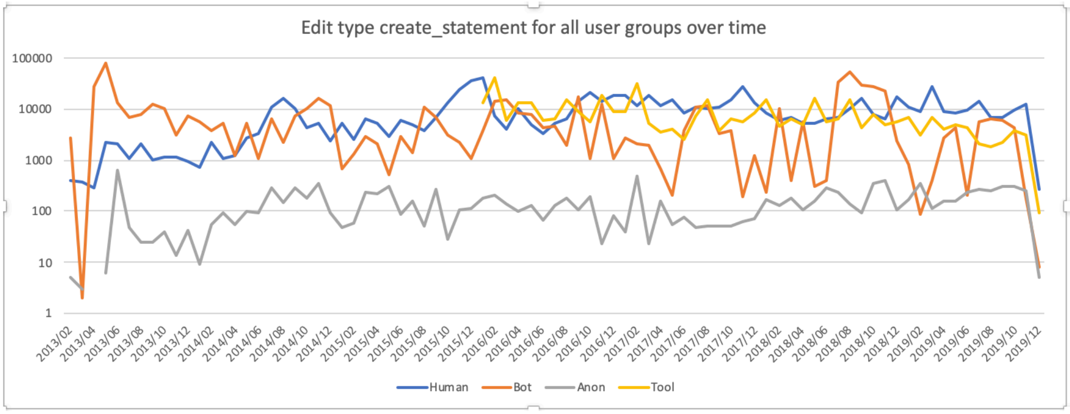


Figure 6.11: An overview of create_statement edit type per user group over time. Source: *Wikidata Revision History Dataset.*

In Figure 6.11, each user group exhibits distinct patterns of behavior. The user group Anonymous maintains a consistent level of edits, remaining below 1,000 revisions per month over time, mostly hovering around 100. The Tool user group began its revisions at the beginning of 2016 and maintained a steady pace of around 10,000 over time. However, when observing the user groups Human and Bot, Figure 6.11

does not provide a clear distinction. Thus, we examine their revisions of the edit type *create_statement* more closely in Figure 6.12. Here, we can observe that humans exhibit a relatively stable editing pattern each month, with some fluctuations but no instances of zero activity. In contrast, bots display significant peaks in certain months, accompanied by periods of very low activity or even no activity at all.

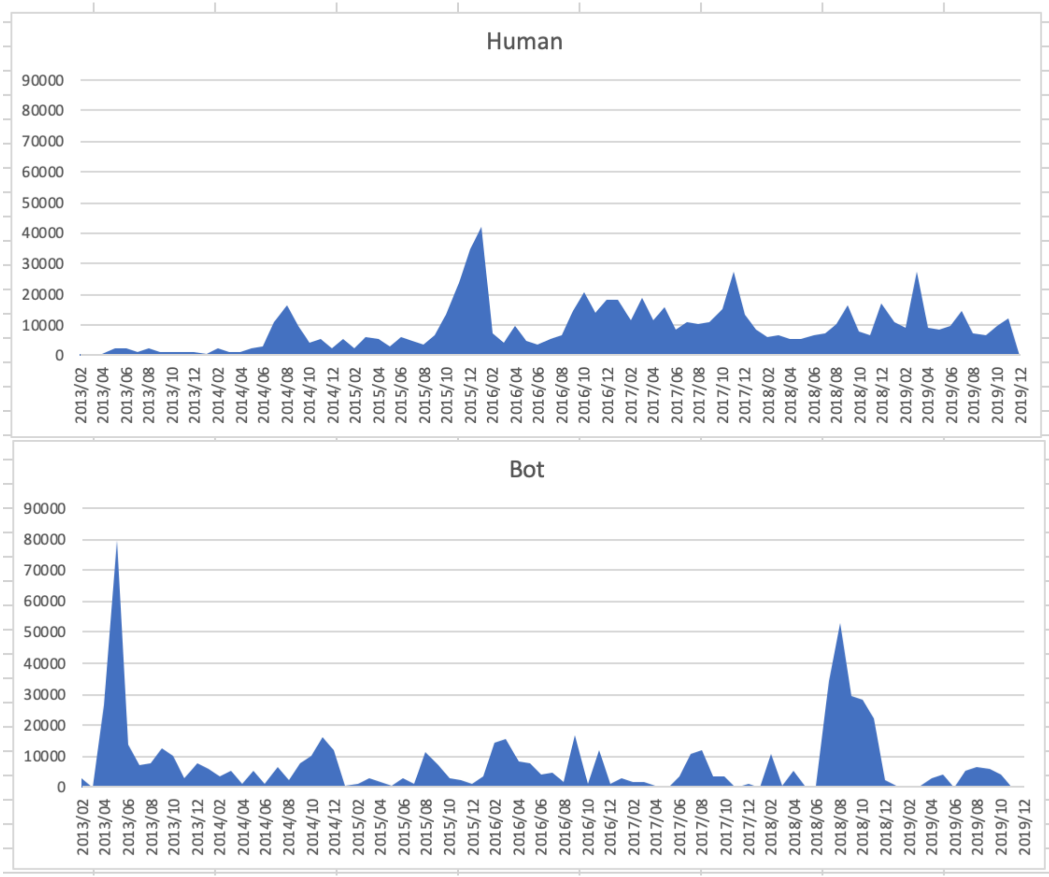


Figure 6.12: An overview of create_statement edit type per user groups human and bot over time. Source: *Wikidata Revision History Dataset.*

Although the edit type of *create_statement* has similar levels of activity for humans and bots, Figure 6.12 shows similar results with the Figure 6.10. This means that bots perform batch edits at certain points in time and after that, they are not active until the next time they are supposed to perform a massive editing session. As bots perform the edits that are uniform and run through scripts, these massive editing sessions perform the same edits for all items. Even the rate of introducing errors is high in such automated edits and could cause the entire edits to revert. Conversely, the regular manual edits carried out by humans each month, despite their lower editing volume, could suggest that humans exhibit a more open and consistent editing behavior. For instance, we notice three notable peaks in human edits, occurring around the months of December and January, likely coinciding with holiday periods when contributors have more available time for engagement.

6.3.3 Bots in Wikidata Context

In the Wikidata community, both human and bot users are actively contributing to Wikidata and performing shared edits. This highly active contribution of bots in Wikidata makes the Wikidata community unique and provides the opportunity to look into how machines could influence a system when performing shared tasks with humans. Machines were developed to solve certain issues and their success is measured by how well they fulfill their tasks, nevertheless, these machines might have unintended impacts on the environment they operate [249]. Since these machines are part of the human ecosystem, it's inevitable that humans are not affected by machine behavior, so to fully understand machine behavior and their possible side effects on an environment, machines need to be studied from the context of their operation and social environment [249].

In this section, we have offered a comprehensive description of bots within the context of Wikidata. As previously discussed, our findings indicate that the Wikidata community is highly receptive to bots, with requests for bot accounts being rarely denied. However, sometimes the bot account privileges are granted after long discussions which makes the requesting side lose interest in operating the bot and decide to withdraw. This shows that the community wants to ensure the integrity of data in Wikidata when allowing automation, as automation can create a mess if not handled properly and would need extra efforts to revert any unwanted changes.

Although the Wikidata community is a group of volunteer editors who can come from any origin with any background, discussions in the bot approval process show that sometimes requests are approved without detailed enough information regarding the bot tasks[29] or any discussions by the community[30] because decision-makers know the operator[31]. On the other hand, some requests are discussed in very detail when the community is not sure of operators' technical abilities to handle the bot[32] or rejected when the operators could not prove to be trusted for their abilities[33]. Hence, it seems that the community only allows experienced programmers to operate bots because operating a bot requires programming skills. In essence, bot operators are required to possess advanced programming skills. They not only create and deploy their bots but must also troubleshoot them in case of unexpected issues. Consequently, this implies that the majority of content added to Wikidata is determined by a limited group of operators. Most of these operators are computer scientists hailing from Western societies where Wikidata enjoys popularity. Thus, while the Wikidata community welcomes bots, it imposes certain criteria on their operators. This suggests that bot-generated edits tend to reflect the perspectives of their professional operators and may be less diverse compared to contributions from human users with a wider range of backgrounds.

Theoretically, Wikidata should be diverse enough to reflect world knowledge because it has the ability to store all of the possible statements related to every single item. However, relying on the ability only doesn't ensure the diversity of knowledge in

[29] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Svenbot

[30] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Svenbot

[31] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/BotMultichill

[32] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/JWbot

[33] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/PokestarFanBot_5

Wikidata since data comes in a collaborative manner from a volunteer community. This means the knowledge in Wikidata is heavily dependent on the editors who contribute this data to Wikidata. The more editors with diverse backgrounds, the more diverse topics, and statements to expect in Wikidata. Nevertheless, the existing little knowledge about the Wikidata community makes it challenging to get an overview of the Wikidata diversity status. Further, the clues of bot edits' dominance in the existing research also raise concern regarding diversity in Wikidata as less than 500 automated accounts (i.e., bots) outshines the edits of more than 20K active human users. In other words, views of less than 500 bot operators overshadow the views or topics of interest of more than 20K users in Wikidata. Questioning the diversity of knowledge in Wikidata means questioning the success of Wikidata in achieving its overall goal of becoming the world KB. Answering this question requires a thorough understanding of the Wikidata community, especially the most active editors with the majority of edits which in the case of Wikidata are bots.

Upon examining our sample from the Wikidata revision history, we find that bots execute their designated tasks as requested. This indicates that the Wikidata community closely monitors bot edits and can prompt reverts if any malicious activities are detected. In contrast to Wikipedia, where bots primarily engage in maintenance tasks while human users handle edits, the dynamics are somewhat different in Wikidata. Here, bots tend to undertake content editing tasks, whereas humans focus more on maintenance activities. This trend can be attributed to the structured format of data in Wikidata, which lends itself well to bot-driven data entry. As a result, the convenience of this approach has led operators to increasingly utilize bots, resulting in their dominance over edits originating from human users.

Bots' focus on creating, updating items, and adding statements and references indicates their positive role in populating Wikidata with fundamental content, preparing it for a broader range of edits from various user groups, especially humans. Some tasks are exclusive to human users due to their nature requiring manual updates, like the *update_rank* edit type depicted in Figure 6.7. The automation of edits holds the potential to reshape the diversity and balance of items and their contents, potentially influencing the overall data diversity within Wikidata.

Bot edits are significantly different from human users, so we can expect that their edits leave a different impact on the diversity of data in Wikidata in comparison to human users. If used with the purpose of overcoming the data gaps, automation can serve to increase diversity in domain/class and item levels by adding a variety of items and statements for a higher variety and more balanced data across Wikidata domains and classes. Alternatively, automation could lead to an increased concentration of data in specific areas, even if those areas aren't universally relevant or reflect the interest of people worldwide, as is currently the case.

Thus, we can infer that bots might have contributed to the concentration of data in specific domains/classes, resulting in an imbalanced data coverage across various Wikidata domains. With this in mind, we delve into bot edits from a diversity perspective, investigating their involvement in Wikidata domains to determine if bots are indeed contributing to the low diversity observed in these domains.

6.4 Bots and Diversity in Wikidata

In this study, we observed that Wikidata is a collaborative KB with the goal of reflecting world knowledge. However, we noticed that the Wikidata community heavily relies on algorithms for importing large amounts of data into the platform. Our earlier findings unveiled an uneven distribution of content among different topical domains. Concurrently, existing research highlights the prevalence of specific languages over others within Wikidata. Interestingly, our findings illustrate that these dominant language versions in Wikipedia are frequently referenced in Wikidata bot requests. This implies that bots may be contributing to the observed language imbalance in Wikidata, further underscoring the potential influence of bots on the platform. While bots have successfully carried out their intended tasks and contributed substantial amounts of data, it is crucial to investigate the unintended side effects that extensive automated edits may have had thus far. We need to determine whether bots are responsible for the aforementioned data imbalances in Wikidata and understand the potential consequences of their actions.

With a huge amount of edits coming from bot accounts, we get the impression that bot operators' edits are dominating other human user account edits, while, bot accounts should have mainly been used to perform repetitive or time-consuming tasks to enhance human efficiency. Despite the fact that bots are run by human operators and ultimately humans are responsible for bots' behavior, the rights these accounts have and the way they edit Wikidata make them different from other human users. Therefore, we saw a significant difference between humans' and bots' edits in the Wikidata edit history. This could influence the goal of Wikidata for representing world knowledge, as the edits of a very small group of users (i.e., bots/ bot operators) can cause the edits of thousands of human users to get overlooked. In addition, the reason behind such imbalanced contributions can be sought in the contributing community and their contribution patterns. Especially, when concerns from low diversity of bot edits [235] already exist.

In our proposed concept for measuring diversity in Wikidata, we have mentioned that editor diversity can be measured considering different factors like user background, roles, experiences, usage of edit types, and contribution in topical domains as some possible differentiating factors for diversity measurement in Wikidata. Choosing one factor over the other depends on the research question and the angle from which we want to measure editor diversity. In the Wikidata literature, editor diversity remains a blind spot and we know very little about Wikidata editors in general and in particular from a diversity angle; despite the unique community Wikidata has with very active contributions coming from bots. Since we are interested in understanding the effect of bot edits on diversity, we look at the editing behavior of Wikidata editors in the Wikidata edit history. We examine editor contributions in the topical domains to find out the answer for who is responsible for most of the edits in the topical domains. We, additionally, look into how humans and bots use edit types in the Wikidata domains to know more about bot behavior and their effect on Wikidata domains. Bots have the right to perform batch edits at higher speeds than humans, therefore, we want to explore bots' potential to impact data diversity in Wikidata domains. If we can confirm bots' effect on diversity in Wikidata, we can then look to find a solution on how to use bots to improve diversity in Wikidata.

Here, we aim to explore Wikidata's edit history with a focus on bot edits in comparison to human edits. While, we know human and bot edits are significantly different, we want to see if this difference is the cause for such low diversity status in Wikidata. In this section, we measure the diversity of bot edits in comparison to human edits using the *Wikidata Revision History Dataset* 5.3.

6.4.1 Diversity of Bots Edits in Wikidata

Anyone contributing data through performing edits like adding, updating, or removing data is called an editor. Bots have been the most active editors of the Wikidata community since its inception. They perform edits similar to human users because Wikidata is a structured KB and dealing with structured data is ideal for machines. In our proposed concept for measuring diversity in Wikidata, we have mentioned an approach for measuring editors' diversity. Using this approach we provide insight into the existing editor diversity in Wikidata through the variety of edit types (i.e., edit activity on edit focus) they perform and the variety of domains they contribute. Our findings are based on the *Wikidata Revision History Dataset* explained in Section 5.3. As we are interested in understanding the impact of bot edits on diversity in Wikidata, we only consider edit types usage and domain contributions. We skip the other metrics we have mentioned for this approach in Table 3.4 because our focus is on understanding the bot behavior in Wikidata domains. In addition, the metric background does not seem to be directly related to bots at this stage. Bots are automated scripts that can easily be altered to add content in another language or can run for any long period without reflecting any age-related issues like human users whose interest in topics or opinions are influenced by age, language, or country of origin. Further, finding the background of the bot operators is not a straightforward task. Some bots like ProteinBoxBot[34] are operated by more than one operator, while, in some cases, multiple bots are operated by one operator[35].

Here, we discuss our findings regarding the diversity of edits among Wikidata user groups (i.e., Anonymous, Bot, Human, Tool) with a focus on bots.

6.4.1.1 Diversity of Edit Domains

One aspect of measuring editor diversity is to determine how balanced or concentrated the contributions are among the users of Wikidata user groups, i.e., bots, tools, anonymous, and humans, across various topical domains within Wikidata. This would also help in understanding if all items are edited collaboratively by all user groups, or if there are items or classes which show contributions from a single user group and are neglected by others. This way we could see which topics or classes are more popular among the Wikidata community and which ones are not given much attention.

Additionally, how each user group has edited each class that has brought these classes and their domains to different levels of diversity. In particular, it would be interesting to see how the high volume of bot edits affects the balance of class and item contents. In other words, do large amounts of data by bots in certain items

[34] User page for ProteinBoxBot https://en.wikipedia.org/wiki/User:ProteinBoxBot[Accessed: 27.04.2023]

[35] For example Magnus Manske runs five bots: QuickStatementsBot, CommonsDelinker, ListeriaBot, SourcererBot, Reinheitsgebot.

or classes have caused the concentration of data? Where do the classes and items with the absence of bot edits stand? If bots cause imbalance and result in a lower diversity status of Wikidata, how can we fix this issue and prevent it in the future?

To find answers to the above-mentioned questions, we first take a general look at how the edits of each user group (also called revisions) are distributed across the Wikidata domains in order to find the most popular domains that have received the most revisions. Next, we dig deeper into the class level and examine the edits of each user group in Wikidata classes for a more detailed picture of user contributions in Wikidata domains.

In Section 4.2.1 where we measured domain diversity using the number of items per class and the number of properties per item, we could see that the domains of Media and Geography are the most diverse ones. Since, we want to know if user groups had focused edits on some domains, here, we look at revisions because revisions contain any kind of edit, including the removals and reverts which are not available when measuring the actual data in Wikidata.

Measuring the diversity of Wikidata domains based on the number of revisions/ edits each item and class in a domain has received in *Wikidata Revision History Dataset*, we get similar results. The results of Table 6.7 once more confirm that the domain Geography is the most diverse based on the results of the diversity measures Entropy and Simpson's Index, while, the domain Media is the most diverse according to the Rao-Stirling Index. In short, the most diverse domains from both, the number of items based on Table 4.1 and the number of revisions based on Table 6.7, are *Geography* and *Media*, while, the least diverse ones are the domains *Person* and *Biology*. Thus, not all Wikidata domains represent equally diverse contents, and not all domains are given equal attention from the user groups. To get a closer picture of the user contributions in each domain, we look at the number of edits each user group has in each of the domain classes.

Table 6.7: Domain Diversity based on the number of revisions from *Wikidata Revision History Dataset*. (Note: Higher numbers indicate greater diversity and bold values represent the highest diversity for a domain calculated from the diversity measure in each column.)

	Entropy	**Simpson Index**	**Rao-Stirling Index**
Biology	0.858	0.540	840.714
Geography	**1.388**	**0.728**	125089.777
Media	1.198	0.609	**266193.664**
Organization	1.225	0.635	28318.861
Person	0.437	0.188	5011.266

Looking at the details of user group contributions in the classes of Wikidata domains, we are interested in a comparative view of user group contributions in these classes.

To analyze user contributions in the mentioned classes, a heatmap is generated using the 'diverse' package of the R programming language in RStudio software. The heat map represents the relative concentration/ focused edits of each user group's contributions in Wikidata classes, where colored cells represent classes that have received relatively more edits from the user group than other classes represented with white cells. In other words, the visualization here serves to provide insight into which of the classes each user group has relatively more focused contributions than others, in particular the bot user group.

The data used to generate this heat map comes from the *Wikidata Revision History Dataset*. With raw data, we were not able to generate a heat map that could display the focused contributions. This was due to big differences in editing amounts these classes had received from user groups, hence, lower values were not clearly distinguishable from zero. For this reason, we used the approaches proposed by [105] for the normalization of datasets when analyzing diversity and used the Revealed Comparative Advantages (RCA) [17] mechanism for normalizing our data. The normalized data displays the degree of each user group's relative contribution in each class. The degree of a user group's contribution in one class is calculated in reference to the overall contributions of that user group in all of the classes. This way we can see the relative contributions of user groups across these Wikidata classes. At this stage, we performed a further filter on the RCA data by dropping the values below 1 which are shown by empty cells and provide a clearer view of the classes with relatively more attention, as can be seen in Figure 6.13.

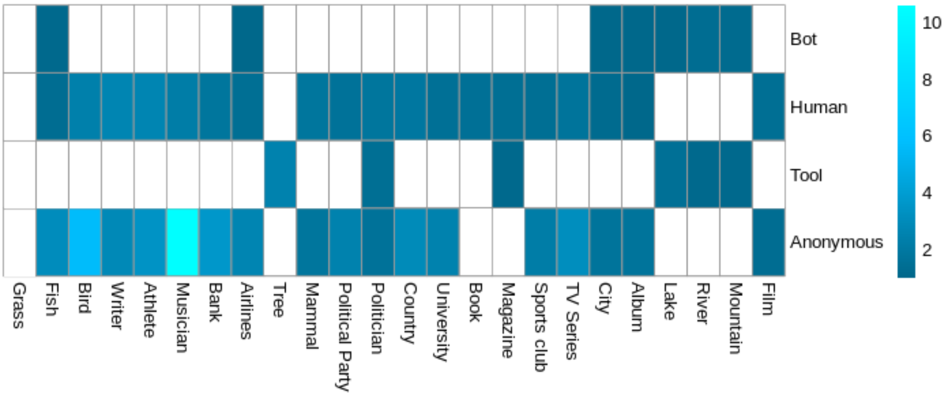


Figure 6.13: Heat map of Editor contribution to Wikidata classes. Higher values are represented with lighter turquoise color shades. Data is ordered with the highest values in the top right corner and the lowest values in the left bottom corner of the Figure. In the top right we can see the bot user group followed by humans which had the highest contributions so far, and on the right side are the classes of Media and Geography domains with higher diversity levels, and on the left side classes of Person and Biology domains which have the lowest diversity levels. Source: *Wikidata Revision History Dataset*.

The heat map reveals that each user group had different participation levels among the classes in Wikidata. We can see that user groups of Bot and Tool show similarity in having an interest in a rather smaller number of classes and both of these classes perform automated edits. In the same manner, between the user groups Human and Anonymous, which perform manual edits, akin patterns can be seen. The two user groups have shown attentiveness to a rather more diverse range of Wikidata classes. Overall, we can see differences in patterns between automated and manual edits. Hence, we can conclude that human and anonymous user groups show contributions to a higher number of classes, so have more diverse class contributions, while, bot and tool user groups have focused edits on a limited number of classes, thus have more concentrated contributions.

In addition, it is particularly interesting to see that the classes Mountain, River, and Lake that belong to the domain of Geography have received focused attention from the automated and semi-automated editors only. The concentrated efforts of bots in the majority of Geography classes indicate their positive influence on this domain through enhancing data variety and balance. One of the reasons why geographical data was added through automation could be the availability of geographical data in formats that are easy to import through bots, e.g., tables and datasets. In the domain media, we only see bot-focused contributions in the class Album, and in the rest of the classes Humans have had focused edits. Although Figure 6.13 shows that bots didn't have focused contributions in the class Film, the order of classes in the heat map shows that bots still had very high contributions in the class Film.

Above, our aim was to see if user groups treat all classes in the same way, or show interest in some and neglect others. Here, we want to find out which user group has performed more diverse edits based on participation in the number of classes and the number of revisions in those Wikidata classes. Although Figure 6.13 indicates that the Human user group had rather even contributions to the majority of Wikidata classes, we want to further confirm it through the usage of diversity measures on the revisions performed by user groups in Wikidata classes. Looking at the user edits in each class from an angle of the user groups also reveals a similar result.

Table 6.8: Descriptive statistics of user group diversity based on their contributions to Wikidata classes. (Higher values show higher diversity). Source: *Wikidata Revision History Dataset*.

