THE APPLICATION OF SEMANTIC WEB TECHNOLOGIES TO CONTENT ANALYSIS IN SOCIOLOGY

MASTER THESIS

TABEA TIETZ

Matrikelnummer: 749153

Faculty of Economics and Social Science
University of Potsdam

Erstgutachter: Alexander Knoth, M.A.
Zweitgutachter: Prof. Dr. rer. nat. Harald Sack

Potsdam, August 2018

ABSTRACT

In sociology, texts are understood as social phenomena and provide means to analyze social reality. Throughout the years, a broad range of techniques evolved to perform such analysis, qualitative and quantitative approaches as well as completely manual analyses and computer-assisted methods. The development of the World Wide Web and social media as well as technical developments like optical character recognition and automated speech recognition contributed to the enormous increase of text available for analysis. This also led sociologists to rely more on computer-assisted approaches for their text analysis and included statistical Natural Language Processing (NLP) techniques. A variety of techniques, tools and use cases developed, which lack an overall uniform way of standardizing these approaches. Furthermore, this problem is coupled with a lack of standards for reporting studies with regards to text analysis in sociology. Semantic Web and Linked Data provide a variety of standards to represent information and knowledge. Numerous applications make use of these standards, including possibilities to publish data and to perform Named Entity Linking, a specific branch of NLP.

This thesis attempts to discuss the question to which extend the standards and tools provided by the Semantic Web and Linked Data community may support computer-assisted text analysis in sociology. First, these said tools and standards will be briefly introduced and then applied to the use case of constitutional texts of the Netherlands from 1884 to 2016. It will be demonstrated how to generate RDF data from text and how to publish and access these data. Furthermore, it will be shown how to query the local data on its own as well as though the enrichment of existing data with external knowledge from DBpedia. A thorough discussion of the presented approaches will performed and intersections for a possible future engagement of sociologist in the Semantic Web community will be elaborated.

ZUSAMMENFASSUNG

In der Soziologie werden Texte als soziale Phänomene verstanden, die als Mittel zur Analyse von sozialer Wirklichkeit dienen können. Im Laufe der Jahre hat sich eine breite Palette von Techniken in der soziologischen Textanalyse entwickelt, du denen quantitative und qualitative Methoden, sowie vollständig manuelle und computergestützte Ansätze gehören. Die Entwicklung des World Wide Web und sozialer Medien, aber auch technische Entwicklungen wie maschinelle Schrift- und Spracherkennung tragen dazu bei, dass die Menge an verfügbaren und analysierbaren Texten enorm angestiegen ist. Dies führte in den letzten Jahren dazu, dass auch Soziologen auf mehr computergestützte Ansätze zur Textanalyse setzten, wie zum Beispiel statistische 'Natural Language Processing' (NLP) Techniken. Doch obwohl vielseitige Methoden und Technologien für die soziologische Textanalyse entwickelt wurden, fehlt es an einheitlichen Standards zur Analyse und Veröffentlichung textueller Daten. Dieses Problem führt auch dazu, dass die Transparenz von Analyseprozessen und Wiederverwendbarkeit von Forschungsdaten leidet. Das 'Semantic Web' und damit einhergehend 'Linked Data' bieten eine Reihe von Standards zur Darstellung und Organisation von Informationen und Wissen. Diese Standards werden von zahlreichen Anwendungen genutzt, darunter befinden sich auch Methoden zur Veröffentlichung von Daten und 'Named Entity Linking', eine spezielle Form von NLP.
Diese Arbeit versucht die Frage zu diskutieren, in welchem Umfang diese Standards und Tools aus der Semantic Web- und Linked Data- Community die computergestützte Textanalyse in der Soziologie unterstützen können. Die dafür notwendigen Technologien werden kurz vorgsetellt und danach auf einen Beispieldatensatz der aus Verfassungstexten der Niederlande von 1883 bis 2016 bestand angewendet. Dabei wird demonstriert wie aus den Dokumenten RDF Daten generiert und veröffentlicht werden können, und wie darauf zugegriffen werden kann. Es werden Abfragen erstellt die sich zunächst ausschließlich auf die lokalen Daten beziehen und daraufhin wird demonstriert wie dieses lokale Wissen durch Informationen aus externen Wissensbases angereichert werden kann. Die vorgestellten Ansätze werden im Detail diskutiert und es werden Schnittpunkte für ein mögliches Engagement der Soziologen im Semantic Web Bereich herausgearbeitet, die die vogestellten Analysen und Abfragemöglichkeiten in Zukunft erweitern können.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

1 INTRODUCTION

Writing was developed independently in Mesopotamia around 3100 B.C., in China around 1500 B.C. and in Mesoamerica around 300 B.C. (Silberman, 2012). Since then, writing served as a means of human to human communication and is firmly established in human cultures. Text as "written or printed words, typically forming a connected piece of work"[1] has been crucial in the human development. Texts have proven essential to human societies when forming a government or laying down fundamental laws. But also basic interactions which affect the day-to-day life of humans is highly influenced by the written text they produce, share and consume. The development of the Internet and later the Web increased the amount and diversity of text available to humans. The rise of technologies (e.g. optical character recognition (OCR) enable to digitize vast amounts of text as currently attempted by Google Books[2] and automated speech recognition (ASR) enables to extract spoken words in audiovisual material into text. An unthinkable expanse of information about anything humans do, think and know is captured in form of text through Emails, Blogs, social media platforms, chatforums, and so on. Alone on Twitter, one of the leading social networks worldwide with more than 300 million monthly active users, more than 58 million "tweets" are posted every day[3]. Ever since humans began to write and later to print textual documents, scientists of numerous research fields and professions started to analyze these writings to research the works' authors themselves, the context a work was written in (e.g. the time period) or the impact a written work had on a society.

In sociology, text grants a researcher access to social reality, it provides means to the realization of society or certain aspects of society (Lemke and Wiedemann, 2015). Already during the 18th century an analysis of religious symbols in songs was performed. This depicts the first well-documented account of the analysis of a quantitative analysis of printed material. Until the late 1950s, text analysis was mainly used to describe text, e.g. through a word frequency analysis. Also simple valence analyses were developed in which researchers attempted to determine whether specific words were more positively or negatively valued. Intensity analyses helped to grant certain words or phrases more weight than others to enable more precise analysis results (Popping, 2000) . During the 1960s researchers began using the computer for text analysis and especially the development of *The General Inquirer*, a mainframe program to classify and count words or phrases within certain categories created a milestone in social scientific research (Stone, Dunphy, and Smith, 1966). Everything that had previously been accomplished with pen and pencil e.g. data coding, could now be achieved with the computer. From that moment text files which have been available as data files on computer systems could

1 https://en.oxforddictionaries.com/definition/text, last visited: June 20, 2018
2 https://books.google.com/intl/en/googlebooks/about/index.html, last visited: June 20, 2018
3 https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, last visited: June 5, 2018

be entered into programs and automatically analyzed (Popping, 2000). During the last decade, statistical Natural Language Processing (NLP) approaches have been more and more used in social scientific text analysis and the algorithms have simultaneously become more accurate and efficient in a way that they supported to uncover linguistic structures as well as semantic associations (Evans and Aceves, 2016).

While the analysis of natural language text has become an important component of research in sociology, numerous interesting methods of computer assisted data acquisition and analysis have established. However, Mayring, (2015) criticizes that especially in social scientific research, no standardized and systematic means of the analysis of complex text material has emerged. Also according to Lemke and Wiedemann, (2015) it is still absolutely necessary to establish universal standards for a sustainable computer assisted text mining in sociology which will enable researchers to focus more on the actual research work and less on the development of new methods. Furthermore, in sociology, data sharing and publishing is to this day widely un-standardized and often not practiced at all (Herndon and O'Reilly, 2016; Zenk-Möltgen and Lepthien, 2014). According to Büthe et al., (2014), this lack of transparency lowers the integrity and interpretability of the performed research. Another widely discussed issue in sociology is the re-use of research data, especially qualitative data (Moore, 2007). A study by Curty, 2016, suggests that sociologists generally welcome re-using research data in sociology, but certain aspects which includes the difficulty of finding and accessing these data often prevents them to do so.

This thesis builds on the idea to develop standards for sociological text analysis based on Semantic Web and Linked Data technologies. While working at CERN, Tim Berners-Lee established the general idea of the World Wide Web (WWW) in 1989. While previously and without the Web, the Internet was mostly accessible to experts only, the Web was developed in a way which enabled everyone to create and consume content. Today, social media allows anyone to post text, audiovisual content and share locations with user friendly interfaces. However, the development of the Web also called for numerous standards and formats to be able to assure that new applications can be created by anyone in the existing framework of the traditional Web. During the early 2000s, the Semantic Web started to evolve which brought even more sophisticated standards, recommended by the World Wide Web Consortium (W3C)[4]. "The Semantic Web is an extension of the traditional Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee, Hendler, and Lassila, 2001). The Semantic Web provides "a common framework for the liberation of data" (Berners-Lee et al., 2006) by giving data an independent existence which is free from the constraints of the document in which they appear (Halford, Pope, and Weal, 2013). Several domains have already not only firmly established methods to utilize the possibilities provided by Semantic Web and Linked Data technologies and standards, they have also found ways to take part in the development, providing new applications based on the general idea. To these domains belong initiatives in the Life Sciences (e.g. cancer research by McCusker et al., (2017)), the media and

4 https://www.w3.org/, last visited: June 5, 2018

film industry (e.g. by the BBC (Kobilarov et al., 2009) or film production in general (Agt-Rickauer et al., 2016)) and many more (Schmachtenberg, Bizer, and Paulheim, 2014). However the field of sociology has so far not contributed immensely to the Semantic Web even though there are many points of intersection, especially in the field of text analysis. On this foundation, the following research question has been developed:

> **Research Question:**
> To which extend can state-of-the-art Semantic Web and Linked Data technologies, standards and principles support computer-assisted text analysis in sociology to improve research transparency and data re-usability.

This thesis attempts to show and discuss these intersection points. It will be elaborated how different technologies and standards part of the Semantic Web help to make the research process more transparent and reproducible, make data re-usable and to support text analysis with these technologies. Thereby it should be made clear, that the goal of this thesis is not at all to show how to replace but rather how to support the traditional NLP techniques as introduced by Evans and Aceves, (2016), and utilized by e.g. Knoth, Stede, and Hägert, (2018).

Halford, Pope, and Weal, (2013) furthermore discuss that sociologists intending to work with computer-assisted methods should not simply wait until the perfect method or perfect data suddenly appears. Using their domain knowledge, sociologists should engage in the way data can be represented on the Web and analyzed according to their needs. This thesis furthermore attempts to discuss intersection points for sociologists to engage in future work by emphasizing the imperfections of current tools and standards provided by the Semantic Web community.

The contributions of this thesis include:

1. A summarization of the foundations of Semantic Web technologies with focus on the target audience of sociologists and the field of text analysis in sociology
2. The integration of Semantic Web technologies in sociological text analysis on real world examples and topics. This includes ontological engineering, Named Entity Linking approaches, and data querying with SPARQL
3. The utilization and discussion of state-of-the-art tools used in Semantic Web research as well as the development of Python scripts for format conversion
4. A discussion of future interdisciplinary work to integrate sociological domain expertise in the field of Semantic Web

The thesis is structured as follows:

Chapter 2 will introduce the theoretical and methodological implications of text analysis in Sociology. Chapter 3 will very briefly summarize the technological foundations and basic principles of the Semantic Web and Linked Data needed for the

semantic analysis of natural language text. The following chapter 4 will demonstrate the utilization of these technologies and principles on real world research examples. In chapter 5, an in-depth discussion of the results and presented approaches will be provided. The final chapter summarizes and concludes this thesis.

2 THEORETICAL AND METHODOLOGICAL IMPLICATIONS

In this chapter, theoretical and methodological concepts important for computer-assisted text analysis in social science are briefly introduced.

2.1 SOCIAL REALITY IN THE CONTEXT OF TEXT ANALYSIS

In sociology, text grants a researcher access to social reality, it provides means to the access the reality of society or certain aspects of society. According to Luhmann, (1993), texts are social phenomena and social reality is constituted in communication, cf. (Lemke and Wiedemann, 2015, p 35). Therefore text analysis is a legitimate method to research social reality. This understanding of social reality is based on Berger and Luckmann, (1966). It describes that the social order in which the people in a society live cannot be merely understood as something that is necessary and a piece of objectively created history. Instead, this social order is more of a contingent process that is produced by the people themselves.

Two major disciplines which reflect the relationship between text and social reality are the *hermeneutic perspective* and the *perspective of the sociology of knowledge* (Lemke and Wiedemann, 2015, p 3). Both theoretical perspectives take a different look at the context in which texts emerge and the contexts of validity that shape the articulated meaning. The idea of such contextuality of text allows two directions of knowledge: (1) from context to the text it interprets, (2) from text to the context of which the traces are visible in the text. The consequence of the second direction is that text can be interpreted as a manifestation of a social being or as a context that determines them. Thus, text can in theory function as a medium for analyzing social reality, a task especially entrusted to the social sciences (Lemke and Wiedemann, 2015, pp 18-21). The hermeneutic perspective and the perspective of the sociology of knowledge have different approaches to analyze text in relation to social reality. As shortly discussed above, the hermeneutic perspective brings the context of the text to focus. As Lemke and Wiedemann, (2015) conclude, the information about a text and its respective context may be derived from the complete works of an author or debates of other authors on the same topic. A text can also be viewed in the context of contemporary events or cultural background. However, a text can only be viewed in its context, if it contains traces which enable to bring together text and context, e.g. mentions of certain names, locations or events known to the reader. Hence, the hermeneutic perspective of text analysis focuses on the text's context to analyze social reality (ibid. pp 21-23). The Sociology of knowledge on the other hand focuses on the relation between knowledge and social reality, to be analyzed in the text. The essence of both perspectives' differences of both is defined by their distance and proximity to an object of knowledge. While the hermeneutic perspective focuses on the a close proximity to the object of knowledge, the sociology of knowledge keeps distance to the object to enable a broader view of

the matter. These perspectives establish a proximity to the analysis methods of *distant reading* and *close reading* (ibid. pp 30). While both perspectives provide certain assets and drawbacks, Luhmann, (1984) warns that distance can only be kept, if researchers can rely on their utilized instruments. Lemke and Wiedemann, (2015) suggest that bringing together both perspectives in the modular analysis process entitled as *blended reading* would enable to deliver the best results for text analysis in sociology. The authors do not understand the hermeneutical approach and the approach of the sociology of knowledge as oppositional but assume that in blended reading both methods enable to generate synergetic effects if algorithms and humans work together in semi-automatic methods and optimally combine their respective competencies (ibid. pp 43 -54).

2.2 TEXT ANALYSIS IN SOCIOLOGY

According to Mayring, (2015, p 11-15), content analysis in social sciences is based on communication which may exist for example in form of text, music, or images. In order to analyze these forms of communication, it has to be fixed in some form of protocol. The analysis process itself should be performed systematically and guided by pre-defined rules and theories. Furthermore, the analysis of this fixed communication should be performed in relation to its context and should not be viewed as a single and independent piece of text. In this thesis, the content to be analyzed focuses on text corpora. It is acknowledged that other forms of communication exist. However, often, these formats are converted into text as well. For example, interviews are often transcribed into text as a pre-process of the analysis (Froschauer and Lueger, 2003) and modern automated analysis methods like OCR and ASR enable to convert large documents into text.

2.2.1 *Qualitative vs. Quantitative Analysis*

Text analysis can furthermore be categorized into qualitative and quantitative approaches. Mayring, (2015) discussed a variety of categories to differentiate between qualitative and quantitative research. The first category is merely a terminological distinction. According to the author, qualitative terms are used to divide objects into classes, e.g. house, car, street) while quantitative terms introduce numerical functions into language. In social scientific research, methods are often categorized according to their scale level. Mayring defines that any analysis that is based on a nominal scale is most likely a qualitative approach and analyses based on ordinal, interval or ratio scales belong to quantitative research methods. Of course, they may also overlap which makes a clear distinction more difficult. Another method of distinguishing between both methods in sociology is based on the implicit understanding of research. That means, qualitative research analyzes the complexity of a matter and intends to understand it, therefore it is rather inductive. Quantitative research on the other hand isolates an object of analysis into variables and defines the impact of interfering effects. It intends to explain things rather than understanding them and thus tends to be more deductive (ibid. p 17-21).

Mayring clarifies that the qualitative vs. quantitative battle often created by social scientists is unnecessary, because both methods can be used in the process of text analysis in synergy, a view which is shared and acknowledged in this thesis. The author explains that the first step in text analysis is always the definition of the research topic and the clarification what is analyzed. Then, either qualitative or quantitative or both methods may be used for the analysis process, depending on the use case and research question. In the last step, qualitative methods are used to interpret the observations made in the analysis (ibid. 20-22).

2.2.2 *Transparency and Re-usability in Text Analysis*

The research question presented in chapter 1 introduces the importance of transparency and data re-usability in sociological research. Büthe et al., (2014, p 2), refer to transparency in the research process as means to provide a "clear and reliable account of the sources and content of the ideas and information on which a scholar has drawn in conducting her research, as well as a clear and explicit account of how she has gone about the analysis to arrive at the inferences and conclusions presented - and supplying this account as part of (or directly linked to) any scholarly research publication." The authors furthermore emphasize transparency to be a corner stone for the integrity and interpretability of research.

The re-usability of research data in sociology is a widely discussed topic. Moore, (2007), pointed out that one of the major issues is the questions whether data can be re-used apart from the original context in which it was previously collected. While referring especially to qualitative data Heaton, (2004), discussed that data interpretation is believed to be dependent on the primary researcher's knowledge of the particular context of data collection. On the other hand, a study by Curty, 2016, suggests that in general, sociologists welcome re-using research data in sociology. However one of the aspects which often prevent re-use is the limited access. In this thesis, methods and standards will be discussed which enable re-using research data in Linked Data formats. The issues raised by Moore, (2007), and Heaton, (2004), are acknowledged. Furthermore it is acknowledged that re-usability may be challenging when rather restrictive licenses are applied to data, no licenses at all, or when data privacy is in danger. This thesis considers re-usability merely from the technical perspective and it will be assumed that it is the single researcher's responsibility whether the re-use of data in a given context is acceptable or not.

This chapter gave an overview of the considered theoretical and methodological concepts this thesis is built upon. It has been clarified that both, the theoretical concepts of the sociology of knowledge and the hermeneutical concepts can provide benefits in social scientific text analysis. Furthermore, it has been discussed that in the course of this thesis, text analysis is neither understood as a sole qualitative or quantitative approach, but a synergy of both. In the last section, the meaning of research transparency and data re-usability in text analysis has been emphasized. Chapter 3 will briefly introduce the foundations of Semantic Web and Linked Data technologies necessary for social scientific text analysis.

3

A BRIEF INTRODUCTION TO SEMANTIC WEB TECHNOLOGIES AND LINKED DATA

This chapter gives a brief overview of Semantic Web and Linked Data technologies. After short historical introduction of how the Semantic Web evolved in section 3.1, the technologies, principles and standards needed especially for the analysis of text as elaborated in chapter 4 are briefly introduced in section 3.2.

3.1 FROM THE INTERNET TO THE SEMANTIC WEB - A QUICK OVERVIEW

In this section, a brief chronological overview of the technological development from the Internet to the Web of Data will be given, mentioning the most notable figures and advances in the process.

3.1.1 *The Internet - Computer Centered Processing*

The development of the Internet already began during the 1960s. During a meeting of the Advanced Research Projects Agency (ARPA) research directors in 1967 the heads of the Information Processing Techniques Office Joseph Licklider and Lawrence Roberts first raised a discussion about connecting heterogeneous computer networks. As a result, so-called Interface Message Processors (IMP) were developed to connect proprietary computer systems to telephone networks. October 29, 1969 marked the birth of the so-called ARPANET. The first four connected nodes of this newly created network belonged to research departments of the Universities of Santa Barbara, Utah, Los Angeles, and Stanford. The extension of this network soon reached the west coast of the United States and by the early 1970s, 23 hosts were connected via 15 nodes. The first international nodes were connected in 1973 and starting from 1975, the network was not only connected via telephone cables but also via satellite. To the first famous applications of this network belonged an Email program developed by Ray Tomlinson in 1971. Another milestone was reached in 1983 when the communication software of all connected computer systems was adapted to the TCP/IP protocol under the leadership of Vinton Cerf and Robert Kahn. This marks the birth of the Internet (Meinel and Sack, 2011).

3.1.2 *The World Wide Web - Document Centered Processing*

Imagine the Internet without the comfortable Web applications we know today which enable us to access our social media feeds in our browser or on smartphone applications or send large files around the world using user friendly applications. In order to access and use information on the Internet, users were required to connect to a remote system (e.g. using a terminal), retrieve the file system data on said remote system, download the file and read it on a local system. All of

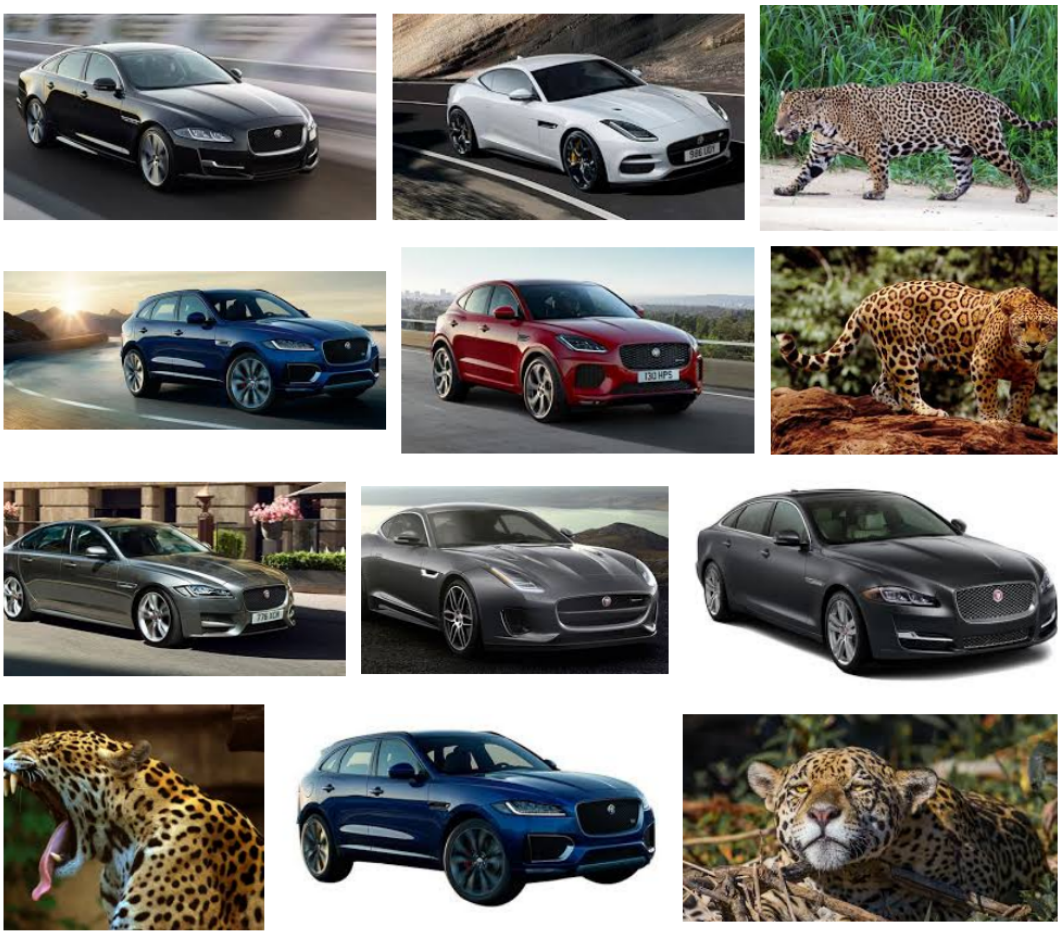

Figure 1: Google image search using the keyword "Jaguar".
Retrieved on June 6, 2018

these steps are accomplished with several command lines. As revolutionary as this method was during the early age of the Internet, simply accessing information via the Internet required high expert knowledge.

While working at CERN, Tim-Berners Lee published "*Information Management: A Proposal*" (1989). In it, the author proposed a decentralized hypertext based document management system with the purpose to administrate the enormous amount of research data and documentations of CERN. Together with Robert Cailliau, Berners-Lee began working on his initial idea using the NeXT computer system. In November 1990, the term WorldWideWeb was coined by Tim Berners-Lee and in 1991 the first Web browser was released. The foundation of the World Wide Web (WWW) is the interlinking of documents via *hyperlinks*. A hyperlink is defined as an explicit reference of one document to another document or within the same document. Text based documents on the Web are referred to as *hypertext* documents (Meinel and Sack, 2011).

3.1.3 *The Semantic Web - Data Centered Processing*

The traditional Web is a document based decentralized network and its numerous applications make it possible for everyone to access information and to publish information on the Web and hence to participate in it's content and development without expert knowledge.

