WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

Bachelor Thesis

Representation of political data in Wikidata

Miklos Kosarszky

Date of Birth: 12.11.1996
Student ID: 01621924

Subject Area: Data Science

Studienkennzahl: 01621924

Supervisor: Dr. Sebastian Neumaier, Prof. Dr. Axel Polleres

Date of Submission: 11.06.2020

Department of Information Systems and Operations, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR INFORMATIONS-VERARBEITUNG UND PROZESS-MANAGEMENT DEPARTMENT OF INFORMATION SYSTEMS AND OPERATIONS

Contents

Abstract

Looking at the massive amount of data which is available on the Web, we humans are in need of open knowledge bases that are consistent, complete, and useful. Wikidata, the data management platform for Wikipedia stores all kinds of data. As such, when we are looking for a specific dataset or if we want to answer a question, we could start by looking at Wikidata. In this paper, we are going to analyze all forms of political data on Wikidata, finding out which current representations exist and develop a method which can be used to integrate the most important missing data about political data on Wikidata articles.

1 Introduction

Governments publish every sort of political data such as the results of parliamentary, regional, and local elections. Most of these data can be found on the data pages of governments but there are many other sources which are beginning to focus on political data, such as the Wikidata knowledge base. However, as the existing data on Wikidata is mainly incomplete, we need to find methods to gather political data from several sources which are reliable. These sources can be found as CSV files, Wikipedia pages,and political statistics resources. Political data on Wikidata is especially inconsistent. Regarding this topic, there are already existing studies, such as the WikiProject Every Politician. The focus of this WikiProject is to have complete and consistently structured data across the world on all elected representatives, from national to local level. WikiProject Every Politician aims to have complete and comprehensive coverage of the current members of at least every national legislature, along with ministers, in a consistent format. According to the project, the properties of political events and persons on Wikidata are not consistently added to the records. In the case of politicians, these properties are called "position held", "parliamentary term", "start time", "end time", "parliamentary group", "electoral district", "elected in", etc.. However, in this thesis, we will not only focus on politicians, but also on elections and parties.

1.1 Research problem

Knowledge bases such as Wikidata, can be incomplete, and they require new data to be more useful for users. Looking at the Wikidata articles of simple political elections, we can clearly notice that most of these are in lack of needed information. However, there are elections where the Wikidata page is complete with data. Some of the elections contain only the point in time and the country of the event, while others are rich in information, and giving details about the results of the election, the parties that took part, the leading politicians or candidates of the given party, the previous winner of the same election, etc. Naturally, this problem does not only occur with the elections, but also with political parties and politicians. In order to solve these inconsistencies, we are going to develop a method to find the most crucial information about elections, politicians and parties which are relevant but missing in several cases, and try to integrate the data on Wikidata. The primary issue is that we are not able to assess which knowledge bases are incomplete and require more information and resources. Wikidata pages vary in their completeness, where one article is filled with hundreds of values, for

example articles of leading politicians, European Parliament Elections and powerful political parties, while in others, we can not assess, how many of this information is present in one article and missing in the other. Our first goal is to find methods which are able to assess the completeness of Wikidata articles and provide us with the necessary details to complete political data on Wikidata. Secondly, there is the difficulty of feeding Wikidata with the necessary resources. Even if we are able to evaluate the completeness on Wikidata, we need to find reliable datasets to impute the missing information on these articles. These datasets are present on several sources, such as the open data portals of governments and the European Data portal, and can be extracted from semi-structured knowledge bases like Wikipedia. Finally, we are going to discuss how to integrate the missing entities and properties to Wikidata. With this statement, we can conclude the research question of this thesis, which is the following:

"How well is political data represented in Wikidata and what existing sources could be leveraged to complete the information presently available in Wikidata?"

1.2 Structure of the thesis

This paper conducts a research on political data in Wikidata and discusses how this kind of data could be best represented in the Wikidata knowledge base to give a complete overview. The following chapter "Background" will give a general outline on the knowledge graphs, including a brief history of this phenomenon from the 1980s until today, possible definitions of knowledge graphs and possible ways to assess the quality of these. The paper will discuss about existing methods on how to integrate data and information in a KG, and where to look for this information. After that, the subsequent chapters will explain what the Wikidata knowledge base is and how it works, including the purpose of this project. For the purposes of discussion, this chapter will include the Wikidata query interface and language called SPARQL, which is a powerful tool to obtain relevant data from the knowledge base. A general overview of this query service will follow, describing how to use it for basic purposes. This will be followed by the chapter "Political data in Wikidata", which will analyze the completeness of the existing information in Wikidata, describing statements which are already available and properties which are still missing in the knowledge base. Moreover, the paper will answer the question to the existing representations of political data on Wikidata, and provide a way on how to look for already existing patterns which could be a good starting point for completing the missing data. Furthermore, the thesis will include a chapter on finding a reasonable data model to represent all the crucial political data in the knowledge base. Accordingly, this will be followed by several possible approaches on how to integrate the missing properties and entities on Wikidata with already existing data sources. To conclude this work, related and past researches on this topic will be mentioned , followed by ideas on how to continue with this work in the future.

2 Background

In order to understand the concept of a knowledge graph (KG) and Wikidata better, this section will outline the definition of a knowledge graph, the main characteristics of it, as well as a short history and possible ways to assess the quality of KGs. This will be followed by an introduction to Wikidata and to its query service, which uses the powerful SPARQL Query Language. The knowledge graph is a very useful tool when it comes to quickly accessing relevant information about certain objects, and in our case political data. Knowledge graphs on the Web are the support of many information systems that require access to structured knowledge [1].

2.1 The Knowledge Graph

The idea of integrating general, formalized knowledge into intelligent systems originates from the 1980s [2]. During the 1980s, the upheaval of computation from industry to homes was a result of the evolution of personal computers. Moreover, the Relational Database industry was developing rapidly in the data management sector such as Oracle, Sybase and IBM [3]. The WEB tended to change the way people communicated and transferred information. Combining data and knowledge gained importance in the 80s. Japanese scientists wanted to develop an Artificial Intelligence hardware and software that could initiate conversations, translate languages and reason like human beings by combining logic and data[3]. As a result, they developed logic programming as the base to combine logic and data [3]. In the 2000s, the Semantic Web Project focused on the combination of data and knowledge. The project concentrated on knowledge representation, especially the work on ontologies from the 1990s [3].

In the 2000s, Linked Data became important for interconnecting data in order to increase knowledge on the web. As a result, the Linked Open Data Project and large RDF-based knowledge bases such as DBPedia and Freebase were introduced. The latter would eventually lead to Wikidata [3]. The Linked Open Data Project was a test of how data could be integrated at web scale. In 2007, several of these ideas were introduced at the International Semantic Web Conference. Schema.org was released to improve the semantic glossary of websites, which was built on the results of the Semantic Web Project [3].

With the growth of Linked Open Data [4], sources like DBpedia, and by the release of the Google Knowledge Graph in 2012, knowledge graphs have aroused a lot of attention. Knowledge graphs can be constructed in several ways. There are knowledge graphs like Cyc [5], ones that can be edited by

people like Freebase [6] and Wikidata, or extracted from semi-structured web knowledge bases such as Wikipedia.

The term Knowledge Graph was invented by Google in 2012, referring to their use of semantic knowledge [7]. This initiated other types of knowledge graphs as well. In 2013 Facebook launched their graph search, which is highly similar to the KG of Google, and is mainly presenting a virtual graph that integrates compiled data on topics and entities [7]. Several knowledge graphs are being developed from other big companies like Microsoft, Facebook, Amazon and Ebay accordingly [7].

2.1.1 Definition of Knowledge Graphs

In the last few years, there has been extensive research into knowledge graphs, especially in the Semantic Web community, and naturally all sorts of definitions have been suggested [8]. According to Paulheim et al., "A knowledge graph mainly describes real world entities and their interrelations, organized in a graph" [9].

Lei Zhang defined in his thesis Knowledge Graph Theory and Structural Parsing: "Knowledge graph theory is a kind of new viewpoint, which is used to describe human language. Knowledge graphs have advantages, which are stronger ability to express, to depict deeper semantic layers, to use a minimum relation set and to imitate the cognition course of mankind etc. Its appearance gave a new way to the research of computer understanding of human language" [10]. Later, M. Nickel, K. Murphy, V. Tresp and E. Gabrilovich defined knowledge graphs in their article A Review of Relational Machine Learning for Knowledge Graph as follows: "[...] a graph structured knowledge bases that store factual information in form of relationship between entities" [10].

Several big companies have been developing their own knowledge graphs in the last decade, but we know very little about them. For instance, Google keeps it well in secret how their knowledge graph is constructed; there are a few articles from external sources trying to figure out the information flow into the knowledge graph [1]. From those, it can be assumed that Google's Knowledge Graph collects data from open, semi-structured sources such as Wikipedia [1]. Google's Knowledge Graph contains 18 billion statements about 570 million entities, 1500 entity types and 35000 relations [11] .

In knowledge graphs, random entities can be interrelated, covering various topical domains. But how are knowledge graphs created? Several instructions have been implemented as to how to create KGs or how to make them available worldwide, and for everyone to use [10]. Initially, knowledge graphs represent some form of knowledge that is manageable for processing by graph algorithms. Knowledge graphs are structured so that entities are related to their attributes and to other entities, and the source of that knowledge is indicated [10]. KGs can be generated through a variety of approaches, using data sources such as text documents, microdata in web pages, large and small databases [10]. They can be used for data processing activities like mapping entities to concepts, extracting relations from text, integrating different data sources, and finding errors through quality assessment [10]. All of the above mentioned methods have their unique pros and cons, and can often create different knowledge graphs that can have important role in answering questions and providing new predictions [10].