	Variety	**Entropy**	**Simpson Index**	**Rao-Stirling Index**
Anonymous	22	2.169	0.825	185720.9
Bot	23	2.100	0.845	164597.4
Human	23	**2.252**	0.828	**196737.3**
Tool	23	2.158	**0.852**	167618.2

As can be seen in Table 6.8 almost all of the classes are edited by all user groups[36]. The highest values of diversity through the calculations of *Entropy* (i.e., 2.252) and *Rao-Stirling Index* (i.e., 196737.3) show human contributions in Wikidata classes to be the most diverse, which is also in line with the results of the Figure 6.13. While *Simpson Index* shows tools with 0.852 value to have the most diverse contributions across Wikidata classes. The reason most probably lies in the focus of each formula on a specific dimension or property of diversity. Earlier, in the example for explaining the diversity measures in Section 3.2 we have learned that Simpson's Index is sensitive towards sample size. It gives more weight to the common categories and rare categories with a small number of representatives don't affect the diversity. Tools use automation to edit Wikidata, so Simpson's Index considers tools as the most diverse user group due to the high volume of edits they perform in the classes with the high volume of content. In the same manner, bots that also perform automated edits are the second most diverse contributors to the Wikidata classes according to Simpson's Index, again due to their massive contributions to the highly edited classes (i.e., common categories) in Wikidata. In the meantime, the results

[36] The revisions related to the items of the class Grass is not part of our dataset because of the low number of items (i.e., only 6) in that class they were not retrieved during sampling process for creating *Wikidata Revision History Database*.

of Entropy and Rao-Stirling Index declare bots as having the least diversity of edits based on their contribution to Wikidata classes, while, according to the Simpson Index the least diverse user group is anonymous. While each measure's preference has generated different results, here, the results of Entropy and Rao-Stirling seem more relevant. In other words, the human user group having performed the most diverse edits in Wikidata is more probable, considering the results of Figure 6.13 and the fact that the number of human users is much higher than all the other three groups and come from much more diverse backgrounds than all others.

Overall, we could say that bots and tools have focused contributions in some classes and lower contribution levels in other ones, while humans and anonymous treat almost all of the classes equally and have similar contribution levels in all of the classes. Additionally, bots-focused contributions in the most diverse classes on Wikidata indicate that they have had a visible impact on making those classes more diverse. This means that bots have the ability to influence the diversity status of Wikidata classes. Although we see that Wikidata domains are imbalanced and, thus, low in diversity, the details of each domain bring us to a different conclusion. Bots have played a crucial role in increasing the variety and balance of content in the domains they were active. On the contrary, Domains with little contribution from bots have lower variety and balance of content coverage and so, have remained less diverse. Hence, bot contributions have created content concentration which is positive when looking from inside domains, but is negative when looking from the outside and the whole domains of Wikidata.

Next, we look more closely at the diversity of edit types to shed light on the diversity of bot edits in terms of the types of edits they make compared to other user groups.

6.4.1.2 Diversity of Edit Types

In the journey of understanding bot behavior in the hybrid Wikidata community of human and bot users, we have seen that bots have more concentrated contributions in some classes across Wikidata domains. Understanding bot editing behavior also requires information about how bots edit Wikidata, what activities they mostly perform, and which parts of an item they mostly edit. In Section 6.2, we demonstrated that bots requested very similar edit types as humans, and here we would like to see which types of edits bots have actually performed so far. We are interested to see if bots are only there to add new content, or if they have also updated existing content or removed unwanted content. It is also possible that bots have performed edits that are not significant for diversity purposes. For example, *sitelink* is an edit type that is used to manage wiki projects, but does not add to the content of Wikidata and is therefore not used in the diversity measurement concept. Thus, editing a sitelink is not an indicator of having an impact on the diversity of Wikidata.

By investigating the edit types we can see which edit focuses received more bot edits and which activities are more popular among bots, and eventually how this might have impacted the diversity of topical domains in Wikidata. As mentioned before, our results have shown that bots and humans use similar edit types (cf. Section 6.2.1.1), however, their edits are significantly different from each other (cf. Section 6.3.2.3). Exploring how each user group uses edit types can shed light on what makes bot edits significantly different from humans. Is it only the editing speed that matters the most or the usage of edit types also differ between them? Exploring this allows us to better understand user community editing behaviors.

In Table 6.9 we see that only human users perform all of the 34 edit types that we have identified[37] in *Wikidata Revision History Dataset*. Bots have not performed four of the edit types that are mostly maintenance-related edits and some are solely performed by humans, while, tools have used the least number of edit types, i.e., 16 out of the available 34 edit types in *Wikidata Revision History Dataset*. Looking at diversity from the variety of edit types angle, we can say that human edits are more diverse than all others. Nevertheless, we see that according to dual concept measures of Entropy and Simpson's Index, the anonymous user group is declared to have used the edit types more diversely despite humans having a higher variety of edit types. We have very little knowledge about anonymous user group and the only information stored about them is their IP address. For this reason, understanding the reason for their higher diverse performance requires more research about anonymous users in Wikidata. Earlier, Figure 6.13 have shown similarity between the user groups of human and anonymous that they have edited Wikidata classes alike. An akin pattern was also shown between bot and tool user groups which perform automated edits. Since, we know that humans edit Wikidata manually, the resemblance of edits between human and anonymous user groups confirms that anonymous users also edit manually. The reason anonymous is considered to have a higher diversity of edit types, is due to the fact that Entropy and Simpson's Index are dual-concept measures and take balance alongside variety into account. It seems that in our dataset anonymous had a more balanced usage of edit types across the classes of Wikidata domains in comparison to other user groups, and so has earned the highest values for Entropy and Simpson's Index. For instance, in Figure 5.3 we can see that the distribution of editing volume over 29 edit types for anonymous is somewhere between 10 and 10.000, while, this distribution for humans varies from <10 to nearly 1 Million edits across 34 edit types which is very similar to the editing range of bots with 30 edit types. Since the anonymous user group has a more balanced distribution of editing volume across its edit types than human and bot user groups and a higher variety of edit types than the tool user group, it is considered the most diverse among all.

Table 6.9: Descriptive statistics of user group diversity based on the usage of edit types. Source: *Wikidata Revision History Dataset*.

	Variety	**Entropy**	**Simpson Index**	**Rao Stirling Index**
Anonymous	29	**2.349**	**0.858**	247121.7
Bot	30	1.978	0.801	**260800.7**
Human	**34**	1.981	0.757	256858.2
Tool	16	1.705	0.746	248317.4

On the contrary, according to Rao-Stirling's Index, bots are declared to have the most diverse usage of edit types in Wikidata classes, while anonymous users are rated as having the lowest diversity, making it challenging to determine the most diverse user group in this context. Earlier, in the comparison of diversity measures in Table 3.2, we observed that Rao-Stirling gives more weight to common categories, and higher variety doesn't have a significant effect on diversity. Since bots have

[37] List of all identified edit types from edit summaries in Wikidata in Table A.3

performed higher volumes of the most edited edit types, they are considered to have performed the most diverse edits, even though the human user group exhibits a higher variety of edit types. Simultaneously, we observe that the human user group has the second-highest values for two of the three measures, along with the highest value for variety in Table 6.9, which adds more consistency. The entropy measure has also identified the human user group as the second most diverse one, yielding the most relevant results so far. To interpret the results obtained from Table 6.9, we present a heatmap depicting the usage of edit types per Wikidata user group.

In Figure 6.14 a comparative view of the edit types usage by Wikidata user groups is exhibited. Here again, we started with the raw values and gradually filtered values to be able to have a visual representation of the values during the process. Recurrently, in the graph using the raw data no useful outcomes were displayed due to high differences among the values and, similarly, data normalized with RCA was not better than the raw values, while, data after filtering the values lower than 1 have shown more clear outcomes as can be seen in Figure 6.14. Here, we can see that not all user groups have shown interest in performing all edit types, e.g., tools are focused on a very limited number of edit types which supports the fact in Table 6.9 that tools performed the least number of available edit types.

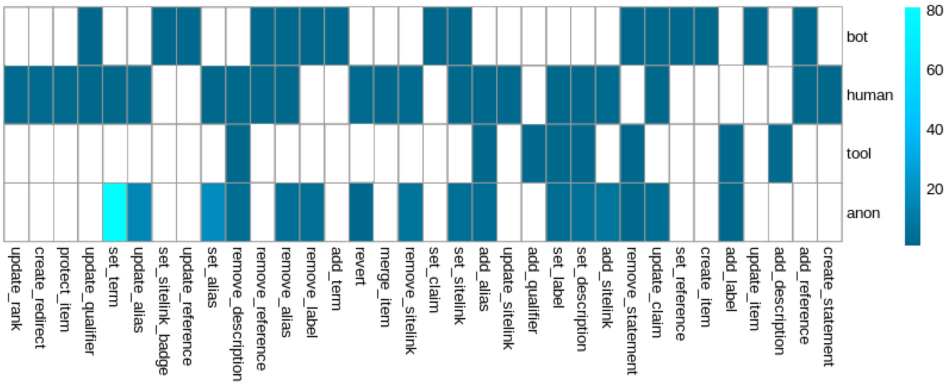


Figure 6.14: Heatmap of edit types performed by user groups in Wikidata revision sample. Higher values are represented with lighter turquoise color shades. Data is ordered with the highest values in the top right corner and the lowest values in the left bottom corner of the Figure. As can be seen, the most used edit types are the ones performing create, add, and update activities on the right side, while, edit types on the left side are the ones mainly used by human users with low amounts. Source: *Wikidata Revision History Dataset.*

Additionally, Figure 6.14 displays that similar to humans, bots also perform a rather diverse set of edit types and are not concentrated in a limited number of edit types like tools, despite sharing the automation behavior with tools. As can be seen, bots are focused on creating and adding new content, as well as, updating and removing the existing content. It means that bots are not only used to add new data but are also vastly used to improve and get rid of any erroneous content. Bots are specifically editing reference, item, and claim edit focuses. Although bots had fewer requests for adding references (cf. Section 6.2.1.1), they are actively editing references, i.e., performing add, set, update, and remove activities on references in Wikidata. Creating new items and adding new claims is another area where bots have concentrated their edits. This shows that bots have contributed to the diversity of items and influenced the variety, balance, and disparity of terms and statements (claims, qualifiers, and references) in Wikidata domains. Bots have used 30 out of 34 total edit types in our dataset. The remaining four edit types are *update_alias*, *set_term*, *update_rank* and *protect_item*. While, *set_term* and *update_alias* are general edit types to edit the Wikidata contents, *Update_rank* is not simple like adding new content or updating the existing ones; it requires decision-making based on some criteria like timeliness of values or trustworthiness of the sources. *Protect_item* is performed on certain item pages to protect[38] them against modification when those items face vandalism and damaging attacks in a large scale. Hence, *Protect_item* has solely been performed by human users who can have administrative rights. Both of the mentioned edit types require human intervention and so far had much lower occurrences (i.e., less than 150 revisions) so they were not automated and performed through bots.

Overall, we see that human users have treated the edit types more evenly than other user groups and neglected fewer edit types. Thus, humans seem to use more diverse edit types when editing Wikidata.

6.4.2 Impact of Bot Edits on Wikidata Domain Diversity

The results of our proposed concept for measuring Wikidata diversity status have shown that Wikidata has imbalanced domain coverage from the angle of the number of items per domain class and the number of statements per their items. The diversity ratings of these classes vary among the *balanced*, *imbalanced*, and *heavily imbalanced* values (cf. Table 4.2). Given that only a limited number of classes are balanced and the rest are designated as imbalanced, some even to a considerable extent, it serves as an indicator of low data diversity levels in Wikidata. This is consistent with our earlier observations that domains such as Geography and Media are more diverse compared to domains like Biology and Person. Here we can see the impact of bots on diversity, e.g., bots have very dedicated contributions in the classes Mountain, River, Lake, Album, and City which could be the reason for these classes to have higher diversity ranking, i.e., higher variety of items and balanced contributions. At this stage, bots do not seem to negatively impact diversity. Further, in classes with lower diversity levels, we see that bots are not focused. This could mean that bots could be used to increase variety and balance the content of Wikidata classes.

Searching for common factors among balanced classes that can differentiate them from others, we find that balanced classes usually have a higher number of items (cf. Table 4.1). In other words, balanced classes have a higher variety of items. They seem to have been added in a defined way so that all of them have a similar number of properties. This looks very similar to bot edits which import data from a data source with a predefined and even number of properties. An exception here is the Class Country with 180 items. Since the number of countries is in the hundreds only and the information about the countries is commonly available, we don't wonder if bots have not actively edited this class. Nevertheless, other balanced classes have at least one dedicated bot that added a large number of items with a defined number of statements into these classes.

[38] Protection policy https://www.wikidata.org/wiki/Q4616470 [Accessed: 27.10.2022]

The classes with balanced contributions as listed in Table 4.2 include Film, Book, Album, Mountain, Country, and Gene. Their respective dedicated bots are Symac bot[39] for movies data, Research Bot1[40] for book-related content, MineoBot_2 [41] and WhidouBot[42] for music album information, PLbot 3[43] for data on mountain and geography related themes and ProteinBoxBot[44] for adding gene and protein details in Wikidata. Thus, a second common factor found among the balanced classes is having received dedicated contributions from specific bots. This dedicated contribution could be due to the availability of data formats that are easily suitable for automated import[45].

Bots' focused contributions in these balanced classes indicate bots' positive impact in making the Wikidata classes balanced. Nevertheless, the classes where bot contributions are not focused, have been left with a low number of items and different levels of property coverage. For example, the class Grass has only six items and even these six items are not well covered (see Table A.2 on page 163). It is possible that Wikidata domains are inherently more diverse than what current measurement methods can accurately capture. This suggests that the instantiation of Wikidata items is not uniform across all cases [46], as becomes evident when comparing items related to *Douglas Adams (Q42)*[47] and *Dany Saadia (Q5221412)*[48]. For this reason, as previously noted, addressing the class hierarchy issue of Wikidata would contribute to a more precise assessment of Wikidata's domain or class diversity and understanding a the actual influence of bots on diversity within Wikidata.

In addition to bots' capacity for adding a substantial amount of well-balanced content, they can also provide more accurate values for specific attributes compared to humans. For instance, they can accurately input information like city or country populations and geographical coordinates. However, bots, like any other tools, have the potential to introduce significant amounts of erroneous data if not managed carefully. Therefore, bots can prove highly advantageous when utilized with a clear purpose, well-structured planning, and human oversight.

At the item level, however, the existing research shows that bots are less diverse when adding labels and references. Our results have shown that most bots tend to import data from Western languages of Wikipedia, where English is at the top of the list (see Figure 6.6 on page 114). Eastern countries like China or India are the

[39] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Symac_bot_3

[40] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Research_Bot

[41] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/MineoBot_2

[42] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/WhidouBot

[43] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/PLbot_3

[44] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/ProteinBoxBot_2

[45] For example Gene database: https://www.ncbi.nlm.nih.gov/gene/, movie database: https://www.themoviedb.org/, and geographical data set: https://data.world/cegomez22/geographic-location-dimension [Accessed: 31.10.2022]

[46] *Douglas Adams (Q42)* the English writer and humorist is declared as an instance of the class *human (Q5)* and has *writer (Q36180)*, *novelist (Q6625963)*, *comedian (Q245068)*, *screen writer (Q28389)*, to mention some, as his occupation. On the other hand, *Dany Saadia (Q5221412)* the Mexican filmmaker and podcaster is not only an instance of *human (Q5)* but also an instance of *film producer (Q3282637)*, *podcaster (Q15077007)* and *screenwriter (Q28389)*. Similarly, he has *film director (Q2526255)*, *podcaster (Q15077007)* and *screenwriter (Q28389)* as his occupation which show redundant information.

[47] https://www.wikidata.org/wiki/Q42[Accessed:28.04.2020]

[48] https://www.wikidata.org/wiki/Q5221412[Accessed:28.04.2020]

highest populated countries and we would expect somewhat similar contributions or data representation from these countries; however, studies have shown no relation between the population of a language and the data representation of that language in Wikidata [142]. On the other hand, we see the Catalan language to be one of the very well-represented languages in Wikidata despite its rather lower population size in comparison to languages like Chinese, Hindi, or Persian with much larger populations than Catalan. Since the Catalan language has received high amounts of bot edits, it can be a clear indicator of how automation could help a language or culture to be well represented and even overshadow other ones on the row.

At the statement level, it's evident that Wikidata items within the same class do not exhibit uniform completeness in terms of the number of properties. Additionally, a majority of items in these classes appear to be in their early stages of creation or inception. Our findings confirm the expected trend that the *create_item* edit type is predominantly carried out by bots, resulting in a significant number of newly created items. These items might originate from lists or be imported from external sources, often existing in a very basic form. Because bots can create a large number of empty items at once, they can easily change the balance in a class or domain. While their influence on domain and class is visible, how bots impact item and statement levels (i.e., plurality) is subject to further study and is yet to be explored.

Currently, mechanisms exist to suggest such empty or incomplete items to Wikidata editors. However, if these basic or empty items are not created with proper care, they can end up being of little utility, as they may require a significant amount of time and effort for human users to populate the incomplete items generated through automation. Once more, bots can exhibit optimal efficacy when strategically employed to import pre-digitized data in a relatively comprehensive state and are subject to community oversight and control.

In summary, bots possess the capability to influence diversity within Wikidata, particularly at the domain/class level. While they have engaged with all of the Wikidata classes in our dataset, their pronounced contribution to the balanced classes serves as clear evidence of their role in achieving balance in these categories.

Moving forward, we will elaborate on how bots can contribute to knowledge diversity in Wikidata using our proposed concept for measuring diversity within the platform.

6.5 Summary

Bots are the most active contributors in the Wikidata community. They have been actively editing Wikidata shortly after the launch of Wikidata in late 2012. Despite their abilities for performing speedy and batch edits, they have remained a rather less explored user group of the Wikidata community. This section has explored bots in the Wikidata context from what they are and how they come into being to what they actually do in Wikidata and how their editing patterns are in comparison to human users.

Bots are generally defined as automated assistants which mainly perform simple and repetitive tasks. In the Wikidata context, bots are user accounts with special rights of high-speed edits and are operated by human users called operators. There is a defined procedure for requesting bot rights in Wikidata. The requests are made through the Wikidata platform named Requests for Permissions (RfP) pages. These requests are discussed by the Wikidata community and then decided upon. The majority of requests have been approved and only a small number of requests were denied. Even the main denial reasons are caused by the unresponsiveness of the bot-requesting operators and not the Wikidata community. This shows the openness of the Wikidata community to bots and their approval for bots to share similar editing tasks as human users like adding and updating data.

Looking at the actual bot edits in Wikidata, we find that bots are performing content editing tasks and are generally dealing with data rather than performing maintenance tasks. Maintenance tasks in Wikidata are mainly done by human users. This is the opposite of how the Wikipedia community works, human users are responsible for editing articles and bots mainly perform maintenance tasks.

While bots and humans may perform similar types of edits in Wikidata, there are notable differences in their editing patterns. Bots, due to their ability to perform high-speed edits and their emphasis on specific types of edits, stand out as distinct editors from human users. These differences in editing behavior between bots and humans provide a potential explanation for the existing domain imbalance in Wikidata. Since bots contribute a significant portion of the content in Wikidata, their unique editing patterns may contribute to the observed imbalances.

To investigate the impact of these differences, we conducted a detailed analysis of bot edits, specifically focusing on their contributions to Wikidata classes and their editing behavior, including the types of edits they perform. By examining the usage of different edit types by bots, we aimed to uncover whether these differences are indeed contributing to the domain imbalance in Wikidata. Through this investigation, we sought to shed light on the relationship between bot editing behavior and the observed imbalance in content distribution.

We have demonstrated that bots have played a significant role in achieving balance within classes that exhibit a high variety of items with balanced content. We observed that these classes typically have at least one approved bot associated with them. This finding suggests that bots can also contribute to bringing balance to imbalanced classes within Wikidata, given proper planning and control by the Wikidata community. By strategically leveraging the capabilities of bots, it is possible to address the diversity gaps and promote a more equitable representation of information across different classes in Wikidata.

Additionally, the experience with diversity measures in our case confirms the effectiveness of employing multiple diversity measures for more insightful results. Relying on a single measure would not provide a comprehensive understanding of the underlying reasons for the observed results. Our data analysis would lack alignment, and the support for a unified result would be diminished without the convergence of multiple measures.

In the next chapter, we present our recommended approach for enhancing diversity in Wikidata domains and classes through the utilization of bots. Drawing from our research findings and analysis, we propose a strategy that leverages the capabilities of bots to address the existing imbalances in content distribution. Our approach focuses on implementing controlled and planned bot interventions to target the imbalanced domains and classes in Wikidata. By strategically deploying bots with specific tasks and considering the diversity goals, we aim to foster a more balanced representation of knowledge across different domains.

Furthermore, we provide an application use case in the next chapter to demonstrate the practical implementation of our proposed approach. This use case highlights how bots can be employed to improve diversity in a specific domain or class within Wikidata, showcasing the potential impact of our recommended approach. Through our proposed approach and the accompanying use case, we aim to contribute to the ongoing efforts to enhance diversity in Wikidata and promote a more inclusive and comprehensive representation of world knowledge.

Chapter 7

RECOMMENDATIONS ON DIVERSITY IMPROVEMENT

Human knowledge and culture have witnessed the transformative impact of the digital world. The availability of knowledge in digital format enables its accessibility and safeguards it from potential loss, such as the destruction of hard printed copies due to fire or other natural disasters. Moreover, digitalization provides a means to preserve and protect languages and cultural heritage from deterioration and manipulation over time.

Wikidata, as a collaborative knowledge base, plays a crucial role in this digital landscape. It embraces the concept of plurality, enabling the storage of diverse perspectives and opinions on various subjects. By allowing users to consider multiple angles and explore existing claims, Wikidata provides a comprehensive and nuanced understanding of a given topic.

Building on this foundation, we propose recommendations for improving diversity within Wikidata and harnessing the potential of bots to bridge the existing diversity gap in domains and classes. Our recommendations are aimed at fostering a more inclusive and representative KB. One that accommodates a wide range of perspectives and cultural nuances.

To illustrate the practical implementation of our recommendations, we present a use case where we apply these strategies to improve data diversity in a specific context. By showcasing the application of our recommendations, we aim to demonstrate their effectiveness and encourage their adoption in broader efforts to increase diversity within Wikidata.

Ultimately, our recommendations emphasize the importance of embracing diversity and leveraging technological advancements, such as bots, to promote inclusivity, preserve cultural heritage, and ensure a comprehensive representation of human knowledge in the digital age.

7.1 Diversity Improvement through Automation

Earlier, we observed that the focused contributions of bots in certain classes resulted in an imbalanced coverage across different domains in Wikidata. This imbalance has led to an overall low diversity status in Wikidata. In a similar way, research has also shown that bots were involved in making some languages dominant leaving other ones overlooked. While this indicates the potential of bot edits to have a negative impact on diversity in Wikidata, this potential can also be used for a positive impact by using bots in classes with low variety and imbalanced content. The use of automation through bots and tools in the low-diversity areas of data in Wikidata can be used to bring the balance back in the underrated languages and overlooked domains/classes. Automation has been a widely used tool by Western communities. In addition to having active human contributors, they have widely taken advantage of automation. For this reason, we see incomparably more content in Western languages in Wikidata. Inspired by this, we recommend using bots in the same manner in overlooked areas of Wikidata to use the potential of automation for data balancing. However, we first see the need for an active community that would control this automation process and make data more plural by augmenting automation with manual edits.

In our recommendations here, we differentiate between social mechanisms and technical mechanisms. Social mechanisms deal with how communities can perform more diverse tasks, or how they can be enabled to contribute content that improves the overall diversity of Wikidata. Social mechanisms are rather implicit indicators of content diversity and do not directly contribute to increasing the diversity of knowledge. Technical mechanisms refer to the steps or activities that need to be performed and provide direct results. However, both of these mechanisms are closely tied together; social mechanisms pave the way for technical mechanisms to be implemented.

Our recommendations for diversity improvement in Wikidata through bots mainly evolve around the topic of domain/ class diversity. Since the contents of domains/ classes come from contributors, we start from the social angle of user diversity improvement and then move to technical aspects and define concrete steps for increasing data diversity.

7.1.1 Improving User Diversity

As mentioned before, in a KB the knowledge is contributed by editors/ contributors and more diverse contributors are indicators of more diverse content. Hence, using mechanisms to move towards a more diverse contributing community means taking a first step in the direction of making Wikidata knowledge more diverse.