However, accessing data on the Web (especially in the context of scientific research) is often difficult due to the variety of standards and formats used. Documents on the Web can be encoded as HTML (Hypertext Markup Language), PDF (Portable Document Format) or proprietary document formats, e.g. Microsoft Word or Excel. Data in these documents are often unstructured and embedded in text or semi-structured in tables. To make use of these data, they have to be semi-automatically extracted which is not only labor and time intense but also error prone (Pellegrini, Sack, and Auer, 2014). Next to these formats, XML evolved as a prominent standard on the Web to encode data syntactically by creating a tree of nested sets of tags. However, proprietary style sheets and parsers are needed to make use of these information efficiently (this will be discussed in more detail in section 4.1.1.1)

In the traditional Web, HTML is the standard markup language to create Web pages and applications. It describes how information is presented and how information is linked (Faulkner et al., 2017). However, HTML cannot describe what the information actually *means*. This becomes clear when initiating a Google image search using the keyword *Jaguar*. As Figure 1 shows, the search engine returns images of the animal as well of the car *Jaguar*. The reason is that natural language is often ambiguous and contains words or phrases with the same spelling but different meanings (e.g. *Jaguar*) as well as words or phrases with a different spelling and the same meaning (e.g. *important, substantial, essential*). The former is defined as a homonym and the latter is defined as a synonym.

But why does it seem to be important to also include the *meaning* of words and phrases into Web applications and to disambiguate natural language? In communication, the *meaning* is necessary to *understand* information which is conveyed in a message using a specific language. Information is understood by the receiver of a message if the receiver *interprets* the information correctly. Hence, if a machine is programmed to not only read data but to also understand it, human-computer as well as computer-computer communication can be significantly improved. The example above shows that without any further given context neither humans nor a computer program can correctly and unambiguously interpret the meaning of the keyword *Jaguar*. This mis-interpretation causes communication problems as the results the computer returns here may differ to the results the user had expected.

The Semantic Web with its underlying technologies enables the development of machine understandable data. In it, the meaning (semantics) is made explicit by formal (structured) and standardized knowledge representations (ontologies). The Semantic Web makes it possible to automatically process the meaning of information, relate and integrate heterogeneous data and deduce implicit information from existing information. The example above helps to understand what the effect of programming a computer to interpret the meaning of information may look

Figure 2: Google image search using the query "Jaguar with two doors". Retrieved on June 6, 2018

like. If the search query is changed from simply "*Jaguar*" to "*Jaguar with two doors*" Semantic Web technologies enable to automatically take into account the context of the given query. *Context* denotes the surrounding of a symbol (concept) in an expression with respect to its relationship with surrounding expressions (concepts) and further related elements. If a human reads a query like "*Jaguar with two doors*" it is immediately clear that the term *Jaguar* it is not about the large cat since an animal does obviously not have two doors. Taking into account the given context (two doors) it becomes immediately clear to humans that *Jaguar* can only be a type of car. Semantic Web technologies enable to make these differentiations as well, as shown in Figure 2. They enable to structure information in a way to make clear that a Jaguar can be a type of car with a specific amount of doors, an engine and four wheels or a type of wild cat species.

In contrast to the traditional Web where documents are interconnected to organize semi-structured information, the Semantic Web enables to structure data, give data a well defined meaning and derive completely new implicit knowledge from explicit knowledge via logical reasoning. With the development of the Semantic Web, a variety of standards, methods and practices have been created to structure information (data) to be machine interpretable.

The following section will give a brief overview of the underlying technological foundations of the Semantic Web, before its potential for social scientific text analysis can be discussed thoroughly.

3.2 BASIC PRINCIPLES OF THE SEMANTIC WEB AND LINKED DATA

The Semantic Web is closely related to the traditional Web of documents. Tim Berners-Lee, who is credited with the invention of the WWW, also coined the term *Semantic Web*. According to Berners-Lee, the "Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee, Hendler, and Lassila, 2001).

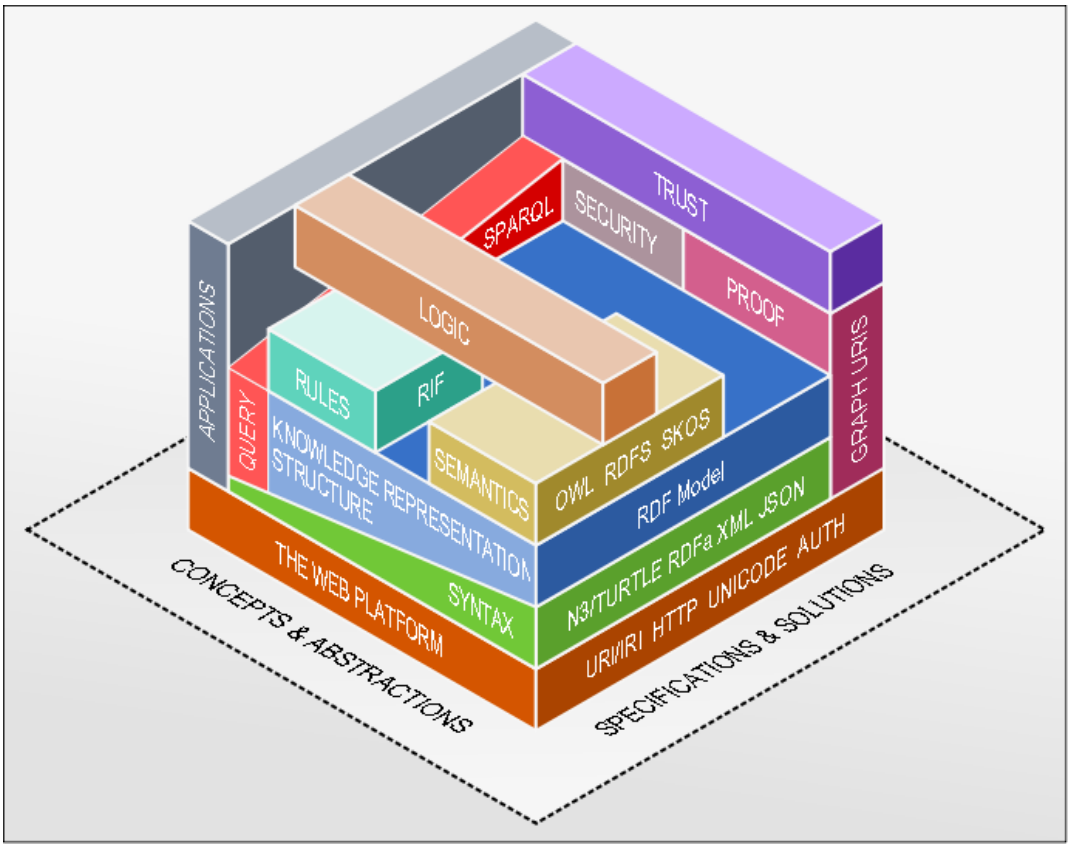

Figure 3: Semantic Web Technology Stack

According to Hitzler, Kroetzsch, and Rudolph, (2009) the Semantic Web shares a number of goals with the traditional Web:

- Make knowledge widely accessible
- Increase the utility of this knowledge by enabling advanced applications for searching, browsing, and evaluation

The Semantic Web further "allows computers to intelligently search, combine, and process Web content based on the meaning that this content has to humans". However, since artificial intelligence on a human level is (not yet) possible, the mentioned intelligence can only be achieved if the meaning (semantics) "of Web resources is explicitly specified in a format that is processable by computers" (Hitzler, Kroetzsch, and Rudolph, 2009). In order to do so, storing data in a machine readable way (e.g. by means of HTML) is not sufficient. To create machine understandable content, it is necessary to make the meaning of information explicit with the help of specified models and standards.

3.2.1 *The Semantic Web Technology Stack*

The basic technologies used in Semantic Web applications are visualized in the so-called Semantic Web technology stack as shown in Figure 3. It gives an overview of the used standardized concepts and abstracts (left side) as well as specifications

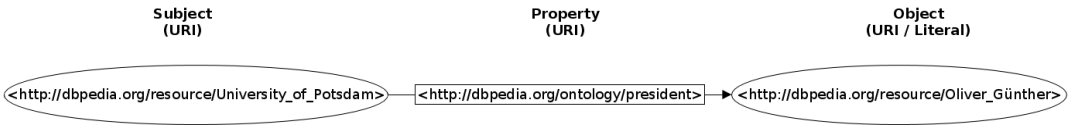

Figure 4: A simple RDF graph

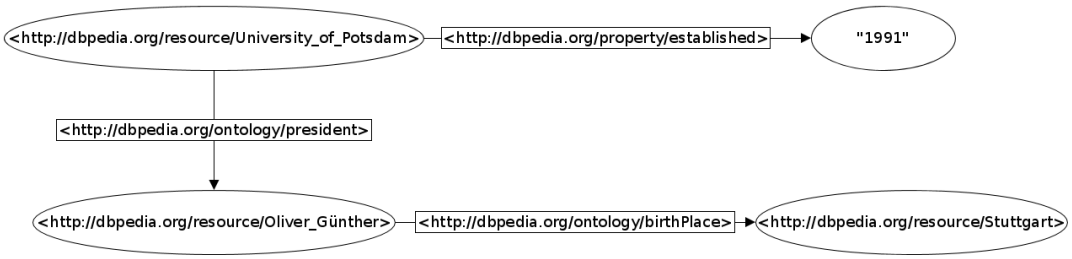

Figure 5: An RDF graph extended by two triples

and solutions (right side). While standards have always played an important role in the Web of Documents, the development of the Semantic Web increased the importance of standardizations even more. Most standardizations in this area have been conducted under the lead of the World Wide Web Consortium (W3C)[1] (Hitzler, Kroetzsch, and Rudolph, 2009). In the course of this section, the stack will be used to visualize the position of the explained technologies and concepts to give a better general overview.

3.2.2 *Uniform Resource Identifiers (URIs)*

The Uniform Resource Identifier (URI) is part of the *Web Platform* layer in the Semantic Web Technology stack (cf. Figure 3). A URI "defines a simple and extensible schema for worldwide unique identification of abstract or physical resources" (Berners-Lee, Fielding, and Masinter, 2005). A resource in that sense can be any object with a clear identity, e.g. a Web page (URL) or a Book (ISBN). In the Semantic Web, URIs are used to uniquely distinguish resources from each other.

3.2.3 *The Resource Description Framework*

In the Web of data, information is encoded in the so-called Resource Description Framework (RDF)[2]. RDF is part of the *Information Exchange* layer in the Semantic Web Technology stack and depicts one of its main building blocks (cf. Figure 3). RDF is a framework to express information about resources. Resources can by anything including documents, objects, concepts or people. RDF is used when information on the Web has to be processed by applications rather than being displayed to humans. The framework can be used to publish and interlink data on the Web.

RDF enables to make statements about resources and one RDF statement expresses a relationship between two resources. One statement (or fact) is repre-

1 https://www.w3.org/, visited: June 11, 2018
2 https://www.w3.org/2001/sw/wiki/RDF, visited: June 11, 2018

Listing 1: RDF N-Triples Serialization

```
<http://dbpedia.org/resource/University_of_Potsdam>
     <http://dbpedia.org/ontology/president>
          <http://dbpedia.org/resource/Oliver_Günther> .
<http://dbpedia.org/resource/University_of_Potsdam>
     <http://dbpedia.org/property/established>
          "1991" .
<http://dbpedia.org/resource/Oliver_Günther>
     <http://dbpedia.org/ontology/birthPlace>
          <http://dbpedia.org/resource/Stuttgart> .
```

sented by a so-called triple consisting of a *subject*, a *property* and an *object*. Thereby, the subject and object are the resources, which are being related and the property defines the nature of this relationship. As shown in Figure 4, an RDF document describes a directed graph. That means a set of nodes is linked by directed edges (arrows) (Schreiber and Raimond, 2014). In the graph depicted in Figure 4 the subject is represented by the URI `<http://dbpedia.org/resource/University_of_Potsdam>`, the property is represented by `<http://dbpedia.org/ontology/president>`, and the object by `<http://dbpedia.org/resource/Oliver_Günther>`. The graph expresses the natural language statement *The president of the University of Potsdam is Oliver Günther*. In an RDF graph, the subject and property use URIs as names, the object uses either a URI or a literal (as shown in Figure 5). Literals enable to name abstract resources, which cannot be represented by a computer. A major advantage of RDF (e.g. in contrast to XML) is that RDF data from multiple sources can be combined and more facts can be added to a graph easily. Figure 5 demonstrates how new facts can be added to the previous graph. In this case the facts that *Oliver Günther was born in Stuttgart* and that *The University of Potsdam was established in 1991* were added, hence the graph now consists of three triples.

According to the W3C, "RDF is designed to represent information in a minimally constraining, flexible way. It can be used in isolated applications, where individually designed formats might be more direct and easily understood, but RDF's generality offers greater value from sharing. The value of information thus increases as it becomes accessible to more applications across the entire Internet" (Schreiber and Raimond, 2014).

3.2.3.1 *RDF Turtle Serializations*

As depicted in Figure 4 and 5, RDF graphs can be easily represented by means of diagrams. While this graphical representation makes it often easier to comprehend by humans, they are not suitable for processing RDF efficiently in computer systems (Hitzler, Kroetzsch, and Rudolph, 2009). There are several ways to represent RDF by means of character strings, which requires to split a graph into several smaller parts to be stored one by one. This process is called *serialization*. N-Triples is a line-based, plain text format for encoding an RDF graph (Beckett,

Listing 2: RDF Turtle Serialization with Prefixes

```
1 @prefix dbr:  <http://dbpedia.org/resource/> .
2 @prefix dbp:  <http://dbpedia.org/property/> .
3 @prefix dbo:  <http://dbpedia.org/ontology/> .
4
5 dbr:University_of_Potsdam dbp:president dbr:Oliver_Günther ;
6      dbp:established "1991" .
7 dbr:Oliver_Günther dbo:birthPlace dbr:Stuttgart .
```

2014). Listing 1 depicts the N-Triples serialization of the graph shown in Figure 5. The syntax directly translates from the graph visualization into triples. URIs are written in angular brackets while literals are written in quotation marks. Each triple is terminated by a full stop. Due to the lengthy names, triples in Listing 1 are spread over several lines. Turtle offers a mechanism to abbreviate URIs using so-called *namespaces* by means of defining *prefixes* as depicted in lines 1 - 3 in Listing 2. The prefix text can be chosen freely by the user, but it is recommended that abbreviations are selected which are easy to read and refer to what they abbreviate. Turtle furthermore provides the possibility to shortcut triples with the same subject or with the same subject and property (Beckett et al., 2014). The lines 1 and 4 in Listing 1 show that `dbr:University_of_Potsdam` is the subject in both triples. In Listing 2 a semicolon was used to indicate that line 6 uses the same subject as line 5.

There are several further ways to serialize RDF triples, including JSON-LD[3] and RDF/XML[4] but for sake of simplicity, they will not be discussed in this chapter.

3.2.4 *Linked (Open) Data*

The term *Linked Data* refers to best practices for publishing data on the Web. These practices or principles have been coined by Tim Berners-Lee in (2006):

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

These principles emphasize that when representing and accessing data on the Web, standards should be used to enable extensibility, reuse and sharing of data as well as interoperability between data sources. The term *Linked Open Data* refers to public Linked Data resources on the Web which are licensed as Creative Commons CC-BY[5]. Tim Berners-Lee created a five star criteria system for Linked Open Data[6]:

3 https://www.w3.org/TR/json-ld/, visited: June 11, 2018
4 https://www.w3.org/TR/rdf-syntax-grammar/, visited: June 11, 2018
5 https://creativecommons.org/licenses/by/2.0/, visited: June 12, 2018
6 http://5stardata.info/en/, visited: June 12, 2018

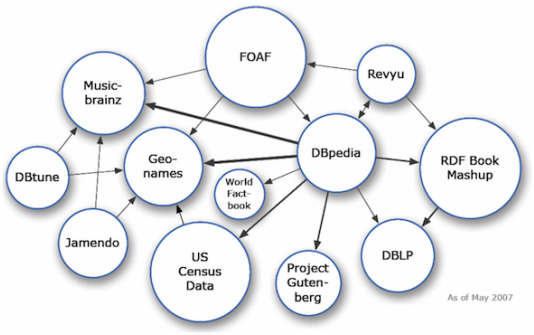

Figure 6: Linked Open Data cloud, version: May 2007

⋆ Available on the Web (whatever format) under an open license

⋆⋆ Available as structured data (e.g., Excel instead of image scan of a table)

⋆⋆⋆ Available in a non-proprietary open format (e.g., CSV instead of Excel)

⋆⋆⋆⋆ Use URIs to denote things, so that people can point at your stuff

⋆⋆⋆⋆⋆ Link your data to other data to provide context

Numerous datasets are available on the Web which fulfill (at least some) of these criteria. The Linked Open Data (LOD) cloud diagram gives an overview of the Linked Datasets that are available on the Web. Not every dataset qualifies for the diagram[7]. For instance a dataset must contain at least 1000 triples to be represented in the LOD cloud. Figure 6 shows the Linked Open Data cloud from May 2007 which contained only 12 datasets which are interlinked with each other. Since then, the amount of Linked Open Data on the Web has increased tremendously and the current diagram shown in Figure 7 contains 1186 datasets. The different colors in the diagram represent specific domains, including life sciences, geography, government or media. To name a few examples, part of the media domain in the LOD cloud is the BBC Music[8] dataset with around 20.000 triples. Part of the geography domain is the LinkedGeoData[9] knowledge base which contains information collected by OpenStreetMap[10]. The LinkedGeoData dataset contains around 3 billion triples and is categorized as five-star Linked Data according to the criteria listed above. As shown in the Figures 6 and 7 many datasets or knowledge bases are interlinked with each other. The knowledge base with the most diverse connections is DBpedia[11]. DBpedia is often referred to as the semantic version of the Wikipedia[12]. DBpedia is similarly to Wikipedia a community effort. Semi-structured information from Wikipedia (e.g. from Wikipedia infoboxes) are extracted and made available on the Web in form of RDF triples (Lehmann et al., 2015). DBpedia currently contains around 9.5 billion triples[13].

7 http://lod-cloud.net/, last visited: June 11, 2018
8 https://lod-cloud.net/dataset/bbc-music, last visited: June 11, 2018
9 https://lod-cloud.net/dataset/linkedgeodata, visited: June 11, 2018
10 https://www.openstreetmap.org, last visited: June 11, 2018
11 https://lod-cloud.net/dataset/dbpedia, last visited: June 11, 2018
12 https://www.wikipedia.org/, last visited: June 11, 2018
13 https://lod-cloud.net/dataset/dbpedia, last visited: June 11, 2018

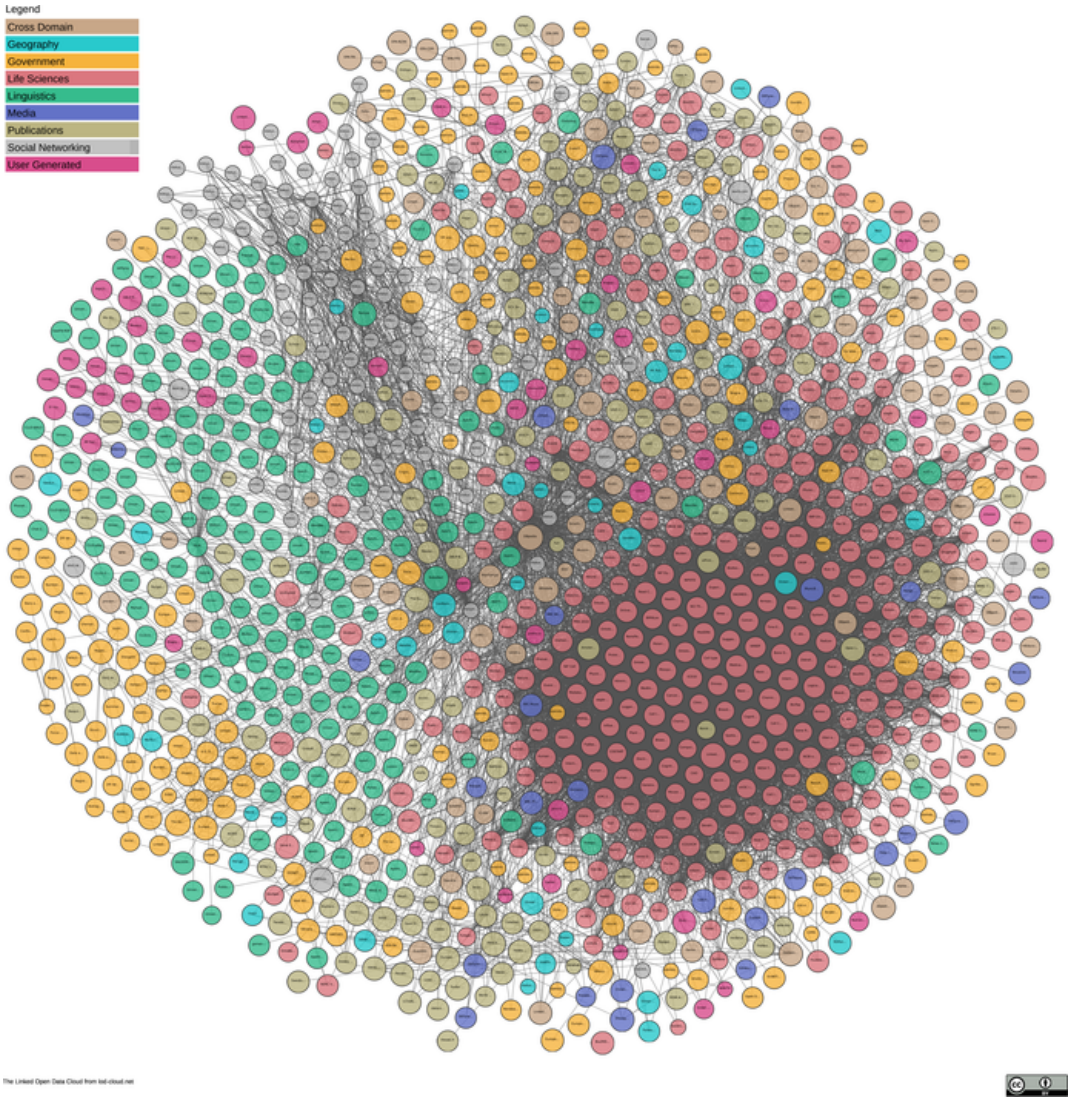

Figure 7: Linked Open Data cloud, version: May 2018

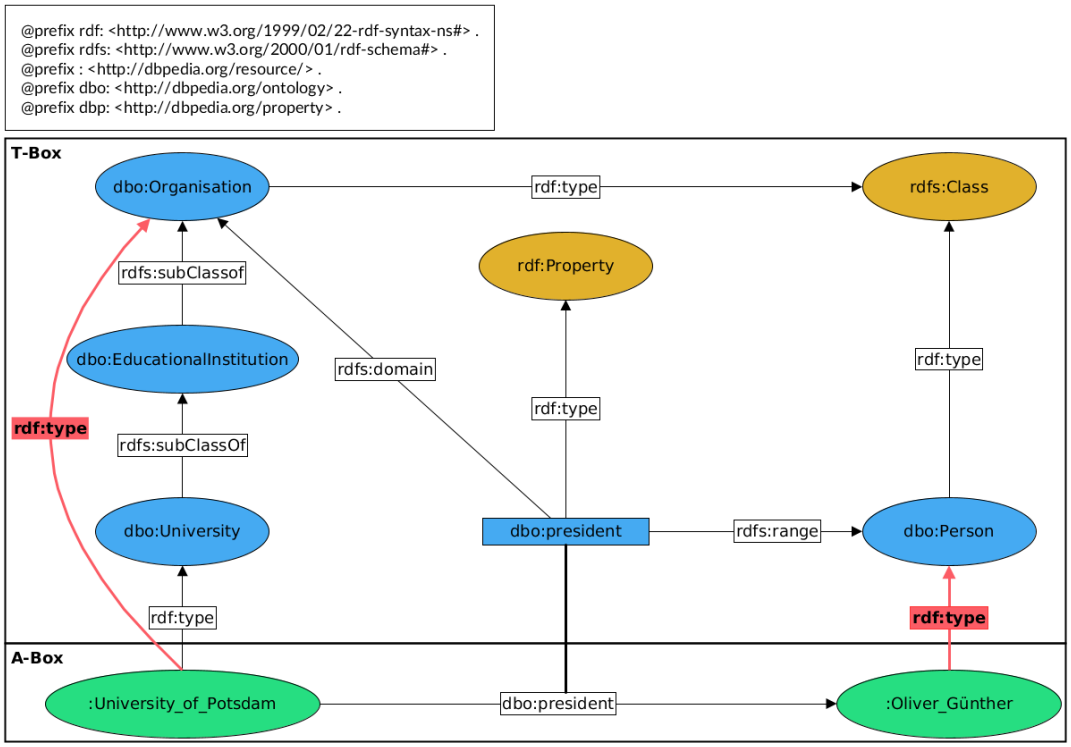

Figure 8: Visualization of terminological knowledge (T-Box) and assertional knowledge (A-Box)

Listing 3: RDFS Limitations

```
1 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2 @prefix ex: <http://example.org/example/> .
3
4 dbr:University_of_Potsdam ex:student ex:Adrian_Schmidt .
5 ex:Adrian_Schmidt rdf:type ex:FulltimeStudent .
6 ex:Adrian_Schmidt rdf:type ex:ParttimeStudent .
```

3.2.5 *RDF Schema and OWL*

In Section 3.1.3 it was stated that Semantic Web technologies enable to embed meaning in data. However, in this section so far, instances (e.g. `dbr:University_of_Potsdam` or `dbr:Oliver_Günther`) and properties which are used to relate instances to each other were merely specified. But where does the meaning introduced in Section 3.1.3 come from? One way to introduce semantics into the provided RDF data is by means of *RDF Schema* (or RDFS) which is part of the *Models* layer in the Semantic Web Technology Stack (cf. Figure 3). RDF Schema allows to specify terminological knowledge, i.e. to express information about the data structure. In order to explain the possibilities of RDF Schema, the concept of classes in RDF has to be defined: Resources may be divided into groups called classes. The members of a class are known as instances of the class. Classes are themselves resources. They are often identified by IRIs and may be described using RDF prop-

Listing 4: Example of OWL definitions in Turtle

```
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix dbr: <http://dbpedia.org/resource/> .
@prefix ex: <http://example.org/example/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

##    Classes
dbo:Person rdf:type owl:Class .
dbo:President rdf:type owl:Class ;
              rdfs:subClassOf dbo:Person .
dbo:University rdf:type owl:Class  .
ex:Fulltime rdf:type owl:Class ;
            rdfs:subClassOf ex:Student ;
            owl:disjointWith ex:Parttime .
ex:Parttime rdf:type owl:Class ;
            rdfs:subClassOf ex:Student .
ex:Student rdf:type owl:Class ;
           rdfs:subClassOf dbo:Person .

##    Properties
dbo:president rdf:type owl:ObjectProperty ,
                       owl:FunctionalProperty ;
              rdfs:domain dbo:University ;
              rdfs:range dbo:President .
ex:student rdf:type owl:ObjectProperty ;
           rdfs:domain dbo:University ;
           rdfs:range ex:Student .

##    Individuals
dbr:Oliver_Günther rdf:type owl:NamedIndividual ,
                            dbo:President .
dbr:University_of_Potsdam rdf:type owl:NamedIndividual ,
                                   dbo:University .
ex:Adrian_Schmidt rdf:type owl:NamedIndividual ,
                           ex:Fulltime .
```

erties. The `rdf:type` property may be used to state that a resource is an instance of a class. The group of resources that are RDF Schema classes is itself a class called `rdfs:Class` (Brickley and Guha, 2014).