2.2 About Wikidata

Wikidata is the knowledge base of Wikipedia, and the central data management platform for Wikipedia [12]. It is a part of the Wikimedia foundation and edited by a community of thousands of users. Since its existence, Wikipedia has been collecting massive amounts of structured data: numbers, dates, coordinates and relationships [12]. Wikipedia's data is stored within 30 million Wikipedia articles in 287 languages, from where it is very difficult to extract [12]. The goal of Wikidata is to overcome these problems by creating new ways for Wikipedia to manage its data [12].

Wikidata contains various data types (e.g. text, images, quantities, coordinates, geographic shapes, dates), which can be modified, copied, and distributed without permission [12]. The data on Wikidata is not showing the real information but it is rather a "knowledge basket" from different sources where users can decide for themselves, whether a statement is valid or not [12]. In the last 8 years the site has gathered data on more than 15 million entities, including 34 million statements, and over 80 million labels and descriptions in more than 350 languages [13]. Forty-thousand registered users have actively contributed since the beginning of Wikidata which can be explained by the close connection with Wikipedia.

One of the main advantages of Wikidata is that it allows every user of the site to extend and edit the stored information, even without creating an account. This way, a large number of users can edit data [12]. All of the information is controlled by the contributor community. Since many facts are simply uncertain, there in no global agreement as to which data is considered as "true". Wikidata allows conflicting data to coexist [12]. Take the population of a city for example. The population is an ever-changing number which is published in a given point in time but it changes every day, so there can't be an exact population of Vienna. Moreover, numbers, dates, and coordinates have universal meaning; labels like Rome and population are translated into many languages [12]. Wikidata is multi-lingual by design. While Wikipedia has independent editions for each language, there is only one Wikidata site. Wikidata's goal is to allow that Wikipedia and external applications can make use of its data [12]. Data is exported through web services in different formats, such as JSON and RDF [12]. Instead of developing a perfect system that is presented to the world in a couple of years, new features are added in steps and as early as possible. These properties characterize Wikidata as a special curated database [12]. Wikipedia's data value has long been obvious, and many attempts have been made to use it.

2.2.1 What query language does Wikidata use?

Wikidata can be queried via a query interface and language called SPARQL[1]. SPARQL is the query language for Resource Description Framework (RDF), a data model used for the portrayal of information about World Wide Web resources [14]. SPARQL was released in 2004 as a query language for RDF and consists of three parts; the pattern matching part, the solution modifier and the output of a SPARQL query [15]. The pattern matching part includes optional parts, union of patterns, nesting, filtering values of possible matching, and the possibility of choosing the data source to be matched by a pattern, all of which are characteristics of pattern matching of graphs [14]. The second part, the solution modifier, permits the modification of the values from the pattern matching part, using operators like projection, distinct, order, limit and offset [14]. Lastly, the output of a SPARQL query can take different forms, varying from yes/no queries and selections of values of the variables which match the patterns to construction of new triples from these values or even descriptions of resources [14]. In order to allow querying for everyone in Wikidata, Wikimedia has launched a query service which runs over an RDF depiction of Wikidata [16].

2.2.2 How to query Wikidata?

The Query Service[2] is one of the highlights of Wikidata. For the basics of querying, we will take a look at the identifiers first. The ones starting with Q stand for items or entities and those beginning with P refer to properties. Entities and relationships are connected with labels, where the example shows labels in English [16]. For purposes of discussion, let us take a look at the following simple query in Figure 1. The query uses properties P31 (instance of), P17 (country) and P585(point in time). Furthermore, there is a single entity Q40231 (election) which can also be called as an item in Wikidata queries. The first row of the query beginning with "SELECT" describes the columns in the resulting query and with "ORDER BY" we are ordering the results by an arbitrary order, in this case in the alphabetic order of the country labels. But what does this simple query mean? Well, in this instance we are asking the Query Service to return all of the elections (Q40231) on Wikidata, in which property country (P17) and point in time (P585) is present.

The table in Figure 2 is the result of the above mentioned query and shows all of the elections in the item columns which contain the statements

[1]https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service

[2]https://query.wikidata.org/

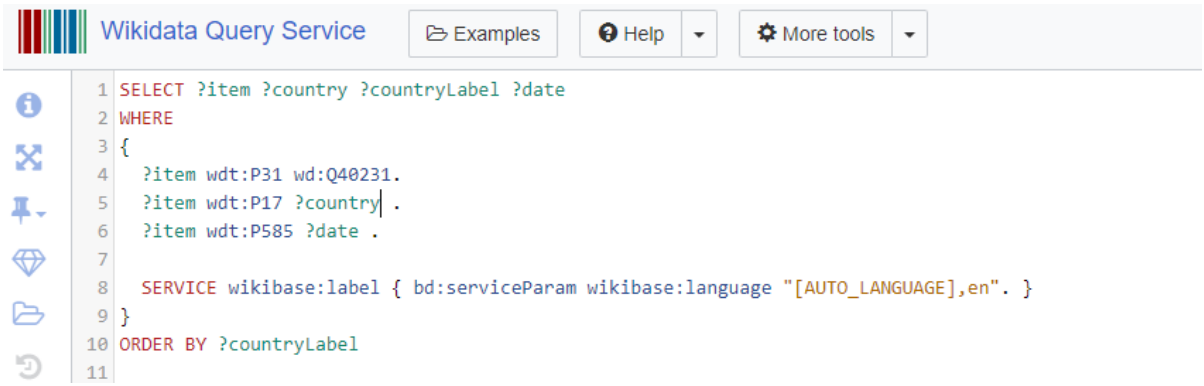
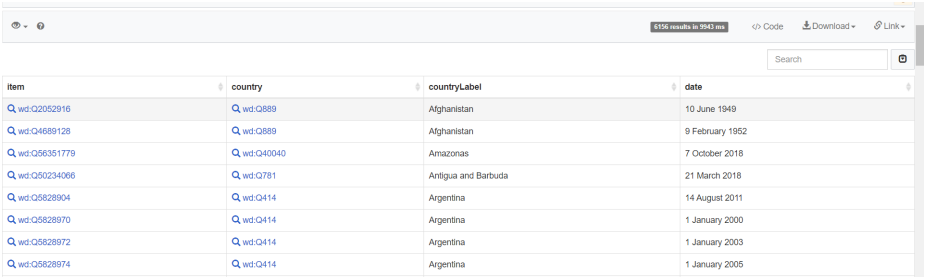


Figure 1: Example query at the Wikidata query interface



item	country	countryLabel	date
wd:Q2052916	wd:Q889	Afghanistan	10 June 1949
wd:Q4689128	wd:Q889	Afghanistan	9 February 1952
wd:Q56351779	wd:Q40040	Amazonas	7 October 2018
wd:Q50234066	wd:Q781	Antigua and Barbuda	21 March 2018
wd:Q5828904	wd:Q414	Argentina	14 August 2011
wd:Q5828970	wd:Q414	Argentina	1 January 2000
wd:Q5828972	wd:Q414	Argentina	1 January 2003
wd:Q5828974	wd:Q414	Argentina	1 January 2005

Figure 2: An example for the result of a query

country and point in time on their Wikidata articles. In the third column, we can observe the "countryLabel" column which is the labeling of its respective entity in the second column. The Label function is a built-in function of the Wikidata Query Service and retrieves the title of the given entity automatically from Wikidata. Wikidata queries can give us relevant information and can extract any kind of data, with a query composed of logical combinations of triples.

Figure 3: Example of a Wikidata statement with a property, a value and a set of qualifiers, *Source: https://www.wikidata.org/wiki/Q22686*

Queries can lead the user to the Wikidata article, where the information of the given subject is stored. Upon examining the Wikidata article of leader politicians, such as in Figure 3, which describes the current position of Donald Trump as President of the United States, we can take a look at a statement on Wikidata. The statement contains Donald Trump (Q22686) as subject, position held (President of the United States) as predicate, and President of the United States of America as object; this relation is associated with pairs of qualifiers (start time, elected in, replaces) and their values (20 January 2017). One can define a Wikidata statement as a primary relation and a set of qualifiers [16]. In an ideal scenario, all of the statements in Wikidata should include values with a reference. However, the knowledge base is incomplete in that some statements may not contain any qualifiers or values and thus the Wikidata query can be incorrect [16].

3 Political data in Wikidata

Political information is conveyed not only through speeches and media reports but also through a variety of data which is available on the web. Which kind of data do we consider political data? Political data consists of all forms of information which is related to political life. Perhaps one of the most common utilization of this data results from elections, politicians and parties. Political elections can differ from general elections through national council election to the European Parliament Elections. Naturally, elections have entrants, a location (country), a time of date, number of votes, a percentage of the result. Political data also comprises of data about politicians. Politicians provide information about the political stance, the party which the given politician represent, but also personal data, such as date of birth, origin, and nationality. Data which can be gathered from political parties give insight about the angle of a party, the representative politicians, received votes in elections, group of people who support a given party.

These events, persons and institutions have produced millions of data which are available to assess on hundreds of platforms. Feasibly, one of the most resourceful sources can be found on open government data pages. These data are free to use without restrictions, available for reuse, and unlicensed. Open government data is useful to improve transparency and accountability by making government data available to all. By making their datasets available, public institutions become more transparent.

As mentioned, political data can vary from election results to the birth place of a leading politician. The data can be qualitative, (e.g: In which country was the certain politician elected?), quantitative (How many votes did the given party receive in a general election?) or any other numeric data which could be gathered through surveys and interviews. In this study, we will mainly focus on data which could be of use to Wikidata articles, and with that in mind we can deduct the relevant elements for this research.