User diversity is mainly measured considering the background of the users. This is because, with different backgrounds, humans have different experiences, interests, values, and beliefs. Their background can influence the data they contribute, so with more humans from different backgrounds, we can expect a wider range of topics of interest and opinions, and, eventually, user diversity results in more diverse content in a KB. In a similar manner, bots can be programmed to contribute more diverse content to the KB. Bots are operated by humans and lack personal backgrounds, so we can measure the diversity of bots based on the tasks they perform. User diversity through bots can be improved in the sense that bots can be created for more diverse tasks, dealing with more diverse sources and languages, and bringing balance to the imbalanced domains of Wikidata. By making bots more diverse from the angle of domain coverage, language contributions, and usage of a higher variety of resources, we will have bots improving the overall diversity of Wikidata. Since diversity is the attribute of a whole system, bots should be programmed with the big picture in mind so that their contributions can have a positive impact.

Since the user community is an influencing factor in a KB, many active communities have used automation for higher efficiency and better results. Although Wikidata has centralized the data in language-independent items, Wikidata communities seem to be following the community structures of Wikipedia which are mainly created based on languages. For this reason, we see different coverage levels among language labels in Wikidata which implies that not all language versions have an active community. In some cases, the language communities might also have sub-communities if the language is covering more than one country. For example, the community contributing to the Persian language is mainly from Iran, and the content related to Afghanistan is not given much attention. This indicates that there is no active community that could add data about Afghanistan in the Persian language.

While there are attempts going on to encourage people from different languages and cultures to start contributing and forming communities like Wikimania[1] and Wikimedia Movement Strategy 2030[2], we suggest the usage of bots in the newly created communities. These communities can use bots to create a basic structure of content and prepare them for further input by human users. For example, bots can create new items and let others fill in the statements. In this way, bots may also attract individuals related to these communities by providing a foundational structure in their language, thereby arousing their interest in contributing. This, in turn, can lead to greater diversity among Wikidata users.

Wikidata is more popular in the West because Western individuals can relate to this knowledge base, which represents their values or topics of interest. When people have a feeling of resonance, they are more inclined to contribute further and share facts they deem accurate. This effect is amplified, particularly when they witness their contributed data being actively utilized by other projects. For attracting more users from different parts of the world we need to develop a sense of connection between them and Wikidata. This connection can be created starting by adding content in their language. Once people find some information in their own languages, they are likely to get interested and build on that content to extend the existing knowledge in their language and gradually move towards making a community. The community can then decide to add further topics from their existing sources. When becoming more experienced and professional, the editors can develop their own bots to share some tasks with bots for more content coverage of their language and topics. This will aid in bringing balance among domains, classes, and language coverage that will eventually make Wikidata's knowledge more diverse.

[1] "Wikimania is the annual conference celebrating all the free knowledge projects hosted by the Wikimedia Foundation." https://wikimania.wikimedia.org/wiki/2023:Wikimania

[2] "By 2030, Wikimedia will become the essential infrastructure of the ecosystem of free knowledge, and anyone who shares our vision will be able to join us." https://www.wikimedia.de/2020/en/themen/movement-strategy/

7.1.2 Improving Data Diversity

Data diversity is an important issue in the Wikidata context since data is the central element of a KB. The data model of Wikidata allows us to look at the data from the angle of its parts. In this study, we have mainly focused on domain-level diversity when elaborating on data diversity. We explain our recommendations on where to use bots for a more balanced coverage of content in the imbalanced classes of Wikidata domains. Bots can also help in adding multilingual labels and descriptions and add content from different sources of knowledge to improve the variety, disparity, and balance of knowledge in Wikidata.

7.1.2.1 Domain/ Class Diversity

Bots can be used to add new items along with their statements and associate these items to the classes of Wikidata domains. New items can increase the variety, and even contents can ensure balance. The more items that are added, the higher the number of classes we can expect. A good example of a balanced class is the class Gene in the biology domain. A glance into Wikidata edit history data shows that in the biology domain, all of the classes are low in variety and balance except for the class Gene which was mainly contributed through automation. Class Gene has balanced content with more than 10K items and is rated as balanced (cf. Table 4.2). This approach can also be applied to the other classes of Wikidata to improve diversity.

7.1.2.2 Item Diversity

We have mentioned that bots can be used to increase domain-level diversity by adding new items. Here, we discuss how items can be edited by bots to improve domain and class diversity in Wikidata. Bots can play an active role in improving the overall diversity by either adding new items or editing the already existing ones.

We recommend that bot operators consider the outcome of their bot edits on diversity when planning to program or run a bot to add new items to Wikidata. It is important that the data being added is checked for the number of statements so that the new items ensure balanced content across all of the newly added items and improve the diversity levels of that class/ domain. This way balanced classes/domains will keep their balance and imbalanced classes/ domains will be a step closer to getting balanced content and improved diversity.

Similarly, when used to edit the existing items, bots can add multilingual terms and statements from a variety of sources to make Wikidata domains more diverse as explained below:

Language Diversity. Although Wikidata has multilingualism by inheritance and users can add labels in over 400 languages[3], we don't see all of these languages being equally represented. The visible reliance of bots on importing data from the Western language versions of Wikipedia could shed light on the reason behind the dominance of Western languages in Wikidata. Our preference for improving diversity here is to reduce the dominance of Western languages by bringing more balance among the Wikidata language terms so that everyone in the world can feel included. One recommendation here is to use bots for importing data in the less

[3] https://dl.acm.org/doi/fullHtml/10.1145/3184558.3191643

covered languages of Wikidata from the existing Wikipedia language versions first, and later from other available sources. In any language that is less covered (in our case Persian), the community can look for a bot to increase the coverage of that language by importing content from its existing Wikipedia language version into Wikidata, for instance, or new topics from other existing sources. The contents from Wikipedia should be added along with their original sources because Wikipedia itself cannot be considered a primary source of knowledge. The usage of bots here could alter the language imbalance and benefit the overlooked languages in the same way that bots have caused some languages to become dominant. In addition, the usage of bots for importing multilingual terms from Wikipedia, in the first step, could also encourage under-rated language speakers to form communities and further enrich their languages. In particular, this will help people who are not familiar with Western languages but can contribute to their own language.

In addition, Wikipedia itself has an imbalanced distribution of language versions, and the data in different language versions are not easily comparable. Importing Wikipedia data into Wikidata, and helping to make under-represented languages visible in Wikidata, not only helps to centralize all these knowledge versions, but also highlights a corner of globally agreed or disagreed knowledge in the Wikimedia sphere and the world.

Source / Reference Diversity. Adding the sources from which the data come from is highly recommended in Wikidata, as it is the way of making the data and statements reliable. Research has shown that Wikidata data is mostly imported from Wikipedia, while Wikipedia itself is a secondary database and cannot be used as a source. In addition, the existing shortcomings of Wikipedia (e.g., gender bias) are also transferred to Wikidata. While the data in Wikipedia is rather easy to import into Wikidata and can serve as a first step in enriching Wikidata with multiple languages, our recommendation would be to also consider enabling or developing bots that can import data from other sources as well. The content in languages other than Western languages is not fully digitalized and ready to be used in Wikidata; for this reason, such languages need to consider another step for resource digitalization as well which is explained in Section 7.2.2.2.

Adding content from Wikipedia can certainly enrich the content in Wikidata, but it is important to ensure proper referencing. We recommend that references include not only the link to the specific Wikipedia article but also, whenever possible, provide the exact source from which the data were taken. Currently, many of the references in Wikidata statements only indicate the language version of the Wikipedia article as the source, without providing a direct link to the article itself. To enhance diversity in Wikidata, it is crucial to define bots that can import data from a wide range of primary sources. This approach helps mitigate biases, concentrated topics, languages, and opinions that may be present when relying solely on Wikipedia.

In the next section, we will apply our recommended approach for improving diversity in Wikidata domains and classes through a use case, which will be explained in detail.

7.2 Use Case: Preserving the Historical Names of Districts in Herat

In this use case, we explain how diversity can play a role in safeguarding or preserving the history of my birthplace, which is currently at risk of history manipulation, one of the motivating factors in this research. Here, we first provide a glance into the history of Herat and how the language of its natives is at risk now. We describe Herat as an example of all the other cities in Afghanistan that have fallen victim to comparable political agendas pursued by Afghanistan's rulers. These agendas are often aimed at securing the continuity of power for their own succeeding generations while minimizing opposition from the people. Then, we present our recommendations for preserving the language, culture, and history of Herat through the use of the diversity concept and bots in Wikidata. This approach is equally applicable to other cities in Afghanistan or anywhere in the world that face challenges similar to those of Herat. Our defined procedure can be advantageous for other communities that are underrepresented on Wikidata, as well as for languages considered underserved, facilitating improvements in their diversity.

7.2.1 Herat at the Risk of History Manipulation

Herat is a city on the Silk roads[4] and a nomination for the UNESCO list of World Heritage[5]. It is located in the northwest region of Afghanistan. Herat is home to many well-known Persian-speaking scholars, writers, artists, and scientists of their time in the region[6] who have all their literature in the Persian language. Prior to becoming a part of Afghanistan, Herat was called Pearl of Khorasan[7] due to serving as a hub for trade and its rich culture and monuments.

Nearly, one and half centuries ago Afghanistan got its current geographical form and Herat was also included in this geography under Afghan/Pashtun rulers by the British to shield British India from Russian attack[8]. Herat and the cities in northern Afghanistan are mainly Persian-speaking people and not native to Pashtu-speaking rulers who came from the southern cities to conquer the cities in the north. To show their dominance over other ethnic groups, Pashtun rulers began distributing lands to Pashtun nomads and changing the names of many areas from Persian to Pashtu [357]. After the fall of Afghanistan to the Taliban, the procedure of language and ethnic cleansing has gained momentum in various ways[9]. Persian is being eliminated from governmental formal letters[10], signboard of government institutions[11], and re-

[4] https://en.unesco.org/silkroad/content/herat[Accessed: 17.10.2022]

[5] Herat city exists on the tentative list of UNESCO World Heritage. On 02.07.2021 media announced Herat city to be listed as a UNESCO World Heritage Site but was not announced on the list of UNESCO World Heritage probably due to the fall of Afghanistan in August 2021.

[6] Fakhruddin Razi a chemist and polymath scientist, Jami a poet and writer, Kamal ud-Din Bihzad a renowned Persian miniature and Khwaja Abdullah Ansari an outstanding figures of the 5th/11th century in Khorasan, as some examples.

[7] Gammell CPW. The Pearl of Khorasan: A History of Herat. London: Hurst; 2016.

[8] Herat History, Medieval Period: https://www.iranicaonline.org/articles/herat-iii

[9] Taliban implicated in mass killings of Tajik men (Tajiks are the majority of Persian-speaking people in Afghanistan/ the Iranians of the East [78]), The Taliban Target Tajiks Yet Again, Afghanistan: Taliban Forced Rift Between Country's Two Main Languages[Accessed: 17.10.2022]

[10] Taliban abolishes the Persian language from Supreme Court bill[Accessed:17.10.2022]

[11] Taliban Group Removes Persian from the Sign Boards at Education Directorate of Herat Province. [Accessed 12.10.2022]

cently from school textbooks. Meanwhile, indigenous inhibitors are being forcefully displaced mainly in Persian-speaking areas in the north, west, and central parts of Afghanistan[12]. Therefore, we can use the power of the Web to preserve the existing history of the region and prevent the elimination of our identity from this geography.

7.2.2 Proposing a Procedure for Preserving the Historical Names of Herat Districts/ Subregions

The World Wide Web is a rather new phenomenon and a Western product, thus, the Western world remains the primary user and contributor of the Web. Persian is an Eastern language, and most of the existing literature in Persian is not yet digitalized and ready to be fully represented on the Web. For this reason, we need to define a proper procedure for adding missing information in the Persian language. Here, we define this procedure through a use case for adding the historical names of the districts in Herat, some of which were renamed from Persian to the Pashtu language [159]. Many other cities like Balk[13] and Kabul[14] have also gone through a similar history and their districts were renamed. Our defined procedure is applicable to all of the cities, however, we here focused on Herat only as an example. Adding the historical names of the sub-regions or districts from reliable sources in Wikidata will not only help preserve the history of the city and language of the indigenous people, but will also provide access to the world and let them look from the lens of the people who have always been silenced for their identity, not the rulers that want to show their ethnic dominance.

In the following, we present our defined procedure for improving Wikidata diversity based on our proposed concept for measuring diversity in Wikidata. Our technical mechanism, in other words, defined steps are:

1. Establishing a community
   (a) Defining the needed roles for the members of this community
   (b) Defining mechanisms for communication/ collaboration among community members
2. Developing a community agenda
   (a) Assessment of language content
   (b) Identification of content gap
   (c) Setting goals
      i. Prioritization of topics to be added
      ii. Digitalization of paper-printed contents
      iii. Automation of tasks/ Creating bots for needed tasks
3. Monitoring the community progress

[12] Forced Displacement under the Taliban, also a Legacy of the Past (?), Afghanistan: Conflict and internal displacement under the Taliban regime, Afghanistan: Taliban Forcibly Displace Civilians

[13] The City of Balkh: Ancient Capital of Bactria and Centre of Buddhism and Zoroastrianism along the Silk Roads [Accessed: 18.10.2022]

[14] https://www.iranicaonline.org/articles/kabul-index [Accessed: 18.10.2022]

(a) Evaluate the goals and achievements

(b) Setting new goals

(c) Repeat step two

Next, we provide the mechanism for preserving the historical names of Herat districts following the above-mentioned steps.

7.2.2.1 Establishing a Contributing Community.

In any collaborative system, contributors are the main actors and have their influence on the data of that system. To the best of our knowledge, no sub-community exists in Wikidata that is focused on adding or updating data related to Afghanistan. We have seen that diversity is not limited to data only and in a KB like Wikidata it is directly related to the users who contribute or consume this data. For this reason, to ensure the diversity of data regarding the history of Afghanistan in Wikidata, a small sub-community should be created that can perform the basic tasks of data contributions. The community can gradually grow and increase its contributions, which can eventually improve the overall diversity of Wikidata.

In the contributing communities where humans and bots collaborate, we see higher content levels. In Wikipedia, for example, the language versions with the highest number of articles are mainly created by human users, e.g., in English[15] Wikipedia 94%, in German[16] and French[17] Wikipedia language versions 87% of articles are manually created by humans and bots create less than 5% of their articles. This is obvious that Wikipedia is a text-based KB and bots might not be very useful in creating content, still, we see a high contribution of bots in the Wikipedia language versions of Dutch[18] with 49%, Arabic[19] with 40% and Persian[20] with 32% of the articles being created by bots. Although the latter three are in a much better position than many other languages, they yet have to compete with the big and very active communities of the dominant languages.

As mentioned earlier, human contributors form the center of a contributing community, and bots are used as assistants to achieve more efficiency. Humans have the brain and bots have the speed, and the result of their combination is visible from the status of well-represented domains and languages in Wikipedia and Wikidata. Bots are operated by humans and when programmed properly, can have positive contributions and improve diversity through variety, disparity, and balance of topics, contents, and languages. However, if not controlled and operated with care, bots can not only cause high error rates but can also create content that would cost humans the time and effort to revert or undo. For this reason, it is important that bots are used under the observation of human users and with proper planning to achieve positive results. We propose the following categories of contributors for a sub-community of our use case:

[15] https://en.wikiscan.org[Accessed: 18.10.2022]
[16] https://de.wikiscan.org [Accessed: 18.10.2022]
[17] https://fr.wikiscan.org[Accessed: 18.10.2022]
[18] https://nl.wikiscan.org [Accessed: 18.10.2022]
[19] https://ar.wikiscan.org[Accessed: 18.10.2022]
[20] https://fa.wikiscan.org [Accessed: 18.10.2022]

- Domain Experts: These individuals possess sufficient computer knowledge to directly access Wikidata and contribute to it. They have the ability to explore Wikidata for existing and missing topics. Additionally, they can search for missing topics or other subjects of interest on the Web.
- Non-technical Domain Experts: There are many people who are experts in the history of Afghanistan and are aware of the existing printed and old resources, however, they are not fully familiar with computers and cannot directly contribute to Wikidata. These experts can serve as consultants to provide feedback on the content coverage and details and introduce reliable sources.
- Technical Experts: The ones with expertise in developing and deploying bots. They can use automation to import the missing content identified by the domain experts.
- Editors: Anyone who knows how to edit and contribute to Wikidata. They can contribute, monitor content, and provide feedback and perspectives.
- Bots: Scripts created and operated by technical experts in order to add the missing topics, contents, and resources identified by domain experts.
- Translators: These contributors can add labels, descriptions, aliases, and statements in multiple languages and translate the content from Persian to other languages, and vice versa. Translators can help in serving data beyond language boundaries and can contribute to language diversity.

Now that we have a draft of the roles of community members, the next thing to do is to define a mechanism to establish communication and collaboration between people with these roles so that they can work together.

We recommend that domain experts, in the *first step*, conduct a basic evaluation of Wikidata regarding the content related to the history of Herat City and identify any gaps. In the *second step*, they can:

- Develop a report to shed light on this evaluation and existing gaps,
- explain how filling these gaps can bring the facts that are only available in paper-printed formats to be digitally conserved, made safe from elimination or manipulation, and made accessible through technology, and
- use this report to inform and motivate other people who have knowledge in this area to contribute.

In the *third step*, they can start building a network of people who have an interest and/or the ability to fill these gaps and become a volunteer contributor. Networking can be done through different means of communication, especially social media platforms. This is because due to the continued instability in Afghanistan, many people have migrated to different parts of the world, and it is only possible to gather all through online platforms or in a hybrid style.

Once a group is formed after one or more meetings and discussions on the importance of filling these gaps and working to prevent history from being forgotten and manipulated, in the *fourth step*, working groups can be established. One working group can take responsibility for contacting Wikimedia Foundation's Community

Development[21] team for their support. Meanwhile, other working groups can start to work on developing a community agenda, setting goals, and defining monitoring mechanisms to measure their progress.

Finally, the social structure is in place, and a small community with the above-mentioned roles can begin by taking inventory of their areas of interest to assess what already exists on Wikidata and what needs to be added. After identifying the missing content and the new topics that need to be added, the technical or concrete steps explained below can be performed to enhance the domain coverage of topics related to Herat history.

7.2.2.2 Improving Domain/ Topical Coverage.

For improvement in diversity at the domain level, we once again refer to our proposed concept for measuring diversity in Wikidata, where data diversity can be dealt with from the two angles of class and item. Diversity can be improved at the class level by covering new topics, or at the item level by dealing with the data more closely considering each part of the item.

As mentioned earlier, domain experts need to first examine the existing topics in Wikidata in comparison to the existing literature and highlight the missing topics or contents in Wikidata. Furthermore, the topics representing the history and culture of Herat are mainly available in the resources which are in the Persian language. For this reason, we look at the topical domain coverage from a language lens and explore how well-represented the Persian language is on the web. We can then decide if we need to look for resources solely in the printed media or we can expect some resources on the web as well. The Persian language is used by 2.6% of the websites around the globe and comes in the eighth position[22] which is a rather good position than many other languages in the world. There are many topics and articles already present on the Web on websites like Iranica[23] Encyclopedia and Wikipedia Persian[24]. These online encyclopedias are contributed by the contributing communities which are mainly Iranians and cover topics mostly related to Iran. The contributions from the Persian-speaking community of Afghanistan are less visible. Most of the content regarding Afghanistan in Wikidata comes from Western sources[25] or don't have any sources mentioned. For this reason, domain experts in our proposed community need to explore the web, in addition to the printed media, and find out what is already present and what is missing, so that the existing contents can be reused and the missing topics can become the focus of the task. For example, the glass-making industry in Herat which produces dishes and decoration pieces out of glass following the glass-making approach from thousands of years ago is currently on

[21] "The Community Development team at the Wikimedia Foundation works to support resilient and growing communities by helping volunteers build the capacities and skills needed to grow their contributions and communities in the free knowledge movement." https://meta.wikimedia.org/wiki/Community_Development [Accessed: 09.05.2023]

[22] Usage statistics of content languages for websites[Accessed: 18.10.2022]

[23] https://www.iranicaonline.org

[24] https://fa.wikipedia.org/

[25] Source for the name Shindand comes from *"L. W. Adamec, Historical And Political Gazetteer Of Afghanistan, Vol. 3, Herat and Northwestern Afghanistan, Akademische Druck-u. Verlagsanstalt, 1972, ISBN 978-3201009423, p. 343"* in Wikipedia page on Shindand district of Herat. [Accessed: 18.10.2022]

the verge of effacement[26]. Similarly, the Shawl weaving or cloth weaving industry which produced hand-made silk shawls from the past centuries is also gradually fading out[27]. Both of the mentioned topics are important parts of the Herat culture and are in need of preservation. There are some news articles and videos regarding these topics available on the web. Domain experts can look for additional printed resources and then plan to preserve these topics on Wikidata.

Here, we explain our defined procedure for improving topical coverage of data regarding the history and culture of Herat.

i. Getting an Overview of the Existing Topics and Sources on the Web. Following the compilation of a list of existing topics concerning the history of Herat available on the web, domain experts can then evaluate the suitability of these materials as reliable sources, ensuring they are appropriately referenced and determining their potential for incorporation into Wikidata. In particular, they can look into Wikipedia or other KBs where data reuse and import into Wikidata is a rather easy task. The prepared list can then be compared with the existing literature and accredited sources that are not yet digitalized. Non-technical domain experts can identify the missing topics, details, and sources after going through the mentioned list and providing suggestions in this regard.

ii. Identifying New Sources. Most of the resources in non-western languages like Persian are not yet digitalized. In order to import them into Wikidata and allow bots to use them, we need to, first, define a procedure that can put these resources into digital format and ready to be utilized on the KBs.

Based on the overview of the resources that already exist on the web and the resources that are not available in digital format, a list should be prepared to contain the resources that need to be imported into Wikidata. The list should contain a complete reference of the resources e.g., books, articles, newspapers, or any other accredited resources, and be categorized into two categories:

- **Digitalized sources:** These sources should be evaluated for how complete and reliable they are. The ones that are missing details or come from unknown sources should be listed separately from the sources that are ready to be used.
- **Non-digitalized sources:** These sources are only available in hard copy or in printed form.

At this stage, the entries of the non-digitalized category should, first, go through the digitalization process to become usable in Wikidata. The very first benefit of storing these sources in a digitalized format would be preserving these sources themselves. A data source file should be created containing the required information for each of the entries that can be added as Wikidata items through a bot.A digital copy of the item should also be created and stored both locally and online on platforms that facilitate data preservation. After being stored as Wikidata items, these digitalized sources can then be used as references to claims and qualifiers. These sources can also provide the possibility to add new topical domains, items, and details to the existing data in Wikidata.

[26] Glass industry in Herat on the verge of breaking https://www.youtube.com/watch?v=L2DHaJYvaXg [Accessed 03.11.2022]

[27] Shawl weaving the art of indigenous people of Herat

iii. Identifying New Content. After performing the basic step of making the sources ready for use in Wikidata, the next step is to compare the existing Wikidata contents regarding Herat districts with the newly identified sources. The result of this comparison can identify new items, inaccurate content in the existing items or missing details in different parts of an item that should be updated, as explained in the following:

- **Items:** The comparison of Wikidata items with the newly identified sources can result in a list of items that are missing and need to be created as new items. These items should be listed and a source file containing adequate details about these new items should be created for a later automated import into Wikidata. It is important that existing items in Wikidata are also identified so that they are not created again. In the existing items, we then take a closer look into each section of the item which is explained next.

- **Terms:** Adding multilingual labels, descriptions, and aliases is essential for a proper representation of an item to the world. When adding new items it's better to include multilingual terms.