RDFS allows to mark a resource as instances of a class, it allows to define hierarchical relationships between classes and between properties, and it allows simple logical inferences. Figure 8 visualizes a possible model based on the examples above. The green nodes in the A-Box depict assertional knowledge with instances, as well as their directed relation to other instances and classes. The blue nodes in the T-Box depict the terminological knowledge with classes, properties and their (hierarchical) relationships. Here, the schema or model in the T-Box defines the structure of the instances in the A-Box. In the visualized example, `dbr:University_of_Potsdam` is of type (`rdf:type`) `dbo:University`, which itself is a `rdfs:subClassOf` `dbo:EducationalInstitution` and so on. The orange nodes in

the T-Box depict whether a resource is defined as a class or a property. In the example, `dbo:Person` is defined as a class and `dbo:president` is defined as a property. Furthermore, constraints on the use of properties and classes in RDF can be defined. The *range constraint* specifies that the values of a property are instances of one or more classes, it is expressed by `rdfs:range`. The *domain constraint* states that any resource that has a given property is an instance of one or more classes and is expressed by `rdfs:domain` (Brickley and Guha, 2014). The visualization in Figure 8 specifies domain and range constraints for the property `dbo:president`. It is defined that the domain of `dbo:president` can only be an instance that belongs to the class `dbo:Organisation` (e.g. `:University_of_Potsdam`) and its range can only be an instance that belongs to the class `dbo:Person` (e.g. `:Oliver_Günther`).

From the human perspective it seems completely obvious that a university is an educational institution and that the president of a university can only be a person. However, for a machine this is far from obvious and without a clear formal definition a machine will never be able to express the hierarchical structures of organizations, educational institutions and universities. That means semantics can only be embedded into data if a formal definition based on logic has been specified. Again, from the human perspective it may seem rational that if the University of Potsdam is a university and a university is a type of educational institution, then the University of Potsdam is also an educational institution. For a machine, this is not completely obvious per se. However, since these explicit formal structures in the T-Box were defined the machine is able to derive implicit knowledge using RDF entailment patterns. For example, it can be automatically derived that if `:University_of_Potsdam` is of `rdf:type dbo:University`, which itself is a `rdfs:subClassOf dbo:EducationalInstitution` and a `rdfs:subClassOf dbo:Organisation`, then `:University_of_Potsdam` is of (`rdf:type`) `dbo:Organisation` as well. Furthermore, it was formally defined that the `rdfs:range` of `dbo:president` is `dbo:Person`. Therefore it can be deduced that the instance `:Oliver_Günther` is of `rdf:type dbo:Person`. If more universities were added to the graph which are connected to the property `dbo:president`, it can be automatically deduced that any instance of that property belongs to the class `dbo:Person`.

Taking into account the previous example visualized in Figure 8 one may want to extend the graph further to model existing university structures. Next to the president Oliver Günther, the University of Potsdam also has students. Some of them are enrolled as full-time students, some of them are enrolled as part-time students. RDFS allows to model these structures as well but it is not possible to state that a student can only be a full-time student OR a part-time student, but not both at the same time. Listing 3 shows that the `dbr:University_of_Potsdam` has an arbitrary student `ex:Adrian_Schmidt`. The instance `ex:Adrian_Schmidt` is modeled to be a full-time student and a part-time student. For humans, this immediately causes a logical contradiction, but a machine does not interpret these classes according to their intended semantics because it can not be specified using RDFS. That means, RDFS does not enable to specify that two classes, e.g. `ex:FulltimeStudent` and `ex:ParttimeStudent` must not contain any common instances. Further, RDFS does not enable to model that the University of Potsdam only has exactly one presi-

dent and not two or three or none (Hitzler, Kroetzsch, and Rudolph, 2009). These examples briefly demonstrate that RDFS lacks semantic expressivity.

The *Web Ontology Language (OWL)* enables to make these differentiations. OWL is an ontology language for the Semantic Web with formally defined meaning and it is based on Description Logic (Welty and McGuinness, 2004). Several different OWL flavors exists. For sake of simplicity, in the following the concept of OWL 2 will be considered. Similarly to RDFS there is a Turtle syntax for OWL. OWL classes are comparable to RDFS classes, individuals can be compared to class instances in RDFS and OWL properties are also comparable to RDFS properties. OWL classes, properties and individuals can be defined in the following way presented in Listing 4. In it, a number of classes are defined in lines 9 - 19. The classes `:Fulltime` and `:Parttime` are both subclasses of `:Student`. However, the `owl:disjointWith` statement in line 15 expresses that any individual (e.g. `:Adrian_Schmidt`) can only be in the class `:Fulltime` or `:Parttime` but never both. Properties are defined in lines 22 - 29. The property `ex:president` is defined as a `owl:FunctionalProperty`. This means that a university can have only one president. If more than one president was added this would cause a logical contradiction. The possibilities of modeling ontologies with OWL and to infer new knowledge are enormous, but for sake of simplicity only a few examples were explained in this section which briefly explore the functionalities of OWL in contrast to RDFS.

3.2.6 *Ontologies in Computer Science*

As discussed in the previous section, it is possible to organize knowledge using RDFS and OWL in a vocabulary and to define how concepts are related to each other. As the examples have further shown, the way these vocabularies are created also allows to easily reuse other peoples vocabularies and also share own vocabularies with the community. When a vocabulary is shared and reused, it is widely referred to as an ontology.

> **Definition:** An ontology is an *explicit*, *formal* specification of a *shared conceptualization* and defines the terms used to describe and represent an area of knowledge. (Gruber, 1993)

In this definition, *conceptualization* refers to the existence of an abstract model about a domain in which concepts and relations between concepts are identified. *Explicit* means that all concepts in the ontologies must be defined. *Formal* denotes that the concepts are expressed in a machine understandable way and *shared* means that there is a consensus about the conceptualization. According to Noy, McGuinness, et al., (2001), the reasons to develop an ontology include:

- Sharing a common understanding of the structure of information among people or agents
- To support the reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from operational knowledge

Listing 5: RDF graph

```
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

:University_of_Potsdam dbo:president dbr:Oliver_Günther ;
     dbo:city dbr:Potsdam ;
     rdf:type dbo:EducationalInstitution ;
     dbo:almaMater dbr:Katherina_Reiche ,
          dbr:Jens_Eisert .
dbr:Humboldt_University_of_Berlin dbo:city :Berlin ;
     dbo:almaMater dbr:Karl_Liebknecht ,
          dbr:Rudolf_Virchow .
dbr:Rudolf_Virchow dbo:knownFor dbr:Cell_theory .
```

Listing 6: SPARQL query based on the RDF graph in Listing 5

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?university ?alumni
WHERE {
     ?university dbo:almaMater ?alumni .
     }
ORDER BY ?university
```

- To analyze domain knowledge

A popular ontology widely used is the FOAF ontology[14] in which people-related terms that can be used in structured data. Another example is the GoodRelations ontology for exchanging information about products, sales, prices and so on for the e-commerce domain (Hepp, 2008). Noy, McGuinness, et al., (2001), have created a detailed and comprehensive guide on how to develop a sophisticated ontology.

3.2.7 *Data Querying with SPARQL*

In the previous sections, RDF was introduced which enables to store data in the form of triples in knowledge bases. In order to utilize the information stored in these knowledge bases, the data has to be queried. Since RDF is stored in a triple format, a query language has to be used that is able to process these patterns. *SPARQL*, short for *SPARQL Protocol and RDF Query Language* allows to query RDF data and thus is part of the *Query* layer of the Semantic Web technology stack, cf. Figure 3. SPARQL is based on the RDF Turtle serialization as well as basic graph pattern matching, i.e. it contains variables at any arbitrary place.

The basic functionalities of SPARQL will be explained on the foundation of the RDF graph displayed in Listing 5, which contains nine triples. In order to make use of the information stored in the graph, a SPARQL query has to be formulated. To find all entities which are connected to another entity with the property `dbo:almaMater`, the query shown in Listing 6 is issued. In line 1, the name space used

14 http://www.foaf-project.org/, last visited: July 3, 2018

university	alumni
<http://dbpedia.org/resource/Humboldt_University_of_Berlin>	<http://dbpedia.org/resource/Karl_Liebknecht>
<http://dbpedia.org/resource/Humboldt_University_of_Berlin>	<http://dbpedia.org/resource/Rudolph_Virchow>
<http://dbpedia.org/resource/University_of_Potsdam>	<http://dbpedia.org/resource/Jens_Eisert>
<http://dbpedia.org/resource/University_of_Potsdam>	<http://dbpedia.org/resource/Katharina_Reiche>

Table 1: Result of the SPARQL query in Listing 6

in the query is defined which is similar to the prefix definition in Listing 5. The SELECT clause in line 2 specifies the output variables `?university` and `?alumni`. In the WHERE clause, a graph pattern is listed. In the example above, the query asks for any `?university` and `?alumni` in the graph which are connected by the `dbo:almaMater` property. The ORDER BY statement in line 5 orders the results by the variable `?university`. The results of the query above are listed in Table 1.

3.2.8 *Metadata and Semantic Annotation*

On the Web, resources are described by metadata, i.e. information about data. Metadata are defined as "structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities" (*Description and Access, Task Force on Metadata* 2000) . *Semantic metadata* are part of the foundations of Semantic Web technologies. The semantics in these metadata are explicitly and formally defined via ontologies and therefore they are machine understandable. Semantic metadata form the basis of *semantic annotation* which describes the process of attaching data to another piece of data. Thereby, a typed relation between the annotated data and the annotating data is established (Handschuh, 2005).

> **Definition:** An annotation A is a tuple (a_s, a_p, a_o, a_c), where a_s is the subject of the annotation (the annotated data) a_o is the object of the annotation (the annotating data) a_p is the predicate (the annotation relation) that defines the type of relationship between a_s and a_o, and a_c is the context in which the annotation is made" (Oren et al., 2006).

Various ontologies have been designed to enable the semantic annotation of textual documents and audiovisual content. The Web Annotation Ontology[15] is one of the most prominent examples for multi-purpose annotations. It "provides an extensible, interoperable framework for expressing annotations such that they can easily be shared between platforms, with sufficient richness of expression to satisfy complex requirements while remaining simple enough to also allow for the

15 https://www.w3.org/ns/oa, last visited: July 15, 2018

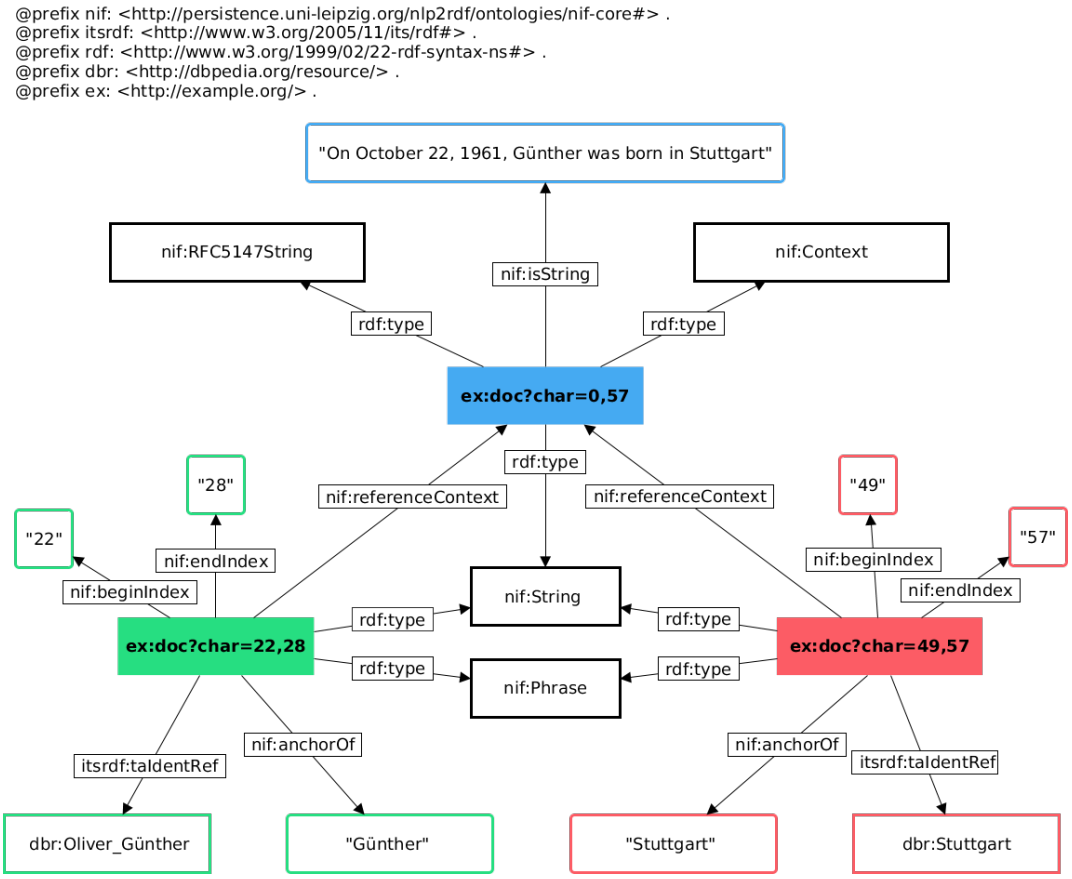

Figure 9: Annotation example using the NIF Core Ontology 2.0

Listing 7: NIF2 Annotation Example

```
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-
     core#> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dbr: <http://dbpedia.org/resource/> .

<http://example.org/doc?char=0,50>
     rdf:type       nif:String , nif:Context , nif:RFC5147String ;
     nif:isString   "On October 22, 1961, Günther was born in
     Stuttgart" .

<http://example.org/doc?char=22,35>
     rdf:type             nif:String , nif:Phrase ;
     nif:anchorOf         "Oliver Günther"^^xsd:string ;
     nif:beginIndex            "22"^^xsd:nonNegativeInteger ;
     nif:endIndex         "35"^^xsd:nonNegativeInteger ;
     nif:referenceContext <http://example.org/doc?char=0,50> ;
     itsrdf:taIdentRef    dbr:Oliver_Günther .