Let's take a look at a Wikidata article of a political election in Figure 4. As the Wikidata article of the 2017 United Kingdom election[3] shows, there are common types of data on Wikidata which occur in most of the cases, such as the country where the election took place, the previous and next election in this category and the date of the election. Some of the articles express more detail about an election, like the 2019 European Parliament election which gives information about the participants, its official website or office contested. There are also leading politicians who have excessive Wikidata articles. The president of France, Emmanuel Macron provides

[3]https://www.wikidata.org/wiki/Q25052149

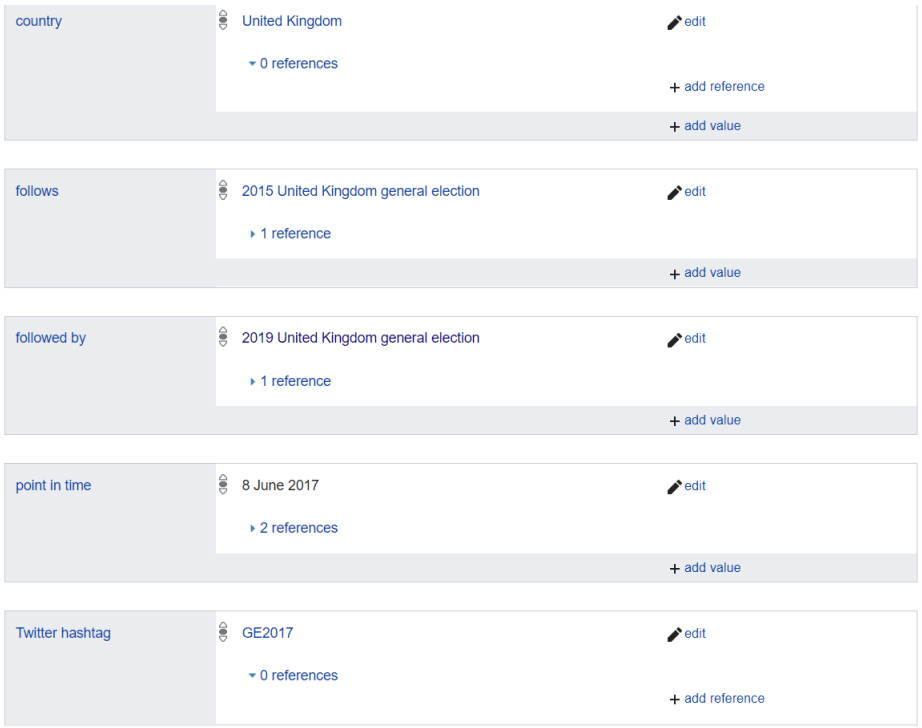


Figure 4: Example of statements about the 2017 UK General election, *Source: https://www.wikidata.org/wiki/Q25052149*

Figure 5: Wikipedia infobox of President Emmanuel Macron,
Source: https://en.wikipedia.org/wiki/Emmanuel_Macron

valuable data from its Wikidata article. If we browse the Wikipedia page of the french president, there is an info box on the top-right corner of the article, containing the most important data of the president. This data is provided by Wikidata. Wikidata acts as a central repository for facts and has been used as a source of information to create info-boxes [17]. The particular info-box for Emmanuel Macron (see Figure 5 and 6) include the current political status as the president of France, his assumed offices, and their start and end time, personal details, such as residence, education and other political affiliations. The Google Result of Macron shows similarities with the Wikipedia info-boxes, because Google info-boxes also refer to Wikidata as primary resource for information. The Figure 6 below consists of a short description about the president, followed by the information provided by Wikidata, which are place of birth and education. These data are the result of knowledge graphs which were shaped by semi-structured knowledge like Wikipedia [1].

Figure 6: Google infobox information of President Emmanuel Macron, *Source: https://www.google.com/search?q=macronoq=macronaqs=chrome/*

3.1 How can we assess completeness?

Wikidata shows a high potential in creating structured information about the world and reveals a high level of correctness [18]. Unfortunately, it lacks the sufficient degree of completeness to be considered a reliable source of information on many different topics, among others political data.

Wikidata articles do not contain more than a few statements in the majority of articles regarding political data. In order to complete data, we need to analyze if the given politician or election contains the relevant information and if it does, we have to assess how much of this information is available on Wikidata. How should we assess the completeness of politicians, and how can we tell if a knowledge base article about a particular politician is complete? For instance, while former UK president, Theresa May can have a handful of properties on Wikidata, the same cannot be said for politicians who are not in a leading position. There are also several giveaways, which hint that the completeness of political data is not complete. With a SPARQL Query in Figure 7, we can find out the share of politicians, who are member of a political party.

In Wikidata, 319149 out of 603707 politicians are member of a party, which is a percentage of 52.9. We know, that this number is incorrect, since most of the politicians do have a party. Furthermore, if politicians are independent and are not member of a party, the property member of political

```
SELECT ?pol ?memberof
WHERE
{

    ?pol wdt:P106 wd:Q82955 .
    ?pol wdt:P102 ?memberof .

    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".}
}
```

Figure 7: Query for politicians who are member of a political party

party will still be included with the value "independent politician", such as in the Wikidata article of Giuseppe Conte[4] . In this case, we can deduce, that data about politicians are incomplete in Wikidata. The incompleteness can be explained by the data coverage of the politicians who are not in leading positions and only contain a few properties. However, since Wikidata does not yet assess the quality of an article, we can't state for sure whether the data is complete or incomplete.

The Recoin (Relative completeness indicator)[5] has been implemented to measure the completeness of a Wikidata article. With the Recoin tool, users can investigate the property and entity completeness. Recoin's goal is to answer the question: "How complete is data about an entity in Wikidata as a whole [19]?" Besides this function, the Recoin tool can also help in improving the quality of the knowledge base. So how does it work exactly? To measure the completeness of data about elections, one should look at the knowledge base contents for other elections worldwide, while the assessment of the completeness of a political party should be done by comparing the data with data on similar parties [19]. Relative completeness is based on two components: the similarity function between subject pairs, and a scoring function that computes a score or rank for the completeness [19]. It looks at the most common properties in the given subject that are missing and shows the average frequency of these missing properties in a comparison set of other subjects in Wikidata [19]. As a result, we get an overall completeness score of five different levels; Level five with a 100-95 percent score, Level four with 95-90 percent score, etc. At the end of the completeness analysis, the results are

[4]https://www.wikidata.org/wiki/Q53844829

[5]https://www.wikidata.org/wiki/Wikidata:Recoin

represented visually in Wikidata by Toolforge, provided by Wikimedia [19]. With the help of the implemented tool, we can analyze the most common entities in similar Wikidata articles, such as elections.

3.2 Which different representations exist and how can we determine preferred representations?

Political data could be one of the highlights in Wikidata, however it is scarcely represented uniform in this knowledge base. For instance, legislative elections tend to provide different information than Wikidata articles of general elections, while parliamentary elections may show a bigger degree of completeness than the former mentioned. In order to investigate the representations of political data on Wikidata, we are going to take a look at the most recent representations of these elections, including similarities and patterns in their structure. The focus of this structure will be on the Wikidata statements, namely the properties and entities of political elections, politicians and parties. The basic issue with the current representation form of these articles on Wikidata is that the certain political objects tend to be often a subclass of an always changing and different class. What does this actually mean? For instance, the 2019 European Parliament election in Austria[6] is a subclass of election to the European Parliament, while the 2018 Salzburg State Election[7] is a subclass of the legislative election. According to that, there are noticeable differences in the statements of these two articles. The Salzburg State Election contains properties : "applies to jurisdiction" and "followed by" which are not represented in the European Parliament election in Austria article, however the latter one contains properties such as office contested, participant, and commons category. This means, that the lack of uniformity on this knowledge base could be a result of the wide variety of classes. If we are to discover the preferred representations of political data in Wikidata, we need to study how classes and subclasses are built up.

The most encountered class of European elections on Wikidata are the legislative election and the presidential election. Legislative election is a subclass of general election which is again a subclass of election, while presidential election is directly a subclass of election . According to the Wikidata Query Service, the election class (Q40231) has 787 subclasses in total.

The problem with these subclasses are that they are represented as "election in the United States" or "election in Romania", and we do not get a clear overview of the different types of elections. Furthermore, in the list of results, only a handful of elections could be identified as political ones. To narrow down the search results, we will focus on general elections[8] . General election contains 328 subclasses, including legislative elections and general elections in more than a 100 countries. Moreover, we could drill-down on legislative

[6]https://www.wikidata.org/wiki/Q43477954
[7]https://www.wikidata.org/wiki/Q28784089
[8]https://www.wikidata.org/wiki/Q1076105

```
SELECT ?item ?itemLabel
WHERE
{
    ?item wdt:P279* wd:Q40231.

    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en"}
}
```

Figure 8: Query for all subclasses of elections

elections which has again 197 subclasses. As previously mentioned, some of the elections are directly subclass of legislative election, but there are some which have a country specific subclass between legislative election and the certain election. Legislative election is the last level of subclasses, which means that further drill-down in subclasses is not possible, since the query result only contains the table of countries where legislative elections exist or have existed. From here on, we are talking about instances. Upon querying all instances of legislative elections in a specific country, for example in Austria, all of the legislative elections in this country will be included in the query result which are available on Wikidata.

The election articles show up statements which can have a typical pattern. These are emerging patterns which are hard to miss if we are looking at the similarity of political data on Wikidata. To investigate these similarities among elections, we will use a query which looks at the properties with the highest occurrence in elections. This query states that the properties "country" and "point in time" are very common items in elections. A single SPARQL Query states that there are 42910 elections in Wikidata, from which 36086 include the property "country".

This means that 84.1 percent of the election have this kind of property. The same can be calculated for the property "point in time", where 37584 elections contain this statement out of the overall 42910 elections. This results in an even higher completeness percentage of 87.6 in all elections. Likewise, "candidate" tend to be present in most of the elections. In 29196 of all election instances (68.0 percent), the property "candidate" is included. Additionally, with a simple query, it is possible to find out the most common properties without querying this information one by one. The query in Table 1 looks for the count of all properties in elections by descending order. This query states that there are a total of 367 distinct properties in the election

```
SELECT ?item ?itemLabel ?country
WHERE
{

    ?item wdt:P31/wdt:P279* wd:Q40231.
    ?item wdt:P17 ?country.

    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en"}
}
```

Figure 9: Query for all elections with the property country

```
SELECT DISTINCT ?p (COUNT(DISTINCT ?item) AS ?count)
WHERE {

  ?item wdt:P31/wdt:P279* wd:Q40231; ?p [] .

}
GROUP BY ?p
ORDER BY DESC (?count)
```

Figure 10: Query to count the number of distinct properties in elections

class. From these properties, table 1 examines the most common properties.