- **Properties:** Currently, there is no unified way to represent facts through Wikidata properties. For example, we see multiple ways for storing the previous names of a place in Wikidata. These names are usually stored in the property *official name (P1448)* as in Constantinople (Q16869)[28], sometimes stored in an alias like Shindand (Q2714337)[29], and there are also cases when different properties are used to convey the fact that this place was previously known by another name[30]. Historical names bear an important role in the culture of a region and should be represented in a proper way. Thus, a unified approach for the usage of properties should be defined after having a glance at the items of similar topics in Wikidata.

- **Values:** Not all values are precise enough in Wikidata, for example, Tahir ibn Husayn[32], the Abbasid caliphate general and governor, was born in the Poshang/ Foshanj[33] district of Herat, and he is famous as Tahir Foshanji. In Wikidata the value for his birthplace is Herat and is referenced from Italian Wikipedia. The value is rather general and not precise enough to mention the district which could show the historical importance of this district. Having precisely mentioning the area of his birthplace provides more accurate data on the history of the region and understanding the reasons for having it renamed. Such values need to be identified and updated.

- **References:** References make the data reliable, especially if cited from an accredited source. Mentioning sources for any data which represent contradictory information is even more important to let the users know how credible these diverse claims are. For this reason, it is important that each claim and qualifier gets a proper reference. Many references in Wikidata contain a

[28] https://www.wikidata.org/wiki/Q16869 [Accessed: 20.10.2022]

[29] https://www.wikidata.org/wiki/Q2714337 [Accessed: 20.10.2022]

[30] For example, an urban square in Berlin is recently renamed from Kaiser-Wilhelm-Platz to Richard-von-Weizsäcker-Platz Q1721568[31]. The older name is visible in the alias and the properties *inception (P571)* and *named after (P138)*.

[32] https://www.wikidata.org/wiki/Q1814900

[33] Currently known as Zendeh Jan(Q2710776)

Wikipedia language version as the value for the property *imported from Wikimedia project (P143)* which doesn't provide a direct link to the Wikipedia article where the claim is taken from. So, these references are better updated, either by adding the link to the actual article through the property *Wikimedia import URL (P4656)* or mentioning an external source.

Now, that we have identified what is missing and what needs updating, the next step will be to add these contents to Wikidata.

iv. Add or Reform New Content. In this step, we first decide on how to add the above-mentioned and identified contents into Wikidata. Since bots are used to perform repetitive tasks and they are more efficient in performing such tasks than humans, we create a data file containing the required information of the missing items to be automatically added through a script or bot[34]. Similarly, new sources can also be added in the same way.

Statements containing property-value pairs and references can also be added through automation if a large number of items miss specific properties. If the task is not simple or repetitive, it is better that it is performed manually with more accuracy. Bots have the potential to add erroneous data in much higher volumes than humans if not dealt with care as a high percentage of references added by bots in Wikidata are not authoritative[35] and not relevant in comparison to much fewer invalid references added by humans [238]. Therefore, we recommend using bots with planning and care and only when the tasks are rather simple and repetitive.

7.3 Summary

Based on our results from the previous chapter, bots have the potential to influence diversity and have thus far played a significant role in high-diversity classes in Wikidata. Therefore, in this chapter, we present our recommendations for enhancing diversity in Wikidata domains/classes through the use of automation and provide an applicable use case for our recommended approach.

Our suggestions follow our concept for diversity measurement approach and cover both aspects of diversity, user, and data. User diversity is concerned with community issues and the social mechanisms of diversity improvement, while, data diversity refers to a rather technical mechanism for the enhancement of diversity in Wikidata. Our main focus, though, is on the topic of domain/ class diversity here.

We apply our recommended approach to a use case to preserve the history of Herat City, one of the motivating factors for this research. The main steps in this recommended approach are a) establishing a contributing community by defining member roles and collaboration mechanisms, b) developing goals and agenda for the community through assessment of the content, identification of missing data, and prioritization of topics for adding new content, and c) monitoring and evaluation of the community progress.

[34] Bots should go through a defined procedure to obtain the right for performing high-speed edits as mentioned in Section 6.2

[35] 'Authoritative sources refers to sources of information that are deemed trustworthy, up-to-date, and free of bias for supporting a particular statement on Wikidata.' https://www.wikidata.org/wiki/Wikidata:Verifiability [Accessed 10.05.2020]

Our proposed steps are rather general and can serve as a rough outline for the creation of any new community or sub-community around the world, in particular with languages where a lot of resources are not yet digitally available. Since community tasks are collaboratively performed and decided, this rather general format can serve as a draft to form the basic structure and let the communities themselves decide on more details.

Chapter 8

CONCLUSION

Here, we present an overview of this study and a summary of our findings which then paves the way for future research directions.

8.1 Summary

Wikidata was developed to serve everyone around the world. On the edge of one decade of its existence, however, it seems to have an imbalanced coverage of global data. While the origins of Wikidata in the West and the presence of a predominantly Western contributor base contribute to this issue, it is essential to recognize that the utilization of automation, specifically through bots, has also played a significant role in shaping the editing patterns and content distribution of Wikidata.

In this research, our primary objective was to investigate the influence of bot edits on the diversity of data within Wikidata. Diversity holds significant value in the context of Wikidata, as it enables the platform to fulfill its mission of serving a global user base by providing information that is relevant and accessible to individuals from all over the world. For this reason, Wikidata was designed with diversity in mind, and all of the design principles of Wikidata implicitly contribute to diversity. Plurality is the design decision that explicitly empowers Wikidata to reflect diversity by allowing multiple statements, which could also be contradictory, to coexist. Despite, the importance of diversity, it is absent as a research topic in the Wikidata literature. Nevertheless, growing numbers of studies on Wikidata, especially from more different countries around the globe, and usage of Wikidata in more application areas, are promising to lead Wikidata toward a more diverse audience.

Diversity, although widely recognized and utilized in various fields, lacks a universally applicable definition that can be uniformly applied across all disciplines. Its interpretation can vary depending on the specific context in which it is used. In general, diversity encompasses three key dimensions that are present in any system or context dealing with diversity and form the basis for measuring diversity. These dimensions include variety, balance, and disparity. In our study, we have developed a conceptual framework for measuring diversity in a KB, specifically within the context of Wikidata, taking into account the fundamental aspects of diversity. In Wikidata, data is organized as items within a hierarchical class structure, which can be further categorized into domains. Accordingly, we propose that data diversity within Wikidata can be assessed by examining its variety, balance, and disparity properties. By considering these dimensions, we aim to provide a comprehensive understanding of diversity in a KB and establish a measurement framework that captures the essence of diversity within this specific KB. Our approach allows for a holistic assessment of diversity and its various facets, allowing us to gain insight into the overall diversity status of Wikidata and identify potential areas for improvement.

We demonstrated that diversity in a KB context needs to be considered as knowledge diversity, where knowledge is the data contributed by users. Thus, to understand the current status of diversity in Wikidata, we need to pay attention to both, data coverage and user participation. Plurality is then measured as part of data diversity and lies at the item level. Thus, we measured the diversity status in Wikidata based on our proposed model to better understand the data and user diversity in Wikidata. The current status of Wikidata diversity based on the existing literature and our collected data from Wikidata domains and classes shows that Wikidata data are at a low diversity level due to the imbalanced distribution of items and contents across Wikidata domains and classes. In the meantime, there existed no information on editor diversity in the existing research and the reason behind this data imbalance could be looked up in the contributing community, in particular, that most of the contributions come from bots.

To answer whether bots' high amounts of contributions are the reason for the data imbalance in Wikidata domains, we first studied bots in detail. Bots are automated accounts run by operators to perform simple and time-consuming tasks. The Wikidata community has approved bot accounts that have mostly been used for data-editing tasks similar to those performed by humans. Bots were intended to import most of the data from Wikipedia language versions with the highest requests, with Western languages such as English at the top of the list.

Looking at the edit history of Wikidata, we have shown that bots have performed their requested tasks. Despite the similarities between the tasks performed by humans and bots, their editing patterns differ significantly, with bot edits differing from human edits at a ratio of one in a thousand. This implies that a large number of bot edits in Wikidata must have a different impact on the data than human edits.

With further exploration of Wikidata's edit history, we have uncovered an important finding: bots have played a significant role in contributing to the balanced classes within Wikidata. Specifically, we observed that classes that received dedicated edits from bots exhibited a more balanced distribution of content. On the other hand, classes that lacked such dedicated bot edits remained imbalanced, with some even heavily skewed. This finding highlights the potential impact of bot edits on data diversity in Wikidata, particularly in achieving content balance within specific classes. However, it is important to note that the mass contributions of bots in certain classes and domains have caused the existing imbalance, further widening the gaps among classes and domains in Wikidata. However, this discovery suggests that bots can be harnessed as a means of addressing the existing imbalance at the domain/class level in Wikidata. By strategically using bots and ensuring focused contributions in underrepresented areas, we can work toward improving diversity and balance within Wikidata. Overall, our research underscores the significant role that bots can play in influencing data diversity and emphasizes their potential for promoting more equitable representation in Wikidata's domains and classes.

We present our recommendations for improving diversity at the domain/class level in Wikidata by leveraging bots. In order to demonstrate the practical application of our recommended approach, we provide a use case that focuses on enhancing diversity in Wikidata through the use of automation. This particular use case addresses the critical issue of preserving historical facts related to specific communities and geographic regions, which are currently facing the risk of manipulation and distortion. This particular issue served as a motivation for conducting this study and exploring ways to improve diversity in Wikidata. By implementing our proposed strategies, we aim to contribute to the broader goal of safeguarding diverse knowledge and ensuring its representation in Wikidata.

In conclusion, our research findings indicate that bots are not a threat to diversity in Wikidata at the domain/class level. On the contrary, they can be utilized as valuable tools to address the existing imbalance in Wikidata domains and classes, contributing to a more diverse and comprehensive KB. Furthermore, our proposed concept for measuring diversity in Wikidata can also serve as a blueprint for assessing diversity in other structured knowledge bases (KB). While customization might be necessary at the item level to align with the specific data models of different KBs, the fundamental principles of diversity measurement can be applied universally. This highlights the potential for applying our approach to promote diversity in other KBs and enhance the representation of diverse perspectives and knowledge.

8.2 Future Work

During this study, we encountered many open questions that were interesting for further investigation but beyond the scope of this study.

Presenting a complete picture of Wikidata's diversity status is a broad question that deals with diversity from both angles of user and data. Drawing a real picture of diversity in Wikidata by looking into the whole content and all of the domains and classes seems complex and challenging at the moment due to the inconsistent class hierarchy structure of Wikidata. In this study, we could only focus on the domain/ class level diversity of the data diversity angle. Item level diversity, in particular measuring plurality or statement diversity in Wikidata, which could provide an overview of the globally agreed vs. disputed statements, is another part of this question that has remained open for further research.

Considering the user diversity angles, the Wikidata community is another topic that needs further investigation efforts. While we investigated editor diversity from the editing patterns angle, editor diversity considering the background of Wikidata contributors is yet to be explored. It could provide an overview of where around the globe Wikidata contributors are, especially, the origins of bot operators could shed light on understanding who contributes the most content now, people with which backgrounds rarely contribute or are missing, and how can this be improved for a more balanced contribution in the future. In addition, a thorough study of consumer diversity can show us which domains/ classes, and topics are more popular among Wikidata consumers and which ones lack attention. We can then look at the reasons for the lower usage of certain topics, languages, or domains and act towards bringing them to attention. Understanding consumer diversity along with data diversity can also shed light on why Wikidata is popular in some countries and languages, but not in the rest, and what needs to be done in this regard.

We have introduced a versatile approach to historic preservation that can be applied to any community or language where digital resources are limited or unavailable. Our approach goes beyond enhancing statement diversity; it also serves as a means of safeguarding non-digitalized resources. We strongly advocate for the adoption of our approach, as it can contribute to the comprehensive preservation of historical knowledge and cultural heritage. By implementing our approach, communities, and languages with limited digital resources can benefit from improved representation and accessibility, ensuring that valuable information is not lost or forgotten.

Moreover, we encourage future research to adopt and apply our proposed concepts for measuring and improving diversity through automation in other KBs, whether structured or unstructured. This is because KBs commonly involve user contributions and data organization into classes and domains, making our concepts applicable across different KB architectures. Furthermore, our framework for measuring diversity is designed to be applicable to all KBs, including those that may not explicitly support plurality similar to Wikidata. By extending the application of our concepts to diverse KBs, we can evaluate the effectiveness, validate the generalizability, and strengthen the overall robustness of our proposed approach.

Another topic of interest that we encountered during the literature review but could not investigate in this research scope is the impact of diversity on data quality. There have been several studies in Wikipedia exploring the impact of diversity on article quality, primarily focusing on editor diversity. In Wikidata, it would also be interesting to find out how diversity can be related to quality and how it might impact it. This is because, from one angle, diversity contributes to data completeness, which is an attribute of quality. From another angle, though, it supports the coexistence of contradictory statements which can question the reliability or trustworthiness of data. For this reason, this topic is another interesting and important future research direction.

In conclusion, we believe that our concepts for diversity measurement and automation-driven diversity enhancement hold broader potential beyond the scope of this study and can contribute to fostering diversity in various KBs, ultimately facilitating inclusive and comprehensive representation of knowledge.

Appendix A

Appendix

This appendix contains descriptive statistics and a codebook for the *Wikidata Revision History Dataset* in the following tables:

Table A.1 displays an overview of Wikidata Topical Domains and Classes showing the distribution of their items and maturity levels. The domains and classes in this table are based on [65].

Table A.2 provides a detailed overview of edit types per user group expressed by Figure 6.7. In this table, we see four main edit focuses, where each edit focus contains further parts: Term consists of the parts *alias*, *description*, *label*, and *term*. Similarly, *statement* contains the parts *claim*, *qualifier*, *rank*, *reference* and *statement*. *sitelink* contains *sitelink* and*sitelink badge* and *item* represents item only. Revert can occur in any edit focus; therefore, it remains stand-alone. We also have a number of revisions marked as unstructured that were dropped before the data analysis phase and make less than one percent of the revisions and they are the revisions that were either empty or not possible to be classified (see Section 5.3.4 on page 100).

Table A.3 contains the codebook developed for mapping edit summaries to edit types in *Wikidata Revision History Dataset*.

Table A.1: Overview of Wikidata Topical Domains and classes showing the distribution of their items and maturity level (20.01.2020)

Topics					**Maturity**			
Domain	**Main-Class**	**Subclass**	**#items**	**#unique items**	**#items 1-10 rev**	**#items 11-100 rev**	**#items 101 - 1,000 rev**	**#items > 1,000 rev**
Person	Person	Musician	617	596	189 (32%)	361 (61%)	46 (8%)	0 (0%)
		Athlete	499	498	128 (26%)	314 (63%)	56 (11%)	0 (0%)
		Writer	398	390	126 (32%)	225 (58%)	39 (10%)	0 (0%)
		Politician	59.661	59.659	45.073 (76%)	14.285 (24%)	300 (1%)	1 (0%)
Media	Show	Film	290.928	289.493	22.075 (8%)	241.596 (83%)	25.773 (9%)	49 (0%)
		TV Series	59.808	59.376	10.828 (18%)	45.976 (77%)	2.558 (4%)	14 (0%)
	Literary Composition	Book	91.647	91.373	43.143 (47%)	47.902 (52%)	327 (0%)	0 (0%)
		Magazine	88.306	88.097	36.443 (41%)	51.542 (59%)	112 (0%)	0 (0%)
	Musical Composition	Album	257.373	256.361	45.003 (18%)	210.926 (82%)	432 (0%)	0 (0%)
Organization	Company	Bank	3.247	3.247	765 (24%)	2.365 (73%)	117 (4%)	0 (0%)
		Airlines	4.781	4.780	567 (12%)	4.083 (85%)	130 (3%)	0 (0%)
	Educational Institution	University	15.839	15.835	1.677 (11%)	12.672 (80%)	1.485 (9%)	1 (0%)
	Social Groups	Sports Club	84.204	84.202	34.235 (41%)	49.099 (58%)	866 (1%)	2 (0%)
		Political Party	19.070	19.070	8.435 (44%)	10.293 (54%)	342 (2%)	0 (0%)
Geography	Topography	Lake	270.907	270.894	28.102 (10%)	242.625 (90%)	167 (0%)	0 (0%)
		River	401.192	401.185	39.200 (10%)	361.212 (90%)	773 (0%)	0 (0%)
		Mountain	526.908	526.898	46.739 (9%)	479.946 (91%)	212 (0%)	1 (0%)
		Country	3.630	3.613	842 (23%)	2.048 (57%)	525 (15%)	197 (5%)
		City	45.776	45.757	7.009 (15%)	27.850 (61%)	10.875 (24%)	23 (0%)
Biology	Animal	Mammal	12.384	12.383	5.830 (47%)	6.472 (52%)	78 (1%)	3 (0%)
		Bird	333	331	157 (47%)	154 (47%)	20 (6%)	0 (0%)
		Fish	76	76	46 (61%)	22 (29%)	8 (11%)	0 (0%)
	Plant	Tree	9.270	9.270	3.059 (33%)	6.201 (67%)	10 (0%)	0 (0%)
		Grass	6	6	3 (50%)	1 (17%)	2 (33%)	0 (0%)

Table A.2: An overview of *Wikidata Revision History Dataset* aggregated by edit focus. (Note: Bold numbers show highest values in a row.)

Target	**Edit Types**	**Anon.**	**Bot**	**Human**	**Tool**	**Sum**
Item	create/merge/ update/protect item	421 (0.00)	**881,850 (0.90)**	53,768 (0.05)	42,577 (0.04)	978,616 (0.18)
	create redirect	0 (0.00)	14 (0.15)	**81 (0.85)**	0 (0.00)	
Term	add/remove/set/ update alias	1,702 (0.04)	6,502 (0.14)	**23,461 (0.52)**	13,301 (0.30)	1,242,146 (0.22)
	add/remove/set description	7,111 (0.01)	50,688 (0.06)	213,080 (0.25)	**584,857 (0.68)**	
	add/remove/set label	5,088 (0.02)	71,836 (0.21)	85,879 (0.26)	**172,044 (0.51)**	
	add/set term	562 (0.09)	**4,756 (0.72)**	1,279 (0.19)	0 (0.00)	
Statement	create/remove statement	12,984 (0.01)	688,194 (0.36)	**773,691 (0.41)**	431,801 (0.23)	3,098,597 (0.56)
	set/update claim	4,318 (0.02)	**108,373 (0.62)**	61,501 (0.35)	0 (0.00)	
	add/update qualifier	18 (0.00)	13,665 (0.24)	12,879 (0.23)	**29,480 (0.53)**	
	update rank	0 (0.00)	0 (0.00)	**3 (1.00)**	0 (0.00)	
	add/set/remove/ update reference	566 (0.00)	**521,937 (0.54)**	318,454 (0.33)	120,733 (0.13)	
Sitelink	add/set/remove/ update sitelink	7,098 (0.04)	38,140 (0.20)	**144,357 (0.74)**	5,104 (0.03)	196,338 (0.04)
	set sitelink badge	3 (0.00)	**1,229 (0.75)**	407 (0.25)	0 (0.00)	
	Revert	213 (0.02)	566 (0.05)	**10,784 (0.93)**	0 (0.00)	11,563 (0.00)
	Unstructured	212 (0.00)	**43,659 (0.87)**	6,050 (0.13)	0 (0.00)	49,921 (0.01)
	Total	40,296 (0.01)	**2,431,409 (0.44)**	1,705,674 (0.31)	1,399,897 (0.25)	5,577,276 (1.00)

Table A.3: Codebook for Mapping Edit Summaries to Edit Types in *Wikidata Revision History Dataset*

No.	Edit Summaries	Edit Types
1	wbsetaliases-add	add alias
2	wbsetdescription-add	add description
3	wbsetlabel-add	add label
4	wbeditentity-update-languages	
5	wbsetclaim-update-qualifiers	add qualifier
6	wbsetqualifier	
7	wbsetqualifier-add	
8	wbsetreference-add	add reference
9	wbsetsitelink-add	add sitelink
10	wbsetentity	add term
11	created	create item
12	special-create-item	
13	wbcreate-new	
14	wbeditentity-create	
15	wbeditentity-create-item	
16	wbcreateredirect	create redirect
17	wbcreateclaim	create statement
18	wbcreateclaim-create	
19	wbsetclaim-create	
20	wbmergeitems-from	merge item
21	wbmergeitems-to	
22	protected	protect item
23	wbsetaliases-remove	remove alias
24	wbsetdescription-remove	remove description
25	wbsetlabel-remove	remove label
26	remove	
27	wbremovereferences	remove reference
28	wbremovereferences-remove	
29	clientsitelink-remove	remove sitelink
30	wbsetsitelink-remove	
31	wbremoveclaims	remove statement
32	wbremoveclaims-remove	
33	wbremoveclaims-update	
34	restore[1]	revert
35	revert[2]	
36	clean[3]	
37	repair[4]	
38	undo	
39	undid	

Continued on next page

[1] Also restored

[2] Also reverted, reverting, rv.

[3] Also cleanup, cleaning, clean’n’repair, cleanup/repair, clean-up.

[4] Also repairing.

Table A.3 – continued from previous page

No.	Edit Summaries	Edit Types
40	wbsetclaimvalue	set claim
41	wbsetaliases-add-remove	set alias
42	wbsetaliases-set	
43	wbsetdescription-set	set description
44	wbsetlabel-set	set label
45	wbsetreferences	set reference
46	wblinktitles-connect	set sitelink
47	wbsetsitelink-add-both	
48	wbsetsitelink-set	
49	wbsetsitelink-set-badges	set sitelink badge
50	wbsetsitelink-set-both	
51	wbsetlabeldescriptionaliases	set term
52	wbsetaliases-update	update alias
53	wbsetclaim-update	update claim
54	wbeditentity-override	update item
55	wbeditentity	
56	wbeditentity-update	
57	wbeditentity-update-languages-and-other	
58	wbremovequalifiers-remove	update qualifier
59	wbsetqualifier-update	
60	wbsetclaim-update-rank	update rank
61	wbsetreference-set	update reference
62	clientsitelink-update	update sitelink

Bibliography

[1] *David Abián, Jorge Bernad, and Raquel Trillo-Lado. Using contemporary constraints to ensure data consistency. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 2303–2310, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359337. doi: 10.1145/3297280.3297509. URL https://doi.org/10.1145/3297280.3297509.

[2] *D. Abián, F. Guerra, J. Martínez-Romanos, and Raquel Trillo-Lado. Wikidata and DBpedia: A Comparative Study. In Julian Szymański and Yannis Velegrakis, editors, *Semantic Keyword-Based Search on Structured Data Sources*, volume 10546, pages 142–154. Springer International Publishing, Cham, 2018. ISBN 978-3-319-74496-4 978-3-319-74497-1. doi: 10.1007/978-3-319-74497-1_14. URL http://link.springer.com/10.1007/978-3-319-74497-1_14.

[3] Waqās Ahmed and Martin Lewis Poulter. Representation of Non-Western Cultural Knowledge on Wikipedia: The Case of the Visual Arts. *Digital Studies / Le champ numérique*, 13(1), January 2023. ISSN 1918-3666. doi: 10.16995/dscn.8078. URL https://www.digitalstudies.org/article/id/8078/.

[4] *Albin Ahmeti, Simon Razniewski, and Axel Polleres. Assessing the Completeness of Entities in Knowledge Bases. In *The Semantic Web: ESWC 2017 Satellite Events*, Lecture Notes in Computer Science, pages 7–11. Springer, Cham, 2017. ISBN 978-3-319-70406-7 978-3-319-70407-4. doi: 10.1007/978-3-319-70407-4_2. URL https://link.springer.com/chapter/10.1007/978-3-319-70407-4_2.

[5] *Vladimir Alexiev, Plamen Tarkalanov, Nikola Georgiev, and Lilia Pavlova. Bulgarian icons in wikidata and edm. In *Digital Presentation and Preservation of Cultural and Scientific Heritage (DIPP 2020)*, volume 10, Burgas, Bulgaria, Sep 2020. Institute of Mathematics and Informatics (IMI BAS), Sofia. URL http://dipp.math.bas.bg/images/2020/045-064_1.2_iDiPP2020-24_v.1c.pdf.