<http://example.org/doc?char=42,50>
     rdf:type             nif:String , nif:Phrase ;
     nif:anchorOf         "Stuttgart"^^xsd:string ;
     nif:beginIndex            "42"^^xsd:nonNegativeInteger ;
     nif:endIndex         "50"^^xsd:nonNegativeInteger ;
     nif:referenceContext <http://example.org/doc?char=0,50> ;
     itsrdf:taIdentRef    dbr:Stuttgart .
```

most common use cases, such as attaching a piece of text to a single web resource" (Sanderson, Ciccarese, and Young, 2016).

NIF Interchange Format 2.0

In this thesis, the annotation task is used in the context of Natural Language Processing. For this use case, Hellmann et al., (2013), developed the NLP Interchange Format 2.0 (NIF2). NIF2 is an RDF/OWL-based format aiming to achieve interoperability between Natural Language Processing tools, language resources and annotations. The format is based on a Linked Data enabled URI scheme for identifying elements in (hyper-)texts that are described by the NIF Core Ontology on a structural layer and a selection of ontologies for describing common NLP terms and concepts on a conceptual layer. The NIF Core Ontology 2.0 provides classes and properties to describe the relations between substrings, text, documents and their URI schemes (Hellmann, 2013). The NIF Core Ontology 2.0 contains two classes which are most important: (1) `nif:Context` is the annotation subject and represents the document containing the text and (2) `nif:String` describes the actual content of the document. Differentiating between these two classes is substantial because there may be two documents with the same content in a corpus. Figure 9 and Listing 7 accordingly demonstrate an annotation example using the sentence '*On October 22, 1961, Günther was born in Stuttgart*'. The rationale here is to identify a word or phrase in the sentence and create a link with the respective text

to the correct DBpedia resource. The URI `http://example.org/doc?char=0,50` is the reference context (`nif:Context`) of `http://example.org/doc?char=22,28` and `http://example.org/doc?char=42,50` . It is further described via the property `nif:isString` and the class `nif:RFC5147String`. The latter means that the URI fragment identifiers have to conform with the syntax of RFC 5147 (Wilde and Dürst, 2008). The URI `http://example.org/doc?char=22,28` (green) is of type `nif:String` and is the anchor of the text fragment surface form '*Günther*'. Each annotation also contains information about it's position in the text. It is encoded as a fragment identifier in the resource IRIs and represented via the properties `nif:beginIndex` and `nif:endIndex`. The `itsrdf:taIdentRef` property holds the annotation object, in this case the DBpedia resource `dbr:Oliver_Günther`. The annotation for the string '*Stuttgart*' works similarly. The URI `http://example.org/doc?char=42,50` (red) refers to the string ranging from character 42 to 50 and the annotation object is the DBpedia resource `dbr:Stuttgart`.

The semantic annotation of data on the Web can be accomplished manually, automatically, and semi-automatically depending on it's use case. For each variant, numerous technologies and tools exist and are part of current research. The process of annotating textual documents with so-called meaningful elements and connect them to specific parts of a knowledge base is part of the field of Natural Language Processing (NLP). The following section will briefly introduce Named Entity Linking (NEL) as part of NLP.

3.2.9 *Named Entity Linking*

In the recent years, Natural Language Processing has become an important means of text analysis in Sociology (Evans and Aceves, 2016; Lemke and Wiedemann, 2015). Named Entity Linking is part of the field of NLP and refers to the task of identifying mentions in a text and linking them to the entity they name in a knowledge base. In that sense, a *named entity* is a real-world object, for instance a person, a location, an organization or a product that is denoted with a proper name. It can be abstract or have a physical existence. In this definition, *named* refers to entities for which *rigid* designators exist, defined by Kripke, (1972). Contrary to rigid designators are *non-rigid* designators which may refer to many different objects in many worlds, e.g. time periods. Rigid designators include mostly proper names or specific terms like biological species, non-rigid designators do not extensionally designate the same object in all possible worlds (Nadeau and Sekine, 2007). In the sentence '*Oliver Günther is president of the University of Potsdam* ', 'Oliver Günther' and 'University of Potsdam' are considered named entities. Both refer to specific objects, while 'president' can refer to many different objects in many worlds. The distinction of rigid and non-rigid designators is not always defined universally in ongoing research and also non-rigid designators (e.g. 'president') may be included in NEL approaches. Assuming that in sociology, rigid as well as non-rigid designators play an important role in the analysis of text, the term named entity will refer to both categories in this thesis. In the process of NEL, mentions in a text are

linked to the entity they name in a knowledge base like DBpedia or Wikidata[16]. An example is given in the following text[17]:

$$\text{Günther}_{\texttt{dbr:Oliver\_Günther}}\ \textit{is}\ \text{president}_{\texttt{dbr:President}}\ \textit{of the}\ \text{University of Potsdam}_{\texttt{dbr:University\_of\_Potsdam}}.$$

The string 'Günther' is annotated with the DBpedia resource `dbr:Oliver_Günther` which is the intended meaning in context with the entities `dbr:University_of_Potsdam` and `dbr:President` in the same sentence. Without the given context the string 'Günther' could relate to many other persons, locations or organizations, e.g. the German soccer player `dbr:Sarah_Günther`. The main challenge of NEL lies in the disambiguation of named entities in natural language text and the identification of the intended entity (out of possibly hundreds of candidates) in a large knowledge base like DBpedia. NEL is part of ongoing research and numerous systems have been established in order to tackle the challenge via automated annotations, manual annotations, as well as hybrid semi-automated annotation approaches. Some of the most recent algorithms have been benchmarked with the General Entity Annotation Benchmark Framework (GERBIL) in order to compare the performance annotation tools with each other given a number of datasets and unified measuring approaches (Usbeck et al., 2015).

3.3 BRIEF SUMMARY

In this chapter, an introduction to a few basic Semantic Web and Linked Data technologies has been given which will be applied in the field of text analysis in sociology in the following chapter.

It has been discussed that the traditional Web offers many possibilities of participation for any user, but an issue of the Web is the availability of information in formats which are often unstructured and do not encode, what the information actually means. It has been shown that the meaning of information on the Web is important for human-computer and computer-computer communication, e.g. in the area of Web search. In the Semantic Web, an extension of the traditional Web, meaning is made explicit by formal and standardized knowledge representations – ontologies. These ontologies are the foundations of Semantic Web and Linked Data applications, which include Named Entity Linking. The goal of this special task of Natural Language Processing is to identify mentions in a text and link them to the entity they name in a knowledge base. This also means to correctly disambiguate the named entities, e.g. to specify whether an entity mention *Günther* refers to the female soccer player Sarah Günther or the university president Oliver Günther.

Embedding these information in text using semantic metadata enables users (humans and machines) to understand the text and its meaning. Linking entities to their representations in a knowledge base also allows to utilize the underlying organization of such knowledge. Through the explicit definition that Oliver Günther is the president of Potsdam University it becomes possible to enrich the original

16 `https://www.wikidata.org/wiki/Wikidata:Main_Page`, last visited: July 15, 2018

17 The prefix dbr: stands for the DBpedia resource URL `http://dbpedia.org/resource/`.

text with additional information about these entities. Thereby the text is given more context about other persons, locations or events.

In the following chapter 4 it will be demonstrated how these technologies and standards can be used in the field of sociology and especially text analysis. First, it will be shown how textual data can be structured using RDF. Then, these data are annotated with semantic entities and queried using SPARQL to exploit not only the corpus on its own but to also use the underlying graph structure to enrich the text with additional context information and aggregate the content in a meaningful way.

4

APPLYING SEMANTIC WEB TECHNOLOGIES AND LINKED DATA IN SOCIOLOGICAL TEXT ANALYSIS

On the bases of the foundations and principles presented in chapter 3 this chapter focuses on the direct applications in the field of sociological text analysis. Section 4.1 first motivates why textual research data in sociology should be published as Linked Data. The section also includes a step by step analysis of the process of publishing data via the use case described in section 4.1.1. Section 4.2 and 4.3 demonstrate how the generated RDF data can be utilized in sociological text analysis. Finally, in section 5 the contributions of this chapter are summarized and all benefits, system limitations and future work will be discussed.

4.1 PUBLISHING SOCIOLOGICAL RESEARCH DATA AS LINKED DATA

The Web has "radically altered the way we share knowledge by lowering the barrier to publishing and accessing documents as part of a global information space" (Bizer, Heath, and Berners-Lee, 2011). Even though the Web provides numerous benefits in the way documents can be shared and accessed, "the same principles that enabled the Web of documents to flourish have not been applied to data" until the rise of the Web of data. On the traditional Web, data has been made available as raw dumps like CSV or XML, or as HTML tables which sacrifices its structure and semantics. The Web evolved from an information space of linked documents to a space where documents and data are interlinked during the last two decades, underpinned by best practices for publishing and connecting structured data, which became known as Linked Data (ibid.). Also in research, initiatives have been promoting to publish data resources, articles, and reviews according to the Linked Data principles (as shortly discussed in 3.2.4) to make research more accessible, transparent, reusable, and therefore more credible. For instance, the Linked Research project[1] promotes to make all articles and resources available in a format that is human and machine-readable and interlinked with other information on the Web.

In sociological research, data sharing and publishing is neither standardized nor is it widely practiced. Studies by Zenk-Möltgen and Lepthien, 2014 and Herndon and O'Reilly, 2016 show that social science journals have just been starting to slowly adapt data sharing policies and most journals which enforce data publishing policies do so mostly in an incomplete and varied way. Bosch and Zapilko, 2015 found that also for social sciences, there are promising applications for Semantic Web technologies, especially concerning publishing and exploring survey and statistical data. The intention of this section is to demonstrate, how textual documents used for the analysis in sociology can be modeled and published by

1 https://linkedresearch.org/, last visited: July 8, 2018

sociologists. It will further be elaborated which level of technical expertise is necessary to accomplish the particular steps.

4.1.1 *Exemplary Use Case*

To demonstrate the feasibility and benefits of storing and publishing documents for sociological research as Linked Data, a corpus of constitutional documents was chosen, building on the work by Knoth, 2016 and Knoth, Stede, and Hägert, 2018. In order to learn about state identities or definitions of affiliations (e.g. citizens, foreigners, heads of state) and their change over time, constitutions provide a decent resource of information (Boli-Bennett, (1979), Go, (2003), and Lorenz, (2005)). According to Heintz and Schnabel, (2006, pp 707 - 708), constitutions can be viewed as a mirror of society and as a self-description of the state in the context of global societies. The research project by Knoth, Stede, and Hägert, 2018 deals with European constitutions (in German language) between 1815 and 2016. In it, NLP techniques are used to semi-automatically analyze the document structure and content with respect to their changes over time. An early task of this research project was the generation of XML files modeling the structure for each constitution version for future analysis (Knoth, Stede, and Hägert, 2018, p 199). This corpus provides a sophisticated use case for publishing sociological texts as Linked Data, because analyzing constitutional documents requires to research its structure as well as content. For a sophisticated analysis of the corpus, it becomes important to reference single parts of the constitution and their development over time with the awareness whether these parts belong to a specific article or section in the document. Publishing these data in RDF gives the (non technical) researcher the most flexibility and control over these texts (Elkins et al., 2014, pp 17 - 18).

4.1.1.1 *XML vs. RDF*

The corpus has been kindly made available as XML. While XML provides a number of benefits regarding the way data can be encoded syntactically, XML in general also has a number of disadvantages in contrast to RDF. XML was designed for markup in documents of any structure and creates a tree of nested sets of tags. With XML, it is possible to read data and get a representation which can be further exploited utilizing an XML parser. However, its major disadvantage over RDF is that XML does not enable to recognize semantic units. It aims at document structure and imposes no common interpretation of data contained in a document (Decker et al., 2000). An RDF document describes a directed graph. RDF enables to easily describe the relationships between resources and allows to combine data from multiple sources (Hitzler, Kroetzsch, and Rudolph, 2009). Furthermore, the potential reuse of RDF data is enormous and goes way beyond the parser reuse offered via XML (Decker et al., 2000). In the specific context of constitution data, Elkins et al., (2014), have pointed out three major reasons on why to use RDF instead of XML. The first regards syntax consistency. When using RDF, it does not matter which syntax is actually chosen as long as the data are modeled as a graph while XML requires to decide upon a schema beforehand to be used to define relationships. The authors have pointed out that constitutions across countries may

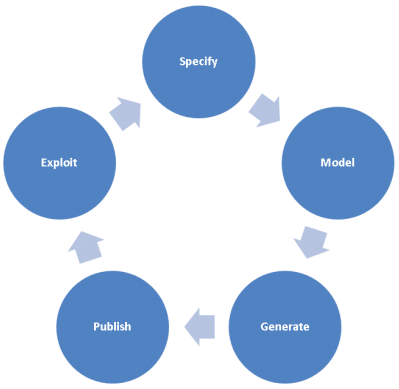

Figure 10: Life Cycle of generating, publishing and maintaining Linked Data by Villazón-Terrazas et al., (2011)

vary in their structure which makes it difficult to provide a schema suitable for all constitutions. Using RDF and modeling underlying ontology eliminates the need of such rigid schema. The second reason regards flexibility, because ontologies allow for changes in the underlying data as well as their architecture, while XML requires to change schema which also includes the re-encoding of each constitutional text. The third reason mentioned by the authors regards the ability to link to other data in the LOD cloud, e.g. DBpedia (Elkins et al., 2014, pp 17 - 18).

In the following section 4.1.2 it will be discussed how the exemplary use case data can be converted to RDF based on best practices developed by the W3C.

4.1.2 *Application of Best Practices*

The W3C has published ten best practices to publishing Linked Data (2014) starting out with creating a common mindset of potential collaborators to selecting and modeling data to converting data and making it accessible to humans and machines for reuse. On the bases of these principles and with respect to the presented use case, it will be demonstrated how textual documents used for the analysis in sociology can be published on the Web as Linked Data.

4.1.2.1 *Prepare Stakeholders*

This step includes the preparation of stakeholders to the process of creating, publishing and also maintaining Linked Data. In principle, this entire chapter can be understood as a preparation of sociologists seeking to publish their research data in the context of text analysis as Linked Data. In order to prepare stakeholders, the overall workflow is usually demonstrated. A popular workflow was published by Villazón-Terrazas et al., (2011). In it, the authors identify a life cycle consisting of five steps to successfully specify, model, generate, publish, and exploit Linked Data in the government domain, as shown in Figure 10. The life cycle also shows that once the data is published, the work is never fully completed. Once new data is modeled, it has to be generated, published and so on. Even though this mostly

Art. 8. Das Recht auf Bildung von Vereinen wird anerkannt. Dieses Recht kann im Interesse der öffentlichen Ordnung durch Gesetz eingeschränkt werden.

Art. 9. (1) Das Recht zur Versammlung und Demonstration wird anerkannt, unbeschadet der Verantwortung jedes einzelnen vor dem Gesetz.

(2) Zum Schutze der Gesundheit, im Interesse des Verkehrs und zur Beseitigung oder Abwehr von Störungen können gesetzliche Vorschriften erlassen werden.

siehe auch Zusatzartikel V.

Art. 10. (1) Jeder hat, unbeschadet der Einschränkungen durch Gesetz oder kraft eines Gesetzes, das Recht auf Wahrung seiner Privatsphäre.

(2) Der Schutz der Privatsphäre wird im Zusammenhang mit der Speicherung und Weitergabe persönlicher Daten durch Gesetz geregelt.

(3) Der Anspruch von Personen auf Einblick in die über sie gesammelten Daten und deren Verwendung sowie auf Berichtigung solcher Daten wird durch Gesetz geregelt.

siehe auch Zusatzartikel VI.

Art. 11. Jeder hat, unbeschadet der Einschränkungen durch Gesetz oder kraft eines Gesetzes, das Recht auf körperliche Unversehrtheit.

siehe auch Zusatzartikel VII.

Art. 12. (1) *Das Betreten einer Wohnung gegen den Willen des Bewohners ist nur den durch Gesetz oder kraft eines Gesetzes bezeichneten Personen in den durch Gesetz oder kraft Gesetzes bezeichneten Fällen erlaubt.*

(2) *Für das Betreten einer Wohnung gemäß Absatz 1 ist die vorherige Legitimation und die Mitteilung des Zwecks des Betretens der Wohnung erforderlich. Der Bewohner erhält einen schriftlichen Bericht über das Betreten der Wohnung.*

Durch Gesetz vom 7. Februar 2002 erhielt der Artikel 12 folgende Fassung:
"**Artikel 12.** (1) Das Betreten einer Wohnung ohne Zustimmung des Bewohners ist nur den durch Gesetz oder kraft Gesetzes bezeichneten Personen in den durch Gesetz oder kraft Gesetzes bezeichneten Fällen erlaubt.
(2) Für das Betreten einer Wohnung gemäß Absatz 1 ist die vorherige Legitimation und die Mitteilung des Zwecks des Betretens der Wohnung erforderlich, unbeschadet der im Gesetz vorgesehenen Ausnahmen .
(3) Der Bewohner erhält schnellstmöglich eine schriftliche Benachrichtigung über das Betreten der Wohnung. Wenn das Betreten der Wohnung im Interesse der nationalen Sicherheit oder der Strafverfolgung erfolgt ist, kann nach durch Gesetz festzustellenden Regeln die Benachrichtigung zurückgestellt werden. In den durch Gesetz zu bezeichnenden Fällen kann die Benachrichtigung unterbleiben, wenn sie dem Interesse der nationalen Sicherheit dauerhaft zuwiderläuft .

Figure 11: Original constitution data collected by Knoth, Stede, and Hägert, (2018) from http://www.verfassungen.eu/, last visited: July 27, 2018

related to government data, the approach is widely generalizable to numerous domains.

4.1.2.2 *Select a Dataset*

According to the W3C, a dataset should be selected that contains uniquely collected or created data that provides benefits for others to reuse and adapt (Hyland, Atemezing, and Villazón-Terrazas, 2014). The dataset selected to publish in RDF in this use case consists of 20 constitution documents of the Netherlands from 1884 to 2016 in the German language. All documents have previously been made available by Knoth, Stede, and Hägert, (2018) in XML. A clear limitation is that the documents are available in German language only. Nevertheless, it will be assumed that the generated data will be valuable to any German speaking researcher intending to study European constitutions based on its structure or content.

4.1.2.3 *Model the Data*

Modeling Linked Data often requires to go from one model to another, e.g. from a relational database to a graph-based representation or from pre-defined XML documents to the intended graph model as intended in this use case. Modeling the data also requires to understand its basic structure. The exemplary dataset used consists of constitution documents which employ a very specific and relatively consistent document structure. The structure of the utilized XML documents has previously been modeled by Knoth, Stede, and Hägert, (2018). The process of generating the original XML dataset was not trivial and is a highly cumbersome task since no machine-readable and chronological dataset of European constitutions is available on the Web. Even though an HTML representation of these constitutions exists on the Web[2], it only consists of the latest version in the constitution with colored text paragraphs on occurring changes of the respective constitution. An example of the original data is given in Figure 11. Constitutions are usually divided into several main chapters which are furthermore divided into paragraphs,

2 http://www.verfassungen.eu/, last visited: July 27, 2018

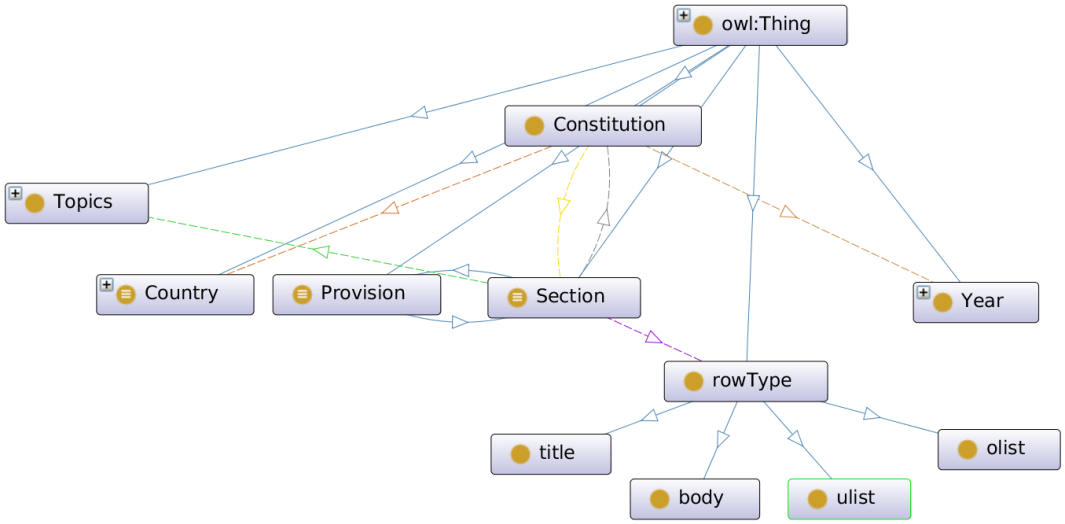

Figure 12: Constitution Ontology as developed by Elkins et al., (2014)

articles and sections. In some cases, articles and sections are directly connected to main chapters, the XML representation of this structure is further discussed by Knoth, Stede, and Hägert, (2018, p 199).

4.1.2.4 *Specify an Appropriate License*

According to the W3C, data reuse is more likely to occur when "there is a clear statement about the origin, ownership and terms related to the use of the published data" (Hyland, Atemezing, and Villazón-Terrazas, 2014). The Creative Commons project provides a sophisticated framework for "free, international, easy-to-use copyright licenses that are the standard for enabling sharing and remix"[3]. All data provided in this use case are published under the Creative Commons 'Attribution-ShareAlike 4.0 International' license[4], which means the data can be copied and distributed in any medium or format, it can be remixed, transformed and build upon the material for any purpose, even commercially. The license appears under the terms that (1) the person or organization to reuse the data must give appropriate credit to the author and (2) can only distribute the contributions under the same license as the original.

4.1.2.5 *Good URIs for Linked Data*

In order to benefit from the value of Linked Data, resources should be identified using HTTP URIs. Furthermore, the URI structure should never contain anything that will change, i.e. sessions or tokens, to give others the possibility to reuse the data.

4.1.2.6 *Use Standard Vocabularies*

The W3C highly promotes the reuse of standardized vocabularies (Hyland, Atemezing, and Villazón-Terrazas, 2014). Especially reusing other peoples' vocabularies has become an important factor in the development of the Semantic Web and Linked Data and it is one of the major aspects that have made the Semantic Web as successful as it is across domains. Building on other people's work, significantly increases its own value, decreases the work load of the person or organization reusing it and enlarges the network of Linked Data on the Web (Noy, McGuinness, et al., 2001). Several services (recommended by the W3C) exist to find existing vocabularies including Linked Open Vocabularies[5] or Prefix.cc[6].

In the context of this use case, a corpus of constitutional documents is used and therefore a vocabulary for this domain was utilized. The Constitute Project[7] already developed an ontology for the domain, which could partially be reused in the context of this corpus. The project aims at creating a platform for professionals drafting constitutions, and thus requiring to read and compare constitutions of various countries with each other (Elkins et al., 2014). The used ontology is freely available on the Web[8]. Figure 12 visualizes parts of the ontology in the protégé editor[9]. The ontology contains a class `co:Constitution`[10] and the subclass `co:Section`. Each section has a `co:rowType` which can be a `title`, `ulist`, `olist`, or `body`[11]. Furthermore, each constitution `co:isConstitutionOf co:Country`. The modeling of all countries in the ontology has been reused by the authors from the FAO Geopolitical Ontology[12]. The way the ontology was created treats all parts of a constitution in the same way, regardless if it is actually an article, section or paragraph. However, in order to query the constitutions for a further analysis in sociology, this information is critical, therefore the ontology has been further extended by these information. The namespace "`https://github.com/tabeatietz/semsoc/`" has been created and used with the prefix `s:`. The classes `s:Part`, `s:Paragraph`, `s:Section` and `s:Article` have been added to the ontology which enables to query for each specific unit separately. The given ontology furthermore models the year the respective constitution was created in. However, the corpus in the use case often contains two constitution versions for a specific year. Therefore the `s:edition` property has been added which accepts values of the type `xsd:date`[13].

Sociologists and Ontology Engineering

This step in the process of publishing Linked Open Data is referred to as ontology engineering or knowledge engineering. Even though ontologies are based on

3 `https://creativecommons.org/`, last visited: July 27, 2018
4 `https://creativecommons.org/licenses/by-sa/4.0/`, last visited: July 27, 2018
5 `https://lov.linkeddata.es/dataset/lov/`, last visited: July 26, 2018
6 `http://prefix.cc/`, last visited: July 26, 2018
7 `https://www.constituteproject.org/`, last visited: July 26, 2018
8 `https://www.constituteproject.org/ontology/`, last visited: July 26, 2018
9 `https://protege.stanford.edu/`, last visited: July 26, 2018
10 The prefix `co:` stands for the namespace `<http://www.constituteproject.org/ontology/>`
11 In the given ontology, the classes title, olist, ulist, body and rowType start out with lower case characters which is against the W3C recommendations
12 `http://www.fao.org/countryprofiles/geoinfo/en/`, last visited: July 26, 2018
13 The prefix `xsd:` stands for the namespace `http://www.w3.org/2001/XMLSchema#`

logic, and not every sociologists is familiar with this complex domain, modeling ontologies can also be achieved by non technicians. Halford, Pope, and Weal (2013), have especially emphasized the possibilities and benefits for sociologists and discussed that creating ontologies is not solely a technical problem, because high domain knowledge is required to create sophisticated models. For instance, a recent project in the domain of film- and TV-production, non-technicians have been creating a filmontology[14] to model the entire production process from the initial idea to costume design to the final editing (Agt-Rickauer, Waitelonis, Tietz, and Sack, 2016). There are numerous tools and guides to support non-technicians in the development and reuse of ontologies. One of the most popular free and open-source editors is called *protégé* (Musen, 2015). The community has build numerous plugins and developed numerous guides to support non-technical users. One of the most widely used guides has been created by Horridge, Knublauch, Rector, Stevens, and Wroe, (2004) and has since been renewed in several editions. If an ontology has to be created from scratch, the PoolParty Thesaurus Management System[15] can help sociologists to collect and describe all necessary concepts, and define relationships to other concepts (Schandl and Blumauer, 2010). The tool was created specifically for domain experts unfamiliar with Semantic Web technologies and without programming skills and also Bosch and Zapilko (2015) have specifically pointed out the usefulness of the Poolparty tool for social scientists.

4.1.2.7 *Convert Data to Linked Data*

Converting the provided XML data to RDF involves mapping source data to RDF statements (Hyland, Atemezing, and Villazón-Terrazas, 2014). The files made available by Knoth, Stede, and Hägert, (2018), have been converted to RDF using a Python script. The script is available on the Web via Google Colaboratory[16]. For sociologists, this step in the process of publishing Linked Data is one of the most crucial and is most likely achieved through an interdisciplinary collaboration, since programming skills are required. Even though several tools exist which provide aids in converting data (e.g. by Lange, (2009), and Heyvaert et al., (2016)) no complete out of the box software exists which runs the process completely automatically and error free. Ideally, this process will become obsolete in the future when adding further constitution data, since the data can be modeled directly in RDF.

4.1.2.8 *Provide Machine Access to Data*

With this best practice, it is made sure that not only humans are able to access and exploit the provided data, but also machines have access to the data. This can be mainly accomplished by providing a RESTful application programming interface (API), by providing a SPARQL endpoint, or by providing a file download in RDF. Which method the sociologist should use highly depends on their skill

14 filmontology.org, last visited: July 26, 2018

15 https://www.poolparty.biz/, last visited: July 26, 2018

16 Python script to convert XML data to RDF at Google Colaboratory http://bit.ly/ConstitutionScript. Colaboratory is a free Jupyter notebook environment that runs entirely in the cloud and requires no setup, though a Google account is required to directly run the scripts. Otherwise the *.py* files can simply be downloaded and executed locally

set, the use case, and the ability to maintain the access point (cf. section 4.1.2.9). The easiest method is to simply provide RDF data dumps. This approach does not require extensive maintenance (as e.g. a SPARQL endpoint) and does not require enormous technical skills. All data generated in this use case are available as a turtle dump file on GitHub[17]. After downloading the dataset, the researcher is able to load it to a triple store of choice and then proceed to formulate queries using SPARQL. Popular triple stores include but are not limited to Apache Fuseki[18], Blazegraph[19], Virtuoso[20].

4.1.2.9 *Announce to the Public and Recognize the Social Contract*

The W3C has furthermore recommended best practices to announce the published data and to recognize the social contract that comes with the published material. The former includes associating an appropriate data license, ensuring data accuracy, planning a persistence strategy as well as a method for people to provide feedback on the data. Recognizing the social contract refers to the responsibilities that comes with maintaining the data as well as its access points. If a SPARQL endpoint is provided to query the data directly, the social contract involves keeping the endpoint available and stable (Hyland, Atemezing, and Villazón-Terrazas, 2014). The way in which data and access points are maintained substantially influences not only the way a specific dataset can be reused and valued by other people, it also contributes significantly to the success or fail of Linked Open Data in the future of the Web (priv. comm.). However, due to the exemplary nature of this provided use case, an extensive announcement of the provided dataset as well as a long term commitment to a social contract has not been planned at this point. This will be further discussed in the future work section in chapter 6.

4.1.3 *Result and Brief Summary*

This section gave an overview of the motivation and process of publishing sociological text documents as Linked Data according to the best practices published by the W3C. Thereby, the exemplary use case of constitutional texts provides a real world scenario. As a contribution, the corpus consisting of 20 constitutional documents has been converted to RDF and made available on the Web, along with the (in section 4.1.2.7 described) Python script converting the files to RDF. A snippet of the generated RDF data is depicted in Figure 13 and Listing 19 (in Appendix A) . In the example, the following information text is depicted:

Die Verfassung des Königreichs der Niederlande
Hauptstück 2 - Regierung
§2. König und Minister
(1) Die Regierung besteht aus dem König und den Ministern

17 https://github.com/tabeatietz/semsoc
18 https://jena.apache.org/documentation/fuseki2/, last visited: July 5, 2018
19 https://www.blazegraph.com/, last visited: July 5, 2018
20 https://virtuoso.openlinksw.com/, last visited: July 5, 2018

Figure 13 shows how these information have been structured in the RDF graph. The entire document (the constitution from 2016) is defined as `co:Constitution` and the rest of the different levels of the document are defined as `co:Section`. To provide a more fine grained analysis, the single section parts have been further divided into `s:Chapter`, `s:Paragraph`, `s:Article` and so on. Thereby, the chapter is `co:parent` of a paragraph and a paragraph is `co:parent` of an article. For each element, it is modeled, whether it is of type `co:title` or `co:body`. Each element further has a `co:sectionID`, which is a sequential number.

The following sections 4.2 and 4.3 will demonstrate how the generated data can be queried and exploited in the context of sociological research.

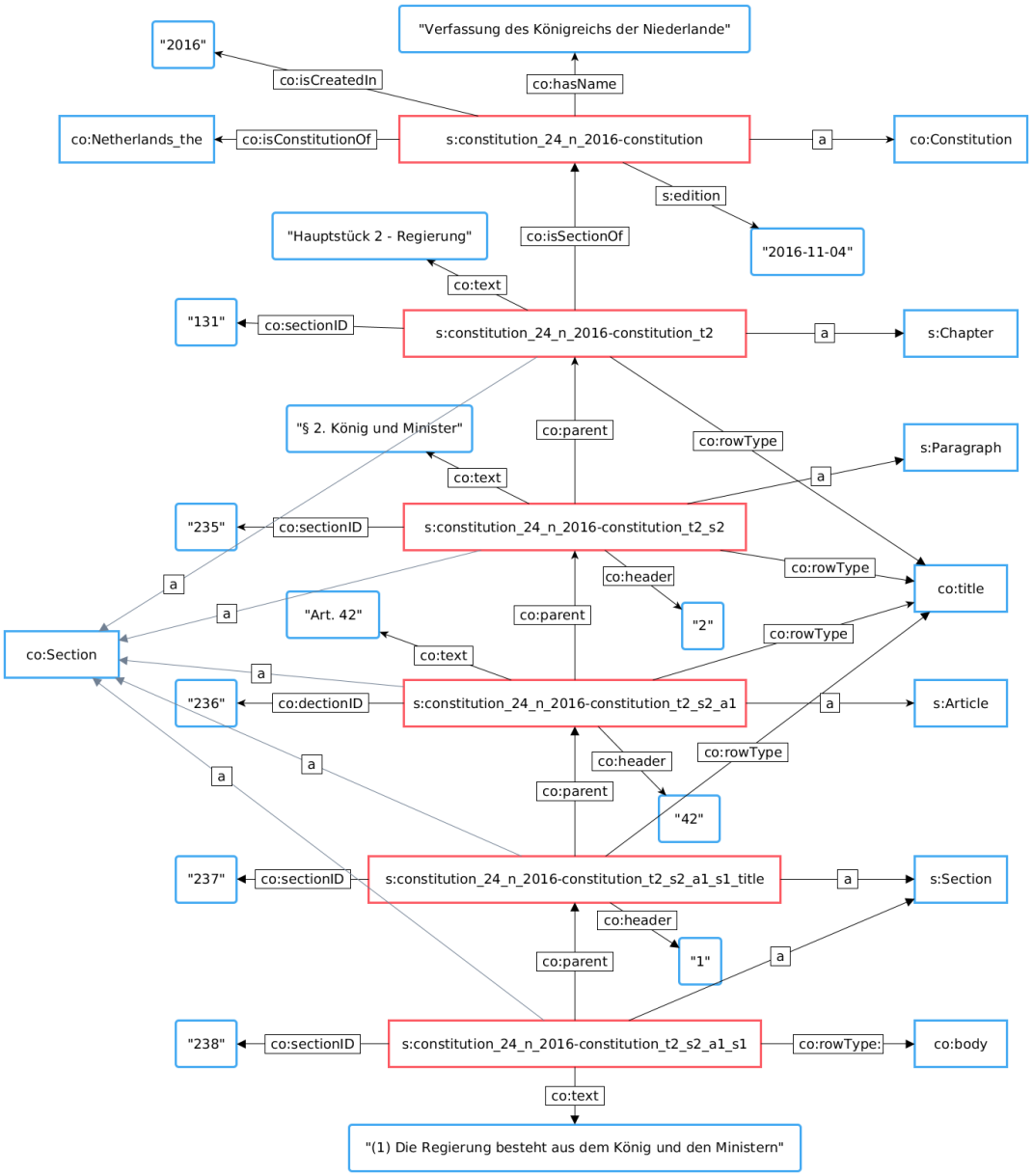

Figure 13: Visualization of a small part of the generated RDF graph

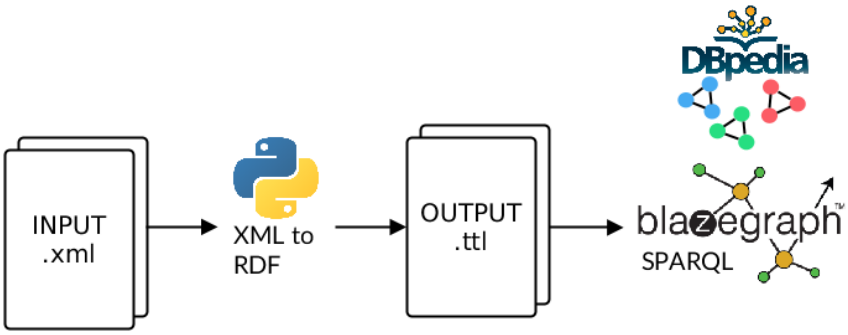

Figure 14: Workflow of converting the provided XML data to RDF and querying the RDF via Blazegraph

4.2 EXEMPLARY STRUCTURE ANALYSIS OF CONSTITUTION TEXTS

In this section, the process of text analysis aided by Semantic Web and Linked Data technologies will be discussed on the foundation of the use case described in section 4.1.1. As described above, the analysis of constitutions in the context of sociological research enables to assess how states model societies. Constitutions are self-descriptions of these states and mirror the different roles in it as well as their relationships with each other. Numerous aspects may be analyzed including affiliations (who belongs to the state or who is considered a foreigner), leadership (who is the head of the state and which and how are the responsibilities allocated), or a more fine grained analysis, e.g. the role of women in the state. Two possible perspectives to analyze these aspects in sociology are the analysis of the document structure as well as their actual content. This section focuses on the structure level before an analysis on content level will be performed in section 4.3.

Constitutional documents follow a strict formal hierarchy. Each document is organized into several units, being the chapters, paragraphs, articles, and sections. Typically, each section belongs to a specific article and each article either belongs to a paragraph or directly to a chapter. When analyzing constitutional documents for specific countries and their changes over time, the document structure may give insights on which chapters, paragraphs or articles have been added, deleted or changed over time. This analysis enables an initial exploration of the corpus before the in-depth content analysis takes place.

4.2.1 *Workflow*

Analyzing the document structure in this use case means to query it using the SPARQL query language as described in section 3.2.7. Figure 14 shows the workflow of this section. The RDF graph generated with the Python script introduced in section 4.1 is taken as the input file here and uploaded to the Blazegraph framework for querying. In order to enrich the documents in the use case with external context information, the DBpedia knowledge base is used via federated querying (cf. section 4.2.2.1).

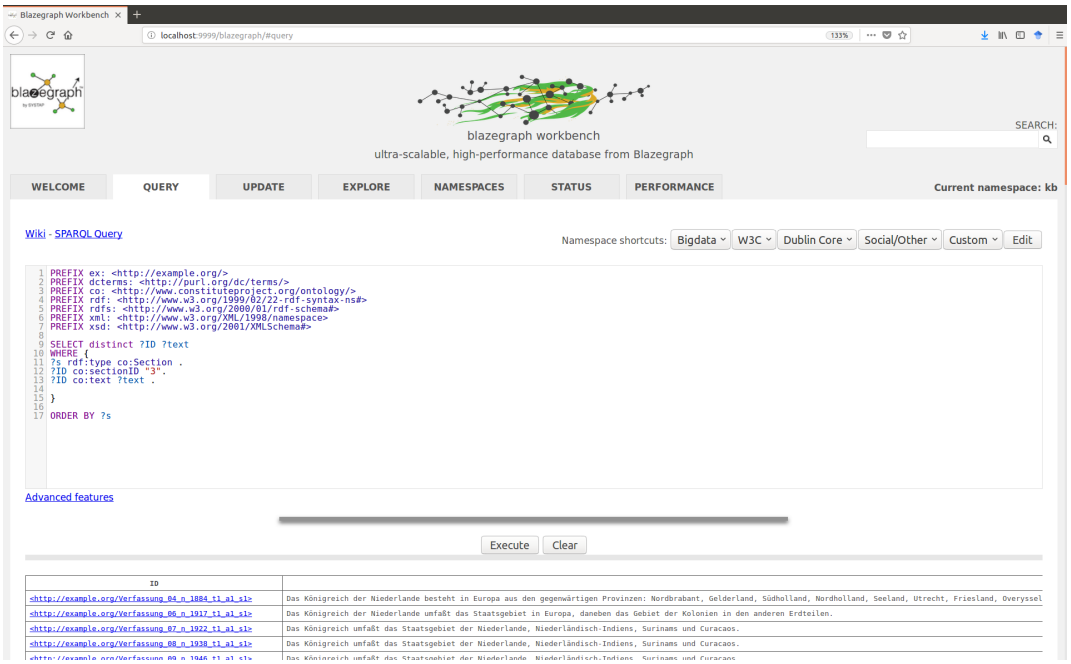

Figure 15: Blazegraph working environment

Listing 8: SPARQL query for the list of constitution documents and their editions