These numbers are helpful to show which statements are truly important to correctly complete and represent the data in political elections. Naturally, the majority of properties don't show such a satisfying percentage. The property "successful candidate" , which identifies the winning party or politician, is only present in 17394 (40.5 percent) of the elections. This number can be misleading however, since not all of the elections use the property "successful candidate" for proclaiming the winner. For instance, the 2014 European Parliament election in the Czech Republic includes the property "winner" instead. Furthermore, there are some elections which are more like referendums and the outcome is rather for or against something, such as the Brexit referendum in 2016, where voters chose between remaining or leaving. However, these differences are pretty well mapped in Wikidata and are subclasses

Property	Count	Percentage
Point in time (P585)	37584	87.6
Country (P17)	36086	84.1
Candidate (P726)	29196	68.0
Office contested (P541)	22123	51.6
Follows (P155)	15473	36.1
Followed by (P156)	15317	35.7
Applies to jurisdiction (P1001)	11178	26.1
Successful candidate (P991)	9859	23.0
Part of (P361)	8398	19.6
Ballots cast (P1868)	4249	9.9

Table 1: Most common properties in elections

```
SELECT DISTINCT ?itemLabel ?candidateLabel ?votes
WHERE
{
  ?item wdt:P31/wdt:P279* wd:Q40231;
        p:P726 [ ps:P726 ?candidate ; pq:P1111 ?votes ] .


SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en"}
}
```

Figure 11: Query for all "votes received" qualifier in Wikidata which is used on "candidate" in elections

of e.g: referendums and not of elections.

Moreover, one of the most important property when talking about election results, "votes received" (P1111), is missing from the most common properties table. This property displays the number of votes that a candidate received in an election. As the property is used as a qualifier on candidate (P726) and successful candidate (P991), it won't be recognized by the previous query. To find out the completeness of "votes received", we need to query it directly. The following query in Figure 11 shows all of the Wikidata candidates in elections who have the "votes received" qualifier. It seems that 26852 of 42904 elections use this qualifier which yields a completeness score of 62.6 percent.

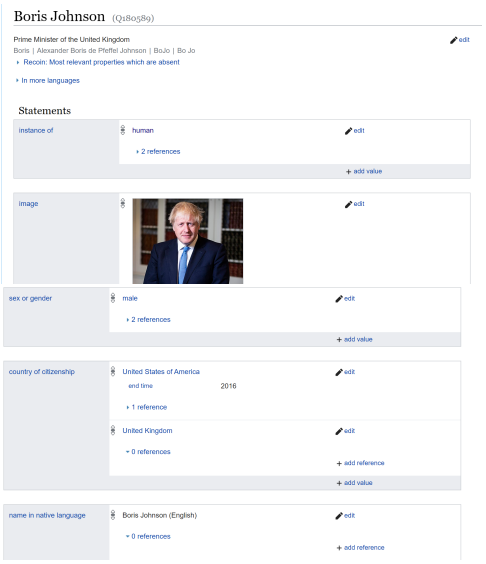


Figure 12: Wikidata item of Boris Johnson,
Source: https://www.wikidata.org/wiki/Q180589

Another important component of political data are politicians themselves. The structure of politicians are highly different from political elections. The Wikidata article of Boris Johnson in Figure 12 gives a general overview how politicians are represented in the knowledge base.

Unlike elections, politicians do not have any superclasses related to politics, a politician in Wikidata is the instance of human (Q5)[9]. For defining a query however, we will have to identify and to disassociate politicians from other natural persons. Politicians can be identified by the occupation (P106) property, in which the value "politician" (Q82955) occurs. The query in Figure 13 below will search for all politicians in Wikidata who have this value in the occupation property. According to the query, there are 603162 politicians registered in the knowledge base.

Moreover, properties could be further analyzed by narrowing the query search. If one would be interested in all politicians in the EU region, the query in Figure 14 would return the desired results.

From these 603162 politicians, 217067 have an EU citizenship, which is a surprisingly high share of 35.99 percent. Naturally, the interpretation of this number can be defined in multiple ways. As politicians in Wikidata don't include the property "country", the best way to identify the nation in which they are politically active, is by the property "country of citizenship".

[9]https://www.wikidata.org/wiki/Q5

```
SELECT ?pol
WHERE
{

    ?pol wdt:P106 wd:Q82955 .

    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".}
}
```

Figure 13: Query for all politicians in Wikidata

```
SELECT ?pol ?citizen
WHERE
{
    ?pol wdt:P106 wd:Q82955 .
    ?pol wdt:P27 ?citizen.
    ?citizen wdt:P463 wd:Q458.

    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".}
}
```

Figure 14: Query for politicians who are citizen of an EU country

```
SELECT ?president ?presidentLabel ?start WHERE {

   ?president wdt:P31 wd:Q5 .
   ?president p:P39 ?positionheld .
   ?positionheld ps:P39 wd:Q1006398 .
   ?positionheld pq:P580 ?start .

   SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en" . }

 } ORDER BY ?start
```

Figure 15: Query for chancellors in Austria with the start date of taking office

This statement, however, can be misleading as the citizenship of a politician is not always the same country where their political position is held. For instance, American senator Ted Cruz[10] has a Canadian citizenship, but is a senator in the United States. The above mentioned Boris Johnson had a USA-United Kingdom dual citizenship, and consequently can be identified as an American politician as well. Another highlight of politicians in Wikidata is the position held (P39) property. In this property, one could look for all roles that were fulfilled by a particular politician in his or her career. For example, Angela Merkel has fulfilled 15 positions according to her Wikidata page, including Federal Chancellor of Germany, member of the German Bundestag, Chairman of the CDU, etc. This property can give access to a number of possibilities upon querying, such as looking for all presidents in a given country or search for all past mayors of a city. For instance, a single query below enables to find all Chancellors of Austria with the start time of taking office.

Since the roles of politicians (position held) are pretty well mapped in the knowledge base, the query results are highly sufficient and complete. Similarly to the example above, one could query for every aspects of political positions such as mayors, prime ministers, foreign ministers, minister for national defence, or representatives of the European Parliament, as well as hundreds of similar roles. Previously, we have managed to count all distinct properties of elections, and identify the most common ones. By politicians, this is somewhat problematic since we will not be able to look for "instance

[10]https://www.wikidata.org/wiki/Q2036942

```
SELECT DISTINCT ?p (COUNT(DISTINCT ?item) AS ?count)
WHERE {
  ?item wdt:P106 wd:Q82955; ?p [] .

}
GROUP BY ?p
ORDER BY DESC (?count)
```

Figure 16: Query to count the number of distinct properties of politicians

```
SELECT ?item
WHERE
{
  ?item wdt:P279* wd:Q7210356.

  SERVICE
  wikibase:label { bd:serviceParam
  wikibase:language "[AUTO_LANGUAGE],en"}
}
```

Figure 17: Query for subclasses of organizations

of" politicians, but to use the property "occupation" (P106). The query in Figure 16 results in a query timeout error, since the query does not use the common "instance of" property and Wikidata may have better support for the "instance of" (P31) property, while P106 takes much longer to load.

The third main aspect of political data are political parties. Political party is the subclass of political organization and non-governmental organization, and the former ones are subclass of organization. Organizations (Q43229) have a total of 13081 subclasses, and only a handful of them are political organisations.

Political organisations have 2008 subclasses, including political associations, parliamentary groups, terrorist organisations, governments, and political parties. According to the Wikidata query in Figure 18, the knowledge base consists of 19198 parties from all over the world.

Following the routine which we performed by elections, we will identify the most common properties in political parties. From the comprehensive 19198 parties, 14138 objects contain the property "country", which results

```
SELECT ?item ?itemLabel
WHERE
{
  ?item wdt:P31/wdt:P279* wd:Q7278.

  SERVICE
  wikibase:label { bd:serviceParam
  wikibase:language "[AUTO_LANGUAGE],en"}
}
```

Figure 18: Query for all political parties in Wikidata

```
 SELECT DISTINCT ?p (COUNT(DISTINCT ?item) AS ?count)
WHERE {
  ?item wdt:P31/wdt:P279* wd:Q7278; ?p [] .

}
GROUP BY ?p
ORDER BY DESC (?count)
```

Figure 19: Query to count the number of distinct properties of political parties

in a 73.6 completeness score. It is followed by "inception", which is the date or point in time when the subject came into existence as defined. "Political ideology" tends to be present in most of the cases, as well as "headquarters location", "official website", and "dissolved, abolished or demolished" (in case the party ceased its existence). Table 2 below provides an overview of the most common properties with their count and completeness percentage in among political parties.