[6] *Kholoud Alghamdi, Miaojing Shi, and Elena Simperl. Learning to recommend items to wikidata editors. In Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam M. Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, editors, *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*, volume 12922 of *Lecture Notes in Computer Science*, pages 163–181. Springer, 2021. doi: 10.1007/978-3-030-88361-4_10. URL https://doi.org/10.1007/978-3-030-88361-4_10.

[7] *Stacy Allison-Cassin and Dan Scott. Wikidata: a platform for your library's linked open data. *Code4Lib Journal*, (40), 2018.

Note: References marked with an asterisk (*) at the beginning denote studies included in the mapping study conducted in this dissertation.

[8] *Paulo Dias Almeida, Jorge Gustavo Rocha, Andrea Ballatore, and Alexander Zipf. Where the Streets Have Known Names. In *Computational Science and Its Applications – ICCSA 2016*, Lecture Notes in Computer Science, pages 1–12. Springer, Cham, July 2016. ISBN 978-3-319-42088-2 978-3-319-42089-9. doi: 10.1007/978-3-319-42089-9_1. URL https://link.springer.com/chapter/10.1007/978-3-319-42089-9_1.

[9] *Gabriel Amaral, Alessandro Piscopo, Lucie-Aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl. Assessing the quality of sources in wikidata across languages: A hybrid approach. *ACM J. Data Inf. Qual.*, 13(4):23:1–23:35, 2021. doi: 10.1145/3484828. URL https://doi.org/10.1145/3484828.

[10] *Gary Ang and Ee-Peng Lim. Learning knowledge-enriched company embeddings for investment management. In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391481. doi: 10.1145/3490354.3494390. URL https://doi.org/10.1145/3490354.3494390.

[11] Adam Angelika, Felix Keppmann, and Delia Rusu. Renderer.pdf. Technical Report D5.1.21, 2013. URL http://render-project.eu/wp-content/uploads/2012/11/D5.1.21.pdf.

[12] Ofer Arazy and Oded Nov. Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 233–236, Savannah, Georgia, USA, February 2010. Association for Computing Machinery. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718963. URL https://doi.org/10.1145/1718918.1718963.

[13] 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. doi: 10.1287/isre.2016.0647. URL https://doi.org/10.1287/isre.2016.0647.

[14] *Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. Negative knowledge for open-world wikidata. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 544–551. ACM / IW3C2, 2021. doi: 10.1145/3442442.3452339. URL https://doi.org/10.1145/3442442.3452339.

[15] *Nils Axelsson and Gabriel Skantze. *Using Knowledge Graphs and Behaviour Trees for Feedback-Aware Presentation Agents.* Association for Computing Machinery, New York, NY, USA, 2020.

[16] *Vevake Balaraman, Simon Razniewski, and Werner Nutt. Recoin: Relative Completeness in Wikidata. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1787–1792, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191641. URL https://doi.org/10.1145/3184558.3191641.

[17] Bela Balassa. Trade liberalisation and "revealed" comparative advantage 1. *The manchester school*, 33(2):99–123, 1965.

[18] *Steven J Baskauf and Jessica K Baskauf. Using the w3c generating rdf from tabular data on the web recommendation to manage small wikidata datasets. *Semantic Web*, (Preprint):1–23, 2021. doi: 10.3233/SW-210443.

[19] *Seyed Amir Hosseini Beghaeiraveri, Alasdair J. G. Gray, and Fiona McNeill. Reference statistics in wikidata topical subsets. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-3.pdf.

[20] *Seyed Amir Hosseini Beghaeiraveri, Alasdair J. G. Gray, and Fiona Jennet McNeill. Experiences of using wdumper to create topical subsets from wikidata. In David Chaves-Fraga, Anastasia Dimou, Pieter Heyvaert, Freddy Priyatna, and Juan F. Sequeda, editors, *Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference (ESWC 2021), Online, June 6, 2021*, volume 2873 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2873/paper13.pdf.

[21] Wolfgang H. Berger and Frances L. Parker. Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168(3937):1345–1347, 1970. doi: 10.1126/science.168.3937.1345. URL https://www.science.org/doi/abs/10.1126/science.168.3937.1345.

[22] *Preeti Bhargava, Nemanja Spasojevic, Sarah Ellinger, Adithya Rao, Abhinand Menon, Saul Fuhrmann, and Guoning Hu. Learning to map wikidata entities to predefined topics. In Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1194–1202. ACM, 2019.

[23] *Carlo Bianchini and Stefano Bargioni. Automated classification using linked open data. a case study on faceted classification and wikidata. *Cataloging & Classification Quarterly*, 59(8):835–852, 2021. doi: 10.1080/01639374.2021.1977447. URL https://doi.org/10.1080/01639374.2021.1977447.

[24] *Carlo Bianchini and Lucia Sardo. Wikidata: a new perspective towards universal bibliographic control. *JLIS.it*, 13(1):291–311, Jan. 2022. doi: 10.4403/jlis.it-12725.

[25] *Adrian Bielefeldt, Julius Gonsior, and Markus Krötzsch. Practical Linked Data Access via SPARQL: The Case of Wikidata. In *Workshop on Linked Data on the Web co-located with The Web Conference 2018, LDOW@WWW 2018, Lyon, France April 23rd, 2018*, volume 2073 of *CEUR Workshop Proceedings*, page 10, Lyon, France, April 2018. URL ceur-ws.org/Vol-2073/article-03.pdf.

[26] Peter Michael Blau. *Inequality and heterogeneity: A primitive theory of social structure*, volume 7. Free Press New York, 1977.

[27] *Angela Bonifati, Wim Martens, and Thomas Timm. Navigating the maze of wikidata query logs. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 127–138. ACM, 2019.

[28] *Armand Boschin and Thomas Bonald. Enriching wikidata with semantified wikipedia hyperlinks. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-6.pdf.

[29] *Styliani Bourli and Evaggelia Pitoura. Bias in knowledge graph embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 6–10, 2020. doi: 10.1109/ASONAM49781.2020.9381459.

[30] *Freddy Brasileiro, João Paulo A. Almeida, Victorio A. Carvalho, and Giancarlo Guizzardi. Applying a multi-level modeling theory to assess taxonomic hierarchies in Wikidata. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 975–980. International World Wide Web Conferences Steering Committee, 2016.

[31] Leon Brillouin. Science and information theory. *Science*, 124(3220):492–493, 1956. doi: 10.1126/science.124.3220.492.b. URL https://www.science.org/doi/abs/10.1126/science.124.3220.492.b.

[32] *Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Elvira Mitraka, Julia Turner, Tim E. Putman, Justin Leong, Chinmay Naik, Paul Pavlidis, Lynn M. Schriml, Benjamin M. Good, and Andrew I. Su. Wikidata as a semantic framework for the gene wiki initiative. *Database: The Journal of Biological Databases and Curation*, 2016, 2016. doi: 10.1093/database/baw015. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4795929/.

[33] *Marçal Mora Cantallops, Salvador Sánchez-Alonso, and Elena García Barriocanal. A systematic literature review on wikidata. *Data Technol. Appl.*, 53 (3):250–268, 2019.

[34] *Alberto Cetoli, Stefano Bragaglia, Andrew D. O'Harney, Marc Sloan, and Mohammad Akbari. A neural approach to entity linking on wikidata. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part II*, volume 11438 of *Lecture Notes in Computer Science*, pages 78–86. Springer, 2019.

[35] *Niel Chah and Periklis Andritsos. Wikimetadata studio: Dashboards from data profiling the languages, properties, and items of wikidata. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd*

Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-13.pdf.

[36] *Hans Chalupsky, Pedro A. Szekely, Filip Ilievski, Daniel Garijo, and Kartik Shenoy. Creating and querying personalized versions of wikidata on a laptop. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-4.pdf.

[37] *Melisachew Wudage Chekol and Heiner Stuckenschmidt. Towards Probabilistic Bitemporal Knowledge Graphs. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1757–1762, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191637. URL https://doi.org/10.1145/3184558.3191637.

[38] *LP Coladangelo and Lynn Ransom. Semantic enrichment of the schoenberg database of manuscripts name authority through wikidata. 2021.

[39] *Davide Colla, Annamaria Goy, Marco Leontino, and Diego Magro. Wikidata support in the creation of rich semantic metadata for historical archives. *Applied Sciences*, 11(10), 2021. ISSN 2076-3417. doi: 10.3390/app11104378. URL https://www.mdpi.com/2076-3417/11/10/4378.

[40] *Nancy Cooey. Leveraging wikidata to enhance authority records in the ehri portal. *Journal of Library Metadata*, 19(1-2):83–98, 2019. doi: 10.1080/19386389.2019.1589700. URL https://doi.org/10.1080/19386389.2019.1589700.

[41] *Rafael Crescenzi, Marcelo Fernandez, Federico A. Garcia Calabria, Pablo Albani, Diego Tauziet, Adriana Baravalle, and Andrés Sebastián D'Ambrosio. A Production Oriented Approach for Vandalism Detection in Wikidata. In *WSDM Cup 2017 Notebook Papers*, Cambridge, UK, February 2017. arxiv.org. URL https://arxiv.org/ftp/arxiv/papers/1712/1712.06919.pdf.

[42] *To Tu Cuong and Claudia Müller-Birn. Applicability of Sequence Analysis Methods in Analyzing Peer-Production Systems: A Case Study in Wikidata. In *Social Informatics*, Lecture Notes in Computer Science, pages 142–156. Springer, Cham, November 2016. ISBN 978-3-319-47873-9 978-3-319-47874-6. doi: 10.1007/978-3-319-47874-6_11.

[43] *Paolo Curotto and Aidan Hogan. Suggesting citations for wikidata claims based on wikipedia's external references. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[44] *Atílio A. Dadalto, João Paulo A. Almeida, Claudenir M. Fonseca, and Giancarlo Guizzardi. Type or individual? evidence of large-scale conceptual

disarray in wikidata. In Aditya K. Ghose, Jennifer Horkoff, Vítor E. Silva Souza, Jeffrey Parsons, and Joerg Evermann, editors, *Conceptual Modeling - 40th International Conference, ER 2021, Virtual Event, October 18-21, 2021, Proceedings*, volume 13011 of *Lecture Notes in Computer Science*, pages 367–377. Springer, 2021. doi: 10.1007/978-3-030-89022-3_29. URL https://doi.org/10.1007/978-3-030-89022-3_29.

[45] *Sarah Dahir, Abderrahim El Qadi, and Hamid Bennis. Query expansion using wikidata attributes' values. EAI, 5 2019. doi: 10.4108/eai.24-4-2019.2284070.

[46] *Sarah Dahir, Jalil Elhassouni, Abderrahim El Qadi, and Hamid Bennis. Medical query expansion using semantic sources dbpedia and wikidata. In Sarika Jain and Sven Groppe, editors, *Proceedings of the International Semantic Intelligence Conference 2021 (ISIC 2021), New Delhi, India, February 25-27, 2021*, volume 2786 of *CEUR Workshop Proceedings*, pages 195–201. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2786/Paper26.pdf.

[47] *Fariz Darari. Coviwd: Covid-19 wikidata dashboard. *Jurnal Ilmu Komputer dan Informasi*, 14(1):39–47, 2021.

[48] *Antonin Delpeuch. Opentapioca: Lightweight entity linking for wikidata. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[49] *Antonin Delpeuch. Running a reconciliation service for wikidata. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[50] *Gianluca Demartini. Implicit bias in crowdsourced knowledge graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 624–630, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317307. URL https://doi.org/10.1145/3308560.3317307.

[51] *Lena Denis. Using wikidata to extract cartographic resources from archival collections. *e-Perimetron*, 16(1):27–38, 2021.

[52] *Abdelmoneim Amer Desouki, Michael Röder, and Axel-Cyrille Ngonga Ngomo. Ranking on very large knowledge graphs. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, page 163–171, 2019. doi: https://doi.org/10.1145/3342220.3343660.

[53] Wikimedia Deutschland. Wikidata community survey 2021, June 2021. URL https://commons.wikimedia.org/wiki/File:Wikidata_Community_Survey_2021.pdf.

[54] *Dennis Diefenbach, Kamal Singh, and Pierre Maret. WDAqua-core0: A Question Answering Component for the Research Community. In *Semantic*

Web Challenges, Communications in Computer and Information Science, pages 84–89. Springer, Cham, May 2017. ISBN 978-3-319-69145-9 978-3-319-69146-6. doi: 10.1007/978-3-319-69146-6_8. URL https://link.springer.com/chapter/10.1007/978-3-319-69146-6_8.

[55] Dennis Diefenbach, Thomas Pellissier Tanon, Kamal Deep Singh, and Pierre Maret. Question answering benchmarks for wikidata. In *Proceedings of the ISWC 2017 Posters and Demonstrations and Industry Tracks*, volume Vol-1963, 2017.

[56] *Dennis Diefenbach, Kamal Singh, and Pierre Maret. WDAqua-core1: A Question Answering Service for RDF Knowledge Bases. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1087–1091, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191541. URL https://doi.org/10.1145/3184558.3191541.

[57] *Paula Dooley and Bojan Bozic. Towards linked data for wikidata revisions and twitter trending hashtags. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, iiWAS 2019, Munich, Germany, December 2-4, 2019*, pages 166–175. ACM, 2019.

[58] Robert Dorfman. A formula for the gini coefficient. *The Review of Economics and Statistics*, 61(1):146–149, 1979. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/1924845.

[59] Marina Drosou, H. V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in big data: A review. *Big data*, 5 2:73–84, 2017.

[60] *Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, volume 11779 of *Lecture Notes in Computer Science*, pages 69–78. Springer, 2019.

[61] *Alexander M. Elizarov, Polina Gafurova, and Evgeny K. Lipachev. Wikidata in metadata formation methods for documents of digital mathematical library. In Alexander M. Elizarov and Evgeny K. Lipachev, editors, *Proceedings of the 23rd Conference on Scientific Services & Internet (SSI 2021), Moscow (online), September 20-23, 2021*, volume 3066 of *CEUR Workshop Proceedings*, pages 23–33. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-3066/paper3.pdf.

[62] *Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. Introducing Wikidata to the linked data web. In *International Semantic Web Conference*, pages 50–65. Springer, 2014.

[63] *Muhammad Faiz, Gibran M. F. Wisesa, Adila Alfa Krisnadhi, and Fariz Darari. OD2WD: from open data to wikidata through patterns. In Krzysztof Janowicz, Adila Alfa Krisnadhi, María Poveda-Villalón, Karl Hammar, and

Cogan Shimizu, editors, *Proceedings of the 10th Workshop on Ontology Design and Patterns (WOP 2019) co-located with 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 27, 2019*, volume 2459 of *CEUR Workshop Proceedings*, pages 2–16. CEUR-WS.org, 2019.

[64] *Ghazal Faraj and András Micsik. Enriching wikidata with cultural heritage data from the COURAGE project. In Emmanouel Garoufallou, Francesca Fallucchi, and Ernesto William De Luca, editors, *Metadata and Semantic Research - 13th International Conference, MTSR 2019, Rome, Italy, October 28-31, 2019, Revised Selected Papers*, volume 1057 of *Communications in Computer and Information Science*, pages 407–418. Springer, 2019.

[65] *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, jan 2018. ISSN 1570-0844. doi: 10.3233/SW-170275. URL https://doi.org/10.3233/SW-170275.

[66] Mariam Farda-Sarbas. Wikidata research articles dataset. http://dx.doi.org/10.17169/refubium-40231, 2023.

[67] *Mariam Farda-Sarbas, Hong Zhu, Marisa Frizzi Nest, and Claudia Müller-Birn. Approving automation: Analyzing requests for permissions of bots in wikidata. In *Proceedings of the 15th International Symposium on Open Collaboration*, OpenSym '19, pages 15:1–15:10, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6319-8. doi: 10.1145/3306446.3340833. URL http://doi.acm.org/10.1145/3306446.3340833.

[68] Mariam Farda-Sarbas, Marisa Nest, and Hong Zhu. Wikidata revision history dataset. http://dx.doi.org/10.17169/refubium-40243, 2023.

[69] Mariam Farda-Sarbas, Marisa Frizzi Nest, and Hong Zhu. Wikidata-requests-for-permissions-dataset. http://dx.doi.org/10.17169/refubium-40234, 2023.

[70] Michael Felderer and Jeffrey C. Carver. Guidelines for Systematic Mapping Studies in Security Engineering. *arXiv:1801.06810 [cs]*, January 2018. URL http://arxiv.org/abs/1801.06810. arXiv: 1801.06810.

[71] *Sebastián Ferrada, Nicolás Bravo, Benjamin Bustos, and Aidan Hogan. Querying Wikimedia Images Using Wikidata Facts. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1815–1821, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191646. URL https://doi.org/10.1145/3184558.3191646.

[72] *Mohamed Amine Ferradji and Fouzia Benchikha. Enhanced metrics for temporal dimensions toward assessing linked data: A case study of wikidata. *Journal of King Saud University - Computer and Information Sciences*, 34 (8, Part A):4983–4992, 2022. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2021.05.010. URL https://www.sciencedirect.com/science/article/pii/S1319157821001257.

[73] Emilio Ferrara, Onur Varol, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, 2016.

[74] A. Field, Jeremy Miles, and Z. Field. *Discovering statistics using R*. SAGE Publications, London, England, 2012. ISBN 9781446200469.

[75] *Dominik Filipiak, Anna Fensel, and Agata Filipowska. Mapping of imagenet and wikidata for knowledge graphs enabled computer vision. In Witold Abramowicz, Sören Auer, and Elzbieta Lewanska, editors, *24th International Conference on Business Information Systems, BIS 2021, Hannover, Germany, June 15-17, 2021*, pages 151–161, 2021. doi: 10.52825/bis.v1i.65. URL https://doi.org/10.52825/bis.v1i.65.

[76] Fabian Flöck and Andriy Rodchenko. Whose article is it anyway?–detecting authorship distribution in wikipedia articles over time with wikigini. *Online proceedings of the Wikipedia Academy*, 2012.

[77] Fabian Flöck, Denny Vrandecic, and Elena Simperl. Towards a diversity-minded Wikipedia. In *Proceedings of the 3rd International Web Science Conference on - WebSci '11*, pages 1–8, Koblenz, Germany, 2011. ACM Press. ISBN 978-1-4503-0855-7. doi: 10.1145/2527031.2527063. URL http://dl.acm.org/citation.cfm?doid=2527031.2527063.

[78] R. Foltz. *A History of the Tajiks: Iranians of the East*. Bloomsbury Academic, 2023. ISBN 9780755649648. URL https://books.google.de/books?id=lTTLEAAAQBAJ.

[79] *Hayden Freedman, André van der Hoek, and Bill Tomlinson. Improving wikidata with student-generated concept maps. In Ceren Budak, Meeyoung Cha, and Daniele Quercia, editors, *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 205–215. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19285.

[80] *Nuno Freire and Antoine Isaac. Technical usability of wikidata's linked data. In Witold Abramowicz and Rafael Corchuelo, editors, *Business Information Systems Workshops - BIS 2019 International Workshops, Seville, Spain, June 26-28, 2019, Revised Papers*, volume 373 of *Lecture Notes in Business Information Processing*, pages 556–567. Springer, 2019.

[81] *Nuno Freire and Diogo Proença. RDF reasoning on large ontologies: A study on cultural heritage and wikidata. In Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations - 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5-7, 2020, Proceedings, Part I*, volume 583 of *IFIP Advances in Information and Communication Technology*, pages 381–393. Springer, 2020.

[82] *Johannes Frey, Marvin Hofer, Daniel Obraczka, Jens Lehmann, and Sebastian Hellmann. Dbpedia flexifusion the best of wikipedia > wikidata > your data. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, volume 11779 of *Lecture Notes in Computer Science*, pages 96–112. Springer, 2019.

[83] Earnest Friday and S.Shawnta Friday. Managing diversity using a strategic planned change approach. *Journal of Management Development*, 22(10):863–880, 2003. doi: 0262-1711.

[84] *Kazufumi Fukuda. Using wikidata as work authority for video games. In *Proceedings of the 2019 International Conference on Dublin Core and Metadata Applications*, DCMI'19, page 80–87. Dublin Core Metadata Initiative, 2019.

[85] *Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. Exception-Enriched Rule Learning from Knowledge Graphs. In *The Semantic Web – ISWC 2016*, Lecture Notes in Computer Science, pages 234–251. Springer, Cham, October 2016. ISBN 978-3-319-46522-7 978-3-319-46523-4. doi: 10.1007/978-3-319-46523-4_15. URL https://link.springer.com/chapter/10.1007/978-3-319-46523-4_15.

[86] *Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M. Suchanek. Predicting Completeness in Knowledge Bases. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 375–383, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4675-7. doi: 10.1145/3018661.3018739. URL http://doi.acm.org/10.1145/3018661.3018739.

[87] *Lorenzo Gatti, Chris van der Lee, and Mariët Theune. Template-based multilingual football reports generation using wikidata as a knowledge base. In Emiel Krahmer, Albert Gatt, and Martijn Goudbeek, editors, *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 183–188. Association for Computational Linguistics, 2018.

[88] R. Stuart Geiger. The social roles of bots and assisted editing programs. In *Int. Sym. Wikis*, 2009.

[89] R Stuart Geiger. Bots are users, too! Rethinking the roles of software agents in HCI. *Tiny Transactions on Computer Science*, 1:1, 2012.

[90] R. Stuart Geiger and Aaron Halfaker. When the levee breaks: without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, pages 6:1–6:6, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1852-5. doi: 10.1145/2491055.2491061. URL http://doi.acm.org/10.1145/2491055.2491061.

[91] R Stuart Geiger and Aaron Halfaker. Operationalizing conflict and cooperation between automated software agents in Wikipedia: a replication and expansion of "even good bots fight". *ACM Hum.-Comput. Interact.*, 1(2):33, November 2017. doi: https://doi.org/10.1145/3134684.

[92] R. Stuart Geiger and David Ribes. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 117–126, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718941. URL http://doi.acm.org/10.1145/1718918.1718941.

[93] *Johanna Geiß, Andreas Spitz, and Michael Gertz. NECKAr: A Named Entity Classifier for Wikidata. In *Language Technologies for the Challenges of the*

Digital Age, Lecture Notes in Computer Science, pages 115–129. Springer, Cham, September 2017. ISBN 978-3-319-73705-8 978-3-319-73706-5. doi: 10.1007/978-3-319-73706-5_10. URL https://link.springer.com/chapter/10.1007/978-3-319-73706-5_10.

[94] Conrado Gini. On the measure of concentration with espacial reference to income and wealth. *Cowles Commission*, 2(3), 1936.

[95] Corrado Gini. Variabilità e mutabilità. *vamu*, 1912.

[96] Fausto Giunchiglia, Vincenzo Maltese, Devika Madalli, Anthony Baldry, Cornelia Wallner, Paul Lewis, Kerstin Denecke, Dimitris Skoutas, and Ivana Marenzi. Foundations for the representation of diversity, evolution, opinion and bias. Technical Report DISI-09-063, University of Trento, Trento (Italy), November 2009.

[97] Fausto Giunchiglia, Vincenzo Maltese, and Biswanath Dutta. Domains and context: First steps towards managing diversity in knowledge. *Journal of Web Semantics*, 12-13:53–63, April 2012. ISSN 1570-8268. doi: 10.1016/j.websem.2011.11.007. URL http://www.sciencedirect.com/science/article/pii/S1570826811000989.

[98] *Lars Christoph Gleim, Rafael Schimassek, Dominik Hüser, Maximilian Peters, Christoph Krämer, Michael Cochez, and Stefan Decker. Schematree: Maximum-likelihood property recommendation for wikidata. In Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez, editors, *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings*, volume 12123 of *Lecture Notes in Computer Science*, pages 179–195. Springer, 2020. doi: 10.1007/978-3-030-49461-2_11. URL https://doi.org/10.1007/978-3-030-49461-2_11.