```
PREFIX co: <http://www.constituteproject.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?constitution ?edition
WHERE  {
?constitution a co:Constitution ;
     co:isConstitutionOf co:Netherlands_the ;
     s:edition ?edition .
}
```

4.2.2 *Querying*

In order to get a first and broad overview of all documents in the corpus, a first query as shown in Listing 8 asks:

1. Which documents are in the corpus that contain the constitution of the Netherlands?
2. What year were these documents created?

The query in Listing 8 selects anything in the knowledge base that is of `rdf:type co:Constitution` and belongs to the Netherlands. Furthermore, the edition of each constitution is selected via the `s:edition` property. Figure 16 visualizes the query result in a timeline overview, generated via TimeGraphics[21]. It is shown that there

21 https://time.graphics/, last visited: August 12, 2018

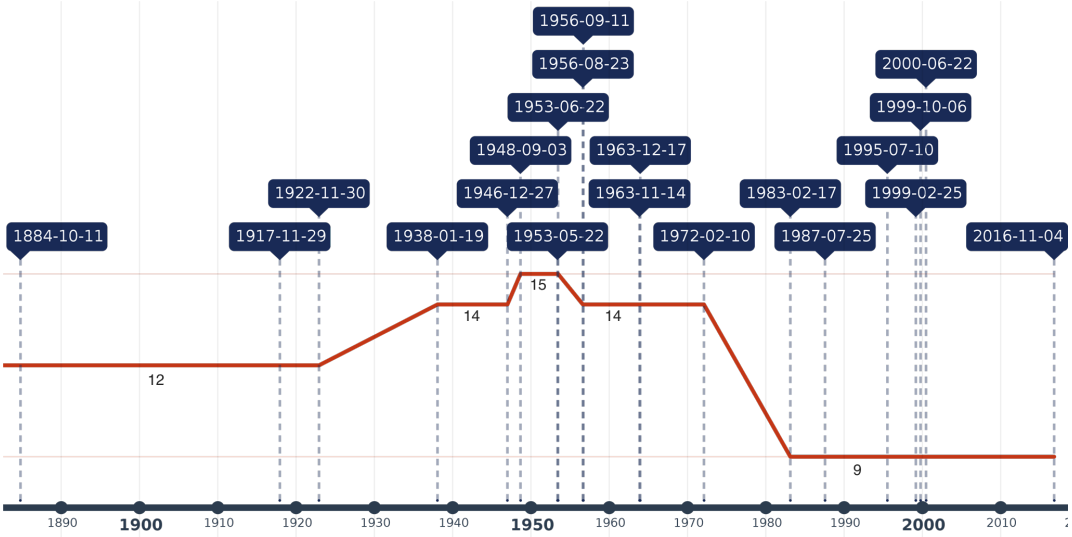

Figure 16: Timeline of constitution editions and chapter numbers

are 20 different editions of the constitution of the Netherlands[22], ranging from 1884 to 2016. The query reveals that in the time span of 54 years between 1884 and 1938, only 4 edited editions of the constitution appeared while during a similar time span of 49 years between 1938 and 1987 the constitution was edited 12 times. In the years 1953, 1956, 1963 and 1999, two editions appeared in the same year. It could be deduced that great political disturbances occurred in the years with a greater editing ratio which could be inner political or a reaction to international affairs. However, more information is needed to research this in detail and possibly create hypotheses.

Chapter Level

The constitutions' most top level elements are the chapters. The chapters set the entire framework of the constitutions, they dictate the main topics as well as their order. Therefore, it is assumed that structural changes on chapter level tend to have a higher impact to the entire constitution structure and content than changes on a lower article or section level. How can changes in a document structure be assessed? Even though simply counting the chapters in each edition is a simple and low effort approach, it is assumed that it delivers a promising entry point into top level document changes. The query in Listing 9 returns results to the question:

- How many chapters does each constitution edition consist of?

The results to the query above are also visualized in Figure 16. The red line in the timeline view shows how many chapters are included in each constitution edition. While the amount of chapters stayed constant until the 1922 edition, there have been a number of changes between 1938 and 1972. Due to the fact that the chapter number dropped from 14 to 9 from 1972 to 1983, it is assumed that significant changes in the constitution occurred. Between 1948 and 1953, one more

22 While currently there are only documents of the Netherlands in the corpus, this was included in the query to increase generalizability

Listing 9: Query to count all chapter numbers per constitution edition

```
PREFIX s: <https://github.com/tabeatietz/semsoc/>
PREFIX co: <http://www.constituteproject.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?edition (COUNT(?chapter) AS ?numchapter)
WHERE  {
  ?chapter a s:Chapter ;
     co:isSectionOf ?constitution .
  ?constitution a co:Constitution ;
     s:edition ?edition .
}
GROUP BY ?edition
```

Listing 10: Query for all chapter 1 headers per edition

```
PREFIX s: <https://github.com/tabeatietz/semsoc/>
PREFIX co: <http://www.constituteproject.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?edition ?text
WHERE  {
  ?chapter a s:Chapter ;
      co:isSectionOf ?constitution .
  ?constitution a co:Constitution ;
      co:isConstitutionOf co:Netherlands_the ;
      s:edition ?edition .
  ?chapter co:header "1";
     co:text ?text .
}
```

chapter was first added and later removed. As simple as these numbers seem, they give initial insights on the extent of the changes from one edition to another. The observations made here provide evidence of significant changes structural (and most likely content) wise and give insights on where to begin a further in-depth investigation. Due to the given observations, further content analysis may probably focus on the editions with more severe changes, (e.g. 1938 and 1983) than on editions with obviously less severe changes.

To get more information of these changes, the next level to investigate are the chapter's titles and it can be asked:

- For each edition of the Dutch constitution, what are the names of the single chapters?

Listing 10 queries all chapter 1 headers per constitution edition. The results show that the title of the constitution of the Netherlands from 1884 to 1972 was '*Erstes Hauptstück - Vom Reich und seinen Einwohnern*' (Of the empire and its inhabitants) and starting from 1983, the first chapter was entitled '*Grundrecht*' (Fundamental Rights). That means the Netherlands first included the Fundamental Rights in

Listing 11: Query to count all sections for the first chapter of the 2016 constitution edition

```
PREFIX s: <https://github.com/tabeatietz/semsoc/>
PREFIX co: <http://www.constituteproject.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT  ?articletext (COUNT(?section) AS ?numsec)
WHERE  {
  ?chapter a s:Chapter ;
     co:isSectionOf ?constitution .
  ?constitution a co:Constitution ;
     co:isConstitutionOf co:Netherlands_the .
  ?chapter co:header "1".
  ?article a s:Article ;
     co:parent ?chapter ;
     co:text ?articletext .
  ?sectiontitle a s:Section ;
     co:parent ?article .
  ?section a s:Section ;
     co:parent ?sectiontitle ;
     co:text ?sectiontext .
  ?constitution s:edition ?edition .
  FILTER(?edition = "2016-11-04"^^xsd:date)
  }
GROUP BY ?articletext
```

their constitution in 1983. The fundamental rights of a constitution represent one of their most important chapters and define the rights which members of a society are guaranteed towards states as stable, permanent and enforceable. They consist of the citizens' rights of defense against a state and define relationships between the citizens. The fundamental rights are often envisioned as the framework for the entire judicial system of a state (Alexy, 1999). Therefore, analyzing the corpus in the context with the background knowledge that a specific article or section was valid under the constraint that the fundamental rights had already been defined in that specific constitution edition is assumed to be crucial.

Article and Section Level

Due to the significance of the Fundamental Rights in the constitution it is furthermore worthwhile to ask:

1. Have there been changes on article and section level in the Fundamental Rights chapter of the constitutions between 1983 and 2016?
   a) What is the amount of articles and sections for each constitution edition?
   b) If changes are detected between both editions, what are they exactly?

The editions of 1983 and 2016 have been chosen because they represent the first and the most current year the Fundamental Rights were included in the constitution. Since the data have been modeled to reference each single section separately, it is possible to query them. The query in Listing 11 counts the amounts of chapter 1 sections for the respective edition chosen for comparison. After comparing

Listing 12: Query to list all section texts for article 12 in the constitution editions of 1983 and 2016

```
PREFIX s: <https://github.com/tabeatietz/semsoc/>
PREFIX co: <http://www.constituteproject.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT  DISTINCT ?edition ?articletext ?sectiontext
WHERE  {
  ?chapter a s:Chapter ;
      co:isSectionOf ?constitution .
  ?constitution a co:Constitution ;
      co:isConstitutionOf co:Netherlands_the .
  ?chapter co:header "1".
  ?article a s:Article ;
      co:parent ?chapter ;
      co:text ?articletext ;
      co:header "12" .
  ?sectiontitle a s:Section ;
      co:parent ?article .
  ?section a s:Section ;
     co:parent ?sectiontitle ;
     co:text ?sectiontext .
  ?constitution s:edition ?edition .
  FILTER(?edition = "2016-11-04"^^xsd:date ||
     ?edition = "1983-02-17"^^xsd:date)
  }
ORDER BY ?sectiontext
```

both editions it becomes clear that the Fundamental Rights did not significantly change on structure level with the exception of article 12 where one section was added to the constitution. To have a closer look at article 12, the query in Listing 10 has been created which returns the text of all article 12 sections. The output as shown in Table 2 allows to compare the single sections of both editions with each other. Article 12 defines rules for the right if entering a person's home. While section 1 of article 12 was only slightly changed, the changes in section 2 were much bigger. It can be seen that the last sentence in the 1983 edition which reads "The resident receives a written report on entering the apartment" was removed in the 2016 edition. Even though it initially seems that this sentence was removed entirely, the last row in the table reveals that this topic was simply defined in greater detail and it now also defines this topic in relation to the national security. Further investigation shows that article 12 in chapter 1 was not edited at all from 1983 to 2000. The changes which now include a paragraph about the national security were only recently added in the 2016 edition.

From this exploration of the document structure, it can be learned that from a top level perspective, after the Fundamental Rights have been first included into the constitution of the Netherlands, there were significant changes in article 12, which deals with the entering of a person's home. In 2016, a paragraph was added to the article that regulates the notification procedure in case the interests of na-

tional security are affected. It is assumed that adding a section that regards the national security is worthy of a further analysis on content level. If there exists further evidence that more paragraphs related to the national security of the Netherlands were added to the constitution, future work may research when this factor of national security became important in the constitution and to what extend. Furthermore it may be studied which entities of the state (citizens, foreigners, civil servants, ministers) are affected by regulations on national security in the corpus.

edition	articletext	sectiontext
1983-02-17	Art. 12.	(1) Das Betreten einer Wohnung gegen den Willen des Bewohners ist nur den durch Gesetz oder kraft eines Gesetzes bezeichneten Personen in den durch Gesetz oder kraft Gesetzes bezeichneten Fällen erlaubt.
2016-11-04	Art. 12.	(1) Das Betreten einer Wohnung ohne Zustimmung des Bewohners ist nur den durch Gesetz oder kraft Gesetzes bezeichneten Personen in den durch Gesetz oder kraft Gesetzes bezeichneten Fällen erlaubt.
1983-02-17	Art. 12.	(2) Für das Betreten einer Wohnung gemäß Absatz 1 ist die vorherige Legitimation und die Mitteilung des Zwecks des Betretens der Wohnung erforderlich. Der Bewohner erhält einen schriftlichen Bericht über das Betreten der Wohnung.
2016-11-04	Art. 12.	(2) Für das Betreten einer Wohnung gemäß Absatz 1 ist die vorherige Legitimation und die Mitteilung des Zwecks des Betretens der Wohnung erforderlich, unbeschadet der im Gesetz vorgesehenen Ausnahmen .
2016-11-04	Art. 12.	(3) Der Bewohner erhält schnellstmöglich eine schriftliche Benachrichtigung über das Betreten der Wohnung. Wenn das Betreten der Wohnung im Interesse der nationalen Sicherheit oder der Strafverfolgung erfolgt ist, kann nach durch Gesetz festzustellenden Regeln die Benachrichtigung zurückgestellt werden. In den durch Gesetz zu bezeichnenden Fällen kann die Benachrichtigung unterbleiben, wenn sie dem Interesse der nationalen Sicherheit dauerhaft zuwiderläuft .

Table 2: Result of the SPARQL query in Listing 12

4.2.2.1 *Content Enrichment via Federated Queries*

When analyzing a document corpus through the exploration of its structure and changes over time, findings may be better understood when placed into their historical and societal context. The exemplary use case utilized in this thesis provides historical textual data about the constitutions of the Netherlands. But what does it actually mean, that a constitution was created in a specific year and was valid for a certain time period? Without any background information, e.g. about historical events, the respective leading party, or ruling monarch, these constitutional versions are only data without any meaning and can hardly be understood thoroughly.

Accessing Linked Data from External Knowledge Bases

So far in this section, all content was queried natively in a local triple store using the RDF graph originally generated. However, Linked Data also provides technologies and standards to include external context information into the corpus. This

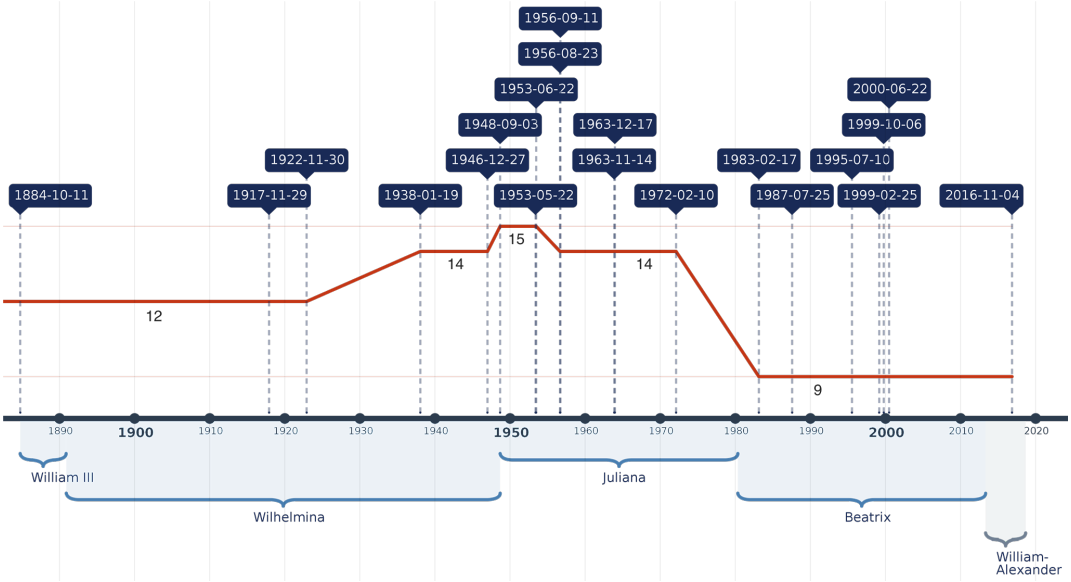

Figure 17: Timeline of constitution editions and chapter numbers

process is called enrichment and it is one of the most important applications of Linked Data. There are several possibilities to access external Linked Data in order to enrich the given content. Depending on the intended application and the available data, they have a number of advantages and disadvantages. The following list briefly introduces some of the most prominent options and discusses their pros and cons with regard to the presented use case.

1. **Federated Queries:** SPARQL not only enables to query RDF in local graph databases, but also allows to "express queries across diverse data sources" (Prud'hommeaux and Buil-Aranda, 2013). This feature is executed via the `SERVICE` keyword and allows to merge data distributed across the Web via an endpoint. A *SPARQL endpoint* enables humans and machines to query a specific knowledge base via SPARQL[23]. In the case of DBpedia, the existing SPARQL endpoint[24] can be used to query very specific information out of the large knowledge base easily without having to download an entire dataset. One of the advantages is that the data does not need to be updated by the user, since its the data provider who has this responsibility. While this method is easy to use, the major disadvantage is the dependence on third party servers. SPARQL endpoints are often not available due to maintenance or other reasons which is a crucial factor for Web applications (Verborgh et al., 2014).

2. **Data dump:** Instead of a specific query of the needed triples, data dumps contain an entire dataset which is entirely downloaded and integrated into the existing data. For instance, DBpedia provides a large number of datasets for eight different language versions in turtle and quad-turtle format[25]. This

23 http://semanticweb.org/wiki/SPARQL_endpoint.html, last visited: August 2, 2018
24 http://dbpedia.org/sparql, last visited: August 2, 2018
25 https://wiki.dbpedia.org/develop/datasets/downloads-2016-10, last visited: August 2, 2018

method is often used, because this way applications are completely independent of the availability of third party severs and solely rely on their own systems. However, querying data in this way is not considered querying on the Web, which is anticipated by the Semantic Web community. Using data dumps for querying means to include a large amount of triples into a graph database, even if only a small part of these data are actually used in the end. Furthermore, the data have to be updated manually by the user, contrary to the federated querying method.

3. **Triple Pattern Fragments:** The methods described above have in common that they enable the access to certain fragments of a Linked Data dataset. The result of each request, e.g. a SPARQL query result or a data dump, can be referred to as a *Linked Data Fragment* (LDF). Ongoing research in the area of Semantic Web attempts to overcome the disadvantages of the traditional access methods mentioned above. One solution is to use so called *Triple Pattern Fragments* (TPF) which are defined as certain types of fragments that can be generated with minimal effort by servers, while still enabling efficient querying (Verborgh et al., 2014). Next to the traditional triples of a dataset that match a selector, the triple pattern fragment contains metadata, triples which describe the dataset of LDF, and controls, hypermedia links or forms which lead to other LDFs (Verborgh et al., 2016). A simple example which demonstrates TPFs based on a query about women in Greek mythology is available on the Web[26].

The exemplary use case in this thesis will focus on federated querying. The advantages have been discussed above. The main disadvantage of federated queries is the low availability of endpoints on the Web. This is a crucial aspect, especially commercial applications. In this use case, it will be assumed that a high availability at all times is not the most crucial factor. Therefore, a federated query seems to be the best option to integrate few and very specific external data in the corpus.

Query Planning

The Netherlands is a constitutional monarchy. The monarch is the head of state and the constitution defines the monarch's position, power and responsibility in the state as well as his or her relationship with the rest of the government. The provided document corpus contains the Dutch constitution from 1884 to 2016. In these 132 years, several monarchs ruled in the Netherlands. When analyzing constitutional editions and their changes over time, the background information which monarch ruled the country in which specific constitution edition in the corpus gives the data meaning and context to understand the dataset. The goal of the query is:

1. List all editions of the constitution of the Netherlands contained in the document corpus and

2. for each edition, present the respective ruling monarch along with the starting and ending year on the throne.

26 http://bit.ly/TPF_Greek_myth, last visited: August 2, 2018

The query is divided in two parts. The first part is a simple query of content already included in the corpus. The second part is not part of the corpus and has to be queried from an external Linked Data knowledge base.

Query Building

To enrich the existing data with external knowledge, a dataset has to be selected. In this case, the DBpedia will be used via its SPARQL endpoint, because it is assumed that (since the data is generated from Wikipedia content) that it covers information about countries and their governments well. An example representation of a Dutch monarch in the DBpedia is the HTML page for the resource `dbr:Beatrix_of_the_Netherlands`[27]. All triples to the subject that are contained in DBpedia are visualized in this HTML page.

Listing 13 shows the entire query necessary to present the constitution editions and their respective monarch in one table. It consists of an inner and an outer query. The outer query makes use of the triples in the local knowledge base and collects all constitutions and their editions (lines 10 - 15). The inner query begins at line 17. The `SERVICE` keyword followed by `<http://dbpedia.org/sparql>` indicates that the DBpedia SPARQL endpoint will be queried for the triples requested in the following curly braces. The lines 18 to 28 ask for something (`?monarch`) that belongs to the class `dbc:Dutch_monarchs` and has an `rdfs:label`. As can be seen in the HTML representation for the entity Queen Beatrix, there are multiple values for `rdfs:label` in several languages. The `FILTER` constraint in line 20 specifies that it should only query for labels in the English language. Furthermore, the query specifies in line 21 that all monarchs have to have a starting year of reign to be included in the results. This is followed by two `OPTIONAL` statements in line 22 and 26. Optional means that the query collects the triples matching the patterns in the curly brackets, but only if they exist in the knowledge base. They are no mandatory specification. The first statement queries for all monarchs who have a successor (`?suc`) and the beginning of the successor's reign is defined as the previous monarch's end of reign. The second optional statement simply queries the end year of the monarch's reign. The reason to include these statements as optional is that it is unknown when the reign of the current King of the Netherlands will end. The filter constraints in line 31 and 32 make sure that only the monarchs are selected that are related to the edition years of the documents. The `BIND` statements create a dummy variable present in both, the inner and outer query to join the results of both queries in the resulting table.

Results

Table 3 shows the result to the query in Listing 13. Regarding the data itself, there are two aspects to observe: (1) the rows 7 and 8 refer to the same edition of the constitution and list two monarchs. The reason is that in 1948 Wilhelmina of the Netherlands died and the throne was inherited by Juliana of the Netherlands. (2) the current King, Willem-Alexander of the Netherlands is not listed, even though he succeeded Beatrix in 2013 and should theoretically be in the result table. The reason is simply an error in DBpedia. The correct resource URL for the ruling King is `dbr:Willem-Alexander_of_the_Netherlands`. Unfortunately, the URL listed as

27 `http://dbpedia.org/page/Beatrix_of_the_Netherlands`, last visited: August 2, 2018

Listing 13: Federated Query for all Dutch monarchs from 1884 to 2016 and the respective constitution editions