Similarly to elections, several properties are missing in the majority of cases, although they belong to the basic information of a political party. "Chairperson" (P488) is only present in 2470 of all political parties and with that have a completeness of 12.9 percent. This modest rate, however, could be explained with the fact that some parties don't have one chairperson, but rather a council with multiple politicians. One of the property which was mentioned above, called "dissolved, abolished, or demolished" could also explain the scarce rate of the chairperson property. In some cases, some

Property	Count	Percentage
Country (P17)	14138	73.6
Inception (P571)	12769	66.5
Political ideology (P1142)	7054	36.7
Headquarters location (P159)	5480	28.5
Official website (P856)	5444	28.4
VIAF ID (P214)	3812	19.9
Dissolved, Abolished or Demolished (P576)	3240	16.9
WorldCat Identities ID (P7859)	2485	12.9
Chairperson (P488)	2479	12.9
Logo image (P154)	2385	12.4

Table 2: Most common properties among political parties

```
SELECT ?party ?partyLabel ?country ?countryLabel
WHERE
{
  ?party wdt:P31/wdt:P279* wd:Q7278 .
  ?party wdt:P17 ?country.
  ?country wdt:P17 wd:Q29.


SERVICE wikibase:label
{ bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Figure 20: Query for the political parties in Spain

older parties do not contain the data of their chairpersons.

Parties can be also easily queried regionally if we were only interested in European Union parties or explicitly looking for parties in a country. For this query, one has to look how the party's location or headquarters is represented in the knowledge base. Politicians were described by their citizenship, but parties, in the contrary, include the property country. The query in Figure 20 will result in a total of 1079 parties in the country of Spain.

Moreover, political parties can be analyzed regarding their political ideology or political alignment. With the combination of "country" and "political ideology", the query service is able to return all parties with the desired political ideology and country. For instance, if someone was looking for the

```
SELECT ?party ?partyLabel ?countryLabel ?politicalidLabel
WHERE
{
  ?party wdt:P31/wdt:P279* wd:Q7278 .
  ?party wdt:P17 ?country.
  ?country wdt:P17 wd:Q27.
  ?party wdt:P1142 wd:Q6216.
  ?party wdt:P1142 ?politicalid


  SERVICE wikibase:label
  { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".}
}
```

Figure 21: Query for political parties in Ireland which have a liberal political ideology

liberal parties in Ireland, the query should look like the one in Figure 29.

The possibilities with Wikidata queries are not limitless, but this mainly relies on the incompleteness of data in the knowledge base. As discussed in this chapter, information about a certain object is scarcely complete, and even if it was, the properties are so heterogeneous that obtaining complete answers about political data with queries is very difficult.

3.3 What is a reasonable data model for representation in order to cover all relevant aspects of political data?

In the previous chapter, we have examined how political data is represented in Wikidata, and how we can obtain detailed answers with different queries. As we have seen, the queries can only provide us with sufficient results if the data in the knowledge base is consistent and more or less complete. Since the representation of data, and of political data in particular is not uniform across the knowledge base, the query results can be incomplete and incorrect. But then again, the consistency of the data depends on how the information in Wikidata is represented. We have seen in several cases, that Wikidata use various properties for the same object, such as "successful candidate" and "winner" for the winner of an election, or "country" or "citizenship" for the definition of a country of an object.

One way to identify a good model for political data is to find properties which are consistent and already present in most of the objects in Wikidata. We have already examined the most common properties in the different aspects of political data, such as elections, parties and politicians in the previous chapter. These properties are a good start for a consistent representation of political data.

In the case of elections, the basic information of the date and time ("point in time"), and location ("country") is more or less complete. The candidates who were looking for taking office in an election are represented with the property "candidate". This property indicates a good completeness as well and can be used for the possible candidates in an election. However, Wikidata is not consistent in how the winning candidate is represented and there are a few properties which try to model this phenomenon. For example, "successful candidate" and "winner" represent the same thing, where only the name of the property differs. With 40.5 percent of completeness, "successful candidate" tends to be a more consistent property than "winner". To be able to get better results upon querying, these properties should be unified and only the property with the highest completeness score should be used to identify one particular entity. Furthermore, in some cases successful candidate is present while candidate is absent and vice versa. There are multiple cases where the problem with heterogeneous properties occurs. Among elections, there are properties called "follows" and "followed by". Although these two properties are widely used in the knowledge base (36.1 and 35.7 percent in all elections), sometimes these properties are embedded in the property "part of the series" (P179), which makes querying difficult. For

```
SELECT ?pol ?partof ?polLabel
WHERE
{
    ?pol wdt:P31/wdt:P279* wd:Q40231 .
    ?pol wdt:P179 ?partof .

    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".}
}
```

Figure 22: Elections which use the property "part of the series"

instance, the Finnish Parliamentary Elections use this property[11] .

According to the query in Fig 22, all of the Italian general, Finnish parliamentary, and Japanese local elections use the property "part of the series" instead of "follows" and "followed by". The property is only included in 91 elections worldwide, so it would be better to use the more common form in terms of consistency. There are also properties which seem to be used in a perplexed way such as the property "candidate" and "participant" (P710). The 2017 German federal election uses participant for the parties which were eligible for vote, while candidate stands for the politicians who represented the party in the particular election. However, the 2018 Hungarian parliamentary election uses candidate for the parties and not for the politicians[12]. In addition, some of the elections use only the property participant instead of candidate.

Politicians, on the other hand, are slightly more different to represent in a consistent way. As already mentioned, politicians have a huge amount of data which are not related to political life. The Wikidata article of the German Chancellor, Angela Merkel consists of properties which are related to a politicians family, spoken languages, awards received, education, current residence, work location and so on. Only a small number of properties can be identified as political, for instance the "member of political party", "candidacy in election", or "affiliation". The previously stated issue with the properties that mean the same but are named differently is also occurring here. Angela Merkel has the property "member of" and "member of political party". This won't be a problem but both of these properties include Angela Merkel's political party. While "member of" displays all memberships includ-

[11]https://www.wikidata.org/wiki/Q41310
[12]https://www.wikidata.org/wiki/Q28723346

```
SELECT ?pol ?memberof ?party
WHERE
{

    ?pol wdt:P106 wd:Q82955 .
    ?pol wdt:P463 ?memberof.
    ?pol wdt:P102 ?party.

    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".}
}
```

Figure 23: Politicians with the property "member of" and "member of political party"

ing parties, "member of political party" consists only of the political parties. Other politicians show a different representation in terms of these properties. French president, Emmanuel Macron possesses both of these properties as well, but "member of" does not include political parties in his case. The head of government in Spain, Pedro Sánchez Pérez-Castejón[13] only has the property "member of political party". According to Wikidata queries, 46972 politicians include the property "member of", 319617 include "member of political party" (which is a 53 % completeness score among politicians regarding there are 604993 politicians in the knowledge base), and 29642 who have both properties. Similarly to elections, we can conclude that properties in this class needs to be represented in a clearer format. While "member of" could rather refer to non-political memberships, "member of political party" should refer to the parties of the politicians, since it is already present in more than half of the articles about politicians.

Political parties feature some properties which can lead to misleading query results. The property "political ideology" represents the political stance of a party. However, since a party can evolve from a liberal to a right wing populist one over time, this property can include all of the political ideologies of a party since its existence. There are parties which can have more than five political ideologies like La République En Marche[14] in France or highly controversial ideologies like FIDESZ[15] in Hungary. As a

[13]https://www.wikidata.org/wiki/Q6070218
[14]https://www.wikidata.org/wiki/Q23731823
[15]https://www.wikidata.org/wiki/Q387006

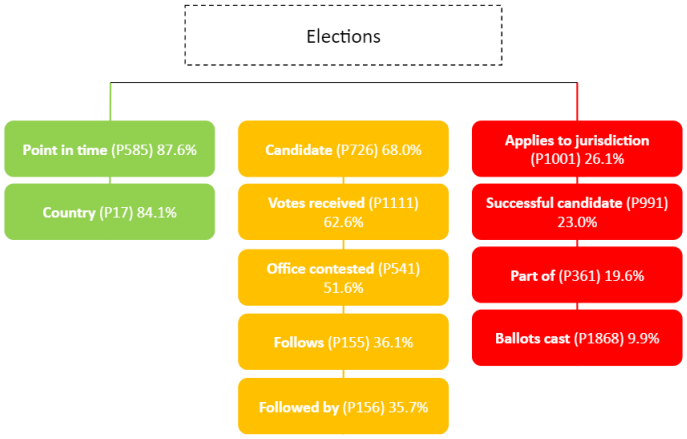


Figure 24: Property Coverage of Elections in Wikidata

result, a query result for a liberal party can provide the users with a party which is not liberal anymore, but rather populist. Moreover, "political ideology" is not to be befuddled with "political alignment" which can describe the political position of a party, such as left- and right wing, centrism and radical alignment. The property "member count" (P2124) is one of the properties which could be significantly extended since the information about the number of members in a party is easily available. The property shows a weak completeness score of 3.2 % (621 out of 19224 parties). However, just as elections, parties already have an emerging schema to represent political data. "Country", "inception", "political ideology", "headquarters location" are already properties which can lay the foundations of a good schema, but their completeness needs to be improved and there are other features of a party which also need to be taken in consideration.

Wikidata's aim is to be as reliable as possible in terms of data consistency. In order to achieve this, Wikidata articles have to be compliant with the preferred data model. The preferred data models are displayed in Figures 24 and 25, and correspond to the most common properties in a given class (in our case elections and political parties).

Observing some of the articles in a specific country is one way to examine the compliance with the model. For instance, the 2019 Austrian legislative election Wikidata article is almost compliant with the preferred data model, since eight out of the ten most common properties are present, only "successful candidate" and "part of" is missing. The same can be observed for

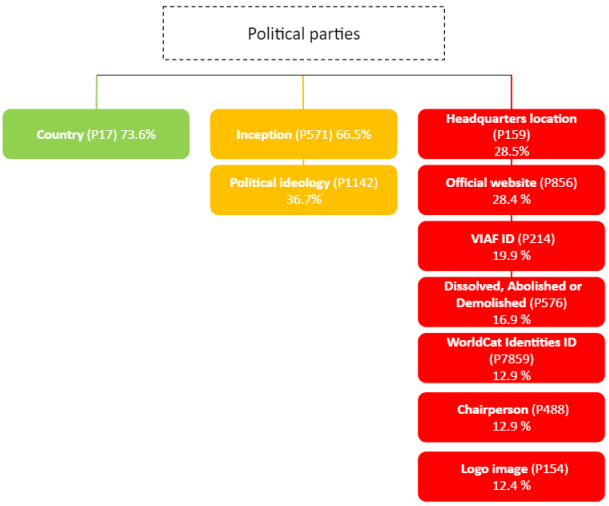


Figure 25: Property Coverage of Political Parties in Wikidata

Austria in the years of 2017 and 2013. The query in Figure 26 checks the Wikidata objects, which are compliant with the preferred model in Austria, and indeed the result yields 2013, 2017 and 2019 elections. Some of the countries are less compliant with the preferred data model, such as Hungary. The 2018 Hungarian parliamentary election includes seven out of the ten most common properties, while the ones in 2014 and 2010 include only six, and the list of candidates are missing. Naturally, the query used previously returns a result of zero elections in Hungary, as none of the articles contain the required properties.

All in all, Wikidata is in a promising status for creating a consistent and uniform model to represent political data, but there is still progress to be made.