[99] *Larry González and Aidan Hogan. Modelling dynamics in semantic web knowledge graphs with formal concept analysis. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1175–1184, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186016. URL https://doi.org/10.1145/3178876.3186016.

[100] *Lino González, Elena García Barriocanal, and Miguel-Ángel Sicilia. Entity linking as a population mechanism for skill ontologies: Evaluating the use of ESCO and wikidata. In Emmanouel Garoufallou and María Antonia Ovalle-Perandones, editors, *Metadata and Semantic Research - 14th International Conference, MTSR 2020, Madrid, Spain, December 2-4, 2020, Revised Selected Papers*, volume 1355 of *Communications in Computer and Information Science*, pages 116–122. Springer, 2020.

[101] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, 1953. ISSN 00063444. URL http://www.jstor.org/stable/2333344.

[102] *Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM*

SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, page 166–175, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330955. URL https://doi.org/10.1145/3292500.3330955.

[103] *Damien Graux, Fabrizio Orlandi, Brian Lynch, Isobel Mahon, Odhran Mullen, Alex Mahon, Flora Molnar, and Lexes Mantiquilla. A real-time visual dashboard for wikidata edits. In Valentina Ivanova, Patrick Lambrix, Catia Pesquita, and Vitalis Wiens, editors, *Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual Conference (originally planned in Athens, Greece), November 02, 2020*, volume 2778 of *CEUR Workshop Proceedings*, pages 41–46. CEUR-WS.org, 2020.

[104] Alexey Grigorev. *Large-Scale Vandalism Detection with Linear Classifiers The Conkerberry Vandalism Detector at WSDM Cup 2017*. SIGIR'15: proceedings of the 38th ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, 2017. ISBN 978-1-4503-3621-5. OCLC: 946557248.

[105] Miguel R Guevara, Dominik Hartmann, and Marcelo Mendoza. diverse: an r package to analyze diversity in complex systems. *RJ*, 8(2):60–78, 2016. ISSN 2073-4859.

[106] *Dhruv Gupta and Klaus Berberich. Optimizing hyper-phrase queries. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, pages 41–48, New York, NY, USA, 2020. Association for Computing Machinery.

[107] *Daria Gurtovoy and Simon Gottschalk. Linking streets in openstreetmap to persons in wikidata. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 294–297, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391306. doi: 10.1145/3487553.3524267. URL https://doi.org/10.1145/3487553.3524267.

[108] *Jonathan A. Gómez, Thomas Hartka, Binyong Liang, and Gavin Wiehl. Context matrix methods for property and structure ontology completion in wikidata. In *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6, 2021. doi: 10.1109/SIEDS52267.2021.9483776.

[109] *Ben Hachey, Will Radford, and Andrew Chisholm. Learning to generate one-sentence biographies from Wikidata. In *Proceeding of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 633–642, valencia, Spain, 2017. LongPress.

[110] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. Burst of the filter bubble? *Digital Journalism*, 6(3):330–343, 2018. doi: 10.1080/21670811.2017.1338145. URL https://doi.org/10.1080/21670811.2017.1338145.

[111] Alexander Halavais and Derek Lackaff. An analysis of topical coverage of wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440, 01 2008. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2008.00403.x.

[112] Aaron Halfaker and John Riedl. Bots and cyborgs: Wikipedia's immune system. *Computer*, 45(3):79–82, March 2012. ISSN 0018-9162. doi: 10.1109/MC.2012.82. URL http://ieeexplore.ieee.org/document/6163451/.

[113] Aaron Halfaker, R Stuart Geiger, Jonathan T. Morgan, and John Riedl. The rise and decline of an open collaboration system: how wikipedia's reaction to popularity is causing its decline. 57(5):664–688, 2013. doi: https://doi.org/10.1177/0002764212469365.

[114] *Andrew Hall, Loren Terveen, and Aaron Halfaker. Bot detection in Wikidata using behavioral and other informal cues. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18, November 2018. ISSN 25730142. doi: 10.1145/3274333. URL http://dl.acm.org/citation.cfm?doid=3290265.3274333.

[115] *Armin Haller, Axel Polleres, Daniil Dobriy, Nicolas Ferranti, and Sergio José Rodríguez Méndez. An analysis of links in wikidata. In Paul Groth, Maria-Esther Vidal, Fabian M. Suchanek, Pedro A. Szekely, Pavan Kapanipathi, Catia Pesquita, Hala Skaf-Molli, and Minna Tamper, editors, *The Semantic Web - 19th International Conference, ESWC 2022, Hersonissos, Crete, Greece, May 29 - June 2, 2022, Proceedings*, volume 13261 of *Lecture Notes in Computer Science*, pages 21–38. Springer, 2022. doi: 10.1007/978-3-031-06981-9_2. URL https://doi.org/10.1007/978-3-031-06981-9_2.

[116] *Lei Han, Alessandro Checco, Djellel Difallah, Gianluca Demartini, and Shazia Sadiq. Modelling user behavior dynamics with embeddings. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 445–454, New York, NY, USA, 2020. Association for Computing Machinery.

[117] *Tom Hanika, Maximilian Marx, and Gerd Stumme. Discovering implicational knowledge in wikidata. In Diana Cristea, Florence Le Ber, and Baris Sertkaya, editors, *Formal Concept Analysis - 15th International Conference, ICFCA 2019, Frankfurt, Germany, June 25-28, 2019, Proceedings*, volume 11511 of *Lecture Notes in Computer Science*, pages 315–323. Springer, 2019.

[118] *Oktie Hassanzadeh. Building a knowledge graph of events and consequences using wikidata. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-12.pdf.

[119] *Hussein Hazimeh, Elena Mugellini, Simon Ruffieux, Omar Abou Khaled, and Philippe Cudré-Mauroux. Automatic embedding of social network profile links into knowledge graphs. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, SoICT 2018, page 16–23, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450365390. doi: 10.1145/3287921.3287926. URL https://doi.org/10.1145/3287921.3287926.

[120] *Regine Heberlein. On the flipside: Wikidata for cultural heritage metadata through the example of numismatic description. IFLA WLIC 2019 - Libraries: dialogue for change, August 2019. URL http://library.ifla.org/id/eprint/2492/.

[121] *Adelheid Heftberger, Jakob Höper, Claudia Müller-Birn, and Niels-Oliver Walkowski. *Opening up Research Data in Film Studies by Using the Structured Knowledge Base Wikidata*, pages 401–410. Springer International Publishing, Cham, 2020. ISBN 978-3-030-15200-0. doi: 10.1007/978-3-030-15200-0_27. URL https://doi.org/10.1007/978-3-030-15200-0_27.

[122] *Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. Vandalism detection in wikidata. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 327–336. ACM, 2016.

[123] *Stefan Heindorf, Martin Potthast, Hannah Bast, Björn Buchhold, and Elmar Haussmann. WSDM Cup 2017: Vandalism Detection and Triple Scoring. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 827–828, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4675-7. doi: 10.1145/3018661.3022762. URL http://doi.acm.org/10.1145/3018661.3022762.

[124] Stefan Heindorf, Martin Potthast, Gregor Engels, and Benno Stein. Overview of the wikidata vandalism detection task at wsdm cup 2017. In *WSDM Cup 2017 Notebook Papers*, 2017.

[125] *Stefan Heindorf, Yan Scholten, Gregor Engels, and Martin Potthast. Debiasing vandalism detection models at wikidata. In Klaus David, Kurt Geihs, Martin Lange, and Gerd Stumme, editors, *49. Jahrestagung der Gesellschaft für Informatik, 50 Jahre Gesellschaft für Informatik - Informatik für Gesellschaft, INFORMATIK 2019, Kassel, Germany, September 23-26, 2019*, volume P-294 of *LNI*, pages 289–290. GI, 2019.

[126] Natali Helberger, Kari Karppinen, and Lucia D'acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207, 2018.

[127] *Sebastian Hellmann, Johannes Frey, Marvin Hofer, Milan Dojchinovski, Krzysztof Wecel, and Wlodzimierz Lewoniewski. Towards a systematic approach to sync factual data across wikipedia, wikidata and external data sources. In Adrian Paschke, Georg Rehm, Jamal Al Qundus, Clemens Neudecker, and Lydia Pintscher, editors, *Proceedings of the Conference on Digital Curation Technologies (Qurator 2021), Berlin, Germany, February 8th - to - 12th, 2021*, volume 2836 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2836/qurator2021_paper_18.pdf.

[128] *Daniel Henselmann and Andreas Harth. Constructing demand-driven wikidata subsets. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-10.pdf.

[129] *Daniel Hernández, Aidan Hogan, and Markus Krötzsch. Reifying RDF What Works Well With Wikidata-SSWS2015_paper3.pdf. In *Proceedings of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems*, volume 1457 of *CEUR Workshop Proceedings*, pages 32–47. CEUR-WS.org, 2015.

[130] *Daniel Hernández, Aidan Hogan, Cristian Riveros, Carlos Rojas, and Enzo Zerega. Querying Wikidata: Comparing SPARQL, Relational and Graph Databases. In *The Semantic Web – ISWC 2016*, Lecture Notes in Computer Science, pages 88–103. Springer, Cham, October 2016. ISBN 978-3-319-46546-3 978-3-319-46547-0. doi: 10.1007/978-3-319-46547-0_10. URL https://link.springer.com/chapter/10.1007/978-3-319-46547-0_10.

[131] *Alejandro González Hevia, Guillermo Facundo Colunga, Emilio Rubiera Azcona, and José Emilio Labra Gayo. Automatic synchronization of RDF graphs representing ontologies and wikibase instances. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[132] Vernon Hilton Heywood, Robert T Watson, et al. *Global biodiversity assessment*, volume 1140. Cambridge university press Cambridge, 1995.

[133] STEPHEN P. HUBBELL. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press, 2001. ISBN 9780691021287. URL http://www.jstor.org/stable/j.ctt7rj8w.

[134] *Filip Ilievski, Pedro A. Szekely, and Daniel Schwabe. Commonsense knowledge in wikidata. In *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[135] *Filip Ilievski, Pedro A. Szekely, Gleb Satyukov, and Amandeep Singh. User-friendly comparison of similarity algorithms on wikidata. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-2.pdf.

[136] *Ali Ismayilov, Dimitris Kontokostas, Sören Auer, Jens Lehmann, and Sebastian Hellmann. Wikidata through the eyes of dbpedia. *Semantic Web*, 9(4): 493–503, 2018.

[137] *Annika Jacobsen, Andra Waagmeester, Rajaram Kaliyaperumal, Gregory S. Stupp, Lynn M. Schriml, Mark Thompson, Andrew I. Su, and Marco Roos. Wikidata as an intuitive resource towards semantic data modeling in data FAIRification. *Semantic Web Applications and Tools for Healthcare and Life Sciences*, 12 2018. doi: 10.6084/m9.figshare.7415282.v2. URL https://swat4hcls.figshare.com/

articles/journal_contribution/Wikidata_as_an_intuitive_resource_towards_semantic_data_modeling_in_data_FAIRification/7415282.

[138] *Isaac Johnson. Analyzing wikidata transclusion on english wikipedia. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[139] Allard Sicco De Jong and Benjamin J. Bates. Channel diversity in cable television. *Journal of Broadcasting & Electronic Media*, 35(2):159–166, 1991. doi: 10.1080/08838159109364114. URL https://doi.org/10.1080/08838159109364114.

[140] *Samaneh Jozashoori, Ahmad Sakor, Enrique Iglesias, and Maria-Esther Vidal. Eablock: A declarative entity alignment block for knowledge graph creation pipelines. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, pages 1908–1916, New York, NY, USA, 2022. Association for Computing Machinery.

[141] Kenneth Junge. Diversity of ideas about diversity measurement. *Scandinavian Journal of Psychology*, 35(1):16–26, 1994.

[142] *Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. A glimpse into babel: An analysis of multilinguality in wikidata. In *Proceedings of the 13th International Symposium on Open Collaboration*, OpenSym '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351874. doi: 10.1145/3125433.3125465. URL https://doi.org/10.1145/3125433.3125465.

[143] *Lucie-Aimée Kaffee and Elena Simperl. Analysis of editors' languages in wikidata. In *Proceedings of the 14th International Symposium on Open Collaboration*, OpenSym '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359368. doi: 10.1145/3233391.3233965. URL https://doi.org/10.1145/3233391.3233965.

[144] *Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frederique Laforest, Jonathon Hare, and Elena Simperl. Learning to Generate Wikipedia Summaries for Underserved Languages from Wikidata. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2:640–645, 2018. doi: 10.18653/v1/N18-2101. URL https://aclanthology.coli.uni-saarland.de/papers/N18-2101/n18-2101.

[145] *Lucie-Aimée Kaffee, Kemele M. Endris, and Elena Simperl. When humans and machines collaborate: cross-lingual label editing in wikidata. In Björn Lundell, Jonas Gamalielsson, Lorraine Morgan, and Gregorio Robles, editors, *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019, Skövde, Sweden, August 20-22, 2019*, pages 16:1–16:9. ACM, 2019.

[146] *Timothy Kanke. Exploring the knowledge curation work of wikidata. *Bull. IEEE Tech. Comm. Digit. Libr.*, 15(1), 2019.

[147] *Timothy Kanke. Knowledge curation work in wikidata wikiproject discussions. *Libr. Hi Tech*, 39(1):64–79, 2021. doi: 10.1108/LHT-04-2019-0087. URL https://doi.org/10.1108/LHT-04-2019-0087.

[148] *Effie Kapsalis. Wikidata: Recruiting the crowd to power access to digital archives. *Journal of Radio & Audio Media*, 26(1):134–142, 2019. doi: 10.1080/19376529.2019.1559520. URL https://doi.org/10.1080/19376529.2019.1559520.

[149] *Lucie-Aimée Kaffee, Hady Elsahar, and Pavlos Vougiouklis. Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 319–334, Greece, 2018. Springer. URL https://2018.eswc-conferences.org/wp-content/uploads/2018/02/ESWC2018_paper_131.pdf.

[150] *Andreas Oskar Kempf. The need to interoperate: structural comparison of and methodological guidance on mapping discipline-specific subject authority data to wikidata. 2018.

[151] *Saransh Khandelwal and Dhananjay Kumar. Computational fact validation from knowledge graph using structured and unstructured information. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, page 204–208, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450377386. doi: 10.1145/3371158.3371187. URL https://doi.org/10.1145/3371158.3371187.

[152] Barbara Kitchenham, Pearl Brereton, and David Budgen. The Educational Value of Mapping Studies of Software Engineering Literature. In *Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering - Volume 1*, ICSE '10, pages 589–598, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-719-6. doi: 10.1145/1806799.1806887. URL http://doi.acm.org/10.1145/1806799.1806887.

[153] Barbara A. Kitchenham, David Budgen, and O. Pearl Brereton. Using Mapping Studies As the Basis for Further Research - A Participant-observer Case Study. *Inf. Softw. Technol.*, 53(6):638–651, 2011. ISSN 0950-5849. doi: 10.1016/j.infsof.2010.12.011. URL http://dx.doi.org/10.1016/j.infsof.2010.12.011.

[154] Aniket Kittur and Robert E. Kraut. Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 215–224, Savannah, Georgia, USA, February 2010. Association for Computing Machinery. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718959. URL https://doi.org/10.1145/1718918.1718959.

[155] *Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu. Monitoring the Gender Gap with Wikidata Human Gender Indicators. In *Proceedings of the 12th International Symposium on Open Collaboration*, OpenSym '16, pages 16:1–16:9, New York, NY, USA, 2016. ACM. ISBN 978-1-

4503-4451-7. doi: 10.1145/2957792.2957798. URL http://doi.acm.org/10.1145/2957792.2957798.

[156] *Nicholas Klein, Filip Ilievski, and Pedro A. Szekely. Generating explainable abstractions for wikidata entities. In Anna Lisa Gentile and Rafael Gonçalves, editors, *K-CAP '21: Knowledge Capture Conference, Virtual Event, USA, December 2-3, 2021*, pages 89–96. ACM, 2021. doi: 10.1145/3460210.3493580. URL https://doi.org/10.1145/3460210.3493580.

[157] *Piotr Konieczny and Maximilian Klein. Gender gap through time and space: A journey through wikipedia biographies via the wikidata human gender indicator. *New Media Soc.*, 20(12), 2018.

[158] *Vladislav Korablinov and Pavel Braslavski. Rubq: A russian dataset for question answering over wikidata. In Jeff Z. Pan, Valentina A. M. Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal, editors, *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, volume 12507 of *Lecture Notes in Computer Science*, pages 97–110. Springer, 2020.

[159] Ghulamhazrat Koshan. *The history of the oppressed Afghan nation in the course of the twentieth century.* Afghan American Association, 1999.

[160] *Tibor Kovács, Gábor Simon, and Gergely Mezei. Benchmarking graph database backends - what works well with wikidata? *Acta Cybern.*, 24(1): 43–60, 2019.

[161] *Bernhard Krabina and Axel Polleres. Seeding wikidata with municipal finance data. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-9.pdf.

[162] Klaus Krippendorff. *Content analysis: An introduction to its methodology.* Sage publications, 2018.

[163] *Markus Krötzsch. Ontologies for Knowledge Graphs? In *Proceedings of the 30th International Workshop on Description Logics*, volume Vol-1879 of *CEUR Workshop Proceedings*, France, July 2017. CEUR-WS. org. URL http://ceur-ws.org/Vol-1879/invited2.pdf.

[164] *Satoshi Kume and Kouji Kozaki. Extracting domain-specific concepts from large-scale linked open data. In *The 10th International Joint Conference on Knowledge Graphs*, IJCKG'21, pages 28–37, New York, NY, USA, 2021. Association for Computing Machinery.

[165] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems: A survey. *Knowledge-Based Systems*, 123:154 – 162, 2017. ISSN 0950-7051. doi: 10.1016. URL http://www.sciencedirect.com/science/article/pii/S0950705117300680.

[166] *Tuan Lai, Heng Ji, and ChengXiang Zhai. Improving candidate retrieval with entity profile generation for wikidata entity linking. In Smaranda Mure-

san, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3696–3711. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-acl.292.

[167] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2nd edition, 2017.

[168] *Mairelys Lemus-Rojas and Yoo Young Lee. Using wikidata to provide visibility to women in stem. In *Proceedings of the 2019 International Conference on Dublin Core and Metadata Applications*, DCMI'19, page 126–131. Dublin Core Metadata Initiative, 2019.

[169] *Mairelys Lemus-Rojas and Jere D. Odell. Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia. *Journal of Librarianship and Scholarly Communication*, 6(1), December 2018. ISSN 2162-3309. doi: 10.7710/2162-3309.2272. URL https://www.iastatedigitalpress.com/jlsc/article/id/12829/.

[170] *Mairelys Lemus-Rojas and Mirian Ramirez Rojas. Wikidata projects in times of covid-19: Iupui libraries' engagement in open knowledge. 2021.

[171] Robert Leonard and George Jones. *Quantifying Diversity in Archaeology*. Cambridge University Press, Cambridge [Cambridgeshire], 1989. ISBN 0521350301.

[172] Brenda Leong and Evan Selinger. *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*. Understanding Dishonest Anthropomorphism. ACM, New York, New York, USA, January 2019.

[173] *Werner Leyh and Homero Fonseca Filho. Interlinking Standardized OpenStreetMap Data and Citizen Science Data in the OpenData Cloud. In *Advances in Human Factors and Systems Interaction*, Advances in Intelligent Systems and Computing, pages 85–96. Springer, Cham, July 2017. ISBN 978-3-319-60365-0 978-3-319-60366-7. doi: 10.1007/978-3-319-60366-7_9. URL https://link.springer.com/chapter/10.1007/978-3-319-60366-7_9.

[174] *Kwan Hui Lim, Shanika Karunasekera, Aaron Harwood, and Lucia Falzon. Spatial-based topic modelling using wikidata knowledge base. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4786–4788, 2017. doi: 10.1109/BigData.2017.8258542.

[175] Randall M. Livingstone. Population automation: An interview with Wikipedia bot pioneer Ram-Man. *First Monday*, 21(1), 4 January 2016. doi: http://dx.doi.org/10.5210/fm.v21i1.6027. URL https://journals.uic.edu/ojs/index.php/fm/article/view/6027/5189.

[176] *Mohamed Lubani and Shahrul Azman Mohd Noah. Building compact entity embeddings using wikidata. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4-2):1437, sep 2018. doi: 10.18517/ijaseit.8.4-2.6831. URL https://doi.org/10.18517%2Fijaseit.8.4-2.6831.

[177] *Michael Luggen, Djellel Eddine Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. Non-parametric class completeness estima-

tors for collaborative knowledge graphs - the case of wikidata. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*, volume 11778 of *Lecture Notes in Computer Science*, pages 453–469. Springer, 2019.

[178] *Michael Luggen, Julien Audiffren, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Wiki2prop: A multimodal approach for predicting wikidata properties from wikipedia. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2357–2366. ACM / IW3C2, 2021. doi: 10.1145/3442381.3450082. URL https://doi.org/10.1145/3442381.3450082.

[179] *Lin Ma and Yuchun Ma. Automatic question generation based on mooc video subtitles and knowledge graph. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, ICIET 2019, page 49–53, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366397. doi: 10.1145/3323771.3323820. URL https://doi.org/10.1145/3323771.3323820.

[180] ROBERT H. MacArthur. Patterns of species diversity. *Biological Reviews*, 40(4):510–533, 1965. doi: 10.1111/j.1469-185X.1965.tb00815.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-185X.1965.tb00815.x.

[181] Anne E Magurran. *Ecological diversity and its measurement*. Princeton university press, 1988.

[182] Anne E. Magurran. *Measuring Biological Diversity*. Blackwell Publishing, 2004. ISBN 978-0-632-05633-0.

[183] *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 Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II*, volume 11137 of *Lecture Notes in Computer Science*, pages 376–394. Springer, 2018.

[184] *M Manske, U Bˆhme, C P¸the, and M Berriman. Genedb and wikidata. *Wellcome Open Research*, 4(114), 2019. doi: 10.12688/wellcomeopenres.15355.2.

[185] *David L. Martin and Peter F. Patel-Schneider. Wikidata constraints on mars. In *Wikidata@ISWC*, 2020.

[186] *Patricia Martín-Chozas, Sina Ahmadi, and Elena Montiel-Ponsoda. Defying wikidata: Validation of terminological relations in the web of data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph

Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5654–5659. European Language Resources Association, 2020.

[187] Cynthia Matuszek, John Cabral, Michael J. Witbrock, and John DeOliveira. An introduction to the syntax and content of cyc. In *AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49. AAAI, 2006. URL http://dblp.uni-trier.de/db/conf/aaaiss/aaaiss2006-5.html#MatuszekCWD06.

[188] *John P. McCrae and David Cillessen. Towards a linking between wordnet and wikidata. In Sonja Bosch, Christiane Fellbaum, Marissa Griesel, Alexandre Rademaker, and Piek Vossen, editors, *Proceedings of the 11th Global Wordnet Conference, GWC 2021, University of South Africa (UNISA), Potchefstroom, South Africa, January 18-21, 2021*, pages 252–257. Global Wordnet Association, 2021. URL https://aclanthology.org/2021.gwc-1.29/.

[189] Daniel G. Mcdonald and John Dimmick. The Conceptualization and Measurement of Diversity. *Communication Research*, 30(1):60–79, February 2003. ISSN 0093-6502, 1552-3810. doi: 10.1177/0093650202239026. URL http://journals.sagepub.com/doi/10.1177/0093650202239026.

[190] Robert P. McIntosh. An index of diversity and the relation of certain concepts to diversity. *Ecology*, 48(3):392–404, 1967. doi: 10.2307/1932674. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1932674.

[191] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

[192] Edward F. Menhinick. A comparison of some species-individuals diversity indices applied to samples of field insects. *Ecology*, 45(4):859–861, 1964. doi: 10.2307/1934933. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1934933.