```
PREFIX s: <https://github.com/tabeatietz/semsoc/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX co: <http://www.constituteproject.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbc: <http://dbpedia.org/resource/Category:>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?edition ?year ?name ?start ?end
WHERE {
  ?constitution a co:Constitution ;
      s:edition ?edition ;
      co:isCreatedIn ?year .
  BIND("1" as ?dummy )

  SERVICE <http://dbpedia.org/sparql> {
     ?monarch dct:subject dbc:Dutch_monarchs ;
        rdfs:label ?name .
        FILTER (LANG(?name)="en")
        ?monarch dbo:activeYearsStartYear ?start.
        OPTIONAL {
               ?monarch dbo:activeYearsEndYear ?end .
               }
        OPTIONAL{
                ?monarch dbo:successor ?suc .
                ?suc dbo:activeYearsStartYear ?end .
        }

         BIND("1" as ?dummy )
  }
  FILTER(str(?year)<= str(?end))
  FILTER(str(?year)>= str(?start))
}
ORDER BY ?edition
```

`dbo:successor` for Queen Beatrix is `dbr:Willem-Alexander`[28]. This creates a mismatch which causes King Willem-Alexander to be missing in the results. This aspect makes clear that any analysis that is executed using Semantic Web and Linked Data can only be as good as the underlying knowledge base. This is one of the limitations of this approach and will be further discussed in section 5.2.3.

The results of this content enrichment are visualized in Figure 17[29]. The results show that the most edits of the constitution took place in the reign of Queen Juliana and Beatrix. Furthermore, the most significant changes which are assumed to have taken place in the 1983 edition occurred under the reign of Queen Beatrix. An in-depth analysis may furthermore focus on the reigns of these two Queens in particular.

28 When visiting the HTML representation for the resource `dbr:Willem-Alexander`, the user is immediately redirected to the correct URL. However, this redirect is not easily accomplished with SPARQL.

29 Even though King Willem Alexander was not found, he was included in gray color in the timeline.

The content enrichment with the reigns of Dutch monarchs is one example of many. In the exact same manner, further context could be provided with the governing Prime Minister in each edition or with periods of war which had significant influences on the country. In this example, it was important to show that, once data is converted to RDF and queried with SPARQL in a local triple store the exploration of the data does not have to end. The possibility to enrich the content automatically with external data and thus, create new context and new knowledge is one of the main achievements of the Semantic Web. I strongly believe that the possibilities that come with the enrichment as demonstrated can be highly beneficial for sociologists when analyzing textual data.

edition	year	name	start	end
1884-10-11	1884	William III of the Netherlands	1849	1890
1917-11-29	1917	Wilhelmina of the Netherlands	1890	1948
1922-11-30	1922	Wilhelmina of the Netherlands	1890	1948
1938-01-19	1938	Wilhelmina of the Netherlands	1890	1948
1946-12-27	1946	Wilhelmina of the Netherlands	1890	1948
1948-09-03	1948	Wilhelmina of the Netherlands	1890	1948
1948-09-03	1948	Juliana of the Netherlands	1948	1980
1953-05-22	1953	Juliana of the Netherlands	1948	1980
1953-06-22	1953	Juliana of the Netherlands	1948	1980
1956-08-23	1956	Juliana of the Netherlands	1948	1980
1956-09-11	1956	Juliana of the Netherlands	1948	1980
1963-11-14	1963	Juliana of the Netherlands	1948	1980
1963-12-17	1963	Juliana of the Netherlands	1948	1980
1972-02-10	1972	Juliana of the Netherlands	1948	1980
1983-02-17	1983	Beatrix of the Netherlands	1980	2013
1987-07-25	1987	Beatrix of the Netherlands	1980	2013
1995-07-10	1995	Beatrix of the Netherlands	1980	2013
1999-02-25	1999	Beatrix of the Netherlands	1980	2013
1999-10-06	1999	Beatrix of the Netherlands	1980	2013
2000-06-22	2000	Beatrix of the Netherlands	1980	2013

Table 3: Result of the federated SPARQL query in Listing 13 with the columns 1-2 queried in the existing dataset and the columns 3-5 queried in DBpedia

4.2.3 *Brief Summary*

The analysis of the document structure enables the sociologist to initially explore the corpus. In this section, it has been demonstrated that analyzing documents in RDF can be accomplished with the help of SPARQL queries using the exemplary use case of constitutional documents. Furthermore, it has been shown how external Linked Data can be integrated into the existing data to create a historical context and to give the data more meaning.

The 20 documents analyzed in this section can be considered a rather small document corpus in which some elements of the structural changes, especially on

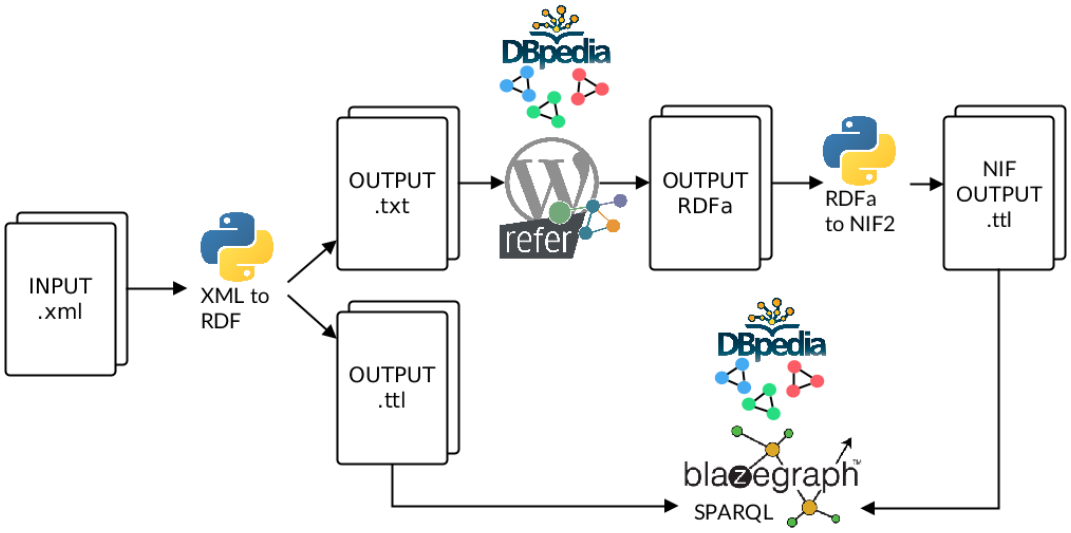

Figure 18: Workflow of semantically annotating the data with DBpedia entities and converting the output document containing RDFa into NIF2

the top chapter level may be surveyed by one researcher quickly without using SPARQL. However, it is easy to imagine that this will not be possible anymore when there are 200 or 2000 constitution documents from numerous different countries to be analyzed. The technologies and standards presented in this section provide means to study these data in the same way this rather small corpus has been analyzed.

The provided queries are merely examples of what is all possible and realistic for sociological research using SPARQL to explore the document structure from the very top level of merely chapter counting to deep into the document structure on section level. This is possible, because the data were modeled in a way that references each section separately with a unique URI. The entire process as discussed is not limited to constitutional documents and therefore generalizable to numerous other document corpora and research topics.

A system limitation is the dependency on available and reliable knowledge bases. Also, it is assumed that another obstacle for users with limited technical skills is the ability to create SPARQL queries. One solution to overcome this problem is to create interactive user interfaces (based on queries as presented) to help researchers explore the content in an easy and fluent manner without having to concentrate on designing the correct queries. Even though many solutions already exist to create generic user interfaces for a broad range of text, this topic is part of ongoing research and will be further discussed in section 5.2.4.

4.3 EXEMPLARY CONTENT ANALYSIS OF CONSTITUTION TEXTS

In the previous sections, the document corpus has been converted to RDF to create unique identifiers for each single unit on section level (cf. 4.1) which enables to query the generated data using SPARQL. Furthermore, RDF enables to enrich the provided content with knowledge from external knowledge bases, which was

demonstrated using DBpedia (cf. 4.2). In this section, it will be shown how Semantic Web technologies and Linked Data enable to extend text analysis in sociology by means of semantic annotation.

4.3.1 *Workflow*

The workflow in this section builds upon the work in the sections 4.1 and 4.2 and is visualized in Figure 18. The Python script converting the XML data into RDF produces two outputs. The `.ttl` file is, as discussed in the previous sections, directly integrated into Blazegraph. The other is a `.txt` file which is used for the semantic annotations. It contains all constitution texts sorted by edition and is imported into a Wordpress editing interface where it is annotated with DBpedia entities, using the *refer* tool. These annotations are stored as a text file containing RDFa and converted into NIF2 via a simple Python script. The resulting *.ttl* file is also integrated into Blazegraph, where both graphs are merged and queried together with SPARQL. As already elaborated in section 4.2.2.1, the existing content can be further enriched via federated queries.

4.3.2 *Annotation*

The rationale and overall methodology of semantically annotating text has been briefly described in section 4.3.2 and will be implemented in the exemplary use case in this section. First, the challenges and functionalities of semi-automated annotation interfaces in general will be introduced followed by a brief description of the *refer* system. Then, the annotation method and criteria are discussed followed by a number of use cases and queries to support the social scientist in the exploration of large textual documents.

4.3.2.1 *Annotation Interfaces*

Semantically annotating text with entities from a large knowledge base like DBpedia requires a well functioning user interface if the annotations are created manually or semi-automatically. The task of the user interface is to suggest possible entity candidates to the annotating user based on an input text. One of the major challenges is to present the entities in a way that users unfamiliar with Linked Data (so called lay-users) are able to make use of the interfaces. Lay-users typically have no further insight about what the content of a knowledge base is or how it is structured, which has to be considered when suggesting the entities the user should choose from (Shneiderman et al., 2016). An example to demonstrate the difficulty of this task is the annotation of the term '*Berlin*'. The entity `dbr:Berlin` as the capital of Germany could be considered as well as the historical reference to the city '*Berlin*' being `dbr:West_Berlin` and `dbr:East_Berlin` or the Person `dbr:Nils_Johan_Berlin`, and many more. Some entity mentions yield to lists of thousands of candidates which a human cannot survey quickly to find the correct one. Therefore, *autosuggestion* utilities are applied to rank and organize the candidate lists according to e. g. string similarity with the entity mention, or

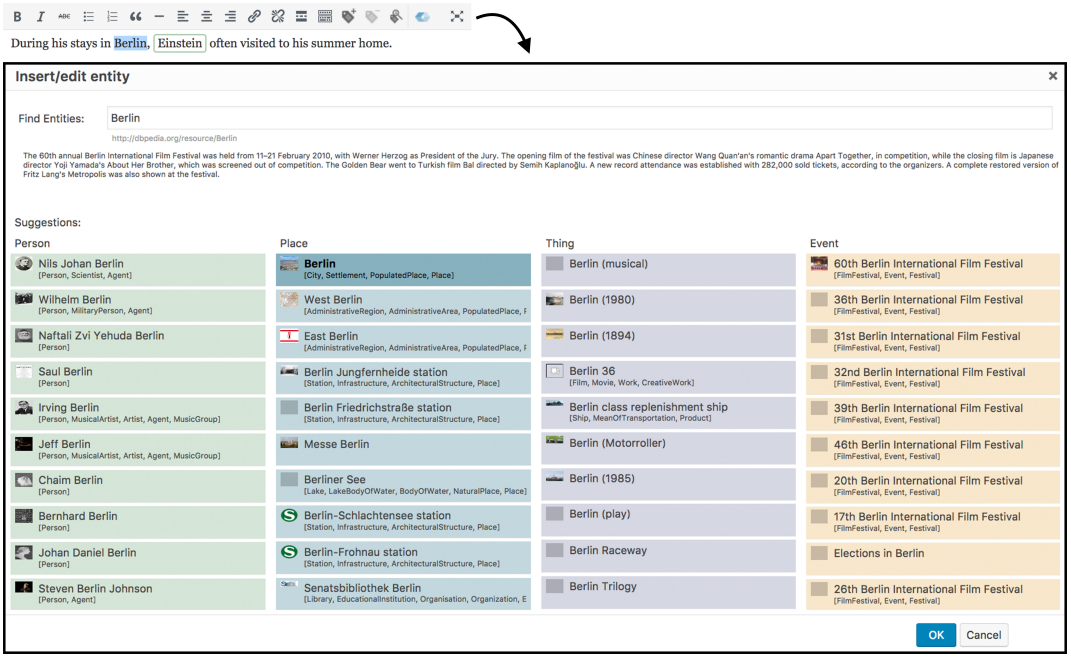

Figure 19: *refer* Inline annotation interface

general popularity of the entity (Osterhoff, Waitelonis, and Sack, 2012). Khalili and Auer, (2013), furthermore defined requirements for semantic content authoring tools. While authoring (and annotating) content, the user should face a minimum level of interruption and entity recommendations should be displayed without (or a minimum) level of distraction. The tool should assist (semi-)automated annotation in a useful manner and provide an easy correction of previous annotations (by another user or an algorithm). The user should be able to distinguish between manually created annotations and automated annotations. Furthermore the user interface should be customizable, depending on the visual layout of the respective publishing environment. There are numerous tools available to semantically annotate text.

The semantic editor and text composition tool *Seed* by Eldesouky et al., (2016), enables automated as well as semi-automated semantic text annotation in real-time. That means, while the author writes a piece of text, the system automatically performs NEL to reduce the work effort for the author. Despite the fact that this feature seems rather useful for blog authors, it is not applicable for this use case, because the text in the documents is already completed. The *Pundit Annotator Pro* by Morbidoni and Piccioli, (2015) also offers to create semantic annotations in text. The tool allows users to define their own properties and knowledge bases. However, it is assumed that in order to define own resources, the sociologist already has to have a profound knowledge about these knowledge bases beforehand. Furthermore, the annotator is not available for free. *dokieli* by Capadisli et al., (2017), is an annotation tool with an integrated support for social interactions. The goal is to employ a tool-agnostic generic format for semantic annotations, which are saved in an HTML+RDFa format. Furthermore, *dokieli* allows to save documents

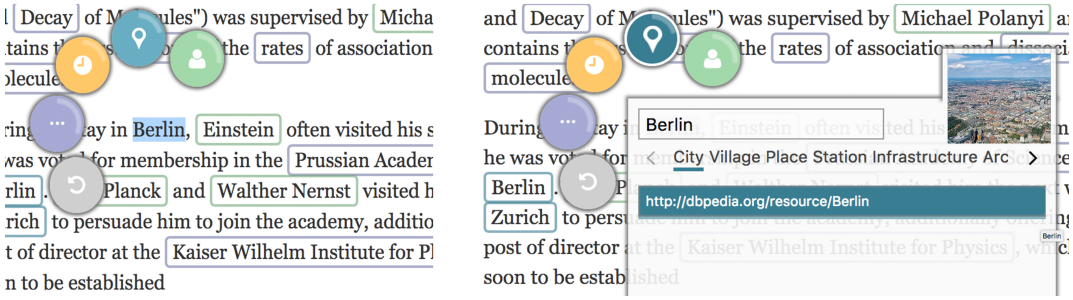

Figure 20: *refer* Modal annotation interface

to a user-defined data storage and includes the management of user authorization, document revisions and social communication.

refer

In order to annotate the constitution documents, the *refer* annotation system is used (Tietz et al., 2016). *refer* consists of a set of powerful tools focusing on NEL. It aims at helping text authors and curators to semi-automatically analyze textual content and semantically annotate it with entities contained in DBpedia. In *refer*, automated NEL is complemented by manual semantic annotation supported by sophisticated autosuggestion of candidate entities, implemented as publicly available Wordpress plugin[30]. Next to content annotation, *refer* also enables to visualize the semantically enriched documents in a navigation interface for content exploration. *refer* is chosen for this task, because it fulfills all of the criteria mentioned by Khalili and Auer, (2013). Furthermore, the entire system is available for download and is easily installed in the wordpress content management system, which increases the reproducibility of this work. A user study focusing on lay-users has shown that the *refer* annotation interface is easy to use and enables a sophisticated annotation process (Tietz et al., 2016).

For automated annotation, *refer* deploys the KEA-NEL (Waitelonis and Sack, 2016), which implements entity linking with DBpedia entities (Usbeck et al., 2015). The user can choose between a manual and automated annotation process. The *refer* annotator includes two configurable annotation interfaces for creating or correcting annotations manually: (1) the Modal annotator, shown in Figure 20 and (2) the Inline annotator, shown in Figure 19. The former builds upon the native TinyMCE editor[31] controls provided by Wordpress to trigger the display of suggested entities in a modal dialog window. The suggestion dialog starts with a text input field, which initially contains a selected text fragment and can be used to refine the search term. Suggested entities are shown below in a table-based layout, divided into four categories Person (green), Place (blue), Event (yellow) and Thing (purple). The window further includes a list of recently selected entities for faster selection of already annotated entities in the same text. The entity's DBpedia abstract and URI are displayed on mouseover. A click selects the entity and encodes the annotation RDFa markup, which is added to the according text fragment. The Inline annotator enables to choose entities directly in the context of a selected text

30 https://www.refer.cx/, last visited: July 28, 2018
31 https://www.tiny.cloud/, last visited: August 2, 2018

and is triggered automatically upon text highlighting. As a user study by Tietz et al., (2016), has shown, the Inline interface provides fast and simple means of semantic text annotation by minimizing the steps for the user. On the other hand, the Modal interface leaves more space for annotations and additional information and provides a parallel view of all available categories. Therefore, the Modal annotation interface has been chosen to annotate the text in the exemplary use case of this thesis.

4.3.2.2 *Annotation Method and Criteria*

As elaborated in the previous section, the *refer* annotation tool has been utilized for this use case. The original corpus provided by Knoth, Stede, and Hägert, (2018), was generated in the German language for a number of reasons. This provides some challenges regarding the annotation process. The automated analysis used with *refer* deploys the KEA-NEL which was created for the analysis of English text. Furthermore, KEA uses the DBpedia to create annotations, which is currently one of the largest Linked Data knowledge bases available. It is mainly generated from Wikipedia infoboxes and therefore provides information about an enormous variety of topics. This leads to the assumption that the automated analysis of this very specific domain of constitutional documents (even if provided in the English language) would be rather error prone, because not only entities related to constitutions are present in the candidate lists but also entities related to any topic part of DBpedia. Therefore the decision was made to manually annotate certain parts of the corpus with entities from the English DBpedia using *refer*. While this method seems rather cumbersome, it was chosen to be the best alternative for this use case. The way the Modal annotation interface was implemented enables an easy adaption. After an initial survey of the corpus, a candidate list of entities has been created and integrated into the interface to improve the annotation process.

Before the actual annotation task can start, annotation criteria have to be defined to ensure a consistent results, especially if the annotations are created collaboratively. First, it has to be defined what a named entity actually is that is worth to be annotated. In the use case of this thesis, it is assumed that rigid as well as non-rigid designators are important for the analysis, as discussed in section 3.2.9. The rationale here is to generate as much knowledge as possible from the text to be able to analyze the data from multiple perspectives. Further entity annotation criteria regard entity specificity and completeness.

Entity Specificity

Another annotation criterion noteworthy in this use case refers to the specificity of entities. While the level of entity specificity may differ for various annotation use cases, the annotations in this thesis are performed with the most specific entity in the knowledge base. If a sentence reads 'In 2018, the Winter Olympics took place in South Korea' the DBpedia entity to annotate the phrase '*Winter Olympics*' with is not `dbr:Winter_Olympic_Games` since it is not the most specific entity in this context in the knowledge base. The context reveals that the sentence refers to the 2018 Winter Olympics, therefore the phrase is annotated with the resource

dbr:2018_Winter_Olympic_Games. Specificity also regards word compounds. Considering the term '*John F. Kennedy Airport*', the entire term should be annotated, e.g. with the DBpedia entity dbr:John_F._Kennedy_International_Airport instead of two separate entities dbr:John_F._Kennedy and dbr:Airport.

Entity Completeness

The aspect of entity completeness means that anything that is named entity (according to the given definition) should be annotated. Even though the 20 given documents are too much content to annotate entirely in the course of this thesis, the articles and sections chosen to annotate have been annotated according to this completeness criterion. In the use case of this thesis, all DBpedia entities have been used for all annotations. Even though the DBpedia is currently one of the largest cross-domain knowledge bases and provides a sophisticated source for semantic annotations, it is clear that it cannot represent all real-world objects or abstract concepts known to humans. DBpedia is generated mainly from Wikipedia infoboxes and therefore depends on the coverage of content in Wikipedia. Unfortunately, Wikipedia suffers a so-called systemic bias which causes an unequally distributed interlinking of entities within the knowledge base. For example, in Wikipedia and thus also in DBpedia, entities about film and music are overrepresented and very well interconnected compared to other domains (Oeberst et al., 2016). The consequence for this use case is to expect that not all named entities related to this specific domain of constitutions are represented in the knowledge base, especially considering that some of the texts have been created in and before the early 20th century. Still, the annotations should be as complete as possible. Furthermore, it is important to measure which of the entities found in the text could not be linked to a DBpedia entity to ensure the validity and informative value of this approach. To overcome these shortcomings, a 'Not In List' (NIL) entity has been created and included in the Modal annotation interface of *refer*. Whenever the annotating user encounters an entity not available in the knowledge base, the NIL entity is used to assess the level of completeness of the annotations.

Temporal Roles

Another factor which requires some discussion, especially in the annotation of persons, is the acknowledgement of the entities' temporal role. That means, if a text in a Dutch constitution document edition from the year 2016 mentions a term like '*der König*' (the King), the term has been annotated with dbr:Willem-Alexander_of_the_Netherlands who was (and currently is) the king of the Netherlands. This task is known as temporal role detection and is part of current research in NLP. Significant advances in this rather young field of research have been accomplished by (Koutraki, Bakhshandegan-Moghaddam, and Sack, 2018), the topic is also tackled in a current research project led by the University of Zurich[32]. Even though the NLP and NEL technologies are constantly improving, this rather difficult task of disambiguation has not yet been solved in a way that it can be easily implemented in any domain. This aspect also affirmed the decision to proceed with a manual annotation process in this use case.

32 http://www.cl.uzh.ch/en/research/completed-research/hist-temporal-entities.html, last visited: July 29, 2018

Description	Count
Triples overall	155.804
Annotations overall	1.175
Distinct Entities overall	218
NIL Annotations overall	242
Annotations - 2016 edition	455
Annotations - 1983 edition	443
Annotations - 1884 edition	277

Table 4: Statistics of all triples and annotations generated in the dataset

4.3.3 *Annotation Statistics*

Parts of three constitutional documents have been semantically annotated with DBpedia entities according to the criteria and method discussed above. Table 4 shows the statistics of generated annotations in the dataset. Overall, 1.175 annotations have been created in three constitution documents using 218 distinct DBpedia entities. This means that on average, each DBpedia entity has been used around five times. Over all documents, 242 NIL annotations have been used, which means that around 20% of all named entities in the documents were not in the knowledge base (or could not be found). It can be concluded that solely using the DBpedia knowledge base is not enough for a profound annotation. The complete list of NIL annotation surface forms is presented in Table 7 in Appendix A. In order to advance in this matter, a next step may involve the analysis of all surface forms that no annotations have been created for. Domain experts may then (1) find another already existing knowledge base or (2) create their own knowledge base to enable a more complete annotation process.

4.3.4 *Content Exploration*

Using the *refer* tool, the texts have been enriched with RDFa annotations. The documents have further been converted to NIF2 and imported into Blazegraph. This section discusses how these annotations enable to explore the generated data. All prefixes used in the SPARQL queries of this section are shown in Listing 14

4.3.4.1 *Locating Entities in the Corpus*

Two datasets have now been imported into Blazegraph, the RDF data representing the entire structure of all documents and the data containing all NIF2 annotations (cf. Figure 18). SPARQL now allows to query both graphs in one query to exploit all data generated so far in the process. Listing 15 shows how to locate a specific entity annotated in the corpus on section level. This is possible because each section, article, paragraph, chapter and constitution have been assigned a unique URI to be references easily.

Listing 14: All prefixes used for the SPARQL queries in this section