```
SELECT DISTINCT ?item ?itemLabel ?countryLabel
WHERE
{

    ?item wdt:P31/wdt:P279* wd:Q40231.
    ?item wdt:P17 ?country.
    ?item wdt:P17 wd:Q40 .
    ?item wdt:P1001 ?d.
    ?item wdt:P155 ?f.
    ?item wdt:P156 ?g.
    ?item wdt:P585 ?h.
    ?item wdt:P541 ?j.
    ?item wdt:P726 ?k.
    ?item wdt:P1868 ?l.


    SERVICE wikibase:label
    { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en"}
}
```

Figure 26: Query to check which elections are complaint with the preferred data model in Austria

4 Existing data sources to complete political data in Wikidata

In order to accomplish the completeness of political data in Wikidata, one has to look for existing data sources on the web. But where could we find relevant data for the Wikidata knowledge base? The first option is to look for datasets from governmental sources such as the dataset for the 2019 Austrian General Election results on the Austrian government web page[16] or obtaining data from the European Election Database[17]. Nevertheless, it is difficult to find consistent datasets for every election which has taken place in Europe. However, the main difficulty is that data about parties and politicians can be represented as a set of letters and numbers instead of a real names, and it takes time to identify that particular entity. Likewise, there could be identifiers for a region where an election took place. Sometimes, these regions tend to be displayed as numbers, and it would be especially hard to identify which region is represented by which number. Furthermore, datasets are scarcely available before a given date. A majority of datasets are only containing information from the 21st century. However, if one were to integrate these information to the Wikidata knowledge base, datasets of open data portals prove to be one of the best resources. Most of the open data portals among elections can be found by looking at the reference option in Wikidata articles, where the "candidate" property is already filled with the overall votes. In cases like this, there is an object beneath every party/candidate which is indicated as reference. This object contains the source from where Wikidata gets the information. The majority of these datasets are CSV files containing numeric and non-numeric data. The following file in Figure 27 demonstrates the result of the 2019 Austrian Legislative Election. The table shows the total votes in Austria, and also the votes from different states and provinces. These datasets are providing us with more detailed information than Wikidata articles in terms of votes received for different political parties but are lacking of general details about the election. These are the information about the previous and the next election, the precise time and date of the event, or the office contested. In terms of completeness, these datasets are a rich resource for the received votes. The data about received votes in Wikidata can be found in the property "candidate" (P726), which again includes the property "votes received" (P1111)[18]. With further analysis, one can see that despite the importance of this property, it is not included in the most com-

[16]https://www.data.gv.at/katalog/dataset/8becadda-124e-43f3-900e-f1ab685574e5
[17]https://nsd.no/european_election_database/
[18]https://www.wikidata.org/wiki/Q19311231

B	C	D	E	F	G	H	I	J	K	L	M	N	
Gebietsname	Wahlberechtigte	Abgegebene	Ungültige	Gültige	ÖVP	SPÖ	FPÖ	NEOS	JETZT	GRÜNE	KPÖ	WANDL	B
Österreich	6396812	4835469	58223	4777246	1789417	1011868	772666	387124	89169	664055	32736	22168	
Burgenland	233182	189911	2850	187061	71566	54965	32448	9130	2351	15113	739	489	
Wahlkarten - Burgenland	0	1	1	0	0	0	0	0	0	0	0	0	
Burgenland Nord	123715	100813	1539	99274	37009	28872	16466	5590	1396	9084	411	292	
Wahlkarten - Burgenland N	0	359	1	358	86	58	57	44	14	87	7	4	
Burgenland Süd	109467	89097	1310	87787	34557	26093	15982	3540	955	6029	328	197	
Wahlkarten - Burgenland Sü	0	359	5	354	86	56	59	39	12	98	1	2	
Eisenstadt Stadt	10798	8574	89	8485	3523	1685	1162	654	142	1220	39	39	
Eisenstadt	10798	7192	78	7114	2966	1450	1023	535	120	944	28	27	
Wahlkarten - Eisenstadt Sta	0	1382	11	1371	557	235	139	119	22	276	11	12	
Rust Stadt	1594	1276	21	1255	424	389	239	69	15	106	5	5	
Rust	1594	1055	18	1037	355	315	209	52	12	82	5	4	
Wahlkarten - Rust Stadt	0	221	3	218	69	74	30	17	3	24	0	1	
Eisenstadt - Umgebung	34307	28264	448	27816	10317	8572	4327	1482	410	2478	114	82	
Breitenbrunn am Neusiedle	1582	1137	27	1110	376	331	149	77	38	132	5	2	
Donnerskirchen	1535	1076	17	1059	470	242	200	60	13	67	6	1	
Großhöflein	1676	1164	25	1139	487	271	183	67	7	115	4	5	
Hornstein	2397	1647	34	1613	515	538	267	89	30	151	13	8	

Figure 27: Result of the 2019 Austrian Legislative Elections in a csv file, *Source: https://www.data.gv.at/katalog/dataset/ergebnisse-der-nationalratswahl-2019-aviso/resource/3865d49b-9d3c-480a-8d0e-959bebddc81a*

mon properties table in the elections class. We have already mentioned that this property is used as a qualifier on candidate (P726) and successful candidate (P991). We know from the previous chapters that candidate shows an overall completeness of 68% in elections, so it requires more data from external sources. In order to be able to integrate this data in Wikidata, one has to map the overall results to the existing property "votes received" (P1111). The data about the missing parties could also be extended since several articles in Wikidata only provide the received votes for the prominent parties and smaller parties which receive a very low percentage, are not indicated. Furthermore, Wikidata could also include the votes in a regional breakdown, including the separate regions with separate vote percentages. Since political elections in Wikidata already consist of the property "candidate", and the respective parties inside this property, there is no need to create a brand new property for this circumstance. Instead of that, the "votes received" property inside of "candidate " should be extended with the overall votes in the country and than with the respective regions and (if needed) the sub-regions of that country. In the case of the example in Fig. 27 the "votes received" would include Austria (Q40) with the respective number of votes, followed by Burgenland (Q43210) and Burgenland's number of votes and so on.

Although these datasets are hard to digest at first, they have a huge advantage. Data can be imported to Wikidata in two ways, either manually, or automatically by a bot. In order to import it automatically, the data has to be publicly reliable and preferably online[19]. The previously mentioned datasets match these criteria and so do all other datasets from open data

[19]https://www.wikidata.org/wiki/Wikidata:Data_Import_Guide/Step_8:_Add_the_data_to_Wikidata

portals. Wikidata also requires that the uploaded data is in some form of tabular data, preferably a spreadsheet, therefore csv datasets tend to be the most useful resources. Upon importing the data, one has to define the structure of the data in Wikidata. We have already analyzed the structure of political data and proposed a preferred model, which could be best used to cover all aspects. The values in the dataset can be mapped to the identified properties in the preferred model, and if none of the existing properties can be used to define the data, one can suggest new properties as well. Once the data has been matched and the data is properly formatted as described in the Wikidata Import Guide, it is ready to be imported to the knowledge base.

On the other hand, not only datasets are available for those who would like to integrate election data to Wikidata. Several other sources are reliable as well. One of them is the Poll of Polls portal of the Politico web page[20], which provides us with information about the most recent elections results of the 27 EU countries, with additional data about entry polls of forthcoming elections, voting intention percentages, and number of seats in the parliament won by the different parties. Figure 28 gives an overview of the two most recent election in Finland on the right, and shows the current status of the distribution of seats in the Finnish parliament on the left.

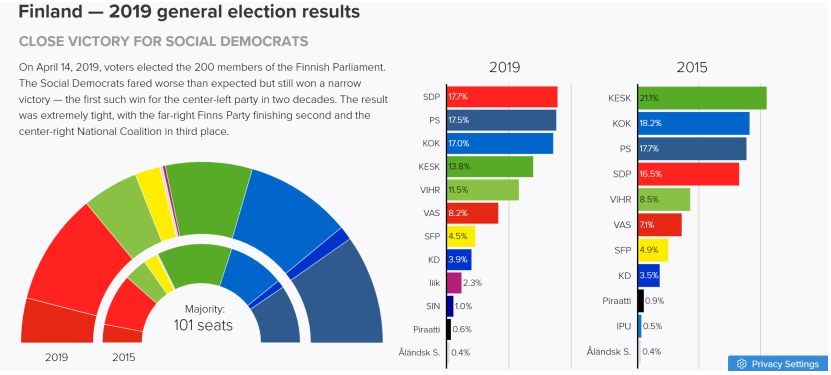


Figure 28: Poll of Polls: Election data in Finland,
Source: https://www.politico.eu/europe-poll-of-polls/finland/

In addition, Poll of Polls also reveals the results of the last European Parliament election with the distribution of seats in the EU Parliament, the overall number of seats for a country and the share of Euroskeptics and Pro-EU countries. An example for Spain can be seen in Figure 29. The above mentioned data are not yet included in Wikidata, but they may be potential