[193] *Daniele Metilli, Valentina Bartalesi, and Carlo Meghini. A wikidata-based tool for building and visualising narratives. *Int. J. Digit. Libr.*, 20(4):417–432, 2019.

[194] *Daniele Metilli, Valentina Bartalesi, Carlo Meghini, and Nicola Aloia. Populating narratives using wikidata events: An initial experiment. In Paolo Manghi, Leonardo Candela, and Gianmaria Silvello, editors, *Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 - February 1, 2019, Proceedings*, volume 988 of *Communications in Computer and Information Science*, pages 159–166. Springer, 2019.

[195] Marc Miquel-Ribé and David Laniado. The wikipedia diversity observatory: A project to identify and bridge content gaps in wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration*, OpenSym 2020, New York, NY, USA, 2020. Association for Computing Machinery. ISBN

9781450387798. doi: 10.1145/3412569.3412866. URL https://doi.org/10.1145/3412569.3412866.

[196] *Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller-Muehlbacher, Lynn M. Schriml, Andrew I. Su, and Benjamin M. Good. Wikidata: A platform for data integration and dissemination for the life sciences and beyond. In *International SWAT4LS Workshop*, 2015.

[197] *Cedric Möller, Jens Lehmann, and Ricardo Usbeck. Survey on english entity linking on wikidata: Datasets and approaches. *Semantic Web*, (Preprint): 1–42, 2022. doi: 10.3233/SW-212865.

[198] *José Moreno-Vega and Aidan Hogan. Grafa: Faceted search & browsing for the wikidata knowledge graph. volume 2180 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[199] E. Kathryn Morris, Tancredi Caruso, François Buscot, Markus Fischer, Christine Hancock, Tanja S. Maier, Torsten Meiners, Caroline Müller, Elisabeth Obermaier, Daniel Prati, Stephanie A. Socher, Ilja Sonnemann, Nicole Wäschke, Tesfaye Wubet, Susanne Wurst, and Matthias C. Rillig. Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecology and Evolution*, 4(18):3514–3524, 2014. doi: https://doi.org/10.1002/ece3.1155. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.1155.

[200] *Mahir Morshed. Modeling syntactic dependency relationships in wikidata lexicographical data. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-7.pdf.

[201] *Hatem Mousselly Sergieh and Iryna Gurevych. Enriching Wikidata with Frame Semantics. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, 2016. doi: 10.18653/v1/W16-1306. URL https://www.researchgate.net/publication/306094169_Enriching_Wikidata_with_Frame_Semantics.

[202] *Alberto Moya Loustaunau and Aidan Hogan. Predicting sparql query dynamics. In *Proceedings of the 11th on Knowledge Capture Conference*, K-CAP '21, pages 161–168, New York, NY, USA, 2021. Association for Computing Machinery.

[203] *Isaiah Onando Mulang', Kuldeep Singh, Akhilesh Vyas, Saeedeh Shekarpour, Maria-Esther Vidal, and Sören Auer. Encoding knowledge graph entity aliases in attentive neural network for wikidata entity linking. In Zhisheng Huang, Wouter Beek, Hua Wang, Rui Zhou, and Yanchun Zhang, editors, *Web Information Systems Engineering - WISE 2020 - 21st International Conference, Amsterdam, The Netherlands, October 20-24, 2020, Proceedings, Part I*, volume 12342 of *Lecture Notes in Computer Science*, pages 328–342. Springer, 2020. doi: 10.1007/978-3-030-62005-9_24. URL https://doi.org/10.1007/978-3-030-62005-9_24.

[204] Claudia Müller-Birn, Leonhard Dobusch, and James D. Herbsleb. Work-to-rule: The emergence of algorithmic governance in wikipedia. In *Proceedings of the 6th International Conference on Communities and Technologies*, C&#38;T '13, pages 80–89, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2104-4. doi: 10.1145/2482991.2482999. URL http://doi.acm.org/10.1145/2482991.2482999.

[205] Claudia Müller-Birn. Bots in wikipedia: Unfolding their duties. Technical reports serie b tr-b-19-01, Freie Universität, Berlin, 2019. URL https://refubium.fu-berlin.de/handle/fub188/24775.

[206] *Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. Peer-production system or collaborative ontology engineering effort: What is Wikidata? In *Proceedings of the 11th International Symposium on Open Collaboration*, page 20. ACM, 2015.

[207] Vivek Nallur, Eamonn O'Toole, Nicolas Cardozo, and Siobhan Clarke. Algorithm diversity: A mechanism for distributive justice in a socio-technical mas. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '16, page 420–428, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450342391.

[208] *Milica Ikonić Nešić, Ranka Stanković, and Biljana Rujević. Serbian eltec sub-collection in wikidata. *Infotheca - Journal for Digital Humanities*, 21(2):60–87, 2022. ISSN 2217-9461. doi: 10.18485/infotheca.2021.21.2.4. URL https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.2.4_en.

[209] *Phuc Nguyen and Hideaki Takeda. Semantic labeling for quantitative data using wikidata. *JSAI Technical Report, Type 2 SIG*, 2018(SWO-045):04, 2018. doi: 10.11517/jsaisigtwo.2018.SWO-045_04.

[210] *Phuc Nguyen, Ikuya Yamada, Natthawut Kertkeidkachorn, Ryutaro Ichise, and Hideaki Takeda. Mtab4wikidata at semtab 2020: Tabular data annotation with wikidata. In Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas, and Vincenzo Cutrona, editors, *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 5, 2020*, volume 2775 of *CEUR Workshop Proceedings*, pages 86–95. CEUR-WS.org, 2020.

[211] *Finn Årup Nielsen. Linking ImageNet WordNet Synsets with Wikidata. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1809–1814, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191645. URL https://doi.org/10.1145/3184558.3191645.

[212] *Finn Årup Nielsen. Danish in wikidata lexemes. In Piek Vossen and Christiane Fellbaum, editors, *Proceedings of the 10th Global Wordnet Conference,*

GWC 2019, Wroclaw, Poland, July 23-27, 2019, pages 33–38. Global Wordnet Association, 2019.

[213] *Finn Årup Nielsen. Ordia: A web application for wikidata lexemes. In Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, editors, *The Semantic Web: ESWC 2019 Satellite Events - ESWC 2019 Satellite Events, Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers*, volume 11762 of *Lecture Notes in Computer Science*, pages 141–146. Springer, 2019.

[214] *Finn Årup Nielsen. Lexemes in wikidata: 2020 status. In Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia, editors, *Proceedings of the 7th Workshop on Linked Data in Linguistics, LDL@LREC 2020, Marseille, France, May 2020*, pages 82–86. European Language Resources Association, 2020.

[215] *Finn Årup Nielsen and Lars Kai Hansen. Inferring visual semantic similarity with deep learning and Wikidata: Introducing imagesim-353. In *Proceedings of the First Workshop on Deep Learning for Knowledge Graphs and Semantic Technologies (DL4KGS)*, volume Vol-2106, Greece, 2018. Department of Applied Mathmatics and Computer Science, Technical University of Denmark. URL http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/7102/pdf/imm7102.pdf.

[216] *Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. Scholia, scientometrics and wikidata. In *The Semantic Web: ESWC 2017 Satellite Events*, pages 237–259, Cham, 2017. Springer International Publishing.

[217] *Finn Årup Nielsen, Katherine Thornton, and José Emilio Labra Gayo. Validating danish wikidata lexemes. In Mehwish Alam, Ricardo Usbeck, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Proceedings of the Posters and Demo Track of the 15th International Conference on Semantic Systems co-located with 15th International Conference on Semantic Systems (SEMANTiCS 2019), Karlsruhe, Germany, September 9th - to - 12th, 2019*, volume 2451 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[218] *Kristian Noullet, Rico Mix, and Michael Färber. KORE 50^{dywc}: An evaluation data set for entity linking based on dbpedia, yago, wikidata, and crunchbase. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2389–2395. European Language Resources Association, 2020.

[219] Eugene P Odum. Organic production and turnover in old field succession. *Ecology*, 41(1):34–49, 1960.

[220] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. The People's Encyclopedia Under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia. *SSRN Elec-*

tronic Journal, 2012. ISSN 1556-5068. doi: 10.2139/ssrn.2021326. URL http://www.ssrn.com/abstract=2021326.

[221] *Antonio Origlia, Silvia Rossi, Sergio Di Martino, Francesco Cutugno, and Maria Laura Chiacchio. Multiple-source data collection and processing into a graph database supporting cultural heritage applications. *J. Comput. Cult. Herit.*, 14(4), jul 2021. ISSN 1556-4673. doi: 10.1145/3465741. URL https://doi.org/10.1145/3465741.

[222] *Natalia Ostapuk, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Sectionlinks: Mapping orphan wikidata entities onto wikipedia sections. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[223] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1):89, 2021. doi: 10.1186/s13643-021-01626-4. URL https://doi.org/10.1186/s13643-021-01626-4.

[224] *Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. From freebase to wikidata: The great migration. In *Proceedings of the 25th international conference on world wide web*, pages 1419–1428. International World Wide Web Conferences Steering Committee, 2016.

[225] *Thomas Pellissier Tanon, Camille Bourgaux, and Fabian Suchanek. Learning how to correct a knowledge base from the edit history. WWW '19, page 1465–1475, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313584. URL https://doi.org/10.1145/3308558.3313584.

[226] *Aleksandr Perevalov, Andreas Both, Dennis Diefenbach, and Axel-Cyrille Ngonga Ngomo. Can machine translation be a reasonable alternative for multilingual question answering systems over knowledge graphs? In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 977–986, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3511940. URL https://doi.org/10.1145/3485447.3511940.

[227] *Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. In *16th IEEE International Conference on Semantic Computing, ICSC 2022, Laguna Hills, CA, USA, January 26-28,*

2022, pages 229–234. IEEE, 2022. doi: 10.1109/ICSC52841.2022.00045. URL https://doi.org/10.1109/ICSC52841.2022.00045.

[228] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic Mapping Studies in Software Engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swindon, UK, 2008. BCS Learning & Development Ltd. URL http://dl.acm.org/citation.cfm?id=2227115.2227123.

[229] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, August 2015. ISSN 0950-5849. doi: 10.1016/j.infsof.2015.03.007. URL http://www.sciencedirect.com/science/article/pii/S0950584915000646.

[230] *Felipe Pezoa, Juan L. Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. Foundations of json schema. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 263–273, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883029. URL https://doi.org/10.1145/2872427.2883029.

[231] *Alexander Pfundner, Tobias Schönberg, John Horn, Richard D. Boyce, and Matthias Samwald. Utilizing the Wikidata system to improve the quality of medical content in Wikipedia in diverse languages: a pilot study. *Journal of medical Internet research*, 17(5), 2015.

[232] *Guangyuan Piao and Weipéng Huáng. Learning to predict the departure dynamics of wikidata editors. In Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam M. Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, editors, *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*, volume 12922 of *Lecture Notes in Computer Science*, pages 39–55. Springer, 2021. doi: 10.1007/978-3-030-88361-4_3. URL https://doi.org/10.1007/978-3-030-88361-4_3.

[233] E. C. Pielou. Shannon's formula as a measure of specific diversity: Its use and misuse. *The American Naturalist*, 100(914):463–465, 1966. doi: 10.1086/282439. URL https://doi.org/10.1086/282439.

[234] *Sini Govinda Pillai, Lay-Ki Soon, and Su-Cheng Haw. Comparing dbpedia, wikidata, and yago for web information retrieval. In Vincenzo Piuri, Valentina Emilia Balas, Samarjeet Borah, and Sharifah Sakinah Syed Ahmad, editors, *Intelligent and Interactive Computing*, pages 525–535, Singapore, 2019. Springer Singapore. ISBN 978-981-13-6031-2.

[235] Alessandro Piscopo. Wikidata: A new paradigm of human-bot collaboration? 10 2018.

[236] *Alessandro Piscopo and Elena Simperl. Who models the world?: Collaborative ontology creation and user roles in wikidata. *Proc. ACM Hum. Comput. Interact.*, 2(CSCW):141:1–141:18, 2018.

[237] *Alessandro Piscopo and Elena Simperl. What we talk about when we talk about wikidata quality: a literature survey. In Björn Lundell, Jonas Gamalielsson, Lorraine Morgan, and Gregorio Robles, editors, *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019, Skövde, Sweden, August 20-22, 2019*, pages 17:1–17:11. ACM, 2019.

[238] *Alessandro Piscopo, Lucie-Aimée Kaffee, Chris Phethean, and Elena Simperl. Provenance Information in a Collaborative Knowledge Graph: An Evaluation of Wikidata External References. In *The Semantic Web ISWC 2017*, Lecture Notes in Computer Science, pages 542–558. Springer, Cham, October 2017. ISBN 978-3-319-68287-7 978-3-319-68288-4. doi: 10.1007/978-3-319-68288-4_32.

[239] *Alessandro Piscopo, Chris Phethean, and Elena Simperl. What makes a good collaborative knowledge graph: Group composition and quality in wikidata. In Giovanni Luca Ciampaglia, Afra J. Mashhadi, and Taha Yasseri, editors, *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I*, volume 10539 of *Lecture Notes in Computer Science*, pages 305–322. Springer, 2017. doi: 10.1007/978-3-319-67217-5_19. URL https://doi.org/10.1007/978-3-319-67217-5_19.

[240] *Alessandro Piscopo, Christopher Phethean, and Elena Simperl. Wikidatians are born: paths to full participation in a collaborative structured knowledge base. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 4354–4363. University of Hawaii, 2017.

[241] *Alessandro Piscopo, Pavlos Vougiouklis, Lucie-AimÃ©e Kaffee, Christopher Phethean, Jonathon Hare, and Elena Simperl. What Do Wikidata and Wikipedia Have in Common?: An Analysis of Their Use of External References. In *Proceedings of the 13th International Symposium on Open Collaboration*, OpenSym '17, pages 1:1–1:10, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5187-4. doi: 10.1145/3125433.3125445. URL http://doi.acm.org/10.1145/3125433.3125445.

[242] *Thomas Ploumis, Isidoros Perikos, Foteini Grivokostopoulou, and Ioannis Hatzilygeroudis. A factoid based question answering system based on dependency analysis and wikidata. In Nikolaos G. Bourbakis, George A. Tsihrintzis, and Maria Virvou, editors, *12th International Conference on Information, Intelligence, Systems & Applications, IISA 2021, Chania Crete, Greece, July 12-14, 2021*, pages 1–7. IEEE, 2021. doi: 10.1109/IISA52424.2021.9555551. URL https://doi.org/10.1109/IISA52424.2021.9555551.

[243] *Adrian Pohl. How we built a spatial subject classification based on wikidata. *Code4Lib Journal*, (51), 2021.

[244] *Jan Portisch, Michael Hladik, and Heiko Paulheim. Finmatcher at finsim-2: Hypernym detection in the financial services domain using knowledge graphs. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 293–297, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383134. doi: 10.1145/3442442.3451382. URL https://doi.org/10.1145/3442442.3451382.

[245] *Radityo Eko Prasojo, Fariz Darari, Simon Razniewski, and Werner Nutt. Managing and Consuming Completeness Information for Wikidata Using COOL-WD. In *Proceedings of the 7th International Workshop on Consuming Linked Data co-located with 15th International Semantic Web Conference (ISWC 2015)*, volume Vol-1666, Kobe, Japan, 2016. CEUR Workshop Proceedings. URL ceur-ws.org/Vol-1666/paper-02.pdf.

[246] *Tim E. Putman, Sebastien Lelong, Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Colin M. Diesh, Nathan A. Dunn, Monica C. Munoz-Torres, Gregory S. Stupp, Chunlei Wu, Andrew I. Su, and Benjamin M. Good. Wikigenomes: an open web application for community consumption and curation of gene annotation data in wikidata. *Database: The Journal of Biological Databases and Curation*, 2017, 2017. doi: 10.1093/database/bax025.

[247] *Hadi Syah Putra, Rahmad Mahendra, and Fariz Darari. Budayakb: Extraction of cultural heritage entities from heterogeneous formats. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, WIMS2019, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361903. doi: 10.1145/3326467.3326487. URL https://doi.org/10.1145/3326467.3326487.

[248] *Wessel Radstok, Melisachew Wudage Chekol, and Mirko T. Schäfer. Are knowledge graph embedding models biased, or is it the data that they are trained on? In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-5.pdf.

[249] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob Crandall, Nicholas Christakis, Iain Couzin, Matthew Jackson, Nicholas Jennings, Ece Kamar, Isabel Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David Parkes, Alex Pentland, and Michael Wellman. Machine behaviour. *Nature*, 568:477–486, 04 2019. doi: 10.1038/s41586-019-1138-y.

[250] William C Rankin, Robert P Markley, and Selby H Evans. Pythagorean distance and the judged similarity of schematic stimuli. *Perception & Psychophysics*, 7(2):103–107, 1970.

[251] *Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. CHOLAN: A modular approach for neural entity linking on wikipedia and wikidata. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 504–514. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.40. URL https://doi.org/10.18653/v1/2021.eacl-main.40.

[252] *Simon Razniewski and Priyanka Das. Structured knowledge: Have we made progress? an extrinsic study of kb coverage over 19 years. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Manage-*

ment, CIKM '20, page 3317–3320, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3417447. URL https://doi.org/10.1145/3340531.3417447.

[253] *Simon Razniewski, Fabian Suchanek, and Werner Nutt. But What Do We Actually Know? pages 40–44. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-1308. URL http://aclweb.org/anthology/W16-1308.

[254] *Simon Razniewski, Vevake Balaraman, and Werner Nutt. Doctoral Advisor or Medical Condition: Towards Entity-Specific Rankings of Knowledge Base Properties. In *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 526–540. Springer, Cham, November 2017. ISBN 978-3-319-69178-7 978-3-319-69179-4. doi: 10.1007/978-3-319-69179-4_37. URL https://link.springer.com/chapter/10.1007/978-3-319-69179-4_37.

[255] Ruqin Ren and Bei Yan. Crowd Diversity and Performance in Wikipedia: The Mediating Effects of Task Conflict and Communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6342–6351, Denver Colorado USA, May 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025992. URL https://dl.acm.org/doi/10.1145/3025453.3025992.

[256] Stephen A. Rhoades. The herfindahl-hirschman index. *Federal Reserve Bulletin*, (Mar):188–189, 1993. URL https://EconPapers.repec.org/RePEc:fip:fedgrb:y:1993:i:mar:p:188-189:n:v.79no.3.

[257] Carlo Ricotta and Giancarlo Avena. On the relationship between pielou's evenness and landscape dominance within the context of hill's diversity profiles. *Ecological Indicators*, 2(4):361–365, 2003. ISSN 1470-160X. doi: https://doi.org/10.1016/S1470-160X(03)00005-0. URL https://www.sciencedirect.com/science/article/pii/S1470160X03000050.

[258] *Daniel Ringler and Heiko Paulheim. One Knowledge Graph to Rule Them All? Analyzing the Differences Between DBpedia, YAGO, Wikidata & co. In *KI 2017: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 366–372. Springer, Cham, September 2017. ISBN 978-3-319-67189-5 978-3-319-67190-1. doi: 10.1007/978-3-319-67190-1_33. URL https://link.springer.com/chapter/10.1007/978-3-319-67190-1_33.

[259] Lionel Robert and Daniel M. Romero. Crowd Size, Diversity and Performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 1379–1382, Seoul, Republic of Korea, 2015. ACM Press. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702469. URL http://dl.acm.org/citation.cfm?doid=2702123.2702469.

[260] Matthew Roth. The wikipedia data revolution, 30 March 2012. URL https://diff.wikimedia.org/2012/03/30/the-wikipedia-data-revolution/. Date of Access: 14.07.2022.

[261] Ronald Rousseau. The repeat rate: From hirschman to stirling. *Scientometrics*, 116(1):645–653, July 2018. ISSN 0138-9130. doi: 10.1007/s11192-018-2724-8. URL https://doi.org/10.1007/s11192-018-2724-8.

[262] *Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphaël Troncy, and Xavier Tannier. Searching news articles using an event knowledge graph leveraged by wikidata. In Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1232–1239. ACM, 2019.

[263] *Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. Falcon 2.0: An entity and relation linking tool over wikidata. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3141–3148. ACM, 2020.

[264] *John Samuel. Collaborative Approach to Developing a Multilingual Ontology: A Case Study of Wikidata. In *Metadata and Semantic Research*, Communications in Computer and Information Science, pages 167–172. Springer, Cham, November 2017. ISBN 978-3-319-70862-1 978-3-319-70863-8. doi: 10.1007/978-3-319-70863-8_16. URL https://link.springer.com/chapter/10.1007/978-3-319-70863-8_16.

[265] *John Samuel. Analyzing and visualizing translation patterns of wikidata properties. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian-Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, volume 11018 of *Lecture Notes in Computer Science*, pages 128–134. Springer, 2018.

[266] *John Samuel. Towards understanding and improving multilingual collaborative ontology development in wikidata. In *Companion of the The Web Conference 2018 on The Web Conference*, pages 23–27, 2018.

[267] *John Samuel. Shexstatements: Simplifying shape expressions for wikidata. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 610–615. ACM / IW3C2, 2021. doi: 10.1145/3442442.3452349. URL https://doi.org/10.1145/3442442.3452349.

[268] *John Samuel. Wdprop: Web application to analyse multilingual aspects of wikidata properties. In Gregorio Robles, Javier Arroyo, Ann Barcomb, Kuljit Kaur Chahal, Sulayman K. Sowe, and Xiaofeng Wang, editors, *OpenSym 2021: 17th International Symposium on Open Collaboration, Virtual Event, Spain, September 15-17, 2021*, pages 10:1–10:12. ACM, 2021. doi: 10.1145/3479986.3479996. URL https://doi.org/10.1145/3479986.3479996.

[269] *Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. Building Automated Vandalism Detection Tools for Wikidata. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1647–1654, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-

1-4503-4914-7. doi: 10.1145/3041021.3053366. URL https://doi.org/10.1145/3041021.3053366.

[270] *Cristina Sarasua, Alessandro Checco, Gianluca Demartini, Djellel 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. *Comput Supported Coop Work*, 28(5):843–882, sep 2019. ISSN 0925-9724, 1573-7551. doi: 10.1007/s10606-018-9344-y. URL http://link.springer.com/10.1007/s10606-018-9344-y.

[271] *Cezar Sas, Meriem Beloucif, and Anders Søgaard. Wikibank: Using wikidata to improve multilingual frame-semantic parsing. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4183–4189. European Language Resources Association, 2020.

[272] *Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Representing mathematical formulae in content mathml using wikidata. In Philipp Mayr, Muthu Kumar Chandrasekaran, and Kokil Jaidka, editors, *Proceedings of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018) co-located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), Ann Arbor, USA, July 12, 2018*, volume 2132 of *CEUR Workshop Proceedings*, pages 46–59. CEUR-WS.org, 2018.

[273] *Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Fast linking of mathematical wikidata entities in wikipedia articles using annotation recommendation. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 602–609. ACM / IW3C2, 2021. doi: 10.1145/3442442.3452348. URL https://doi.org/10.1145/3442442.3452348.

[274] *Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Mathematics in wikidata. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-1.pdf.

[275] *Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Mining mathematical documents for question answering via unsupervised formula labeling. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393454. doi: 10.1145/3529372.3530925. URL https://doi.org/10.1145/3529372.3530925.

[276] *Lukas Schmelzeisen, Corina Dima, and Steffen Staab. Wikidated 1.0: An evolving knowledge graph dataset of wikidata's revision history. In Lucie-Aimée Kaffee, Simon Razniewski, and Aidan Hogan, editors, *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th Interna-*

tional Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-2982/paper-11.pdf.