```
PREFIX s: <https://github.com/tabeatietz/semsoc/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX co: <http://www.constituteproject.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dbc: <http://dbpedia.org/resource/Category:>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX itsrdf: <http://www.w3.org/2005/11/its/rdf#>
```

Listing 15: Query for the location of the entity `dbr:Netherlands` in the corpus

```
SELECT DISTINCT ?edition ?articletext ?contexttext
WHERE
{
  ?phrase itsrdf:taIdentRef dbr:Netherlands ;
          nif:referenceContext ?context .
  ?context co:text ?contexttext ;
          co:parent ?section ;
          co:isSectionOf ?constitution .
  ?section co:parent ?article .
  ?article co:text ?articletext .
  ?constitution s:edition ?edition .
  }
ORDER BY ?edition ?article
```

4.3.4.2 *Frequency Analysis*

A popular query in text analysis to start out with is a frequency analysis (Mayring, 2015). It is assumed that a frequency analysis becomes especially important when comparing document changes over time, as it is the case in this exemplary use case. The analysis can function as a metric to find out when a certain term as first been introduced into a constitution and accordingly if the usage of this term has increased or decreased over a certain time period. The analysis of the representation of marriage and its meaning for the state can be initiated through a frequency analysis. The SPARQL query in Listing 16 counts all annotation of the entity `dbr:Marriage` for all annotated constitutions separately. The results show that the annotation has been used five times in the edition of 1884 and three times each in the editions of 1983 and 2016. Of course, these results are by no means complete, since only small portions of the documents have been annotated. The usage of semantic annotations for a frequency analysis enables to query for specific entities (and also entity categories) regardless of the surface form, which eliminates the problem of covering all possible synonyms in the query.

Listing 16: Query counting all annotations of the entity `dbr:Marriage` in the annotated data

```
SELECT DISTINCT ?constitution ?edition (COUNT(?term) AS ?annocount)
WHERE
     {
     ?term itsrdf:taIdentRef dbr:Marriage ;
          nif:referenceContext ?context .
     ?context co:isSectionOf ?constitution .
     ?constitution s:edition ?edition .
     }
GROUP BY ?constitution ?edition
```

4.3.4.3 *Contingency*

The goal of a contingency analysis is to find out in which relation a specific terms are used in a text document. Osgood, (1959), has been among the first to use this method in content analysis. In the context of the research of constitution documents, a contingency analysis is especially interesting, because it gives insights about how certain topics are modeled. The role of religion in a state has been widely researched (e.g. (Lagler, 2000)). As previously discussed, constitutions mirror how the society of a state is modeled. Therefore, analyzing the development of terms related to religion in constitution documents is plausible. Listing 17 shows what a contingency analysis using SPARQL may look like. The query selects all entities, which have been annotated in the same section as the entity `dbr:Religion`. Thereby, the focus entity itself should not be part of the result list (line 8-10). Table 5 shows the results. According to the annotations made, the entity `dbr:Religion` has been most used in the context of educational topics. Of course, also for this analysis the results are by no means representative since only a small portion of the text has been annotated. Uncommenting the patterns in line 17 and 18 allows to sort the contingency of the entities by constitution edition to explore the changes over time.

4.3.4.4 *Exploiting the Graph Structure in DBpedia*

In section 4.2.2.1, DBpedia has been used to enrich the constitution documents with information about the reigning monarch in each document edition. In this section, the document text has been directly annotated with content from DBpedia. One significant asset of annotating research data in sociological text analysis with semantic entities from an external knowledge base is the possibility to enrich the existing content with external knowledge as will be discussed in this section.

Especially the annotation of text with temporal roles as briefly elaborated in section 4.3.2.2 provides means to use the knowledge in DBpedia for further analysis. One example is the annotation of monarchs in the constitutions with respect to their temporal roles:

Der König$_{\text{dbr:Beatrix_of_the_Netherlands}}$ *ist unverletzlich.*

Listing 17: Query for all DBpedia entities annotated within the same section as the entity `dbr:Religion`

```
SELECT DISTINCT ?entity (COUNT(?entity) AS ?num)
WHERE
{
  ?phrase itsrdf:taIdentRef dbr:Religion ;
          nif:referenceContext ?context .
  ?phrase2 nif:referenceContext ?context ;
          itsrdf:taIdentRef ?entity .
  FILTER NOT EXISTS {
          ?phrase2 itsrdf:taIdentRef dbr:Religion .
          }
  # ?context co:isSectionOf ?constitution .
  # ?constitution s:edition ?edition .
}
GROUP BY  ?entity
ORDER BY DESC(?num)
```

The text above was taken from the 1983 constitution edition. In that year, Queen Beatrix was the current monarch of the Netherlands, as revealed by the query in Listing 13 in section 4.2.2.1. Therefore, the term *König* (King) has been annotated with the respective DBpedia entity. To use DBpedia for knowledge enrichment first requires to know which information the knowledge base holds about the entity in focus and how it is organized. The information is stored in DBpedia in form of triples, a visual representation of these information is provided via a HTML webpage[33]. On the top of the page there is a short abstract about Beatrix, which was automatically retrieved from the Wikipedia page of the former queen[34]. A table is located directly beneath the abstract that lists all of the triples in DBpedia in which Queen Beatrix is the subject. The column on the left represent the property connected to the subject. The right column lists the respective objects or values, which are connected to Beatrix. The DBpedia page lists information like the birth and death dates, family members, religion and succession. The information most vital for the analysis in this chapter is shown in Figure 21. The properties `dct:subject` and `rdf:type` connect the subject to certain classes and categories, which were retrieved from Wikipedia categories or from external knowledge bases, like Wikidata. These classes and categories help to organize each entity in formal structures (ontologies). For instance, the categories show that Beatrix is a Dutch monarch and belongs to the House of Orange-Nassau. Furthermore, Beatrix is of type Person and belongs to a royal family. Clicking on one of the categories, e.g. `dbc:Dutch_monarchs`, reveals a list of all entities, which are also connected to this category, e.g. `dbr:Willem-Alexander_of_the_Netherlands`. This organization of knowledge creates an enormous network of information, which can be exploited via SPARQL queries. In the area of Information Retrieval (IR), these semantic structures are utilized in numerous applications, including semantic search, recommender systems and topic detection (Waitelonis, 2018). Also for

33 http://dbpedia.org/page/Beatrix_of_the_Netherlands, last visited: August 15, 2018
34 https://en.wikipedia.org/wiki/Beatrix_of_the_Netherlands, last visited: August 15, 2018

entity	num
http://dbpedia.org/resource/World_view	6
http://dbpedia.org/resource/Law	4
http://dbpedia.org/resource/Liberty	4
http://dbpedia.org/resource/Education	4
http://dbpedia.org/resource/Freedom_of_thought	2
http://dbpedia.org/resource/Race_(biology)	2
http://dbpedia.org/resource/Sex	2
http://dbpedia.org/resource/Netherlands	2
http://dbpedia.org/resource/Discrimination	2
http://dbpedia.org/resource/State_school	2
http://dbpedia.org/resource/Private_school	2
http://dbpedia.org/resource/Government_spending	2
http://dbpedia.org/resource/Requirement	2

Table 5: Result of the Query in Listing 17

text analysis in sociology, this knowledge can be used to give existing data more meaning.

In constitution texts, the roles of persons (e.g. Prime Ministers or monarchs) are not defined according to their sex, as the example above shows. That means, even if Queen Beatrix was the Queen of the Netherlands in 1983, the respective constitution text does not refer to her as the queen ("die Königin") but in the male form ("der König"). However, when analyzing constitutions especially in the context of sociological gender studies (e.g. (Crawford, 2009)), the information whether the king was actually a king or at the time a queen may be vital. For this purpose, DBpedia's graph structure helps to aggregate the content accordingly to answer the question:

1. Which constitution editions, articles and sections are valid under a reigning female Dutch monarch?
   a) Which edition, article and section has been annotated with an entity ...
   b) under the constraint that this entity belongs to the category of Dutch monarchs in DBpedia ...
   c) and under the constraint that for the entity a gender was specified in the knowledge base which can only be female.

Accordingly, the query in Listing 18 selects all constitution editions, articles and sections that contain a semantic annotation with a female Dutch monarch (line 16). The results of the query are shown in Table 6. For simplicity reasons, only a portion of the results is shown here. This means that the annotations allow to aggregate the previously existing content in a way that exploits the organization of knowledge in an external database.

Of course, in this exemplary use case, only the constitution of the Netherlands has been annotated and analyzed and it may be simply surveyed whether the

edition	parenttext	articletext	monarch	string
1983-02-17	§ 2. König und Minister	Art. 42.	dbr:Beatrix_of_the_Netherlands	(1) Die Regierung besteht aus dem König und den Ministern.
1983-02-17	§ 2. König und Minister	Art. 42.	dbr:Beatrix_of_the_Netherlands	(2) Der König ist unverletzlich; die Minister sind verantwortlich.
1983-02-17	§ 2. König und Minister	Art. 47.	dbr:Beatrix_of_the_Netherlands	Alle Gesetze und Königlichem Erlasse werden vom König und von einem oder mehreren Ministern oder Staatssekretären unterzeichnet.
...	...	...	...	...

Table 6: Shortened result of the query in Listing 18

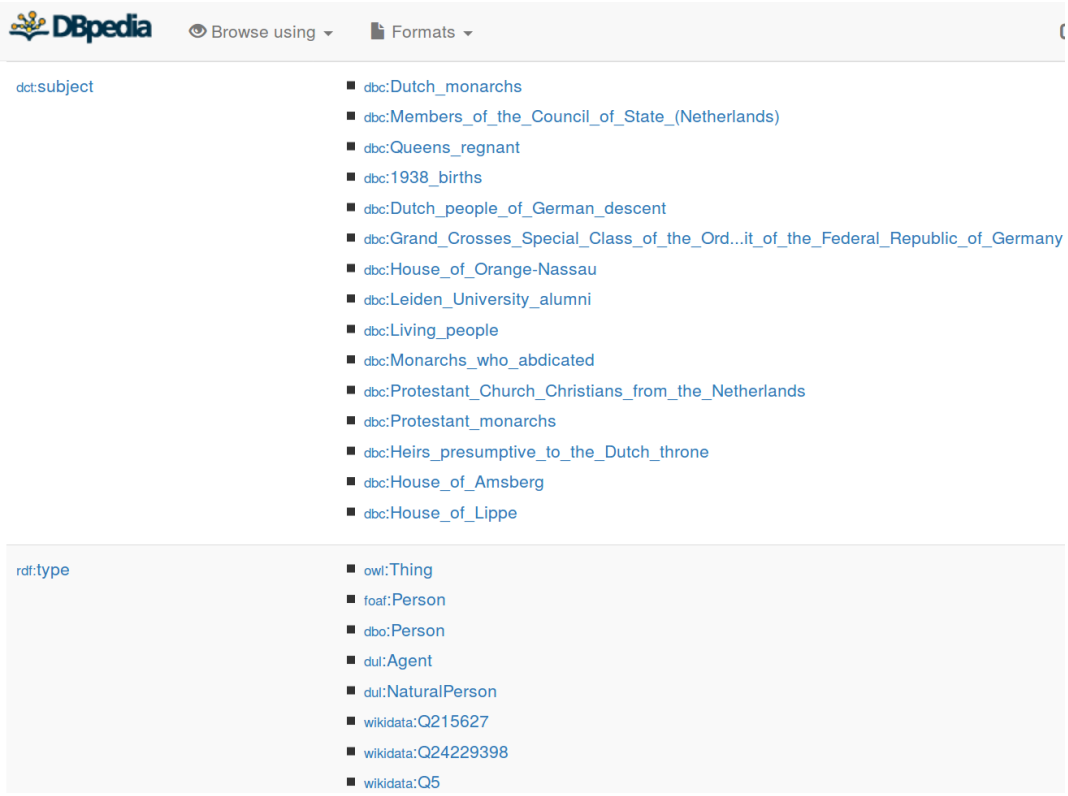

Figure 21: Part of the HTML page about Queen Beatrix in DBpedia

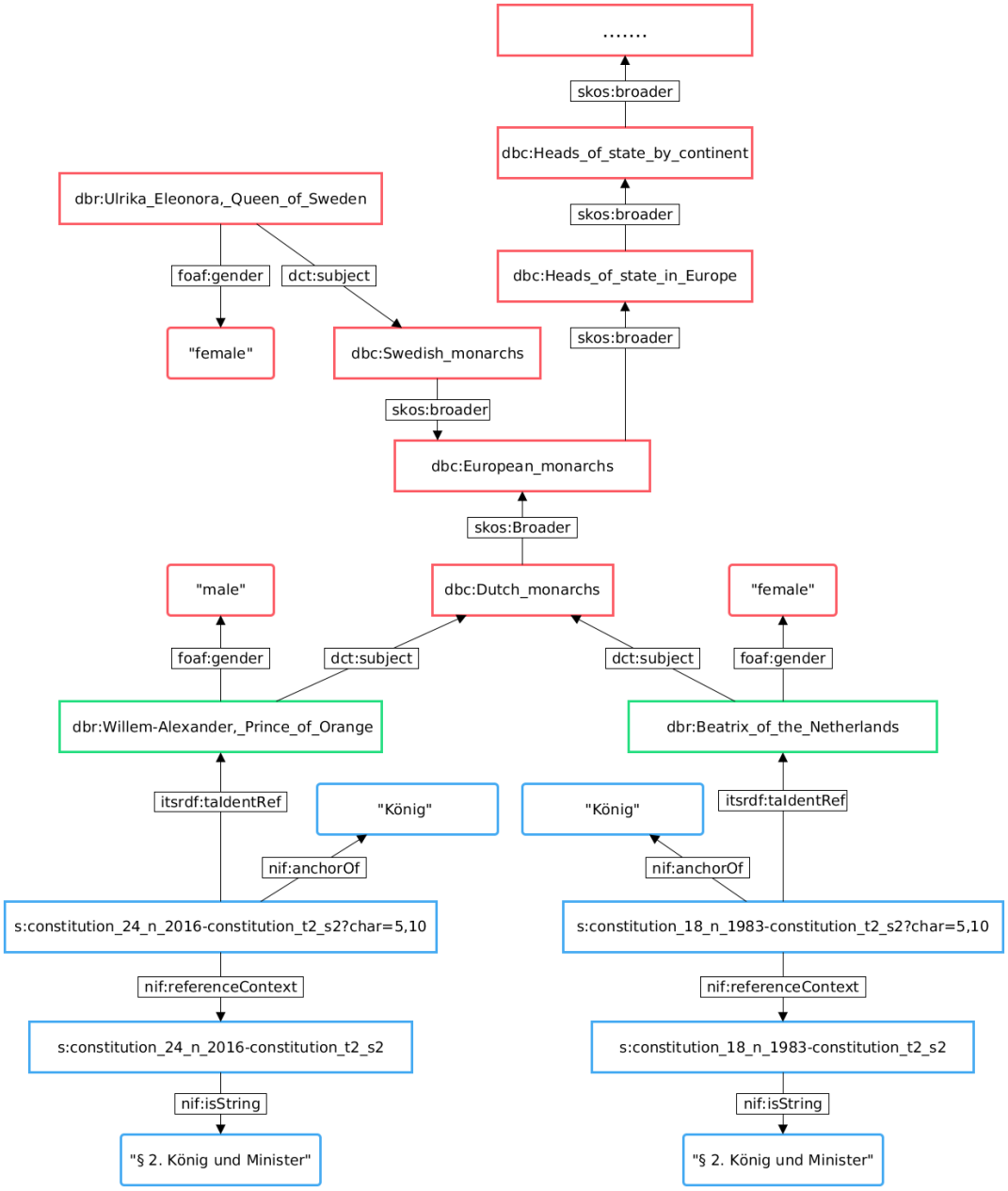

Figure 22: Example use case of the DBpedia category `dbc:Dutch_Monarch`

Listing 18: Federated Query for all constitution editions, articles, and sections that contain a semantic annotation with a female Dutch monarch

```
SELECT DISTINCT ?edition ?parenttext ?articletext ?monarch ?string
WHERE  {
  ?phrase itsrdf:taIdentRef ?monarch .
  ?phrase nif:referenceContext ?context .
  ?context nif:isString ?string .
  ?context co:parent ?section .
  ?section co:parent ?article .
  ?article co:text ?articletext .
  ?article co:parent ?parent .
  ?parent co:text ?parenttext .
  ?context co:isSectionOf ?constitution.
  ?constitution s:edition ?edition .