[20]https://www.politico.eu/europe-poll-of-polls/

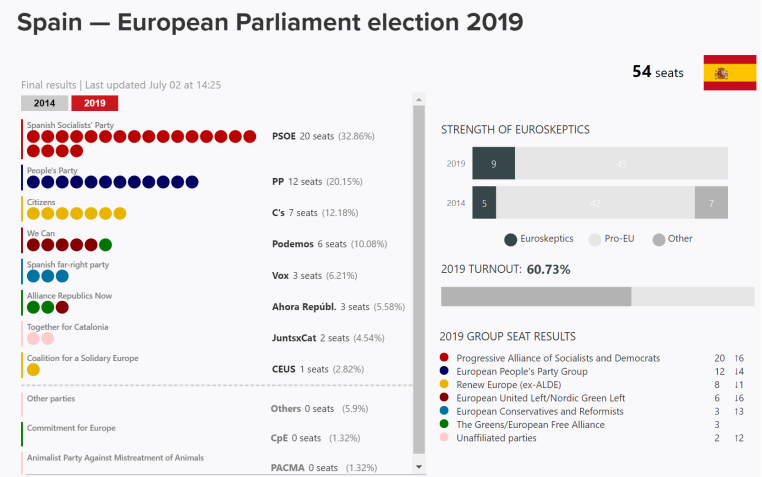


Figure 29: Poll of Polls: European Parliament election in Spain, *Source: https://www.politico.eu/europe-poll-of-polls/spain/*

candidates to complete the knowledge base in the future. Nevertheless, it is impossible to integrate these types of data automatically, as they are not datasets, but merely visualisations of the results.

4.1 How to import the missing data to Wikidata?

In order to complete the missing information, Wikidata offers a way to upload and import data on its website. Data can be added manually by editing the entries in Wikidata items, or they can be uploaded as datasets. If one is choosing the latter version, the first thing to do is to identify a dataset which can be used for the completion of data. For instance, if we were

Gebietsname	Wahlberec	Abgegeber	Ungültige	Gültige	SPÖ	ÖVP	FPÖ	GRÜNE	NEOS	PILZ	GILT	FLÖ	KPÖ	WEIßE
Österreich	6400993	5120881	50952	5069929	1361746	1595526	1316442	192638	268518	223543	48234	8889	39689	9167
Burgenland	232740	196577	1947	194630	64070	63858	49127	3932	5603	5529	1295	273	722	221
Wahlkarten - Burgenland	0	17	17	0	0	0	0	0	0	0	0	0	0	0
Burgenland Nord	122645	103971	1025	102946	34186	33221	25142	2266	3401	3314	730	129	428	129
Wahlkarten - Burgenland Nord	0	333	1	332	70	77	61	24	38	45	4	0	10	3
Burgenland Süd	110095	92589	905	91684	29884	30637	23985	1666	2202	2215	565	144	294	92
Wahlkarten - Burgenland Süd	0	411	2	409	118	89	84	36	33	32	8	0	8	1
Eisenstadt(Stadt)	10628	8722	66	8656	2151	3525	1714	347	419	364	60	9	60	7
Eisenstadt	10628	7546	64	7482	1862	3037	1521	283	351	314	52	9	48	5
Wahlkarten - Eisenstadt(Stadt)	0	1176	2	1174	289	488	193	64	68	50	8	0	12	2
Rust(Stadt)	1564	1299	10	1289	447	371	347	29	44	33	12	0	6	0
Rust	1564	1116	9	1107	369	330	309	25	36	24	11	0	3	0
Wahlkarten - Rust(Stadt)	0	183	1	182	78	41	38	4	8	9	1	0	3	0
Eisenstadt-Umgebung	33905	29325	300	29025	10148	9206	6851	592	877	919	238	34	123	37
Breitenbrunn am Neusiedler See	1598	1199	15	1184	408	356	264	28	40	74	10	1	3	0
Donnerskirchen	1520	1157	12	1145	312	427	310	24	28	28	10	1	3	2
Großhöflein	1661	1293	14	1279	371	484	306	26	37	40	3	6	6	0
Hornstein	2324	1708	20	1688	613	491	387	40	43	71	26	1	13	3
Klingenbach	898	721	10	711	331	188	149	7	11	14	7	1	0	3
Leithaprodersdorf	948	786	8	778	165	394	152	16	27	16	6	0	2	0
Mörbisch am See	1829	1455	11	1444	513	424	406	22	38	22	13	0	5	1
Müllendorf	1139	885	13	872	304	258	208	13	41	38	5	1	4	0
Neufeld an der Leitha	2565	1981	20	1961	804	414	540	27	64	81	17	5	8	1
Oggau am Neusiedler See	1471	1203	13	1190	417	425	278	16	21	24	1	4	3	1

Figure 30: Result of the 2017 Austrian Legislative Elections, *Source: https://www.data.gv.at/katalog/dataset/ergebnisse-der-nationalratswahl-2017/resource/612f3ac8-0c97-47de-8c75-65a917a53e40*

to complete the election results in the 2017 Austrian Legislative Elections, we can find an appropriate dataset on the open data portal of the Austrian Government, which can be observed in Figure 30. For an import of static data like election data, the QuickStatements tool can be helpful. QuickStatements is a tool that can edit Wikidata items, add and remove statements, labels, descriptions, and add statements with optional qualifiers[21]. For the first data import, I'd suggest the following process:

1) Check how the information is structured in Wikidata and what kind of properties are used for election data. In this paper, we have already identified how elections are represented in Wikidata. In the case of this election, we only need to locate a few properties and qualifiers. Since we would like to add additional regional election data to this Wikidata item, we know that the qualifier "votes received" (P1111) in "candidate" (P726) has to be edited. The "votes received" property is added beneath every party and shows the overall votes in Austria. Hence we need additional qualifiers for the 9 Austrian states and their corresponding number of votes. So we will

[21]https://www.wikidata.org/wiki/Help:QuickStatements

have to add Burgenland (Q43210), Lower Austria (Q42497), etc. beneath the overall result.

2) Checking which part of the data may already be in Wikidata, for example with the help of the Query Service. In this scenario, we can skip this step as we already know which properties and qualifiers are included in political elections in Wikidata.

3) Preparing a CSV spreadsheet file to be compatible with the structure of the data in Wikidata. In Figure 30, we can observe that the dataset is overloaded with information which are not necessary for our purposes. The first step of this preparation is to delete all the data which are unrelated to the election results in Wikidata. This means, we can leave out stats such as population of state, number of people who are entitled to vote, invalid votes, and small political parties which are not indicated in the original Wikidata item. As a result, we will get a CSV similar to Figure 31.

	Gebietsname				Gültige	ÖVP	SPÖ	FPÖ	NEOS	JETZT	GRÜNE	KPÖ	WANDL
G00000	Österreich				4777246	1789417	1011868	772666	387124	89169	664055	32736	22168
G10000	Burgenland				187061	71566	54965	32448	9130	2351	15113	739	489
G30000	Niederösterreich				1027276	434783	204679	168565	78760	17751	112607	5347	4784
G40000	Oberösterreich				845096	310611	187173	147873	61853	12451	115983	4750	4092
G20000	Kärnten				311812	108809	81578	61674	21193	5220	29654	1597	1327
G50000	Salzburg				298325	138349	48800	40830	25153	4274	37624	1888	1407
G60000	Steiermark				715170	278228	137494	131999	50754	12156	92799	8953	2787
G70000	Tirol				387159	177351	50393	56867	34300	6751	56947	2159	1447
G80000	Vorarlberg				184433	67540	24232	27121	25017	3921	33462	871	1446
G90000	Wien				820914	202180	222554	105289	80964	24294	169866	6432	4389

Figure 31: Sorted CSV of the election results

4) Converting the CSV file to be compatible with QuickStatements. In order to be able to convert the CSV, we have to indicate the Wikidata Identifier (Entity ID) on every single object in the sorted CSV file such as in Figure 32. Subsequently, we can use the following tool[22] to convert the CSV

Gebietsname	Entity ID			Valid votes	ÖVP (Q186867)	SPÖ (Q179111)	FPÖ (Q131692)	NEOS (Q13564543)	JETZT (Q34198369)	GRÜNE (Q193178)
Austria	Q40			4777246	1789417	1011868	772666	387124	89169	664055
Burgenland	Q43210			187061	71566	54965	32448	9130	2351	15113
Lower Austria	Q42497			1027276	434783	204679	168565	78760	17751	112607
Upper Austria	Q41967			845096	310611	187173	147873	61853	12451	115983
Carinthia	Q37985			311812	108809	81578	61674	21193	5220	29654
Salzburg	Q43325			298325	138349	48800	40830	25153	4274	37624
Styria	Q41358			715170	278228	137494	131999	50754	12156	92799
Tirol	Q42880			387159	177351	50393	56867	34300	6751	56947
Vorarlberg	Q38981			184433	67540	24232	27121	25017	3921	33462
Vienna	Q1741			820914	202180	222554	105289	80964	24294	169866

Figure 32: Sorted CSV of the election results with identifiers

file to a spreadsheet which is compatible with QuickStatements. As a result, we will generate a spreadsheet which will look like Figure 33.

[22]https://tools.wmflabs.org/ash-dev/wdutils/csv2quickstatements.php

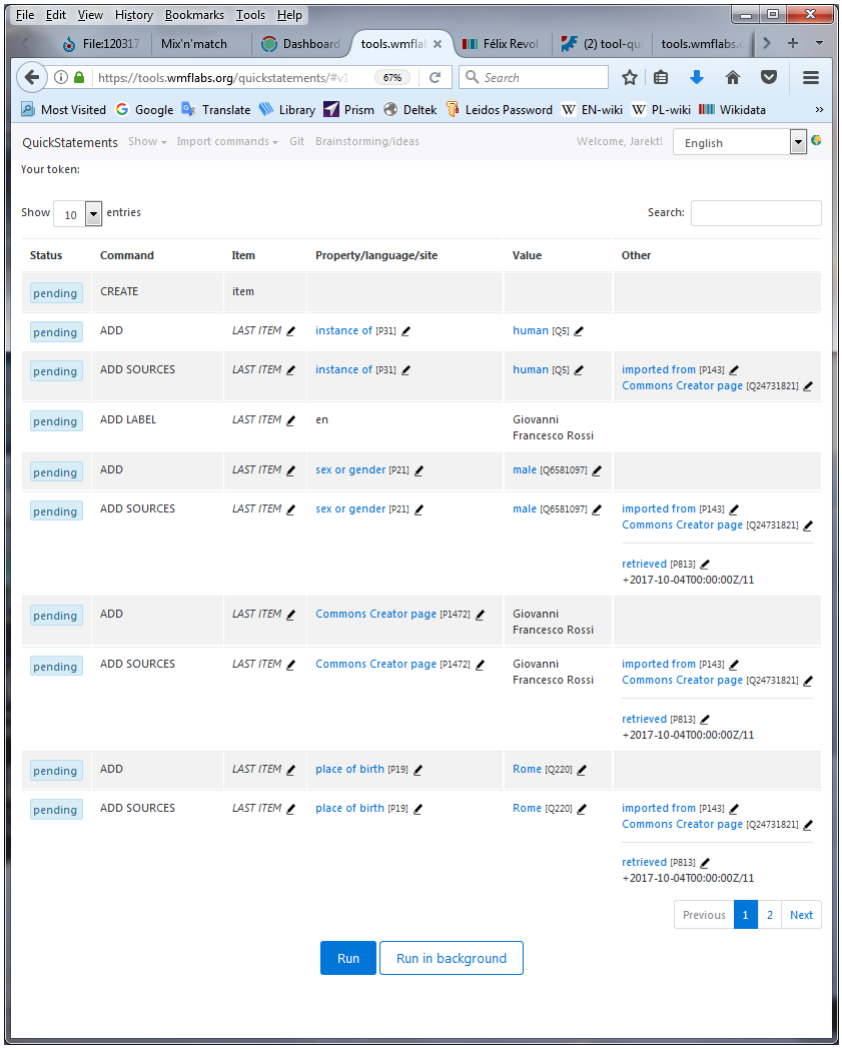


Figure 33: Example of a QuickStatements spreadsheet,
Source: https://www.wikidata.org/wiki/Help:QuickStatements

5) The last step is to run the import with QuickStatements[23] with the already converted file from Step 4.

With this method, we can import sorted datasets to Wikidata. This may seem like a lot of instructions and different tools, but the reason is mostly because Wikidata is a database that has its own structure, doesn't work with "files" and already has a lot of information entered. Naturally, one could integrate the data manually one by one, but that would be significantly slower than by Wikidata QuickStatements import.

[23]https://tools.wmflabs.org/quickstatements/

4.2 Additional resources to integrate missing data

One of the most challenging part of this research is to find data and integrate it on Wikidata. There are several ongoing studies regarding this issue, for instance on how to mine and extract the information from Wikipedia's tables [20]. Knowledge extraction is the formation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images) data [20]. Wikipedia consists of encyclopedic knowledge gathered by millions of users. These articles include factual data stored in yet again millions of tables [20]. The extraction of these tables depends on mining information from Wikipedia info-boxes which are attribute-value tables, located on the top right-hand side of the Wikipedia articles [20]. There are tools which can extract RDF (Resource Description Framework) from these 'info-boxes', such as DBpedia and YAGO2 [20].