[277] *Moritz Schubotz. Generating openmath content dictionaries from wikidata. In *Joint Proceedings of the CME-EI, FMM, CAAT, FVPS, M3SRD, OpenMath Workshops, Doctoral Program and Work in Progress at the Conference on Intelligent Computer Mathematics 2018 co-located with the 11th Conference on Intelligent Computer Mathematics (CICM 2018), Hagenberg, Austria, August 13-17, 2018*, volume 2307 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018. URL http://ceur-ws.org/Vol-2307/paper51.pdf.

[278] *Moritz Schubotz, André Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard S. Cohl, and Bela Gipp. Improving the representation and conversion of mathematical formulae by considering their textual context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 233–242, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351782. doi: 10.1145/3197026.3197058. URL https://doi.org/10.1145/3197026.3197058.

[279] *Kenneth Seals-Nutt and Katherine Thornton. Getting digital preservation data out wikidata. In Marcel Ras, Barbara Sierman, and Angela Puggioni, editors, *Proceedings of the 16th International Conference on Digital Preservation, iPRES 2019, Amsterdam, The Netherlands, September 16-20, 2019*, 2019.

[280] *Eva Seidlmayer, Jakob Voß, Tetyana Melnychuk, Lukas Galke, Klaus Tochtermann, Carsten Schultz, and Konrad U. Förstner. ORCID for wikidata. data enrichment for scientometric applications. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[281] *Shilad Sen, Toby Jia-Jun Li, WikiBrain Team, and Brent J. Hecht. Wikibrain: Democratizing computation on wikipedia. In Dirk Riehle, Jesús M. González-Barahona, Gregorio Robles, Kathrin M. Möslein, Ina Schieferdecker, Ulrike Cress, Astrid Wichmann, Brent J. Hecht, and Nicolas Jullien, editors, *Proceedings of The International Symposium on Open Collaboration, OpenSym 2014, Berlin, Germany, August 27 - 29, 2014*, pages 27:1–27:10. ACM, 2014. doi: 10.1145/2641580.2641615. URL https://doi.org/10.1145/2641580.2641615.

[282] *Asara Senaratne, Pouya Ghiasnezhad Omran, Graham Williams, and Peter Christen. Unsupervised anomaly detection in knowledge graphs. In *The 10th International Joint Conference on Knowledge Graphs*, IJCKG'21, pages 161–165, New York, NY, USA, 2021. Association for Computing Machinery.

[283] Zaina Shaik, Filip Ilievski, and Fred Morstatter. Analyzing race and citizenship bias in wikidata. In *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pages 665–666, 2021. doi: 10.1109/MASS52906.2021.00099.

[284] *Abdul Lathif Fathima Shanaz and Roshan G. Ragel. Named entity extraction of wikidata items. In *14th Conference on Industrial and Information Systems, ICIIS 2019, Kandy, Sri Lanka, December 18-20, 2019*, pages 40–45. IEEE, 2019.

[285] *Abdul Lathif Fathima Shanaz and Roshan G. Ragel. Wikidata based person entity linking in news articles. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 66–70, 2021. doi: 10.1109/ICIAfS52090.2021.9606139.

[286] *Fathima Shanaz and Roshan G. Ragel. Wikidata based location entity linking. In *Proceedings of the 9th International Conference on Software and Computer Applications, ICSCA 2020, Langkawi, Malaysia, February 18-21, 2020*, pages 307–312. ACM, 2020.

[287] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[288] *Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro A. Szekely. A study of the quality of wikidata. *J. Web Semant.*, 72:100679, 2022. doi: 10.1016/j.websem.2021.100679. URL https://doi.org/10.1016/j.websem.2021.100679.

[289] *Renat Shigapov, Philipp Zumstein, Jan Kamlah, Lars Oberländer, Jörg Mechnich, and Irene Schumm. bbw: Matching CSV to wikidata via meta-lookup. In Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas, and Vincenzo Cutrona, editors, *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 5, 2020*, volume 2775 of *CEUR Workshop Proceedings*, pages 17–26. CEUR-WS.org, 2020.

[290] *Mei Si. Infer creative analogous relationships from wikidata. In Gabriele Meiselwitz, editor, *Social Computing and Social Media. Design, Human Behavior and Analytics - 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I*, volume 11578 of *Lecture Notes in Computer Science*, pages 140–151. Springer, 2019.

[291] E. H. Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949. doi: 10.1038/163688a0. URL https://doi.org/10.1038/163688a0.

[292] *Eunah Snyder, Lisa Lorenzo, and Lucas Mak. Linked open data for subject discovery: Assessing the alignment between library of congress vocabularies and wikidata. *International Conference on Dublin Core and Metadata Applications*, pages 12–20, Mar. 2020. URL https://dcpapers.dublincore.org/pubs/article/view/4225.

[293] *Ross Spencer, Katherine Thornton, Richard Lehane, and Euan Cochrane. Wikidata: A magic portal for siegfried and roy. In Zijun Chen, editor, *Proceedings of the 17th International Conference on Digital Preservation, iPRES 2021, Beijing, China, October 19-22, 2021*, 2021. URL https://hdl.handle.net/11353/10.1424926.

[294] *Andreas Spitz, Johanna Geiß, and Michael Gertz. So far away and yet so close: augmenting toponym disambiguation and similarity with text-based networks. pages 1–6. ACM Press, 2016. ISBN 978-1-4503-4309-1. doi: 10.1145/2948649.2948651. URL http://dl.acm.org/citation.cfm?doid=2948649.2948651.

[295] *Andreas Spitz, Gloria Feher, and Michael Gertz. Extracting descriptions of location relations from implicit textual networks. In *Proceedings of the 11th Workshop on Geographic Information Retrieval, GIR 2017, Heidelberg, Germany, November 30 - December 01, 2017*, pages 1:1–1:9. ACM, 2017. doi: 10.1145/3155902.3155909. URL https://doi.org/10.1145/3155902.3155909.

[296] *Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz, and Johanna Geiss. State of the union: A data consumer's perspective on wikidata and its properties for the classification and resolution of entities. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(2):88–95, Aug. 2021. doi: 10.1609/icwsm.v10i2.14832. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14832.

[297] *Ranka Stanković and Lazar Davidović. Infotheca (q25460443) in wikidata. *Infotheca - Journal for Digital Humanities*, 21(1):87–98, 2021. ISSN 2217-9461. doi: 10.18485/infotheca.2021.21.1.5. URL https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.1.5_en.

[298] *Thomas Steiner. Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, pages 25:1–25:7, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3016-9. doi: 10.1145/2641580.2641613. URL http://doi.acm.org/10.1145/2641580.2641613.

[299] Steve Stemler. An overview of content analysis. *Practical Assessment, Research, and Evaluation*, 7(17), 2000. doi: https://doi.org/10.7275/z6fm-2e34.

[300] Andy Stirling. On the economics and analysis of diversity. *SPRU Electronic Working Papers Series*, 28, 01 1998.

[301] Andy Stirling. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719, 2007. doi: 10.1098/rsif.2007.0213. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2007.0213.

[302] Margaret-Anne Storey and Alexey Zagalsky. Disrupting developer productivity one bot at a time. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE 2016, pages 928–931, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4218-6. doi: 10.1145/2950290.2983989. URL http://doi.acm.org/10.1145/2950290.2983989.

[303] *Michael Striewe. Dynamic generation of assessment items using wikidata. In Silvester Draaijer, Desirée Joosten-ten Brinke, and Eric Ras, editors, *Technology Enhanced Assessment - 21st International Conference, TEA 2018, Amsterdam, The Netherlands, December 10-11, 2018, Revised Selected Papers*,

volume 1014 of *Communications in Computer and Information Science*, pages 1–15. Springer, 2018.

[304] *Pero Subasic, Hongfeng Yin, and Xiao Lin. Building knowledge base through deep learning relation extraction and wikidata. In Andreas Martin, Knut Hinkelmann, Aurona Gerber, Doug Lenat, Frank van Harmelen, and Peter Clark, editors, *Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE 2019) Stanford University, Palo Alto, California, USA, March 25-27, 2019., Stanford University, Palo Alto, California, USA, March 25-27, 2019*, volume 2350 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[305] *Yuhan Sun and Mohamed Sarwat. Riso-tree: An efficient and scalable index for spatial entities in graph database management systems. *ACM Trans. Spatial Algorithms Syst.*, 7(3), jun 2021. ISSN 2374-0353. doi: 10.1145/3450945. URL https://doi.org/10.1145/3450945.

[306] Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *Journal of Intelligent Information Systems*, 41(2):109–149, 2013. doi: 10.1007/s10844-013-0239-6.

[307] Marcin Sydow, Katarzyna Baraniak, and PaweA Teisseyre. Diversity of editors and teams versus quality of cooperative work: experiments on wikipedia. *J Intell Inf Syst*, 48(3):601–632, June 2017. ISSN 1573-7675. doi: 10.1007/s10844-016-0428-1. URL https://doi.org/10.1007/s10844-016-0428-1.

[308] *Tomás Sáez and Aidan Hogan. Automatically Generating Wikipedia Infoboxes from Wikidata. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1823–1830, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191647. URL https://doi.org/10.1145/3184558.3191647.

[309] *Thang Hoang Ta and Chutiporn Anutariya. A Model for Enriching Multilingual Wikipedias Using Infobox and Wikidata Property Alignment. In *Semantic Technology*, Lecture Notes in Computer Science, pages 335–350. Springer, Cham, November 2014. ISBN 978-3-319-15614-9 978-3-319-15615-6. doi: 10.1007/978-3-319-15615-6_25. URL https://link.springer.com/chapter/10.1007/978-3-319-15615-6_25.

[310] *Thomas Pellissier Tanon and Fabian M. Suchanek. Querying the edit history of wikidata. In Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, editors, *The Semantic Web: ESWC 2019 Satellite Events - ESWC 2019 Satellite Events, Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers*, volume 11762 of *Lecture Notes in Computer Science*, pages 161–166. Springer, 2019.

[311] *Thomas Pellissier Tanon, Marcos Dias de Assunção, Eddy Caron, and Fabian M. Suchanek. Demoing platypus - A multilingual question answering platform for wikidata. In Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni

Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z. Pan, and Mehwish Alam, editors, *The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers*, volume 11155 of *Lecture Notes in Computer Science*, pages 111–116. Springer, 2018.

[312] *Waran Taveekarn, Chatchanin Yimudom, Supisara Sukkanta, Steven J. Lynden, Wudhichart Sawangphol, and Suppawong Tuarob. DATA++: an automated tool for intelligent data augmentation using wikidata. In *16th International Joint Conference on Computer Science and Software Engineering, JCSSE 2019, Chonburi, Thailand, July 10-12, 2019*, pages 91–96. IEEE, 2019.

[313] *Harsh Thakkar, Kemele M. Endris, Jose M. Gimenez-Garcia, Jeremy Debattista, Christoph Lange, and Sören Auer. Are Linked Datasets Fit for Open-domain Question Answering? A Quality Assessment. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, WIMS '16, pages 19:1–19:12, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4056-4. doi: 10.1145/2912845.2912857. URL http://doi.acm.org/10.1145/2912845.2912857.

[314] *Nishad Thalhath, Mitsuharu Nagamori, Tetsuo Sakaguchi, and Shigeo Sugimoto. Wikidata centric vocabularies and uris for linking data in semantic web driven digital curation. In Emmanouel Garoufallou and María Antonia Ovalle-Perandones, editors, *Metadata and Semantic Research - 14th International Conference, MTSR 2020, Madrid, Spain, December 2-4, 2020, Revised Selected Papers*, volume 1355 of *Communications in Computer and Information Science*, pages 336–344. Springer, 2020.

[315] *Katherine Thornton and Kenneth Seals-Nutt. Science stories: Using IIIF and wikidata to create a linked-data application. In Marieke van Erp, Medha Atre, Vanessa López, Kavitha Srinivas, and Carolina Fortuna, editors, *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018*, volume 2180 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[316] *Katherine Thornton, Euan Cochrane, Thomas Ledoux, Bertrand Caron, and Carl Wilson. Modeling the Domain of Digital Preservation in Wikidata. In *In Proceedings of ACM International Conference on Digital Preservation*, Kyoto, Japan, 2017. iPres'17.

[317] *Katherine Thornton, Kenneth Seals-Nutt, E Cochrane, and C Wilson. Wikidata for digital preservation. *iPRES 2019*, 2019. doi: 10.17605/OSF.IO/6ERN4.

[318] Michail Tsikerdekis. Cumulative Experience and Recent Behavior and their Relation to Content Quality on Wikipedia . *Interacting with Computers*, 29 (5):737–754, 06 2017. ISSN 0953-5438. doi: 10.1093/iwc/iwx010. URL https://doi.org/10.1093/iwc/iwx010.

[319] *H. Turki, D. Vrandecic, H. Hamdi, and I. Adel. Using wikidata as a multilingual multi-dialectal dictionary for arabic dialects. In *2017 IEEE/ACS 14th*

International Conference on Computer Systems and Applications (AICCSA), pages 437–442, 2017. doi: 10.1109/AICCSA.2017.115.

[320] *Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, and Helmi Hamdi. Wikidata: A large-scale collaborative ontological medical database. *Journal of Biomedical Informatics*, 99:103292, 2019. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2019.103292. URL https://www.sciencedirect.com/science/article/pii/S1532046419302114.

[321] *Houcemeddine Turki, Mohamed Ali Hadj Taieb, Thomas Shafee, Tiago Lubiana, Dariusz Jemielniak, Mohamed Ben Aouicha, José Emilio Labra Gayo, Eric A. Youngstrom, Mus'ab Banat, Diptanshu Das, and Daniel Mietchen. Representing COVID-19 information in collaborative knowledge graphs: The case of wikidata. *Semantic Web*, 13(2):233–264, 2022. doi: 10.3233/SW-210444. URL https://doi.org/10.3233/SW-210444.

[322] *Theo van Veen. Wikidata:from "an" identifier to "the" identifier. *Information Technology and Libraries*, 38(2):72–81, Jun. 2019. doi: 10.6017ital.v38i2.10886. URL https://ejournals.bc.edu/index.php/ital/article/view/10886.

[323] *Theo van Veen, Juliette Lonij, and Willem Jan Faber. Linking Named Entities in Dutch Historical Newspapers. In *Metadata and Semantics Research*, Communications in Computer and Information Science, pages 205–210. Springer, Cham, November 2016. ISBN 978-3-319-49156-1 978-3-319-49157-8. doi: 10.1007/978-3-319-49157-8_18. URL https://link.springer.com/chapter/10.1007/978-3-319-49157-8_18.

[324] *Jakob Voß. Classification of Knowledge Organization Systems with Wikidata. In *Proceedings of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016)*, volume Vol-1676, Hannover, 2016. CEUR-WS.org. doi: 10.5281/zenodo.61767. URL ceur-ws.org.

[325] *Denny Vrandecic. The rise of Wikidata. *IEEE Intelligent Systems*, 28(4): 90–95, 2013.

[326] *Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, October 2014. doi: DOI:10.1145/2629489.

[327] *Andra Waagmeester, Gregory Stupp, Burgstaller-Muehlbacher Sebastian, Benjamin M Good, Griffith Malachi, Obi L Griffith, Hanspers Kristina, Hermjakob Henning, Toby S Hudson, Hybiske Kevin, et al. Wikidata as a knowledge graph for the life sciences. *eLife*, 2020.

[328] *Andra Waagmeester, Egon L. Willighagen, Andrew I. Su, Martina Kutmon, Jose Emilio Labra Gayo, Daniel Fernández-Álvarez, Quentin Groom, Peter J. Schaap, Lisa M. Verhagen, and Jasper J. Koehorst. A protocol for adding knowledge to wikidata: aligning resources on human coronaviruses. *BMC Biology*, 19(1):12, 2021. doi: 10.1186/s12915-020-00940-y. URL https://doi.org/10.1186/s12915-020-00940-y.

[329] *Sheeban Wasi, Madhurendra Sachan, and Manuj Darbari. Document classification using wikidata properties. In Milan Tuba, Shyam Akashe, and Amit

Joshi, editors, *Information and Communication Technology for Sustainable Development*, pages 729–737, Singapore, 2019. Springer Singapore. ISBN 978-981-13-7166-0.

[330] Wikidata. Rfp botmultichill, 2013. URL https://m.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/BotMultichill. [Last accessed 01 August 2018].

[331] Wikidata. Rfp elphibot_3, 2013. URL www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/ElphiBot_3. [Last accessed 01 August 2018].

[332] Wikidata. Rfp implicatorbot, 2013. URL www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/ImplicatorBot. [Last accessed 01 August 2018].

[333] Wikidata. Rfp structor, 2014. URL www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Structor. [Last accessed 01 August 2018].

[334] Wikidata. Rfp vlsergeybot, 2014. URL www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/VlsergeyBot. [Last accessed 01 August 2018].

[335] Wikidata. Rfp phenobot, 2016. URL www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Phenobot. [Last accessed 01 August 2018].

[336] Wikidata. Rfp wikilovesesbot, 2016. URL www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/WikiLovesESBot. [Last accessed 01 August 2018].

[337] Wikidata. Rfp mechquesterbot_2, 2017. URL www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/MechQuesterBot_2. [Last accessed 01 August 2018].

[338] *Tom Willaert and Guido Roumans. Nitpicking online knowledge representations of governmental leadership. the case of belgian prime ministers in wikipedia and wikidata. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 30(1):1–41, Dec. 2020. doi: 10.18352/lq.10362. URL https://liberquarterly.eu/article/view/10862.

[339] *Avicenna Wisesa, Fariz Darari, Adila Krisnadhi, Werner Nutt, and Simon Razniewski. Wikidata completeness profiling using prowd. In Mayank Kejriwal, Pedro A. Szekely, and Raphaël Troncy, editors, *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, pages 123–130. ACM, 2019.

[340] *Gerhard Wohlgenannt, Nikolay Klimov, Dmitry Mouromtsev, Daniil Razdyakonov, Dmitry Pavlov, and Yury Emelyanov. Using Word Embeddings for Visual Data Exploration with Ontodia and Wikidata. In *Joint Proceedings of BLINK2017: 2nd International Workshop on Benchmarking Linked Data and NLIWoD3: Natural Language Interfaces for the Web of Data*, volume Vol-1932 of *CEUR Workshop Proceedings*, Vienna, Austria, 2017. CEUR-WS.org. URL http://ceur-ws.org/Vol-1932/paper-03.pdf.

[341] *Melisachew Wudage Chekol, Giuseppe Pirrò, and Heiner Stuckenschmidt. Fast interval joins for temporal sparql queries. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1148–1154, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3314997. URL https://doi.org/10.1145/3308560.3314997.

[342] Tomoya Yamazaki, Mei Sasaki, Naoya Murakami, Takuya Makabe, and Hiroki Iwasawa. Ensemble Models for Detecting Wikidata Vandalism with Stacking - Team Honeyberry Vandalism Detector at WSDM Cup 2017. In *WSDM Cup 2017 Notebook Papers*, Cambridge, UK, December 2017.

[343] *Matthew Y. R. Yang, Siwen Yang, and Jimmy Lin. Integration of text and geospatial search for hydrographic datasets using the lucene search library. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA, 2022. Association for Computing Machinery.

[344] *Xi Yang, Shiya Ren, Yuan Li, Ke Shen, Zhixing Li, and Guoyin Wang. Relation Linking for Wikidata Using Bag of Distribution Representation. In *Natural Language Processing and Chinese Computing*, Lecture Notes in Computer Science, pages 652–661. Springer, Cham, November 2017. ISBN 978-3-319-73617-4 978-3-319-73618-1. doi: 10.1007/978-3-319-73618-1_55. URL https://link.springer.com/chapter/10.1007/978-3-319-73618-1_55.

[345] *Liyang Yu. *Other Recent Applications: data.gov and Wikidata*, pages 551–585. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-662-43796-4. doi: 10.1007/978-3-662-43796-4_12. URL https://doi.org/10.1007/978-3-662-43796-4_12.

[346] Tuo Yu, Yiran Zhao, Xiaoxiao Wang, Yiwen Xu, Huajie Shao, Yuhang Wang, Xin Ma, and Dipannita Dey. Vandalism detection midpoint report—the riberry vandalism detector at wsdm cup 2017. University of Illinois at Urbana?Champaign Student Report, not published., 2017.

[347] *Xue-lu Yu and Lin Qiao. Meronymy Relation Extraction Based on 3-Motif in Wikidata. *DEStech Transactions on Computer Science and Engineering*, 0(cnsce), 2017. ISSN 2475-8841. doi: 10.12783/dtcse/cnsce2017/8915. URL http://www.dpi-proceedings.com/index.php/dtcse/article/view/8915.

[348] *Eva Zangerle, Wolfgang Gassler, Martin Pichl, Stefan Steinhauser, and Günther Specht. An Empirical Evaluation of Property Recommender Systems for Wikidata and Collaborative Knowledge Bases. In *Proceedings of the 12th International Symposium on Open Collaboration*, OpenSym '16, pages 18:1–18:8, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4451-7. doi: 10.1145/2957792.2957804. URL http://doi.acm.org/10.1145/2957792.2957804.

[349] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. *Semantic Web*, 7(1):63–93, March 2015. ISSN 22104968, 15700844. doi: 10.3233/SW-150175. URL http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-150175.

[350] *Charles Chuankai Zhang and Loren Terveen. Quantifying the gap: A case study of wikidata gender disparities. In Gregorio Robles, Javier Arroyo, Ann Barcomb, Kuljit Kaur Chahal, Sulayman K. Sowe, and Xiaofeng Wang, editors, *OpenSym 2021: 17th International Symposium on Open Collaboration, Virtual Event, Spain, September 15-17, 2021*, pages 6:1–6:12. ACM, 2021. doi: 10.1145/3479986.3479992. URL https://doi.org/10.1145/3479986. 3479992.

[351] *Charles Chuankai Zhang, Mo Houtti, C. Estelle Smith, Ruoyan Kong, and Loren Terveen. Working for the invisible machines or pumping information into an empty void? an exploration of wikidata contributors' motivations. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022. doi: 10.1145/3512982. URL https://doi.org/10.1145/3512982.

[352] Haifeng Zhang, Yuqin Ren, and Robert Kraut. Mining and Predicting Temporal Patterns in the Quality Evolution of Wikipedia Articles. 2020. doi: 10.24251/HICSS.2020.485. URL https://hdl.handle.net/10125/64227.

[353] *Xingchen Zhou, Peng Wang, Guozheng Li, Jiafeng Xie, and Jiangheng Wu. Weibo-mel, wikidata-mel and richpedia-mel: Multimodal entity linking benchmark datasets. In Bing Qin, Zhi Jin, Haofen Wang, Jeff Z. Pan, Yongbin Liu, and Bo An, editors, *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction - 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceed-ings*, volume 1466 of *Communications in Computer and Information Science*, pages 315–320. Springer, 2021. doi: 10.1007/978-981-16-6471-7 _27. URL https://doi.org/10.1007/978-981-16-6471-7_27.

[354] Qi Zhu, Hongwei Ng, Liyuan Liu, Ziwei Ji, Bingjie Jiang, Jiaming Shen, and Huan Gui. Wikidata Vandalism Detection - The Loganberry Vandalism Detector at WSDM Cup 2017. In *Proceedings of the WSDM Cup 2017: Vandal-ism Detection and Triple Scoring*, volume abs/1712.06922, Cambridge, UK, December 2017. arxiv. URL http://arxiv.org/abs/1712.06922. arXiv: 1712.06922.

[355] *Arkaitz Zubiaga and Aiqi Jiang. Early detection of social media hoaxes at scale. *ACM Trans. Web*, 14(4), aug 2020. ISSN 1559-1131. doi: 10.1145/3407194. URL https://doi.org/10.1145/3407194.

[356] ابوالقاسم احمد بی جیهانی. اشکال العالم. انتشارات آستان قدس رضوی، ۱۹۸۹ . مترجم: عبدالسلام کاتب

[357] غلام حضرت کوشان. سرگذشت ملت مظلوم افغانستان در مسیر سدهٔ بیستم، افغان امریکن اسوسیشن ،
۱۹۹۹. صص ۱۴۸ تا ۱۸۷.