A popular example for knowledge extraction is the transformation of Wikipedia's data into structured data and also the mapping to existing knowledge. Wikipedia articles are constructed of free text and also contain different types of structured information, such as infobox templates, images, geo-coordinates and links to external web pages and links across different language editions of Wikipedia [21]. By developing an information extraction framework, the DBpedia project focuses on how to integrate Wikipedia information into structured knowledge which converts Wikipedia content to RDF [21]. DBpedia is a source of information for about more than 1.95 million objects, including 80,000 persons, 70,000 places, 35,000 music albums and 12,000 films [21]. It consists of 657,000 links to images, 1,600,000 links to relevant external web pages, 180,000 external links into other RDF datasets and 207,000 Wikipedia categories [21]. Altogether, the DBpedia dataset consists of around 103 million RDF triples [21].

5 Related Work

There have been several researches and projects in the last few years regarding the structure and completeness of political data in Wikidata. In 2017, the WikiProject Heads of state and government[24] was looking for an answer for the following question: "What is the gender breakdown of heads of government across the world?"[25]. This project was particularly interested in finding out how the heads of governments were described in different countries. Since some of the data consisted of blanks and inconsistencies, it was difficult to use SPARQL to answer this question. To solve this problem, the project inspected which data problems were to be blamed for these missing information.

Another project in Wikidata is a more recent one; the WikiProject Every Politician aims to have complete and consistently structured data across the world on all elected representatives, from national to local level. WikiProject Every Politician wants to have complete and comprehensive coverage of the current members of at least every national legislature, along with ministers, in a consistent format. According to the project, the properties of political events and persons on Wikidata are not consistently added to the records. The goal is that the most common of these properties (in the case of politicians are position held, parliamentary term, start time, end time, parliamentary group, electoral district) have to be completed. Moreover, Wikidata editors from all over the world are welcome to enter valuable information about politicians to provide a better overview on the knowledge base. Furthermore, Wikidata also faces the challenge of integrating new data from existing data sources. In this thesis, we are not only focusing on completing missing data about politicians, but also on other aspects of political data, such as the representation of parties and elections in Wikidata, finding an appropriate data model for representing political data, as well as completeness of political data in general, and possible ways to find and integrate data to Wikidata.

[24] https://www.wikidata.org/wiki/Wikidata:WikiProject_Heads_of_state_and_government

[25] https://medium.com/mysociety-for-coders/help-us-find-the-offices-of-heads-of-governments-across-the-world-4558124bcd24

6 Conclusion

In this paper, we have described how different kind of data, and political data in particular, is represented in Wikidata. By analyzing the information with the SPARQL Query, we have been confronted with the challenges of this open knowledge base. As every knowledge base, Wikidata is not complete and several important data are still missing. To find out the relevant missing data, Wikidata provides the users with a built in extension called RECOIN, which is a good way to start if one were to integrate missing data in the knowledge base. In addition, since Wikidata is open for everyone to edit, there is a massive challenge to overcome the difficulty of inconsistency. Wikidata properties tend to be highly inconsistent in the category of political data. The knowledge base uses a wide variety of these properties, and in some cases the schema for the use of properties is not clear. In this work, we have identified the most common methods to describe elections, politicians and political parties by assembling example queries. To identify an emerging best practice for Wikidata to cover all relevant aspects of political data, the data quality and data completeness need to be guaranteed. Accordingly, we have investigated some emerging patterns which could represent political data in a consistent manner. By doing this, we have pointed out the most commonly used properties among political data, and spotted cases where there is no common schema for properties. These are properties like "winner" and "successful candidate" in elections or "follows"/"followed by" and "part of the series" which mean the same but are named and represented differently, "political alignment" and "political ideology" which have a similar name, but their interpretation differs slightly, and properties such as "member of" and "member of political party" which are almost identical properties and are used side by side in several articles. These properties are open for interpretation and are candidates for future work. Moreover, we have investigated existing sources for completing data in Wikidata, namely open data portals, websites of political data providers (such as Politico) and Wikipedia articles. Open data portals provide deep resources for political data, although there is still the challenge of locating and mapping this data to an existing property, or creating brand new properties to correctly represent the data. Those datasets, which are reliable and publicly available, can be imported to the knowledge base by following the importing guidelines which we discussed in this paper. With the contribution of these datasets, the availability of information could be significantly increased, and Wikidata may prevail as the primary resource for data.

References

[1] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

[2] Gerd Brewka. Artificial intelligence—a modern approach by stuart russell and peter norvig, prentice hall. series in artificial intelligence, englewood cliffs, nj. *The Knowledge Engineering Review*, 11(1):78–79, 1996.

[3] Claudio Gutierrez1 and Juan F. Sequeda. A brief history of knowledge graph's main ideas: A tutorial. 10 2019.

[4] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

[5] Douglas B Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[6] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.

[7] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: Lessons and challenges. *Queue*, 17(2):20, 2019.

[8] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In Michael Martin, Martí Cuquet, and Erwin Folmer, editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[9] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

[10] Piero Andrea Bonatti, Stefan Decker, Axel Polleres, and Valentina Presutti. Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[11] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM, 2014.

[12] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[13] 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.

[14] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. Semantics and complexity of sparql. In *International semantic web conference*, pages 30–43. Springer, 2006.

[15] Renzo Angles and Claudio Gutierrez. The expressive power of sparql. In *International Semantic Web Conference*, pages 114–129. Springer, 2008.

[16] Daniel Hernández, Aidan Hogan, Cristian Riveros, Carlos Rojas, and Enzo Zerega. Querying wikidata: Comparing sparql, relational and graph databases. In *International Semantic Web Conference*, pages 88–103. Springer, 2016.

[17] Tomás Sáez and Aidan Hogan. Automatically generating wikipedia infoboxes from wikidata. In *Companion Proceedings of the The Web Conference 2018*, pages 1823–1830. International World Wide Web Conferences Steering Committee, 2018.

[18] Albin Ahmeti, Simon Razniewski, and Axel Polleres. Assessing the completeness of entities in knowledge bases. In *European Semantic Web Conference*, pages 7–11. Springer, 2017.

[19] Vevake Balaraman, Simon Razniewski, and Werner Nutt. Recoin: Relative completeness in wikidata. In *Companion Proceedings of the The Web Conference 2018*, pages 1787–1792. International World Wide Web Conferences Steering Committee, 2018.

[20] Emir Muñoz, Aidan Hogan, and Alessandra Mileo. Using linked data to mine rdf from wikipedia's tables. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 533–542. ACM, 2014.

[21] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.