  SERVICE <http://dbpedia.org/sparql> {
     ?monarch dct:subject dbc:Dutch_monarchs ;
                foaf:gender "female"@en .
     }
  }
ORDER BY ?edition ?articletext
```

monarchy was reigned by a king or queen in a specific time period. One of the benefits of annotating the content in the described way is the easy adaptability in case more content is added to the corpus. It is easy to imagine, that the analysis of constitution text will in the future not only affect one country at a time but especially the comparison with other countries will be of significant value. Using the category `dbc:Dutch_monarchs` helps to do that. The category does not only connect monarchs of the Netherlands with each other but it is also connected to more general categories, e.g. `dbc:European_monarchs` via the property `skos:broader`. This connection enables to bring Queen Beatrix in the context of other monarchs throughout Europe. The example in Figure 22 visualizes how it works. In the figure, the elements existing in the corpus are marked by a blue frame. The elements from DBpedia which are used to extend the existing knowledge are red and the entities connecting both are marked by green frames. It shows that in case the corpus of constitution texts was extended by the Swedish constitution, the category hierarchy allows to easily adapt the query to match not only Dutch monarchs but also Swedish monarchs at the same time.

4.3.4.5 *Infobox Visualization*

The RDFa enrichment created with *refer* in the corpus enables to visualize additional information about annotated entities directly within the context of the document. When the annotated text is published within Wordpress, the annotations are immediately presented in the document's HTML code. Each annotated entity is indicated by thin, semi-transparent, colored lines. The colors indicate whether the entity is of type Person (green), Location (blue), Event (yellow) or Thing (purple). On mouseover, a so-called *infobox* as shown in Figure 23 is displayed below the an-

Figure 23: Infobox visualization of former Prime Minister Ruud Lubbers

notated text fragment. It contains basic information about the entity derived from DBpedia, e.g. a thumbnail and additional data from the entity RDF graph put in a table layout. The visual design and content of infoboxes varies per category and allows the user to gather basic facts about an entity as well as relations to other entities (Tietz et al., 2016).

Sociologists, when exploring a document corpus of interest (which has previously been annotated) can make use of these infobox visualizations to learn more about the data in front of them without having to leave the original context of the text. This can support a better understanding of the text, for instance if a certain term is unknown to them or, as shown in Figure 23, they want to learn about the temporal roles of entities. In the example, the user can learn that in the constitution edition of 1983, Ruud Lubbers was the Prime Minister of the Netherlands.

Of course, this visualization is only a preliminary version of what sociologists could benefit from in the future, because visualizing the document corpus for sociological research in Wordpress as it is done with *refer* may not be sufficient in some cases. Nevertheless, RDFa is versatile and could be embedded in any HTML document. Furthermore, the current system integrates solely content from DBpedia into the infoboxes. It always depends on the use case and document corpus whether this or another knowledge base should be used for this purpose of content enrichment.

4.3.5 *Brief Summary*

In this section, the meaning and potential of semantic text annotations has been evaluated for text analysis in sociology on the foundation of the use case introduced in section 4.1.1. The contributions of this section include 1.175 semantic annotations of three constitution documents, a Python script to convert a text document containing RDFa into NIF2, SPARQL queries and scenarios to exploit the generated data as well as an in-depth discussion of results.

The presented approach is generalizable to a wide range of texts. The analysis of textual documents on the bases of semantic annotations proved to be useful for the given data, especially the exploitation of external knowledge and the external organization of knowledge. The method as presented is highly adaptable and the data model makes it possible to add more data of different countries and time periods without significant changes in modeling or querying the data.

The meaning of these results for research transparency and data re-usability along with lessons learned and a detailed discussion of limitations of this approach will be discussed in chapter 5.

5 DISCUSSION

Chapter 4 provided the methods and possibilities which computer-assisted text analysis in sociology could benefit from when applying Semantic Web and Linked Data technologies and standards. In this chapter, a thorough discussion of the achieved results is presented. In the chapters 1 and 2 it was questioned whether these technologies could also effect research transparency and reproducibility as well as the re-usability of the sociological research process and data when working with text documents, which will be discussed in the following section. Furthermore, system limitations and future work will be highlighted.

5.1 TRANSPARENCY AND RE-USABILITY

As defined in chapter 2, a research process is considered transparent, if the presented content and ideas are based on clear and reliable accounts. In this thesis it has been demonstrated what these accounts of transparency may look like in sociological text analysis by means of Semantic Web and Linked Data technologies. Publishing research data in RDF, as accomplished in section 4.1 allows to reference each single unit of a document separately (e.g. the single sections of constitutions). Transparency also includes the clear definition of explicit models (ontologies) structuring knowledge and defining relationships between concepts and individuals. Semantically annotating text with entities from one or more knowledge bases (as accomplished in section 4.3) furthermore increases the transparency of the process. Thereby it is not only clear which concept has been annotated in which specific section of a document, but NIF2 makes it possible to create this reference up to the specific surface form and characters used.

The re-usability of data in sociological text analysis has been demonstrated in various ways in this thesis. Especially in section 4.1 it has been demonstrated how the used standards enable to easily utilize and build upon other researchers' ontologies. External ontologies can be used and extended by own concepts to fit the intended use case. Thereby, the control over the own data and the responsibility to what the final model will include remains with the researcher. Knoth, Stede, and Hägert, (2018), created the constitutional XML formats used in this thesis from scratch in a cumbersome work, because these important data dealing with European constitutions were simply not available on the Web in a machine readable format. Storing and publishing these data as RDF enables anyone to re-use these data and to create queries (as demonstrated in sections 4.2 and 4.3) using the standardized SPARQL query language as opposed to proprietary XML parsers, for instance.

The semantic annotations made in section 4.3 also contribute to the re-usability of the text data. The annotation criteria discussed in section 4.3.2 help to understand the context in which these annotations have been created. The annotations

may be re-used in form of RDFa, useful for HTML pages, or NIF2 useful for querying and further adaptation. If the annotations have been created thoroughly, they can furthermore function as a gold standard for computer scientists to improve and test NEL systems.

5.2 LIMITATIONS AND FUTURE WORK

In this section, the limitations of the presented approach and future work will be discussed, with an emphasis on automating the annotation process, challenges of especially temporal role annotations, and insufficiencies of existing knowledge bases, and interactive user interfaces.

5.2.1 *Feasibility and Process Automation*

Creating the semantic annotations on German language text as discussed in section 4.3.2 is not feasible for a large corpus. Therefore, it is considered to be absolutely necessary to automate the process in future work. One possibility is to use an automated NEL system to annotate the text first and correct annotation mistakes manually by means of a user interface similarly, which applies the blended reading methodology described in chapter 2. However, as already mentioned, Luhmann, (1984) warned that distance can only be kept if researchers are able to rely on their own instrument. For a fine grained NEL approach on German language text, this requirement is not given, yet. For English language text, this blended reading approach is already possible with decent results, for instance with the *refer* annotation system. The benchmark system GERBIL, created by Usbeck et al., (2015), and extended by Waitelonis, Jürges, and Sack, (2016), acknowledges that different annotation systems do not perform equally well on every text corpus. Some perform better on person entities, some on location entities and so on. The framework enables to detect strengths and weaknesses in every system. However, for German, this is not as easy, because the annotation systems in GERBIL are not configured for German language text. The text used in the document corpus is in German, a real world research example in Sociology. In future work, NEL systems should be improved to also perform better on German language text (and certainly other languages as well) to enable a more efficient use in sociological text analysis. One prominent automated NEL system for German language text is DBpedia Spotlight (Daiber et al., 2013). However, a few initial experiments with the system quickly exposed that the annotation quality is too low to efficiently work with the system. Therefore it was eliminated from the research process. Another reason to continue with manual annotations was the importance of linking temporal roles in the text, which was understood as vital for this work. So far, there is no NEL system available which allows to annotate these temporal roles in German with a decent quality.

Apart from the temporal role disambiguation in this work, one challenge this corpus may provide is the changing style of language in the documents over time. Since the data of this use case deals with the same domain and provides text from 1884 to 2016, it may function as a dataset to analyze these changes in order to

improve NEL systems in the future. The difficulties in annotating the documents within a reasonable amount of time resulted in relatively few annotations. Even though the provided 1.175 annotations have proven enough to perform queries, receive exemplary results and demonstrate the possibilities of these technologies for sociology, it was not possible to perform a representative study on their basis, which is considered a shortcoming of this thesis.

5.2.2 *Annotation Challenges*

During the annotation process of the constitution corpus, a few challenges occurred, worthy of a brief notion. For instance, one annotation criterion has been the acknowledgement of the entities' temporal roles. In the 2016 constitution edition of the Netherlands, the second chapter reads, that the oldest child is next in the line of succession of the king. According to the temporal role criterion, the *oldest child* should be annotated with the actual person entity to be precise (in this case the Princess of Orange Catharina-Amalia). However, in case of a tragic sudden death of the princess the *oldest child* would be her younger sibling and thus, the annotation would become incorrect. This challenges the validity of temporal role disambiguation for future events.

Another challenge in the annotation process was the imprecise definition of some named entities. For example the term *Königswürde* (translated literally: royal dignity). In German, this term is quite ambiguous, it has no real definition nor is it anything tangible. In English texts of the Dutch constitutions, the term has been entitled simply as *throne* or *title of the throne* and so on. But no consistent term was used. Therefore annotating it with a specific entity from DBpedia was not possible.

In other cases, it is assumed that knowledge in the domain of politics and law was needed to provide the correct annotations. The terms in question include (in German) *Gesetzesvorlage*, *eine Vorlage*, *etwas vorlegen*. All of the three terms appeared in the text and could refer to a draft bill, but it was not always clear taking into account the given context.

5.2.3 *Overcoming Knowledge Base Insufficiencies*

Another challenge during the annotation process was to overcome the difficulty that not all concepts could be covered by the knowledge base. These insufficiencies will be discussed in this paragraph.

For this case, the NIL entity has been created and added to the *refer* annotator which has proven to be quite efficient. Table 7 in Appendix A provides a list of all 242 occurrences. The NIL entity enabled to measure the feasibility of DBpedia as a knowledge base for the given use case. As already described in section 4.3.3, about 20% of all annotations used NIL entities. The NIL annotations furthermore allowed to define a clear limit for all automated NEL regarding this given use case. If an entity is not present in the knowledge base, the system will never de-

tect it, therefore the recall can never be at 100 percent[1]. In this use case, the text corpus dealt with constitutions, country specific information and facts about state leaders. These topics are generally well represented in DBpedia and the way these information are structured does not leave much room for discussion. However, the texts analyzed by many sociologists are not always like this. Often, text corpora are analyzed that deal with human to human communication (e.g. conversations) and the topics to investigate are divers. One example is the investigation of family structures and the role of the woman in the family. For cases like this it has to be assumed that no knowledge base available will represent these information sufficiently at the moment. Next to the problem of availability, another issue is how the knowledge is formally defined. That means, a sociologists may have a completely different understanding of the formal definition of the concept *family* than a psychologist or a political scientist. The facts that represent the concept *family* and how they are related may be prioritized in a completely different way in various domains. That means, solely relying on third party knowledge bases from other domains may bring the issue of varying concept definitions into the text analysis than anticipated by the sociologist.

One way to overcome this challenge is to start creating own knowledge bases or even single concepts from the sociological perspective. The methods to achieve this have been described in this thesis. Creating a knowledge base, similarly to creating concepts for coding text in sociology could in fact become part of the research process itself. That is because even though a general sociological perspective on a concept may exist, each researcher may add own ideas. Concepts could also be modeled according to different schools of thought, an interesting possibility for future work.

Modeling own concepts in form of formal and structured ontologies have been widely discussed in this thesis and include first and foremost research transparency and re-usability of the individual research process. Furthermore, this gives sociologists the chance to become a part of the Semantic Web and its future and share their domain knowledge with the community.

5.2.4 *Content Exploration and Interactive User Interfaces*

The sections 4.2 and 4.3 have introduced means for sociologists to query the generated RDF data and to explore its content. However, issuing a SPARQL query for each of these exploration tasks is not really sufficient in the long term. One reason is that a requirement is to learn SPARQL. Even though it can be assumed that when the general idea of RDF has been understood, the query language will not be a major issue, it prevents the researcher so merely focus on the document and its exploration. A solution to enable an easier content exploration is developing interactive user interfaces in which the discussed SPARQL queries merely run in the background. Another problem of using SPARQL (or any other query language) for content exploration that the researcher will only find exactly what they

1 In Information Retrieval, recall is defined as is the proportion of relevant documents retrieved as opposed to precision which is the proportion of retrieved documents which are relevant (Van Rijsbergen, 1979)

are looking for. However, often the most interesting relationships and contexts are explored by means of serendipity, which can be enabled by intelligent interactive user interfaces. *Serendipity* refers to the process of finding valuable information or facts which have not been sought for. This phenomenon is known and widely utilized in information retrieval (Foster and Ford, 2003; Waitelonis and Sack, 2012).

Visualizations and user interfaces enabling the exploration of textual documents is part of current research. Especially using Semantic Web and Linked Data technologies have delivered promising results. Rahman and Finin, (2018), developed an unsupervised method based on deep learning to explore large structured documents like business reports or proposal requests. They created an ontology to capture not only the general purpose semantic structure, but also domain specific semantic concepts. While the method seems promising, it is (so far) missing an exploration feature for non-technical users. Also it is unknown if the method is generalizable to other domains, e.g. constitution documents. Latif, Liu, and Beck, (2018), follow a completely different approach. The authors developed a framework that takes text containing markup, a related dataset, and a configuration file as inputs and produces an interactive document. The result enables to receive further details, visual highlighting, and text comparison. However, the framework does not enable to create aggregations over a large set of documents to receive information on their overall structure.

However, as many solutions exist for this purpose, there is no out-of-the-box one for all solution available at the moment. Another method to make use of but useful interactive interfaces to explore the discussed documents is using simple libraries to create new interfaces, almost from scratch. Data Driven Documents (D3)[2] is a JavaScript library for manipulating documents based on data. The library is well documented and there are numerous examples on the Web which can be reused and further modified. Creating an application to interactively explore promising interdisciplinary research project for future work.

2 https://d3js.org/, last visited: July 15, 2018

6

SUMMARY AND CONCLUSION

Text analysis is an important means in sociology to analyze social reality and has been supported by computers since the 1960s to cope with the ever growing amount of text. During the last decade, Natural Language Processing methods have been introduced in the research field to uncover linguistic structures and semantic associations. Even though various methods to access and analyze these data have established in sociology over the years, no standardized and systematic means of analyzing these complex material have been developed. Furthermore, the problems of research transparency and re-usability have been highlighted. Often, journals do not practice a standardized method to publish sociological research data which negatively affects their integrity and interpretability.

Semantic Web and Linked Data technologies provide a broad set of standards and tools which have been widely utilized in a number of scientific domains. Therefore it could be assumed that they also support text analysis in sociology as well. For this reason the following research question to be answered in this thesis was issued: To which extend can state-of-the-art Semantic Web and Linked Data technologies, standards and principles support computer-assisted text analysis in sociology to improve research transparency and data re-usability.

In order to answer this question, the technologies and standards created within the Semantic Web which relate to text analysis have been discussed briefly in chapter 3. This introduction was followed by the main part of this thesis in chapter 4. On the foundation of the use case dealing with constitutional texts from the Netherlands from 1884 to 2016, it has first been discussed how Linked Data is generated and published on the Web, closely following the recommendations by the W3C. The following two sections 4.2 and 4.3 gave a detailed overview about how the generated data can be exploited in the context of sociological text analysis. This was attempted by first focusing on the structure of the analyzed documents, followed by the analysis of the textual content. The text has been (in parts) semantically annotated with DBpedia entities. Before the annotation process, detailed annotation criteria have been defined. The exemplary analysis itself has been performed by means of SPARQL queries, which also included enriching the existing content with external knowledge via federated querying. The chapter has been concluded with an in-depth discussion of the presented methods in chapter 5. It has been discussed how exactly Semantic Web and Linked Data technologies support research transparency and re-useability. Also, the limitations of the presented approach have been listed, which mainly regard automated Named Entity Linking systems, challenges semantic annotations, insufficiencies of current knowledge bases for social scientists and user interfaces for content exploration. Further contributions of this thesis include two Python scripts for (1) converting XML to RDF and (2) converting an RDFa text file into NIF2.

This thesis closes with a call to action for sociologists, which emphasizes how sociologists can contribute to the effort of the Semantic Web and Linked data community to benefit from its current achievements and future possibilities.

6.1 CALL TO ACTION

A take away message of this thesis for sociologists is that the Semantic Web is a community effort. Researchers (e.g. sociologists in the field of computer-assisted text analysis) who want to benefit from the possibilities, principles and standards and technologies this community offers, have to engage in this effort, which has also been emphasized by Halford, Pope, and Weal, (2013). This thesis has shown that many of the current knowledge bases, interfaces, and analysis tools are not yet mature enough for a sophisticated textual analysis in sociology from start to finish on any text in any language. However, the domain knowledge sociologists can bring into the Semantic Web is immense. It can be assumed that interdisciplinary efforts which also include sociologists more can results in a significant improvement of these insufficiencies. That is because only sociologists know which concepts in knowledge bases are exactly needed to cover important aspects of text analysis and only they know what the requirements for an interactive user interface made for sociological research are, to explore textual content and find even things they have not yet been looking for. One result of this work is the highlighting of various topics, tools and principles that offer sociologists appropriate opportunities for their own sociological contributions to the Semantics Web. The Semantic Web has proven to be highly valuable for life sciences, medicine, and is beginning to be more and more incorporated in digital humanities. Hopefully, in the future sociologists will engage in this ever growing community as well.

APPENDIX

APPENDIX

Listing 19: RDF Turtle depiction of the RDF graph snippet visualized in Figure 13

```
@prefix s: <https://github.com/tabeatietz/semsoc/> .
@prefix co: <http://www.constituteproject.org/ontology/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

s:constitution_24_n_2016-constitution a co:Constitution ;
    co:hasConstName "Verfassung des Königreiches der Niederlande" ;
    co:isConstitutionOf co:Netherlands_the ;
    s:edition "2016-11-04"^^xsd:date ;
    co:isCreatedIn "2016" .
s:constitution_24_n_2016-constitution_t2 a co:Section, s:Chapter ;
    co:isSectionOf s:constitution_24_n_2016-constitution ;
    co:rowType co:title ;
    co:text "Hauptstück 2 - Regierung" ;
    co:header "2" ;
    co:sectionID "131’" .
s:constitution_24_n_2016-constitution_t2_s2 a co:Section, s:Paragraph ;
   co:isSectionOf s:constitution_24_n_2016-constitution ;
   co:rowType co:title ;
   co:parent s:constitution_24_n_2016-constitution_t2 ;
   co:text "2. König und Minister" ;
   co:header "2" ;
   co:sectionID "235" .
s:constitution_24_n_2016-constitution_t2_s2_a1 a co:Section, s:Article ;
    co:parent s:constitution_24_n_2016-constitution_t2_s2 ;
    co:rowType co:title ;
    co:isSectionOf s:constitution_24_n_2016-constitution ;
    co:text "Art. 42." ;
    co:header "42" ;
    co:sectionID "236" .
s:constitution_24_n_2016-constitution_t2_s2_a1_s1_title a co:Section, s:Section ;
    co:isSectionOf s:constitution_24_n_2016-constitution ;
    co:parent s:constitution_24_n_2016-constitution_t2_s2_a1 ;
    co:rowType co:title ;
    co:header "1" ;
    co:sectionID "237" .
s:constitution_24_n_2016-constitution_t2_s2_a1_s1 a co:Section, s:Section ;
    co:isSectionOf s:constitution_24_n_2016-constitution ;
    co:parent s:constitution_24_n_2016-constitution_t2_s2_a1_s1_title ;
    co:rowType co:body ;
    co:text "(1) Die Regierung besteht aus dem König und den Ministern." ;
    co:sectionID "238" .
```

anchor	count	anchor	count
Königswürde	18	Mitglied dieser Organe	2
Amt	18	Einschränkungen	2
Amtes	10	Mitglieder allgemeiner Vertretungsorgane	2
Stimmen	8	schriftlich Gesuche	2
Königlichem Erlaß	8	Veranstaltungen	2
Kammern	7	Sitten	2
Auftrag	6	Störungen	2
Regenten	6	Ministerrat	2
Regent	6	Vorsitzenden der Versammlung	2
Qualität	4	Vorlage	2
Bedingungen	4	ungeborenes Kind	2
Staatsrat	4	Mitglied	2
Ernennung	3	Mitgliedern	2
Betreten	3	Personensteuer	2
Ausnahmen	3	Übertragungs	2
dritten Grade	2	Schenkungssteuer	2
Amtsübernahme	2	Benachrichtigung	2
Einheitlichkeit	2	Münzrecht	1
Königlichem Erlasse	2	Fremde Orden	1
Königlichen Erlasse	2	Erlaubnis	1
Reinigungseid	2	Untertanen	1
Reinigungserklärung	2	obersten Gerichtshofes	1
Reinigungsgelöbnis	2	Kapitalverbrecher	1
Besitzungen	2	Niederschlagung	1
Weltteilen	2	Dispensationen	1
öffentlichen Ordnung	2	Landstreitmacht	1
Eingesessenen	2	kolonialen Finanzbehörden	1
Schwester	2	Kollegien	1
Bewohners	2	Bestimmungen	1
Betretens	2	Derzelver Kreis	1
Bewohner	2	Bürgerschaftsrecht	1
Mitteilung	2	Körperschaften	1
Fernmeldegeheimnis	2	Zuständigkeiten	1
Rechts	2	Einwohnern	1
Verwaltungssachen	2	Karl Georg August	1
Mitbestimmung	2	Repräsentationsrecht	1
Existenzsicherheit	2	Präsentationsrecht	1
Lebensunterhalt	2	Brüder	1
Schutz	2	männlichen absteigenden Linie	1
Umwelt	2	männliche absteigende Linie	1
Förderung	2	ältere weibliche absteigende Linie	1
Wohnraum	2	auswärtigen Angelegenheiten	1
soziale	2	Mitteilungen	1
kulturelle Entfaltung	2	fremden Mächten	1
Unterrichtsarten	2	Abtretung	1
behördlichen Aufsicht	2	Staatsgebiets	1
Gelegenheit	2	See	1
Lehrmittel	2	schriftlichen Bericht	1
vorwissenschaftlichen	2	schriftliche Benachrichtigung	1

Table 7: Surface Forms of NIL Annotations

BIBLIOGRAPHY

Agt-Rickauer, Henning, Jörg Waitelonis, Tabea Tietz, and Harald Sack (2016). "Data Integration for the Media Value Chain." In: *15th International Semantic Web Conference (Posters and Demos)*. CEUR-WS.

Alexy, Robert (1999). "Grundrechte." In: *Enzyklopädie Philosophie* 1, pp. 525–529.

Beckett, David (2014). *RDF 1.1 N-Triples: A line-based syntax for an RDF graph*. W3C Recommendation. W3C, https://www.w3.org/TR/n-triples/.

Beckett, David, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers (2014). *RDF 1.1 Turtle: Terse RDF Triple Language*. W3C Recommendation. W3C, https://www.w3.org/TR/turtle/.

Berger, Peter and Thomas Luckmann (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*.

Berners-Lee, T., R. Fielding, and L. Masinter (2005). *RFC3986: Uniform Resource Identifier (URI): Generic Syntax*. https://www.ietf.org/rfc/rfc3986.txt.

Berners-Lee, Tim, James Hendler, and Ora Lassila (2001). "The Semantic Web." In: *Scientific American* 284.5, pp. 34–43.

Berners-Lee, Tim, Wendy Hall, James A Hendler, Kieron O'Hara, Nigel Shadbolt, Daniel J Weitzner, et al. (2006). "A Framework for Web Science." In: *Foundations and Trends in Web Science* 1.1, pp. 1–130.

Berners-Lee, Timothy J (1989). *Information management: A Proposal*. Tech. rep. CERN.

Bizer, Christian, Tom Heath, and Tim Berners-Lee (2011). "Linked Data: The Story So Far." In: *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. IGI Global, pp. 205–227.

Boli-Bennett, John (1979). "The Ideology of Expanding State Authority in National Constitutions, 1870-1970." In: *National development and the world system*, pp. 212–237.

Bosch, Thomas and Benjamin Zapilko (2015). "Semantic Web Applications for the Social Sciences." In: *IASSIST Quarterly* 38.4, pp. 7–7.

Brickley, Dan and R. V. Guha (2014). *RDF Schema 1.1*. W3C Recommendation. W3C, https://www.w3.org/TR/rdf-schema/.

Büthe, Tim, Alan M Jacobs, Erik Bleich, Robert J Pekkanen, and Marc Trachtenberg (2014). "Qualitative & Multi-Method Research." In: *Journal Scan: January* 2015, p. 63.

Capadisli, Sarven, Amy Guy, Ruben Verborgh, Christoph Lange, Sören Auer, and Tim Berners-Lee (2017). "Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli." In: *International Conference on Web Engineering*. Springer, pp. 469–481.

Crawford, Katherine (2009). *Perilous Performances: Gender and Regency in Early Modern France*. Vol. 145. Harvard University Press.

Curty, Renata Gonçalves (2016). "Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study." In: *IJDC* 11.1, pp. 96–117.

Daiber, Joachim, Max Jakob, Chris Hokamp, and Pablo N. Mendes (2013). "Improving Efficiency and Accuracy in Multilingual Entity Extraction." In: *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, pp. 121–124.

Decker, Stefan, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, and Ian Horrocks (2000). "The Semantic Web: The Roles of XML and RDF." In: *IEEE Internet computing* 4.5, pp. 63–73.

Description and Access, Task Force on Metadata (2000). https://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html, last visited: July 19, 2018. Final Report.

Eldesouky, Bahaa, Menna Bakry, Heiko Maus, and Andreas Dengel (2016). "Seed, an End-User Text Composition Tool for the Semantic Web." In: *International Semantic Web Conference*. Springer, pp. 218–233.

Elkins, Zachary, Tom Ginsburg, James Melton, Robert Shaffer, Juan F Sequeda, and Daniel P Miranker (2014). "Constitute: The World's Constitutions to Read, Search, and Compare." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 27, pp. 10–18.

Evans, James A and Pedro Aceves (2016). "Machine Translation: Mining Text for Social Theory." In: *Annual Review of Sociology* 42, pp. 21–50.

Faulkner, Steve, Arron Eicholz, Travis Leithead, Alex Danilo, and Sangwhan Moon (2017). *HTML 5.2*. W3C Recommendation. W3C, https://www.w3.org/TR/html/.

Foster, Allen and Nigel Ford (2003). "Serendipity and Information Seeking: an Empirical Study." In: *Journal of documentation* 59.3, pp. 321–340.

Froschauer, Ulrike and Manfred Lueger (2003). *Das qualitative Interview: Zur Praxis interpretativer Analyse sozialer Systeme*. Vol. 2418. UTB.

Go, Julian (2003). "A Globalizing Constitutionalism?: Views from the Postcolony, 1945-2000." In: *International Sociology* 18.1, pp. 71–95.

Gruber, Thomas R. (1993). "A translation approach to portable ontology specifications." In: *Knowledge Acquisition* 5, pp. 199–220.

Halford, Susan, Catherine Pope, and Mark Weal (2013). "Digital Futures? Sociological Challenges and Opportunities in the Emergent Semantic Web." In: *Sociology* 47.1, pp. 173–189.

Handschuh, Siegfried (2005). "Creating Ontology-based Metadata by Annotation for the Semantic Web." PhD thesis. Karlsruher Institut für Technologie.

Heaton, Janet (2004). *Reworking Qualitative Data*. Sage.

Heintz, Bettina and Annette Schnabel (2006). "Verfassungen als Spiegel globaler Normen?" In: 58, pp. 685–716.

Hellmann, Sebastian (2013). *NIF 2.0 Core Ontology*. Ontology Description. AKSW, University Leipzig, http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html.

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer (2013). "Integrating NLP using Linked Data." In: *International Semantic Web Conference*. Springer, pp. 98–113.

Hepp, Martin (2008). "Goodrelations: An ontology for describing products and services offers on the web." In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, pp. 329–346.

Herndon, Joel and Robert O'Reilly (2016). "Data Sharing Policies in Social Sciences Academic Journals: Evolving Expectations of Data Sharing as a Form of Scholarly Communication." In: *Databrarianship: The Academic Data Librarian in Theory and Practice*.

Heyvaert, Pieter, Anastasia Dimou, Aron-Levi Herregodts, Ruben Verborgh, Dimitri Schuurman, Erik Mannens, and Rik Van de Walle (2016). "RMLEditor: A gGaph-based Mmapping Editor for Linked Data Mappings." In: *International Semantic Web Conference*. Springer, pp. 709–723.

Hitzler, Pascal, Markus Kroetzsch, and Sebastian Rudolph (2009). *Foundations of Semantic Web Technologies*. CRC press.

Horridge, Matthew, Holger Knublauch, Alan Rector, Robert Stevens, and Chris Wroe (2004). "A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0." In: *University of Manchester*.

Hyland, Bernadette, Ghislain Atemezing, and Boris Villazón-Terrazas (2014). *Best Practices for Publishing Linked Data*. W3C Recommendation. W3C, https://www.w3.org/TR/ld-bp/.

Khalili, Ali and Sören Auer (2013). "User interfaces for Semantic Authoring of Textual Content: A Systematic Literature Review." In: *Web Semantics: Science, Services and Agents on the World Wide Web* 22, pp. 1–18.

Knoth, Alexander Henning (2016). "Staatliche Selbstbeschreibungen Analysieren. Soziologische und computerlinguistische Ansätze der Dokumentenarbeit." In: *Trajectoires. Travaux des jeunes chercheurs du CIERA* Hors série n° 1.

Knoth, Alexander, Manfred Stede, and Erik Hägert (2018). "Dokumentenarbeit mit hierarchisch strukturierten Texten: Eine historisch vergleichende Analyse von Verfassungen." In: *Kritik der digitalen Vernunft. Abstracts zur Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum, 26.02.-02.03. 2018 an der Universität zu Köln, veranstaltet vom Cologne Center for eHumanities (CCeH)*. Ed. by Georg Vogeler. Universität zu Köln, pp. 196 –203.

Kobilarov, Georgi, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee (2009). "Media Meets Semantic Web–How the BBC uses DBpedia and Linked Data to Make Connections." In: *European Semantic Web Conference*. Springer, pp. 723–737.

Koutraki, Maria, Farshad Bakhshandegan-Moghaddam, and Harald Sack (2018). "Temporal Role Annotation for Named Entities." In: *Proceedings of the 14th Int. Conference on Semantic Systems*. (to be published).

Kripke, Saul A (1972). "Naming and necessity." In: *Semantics of natural language*. Springer, pp. 253–355.

Lagler, Wilfried (2000). *Gott im Grundgesetz? Zur Bedeutung des Gottesbezugs in unserer Verfassung und zum christlichen Hintergrund der Grund-und Menschenrechte*.

Lange, Christoph (2009). "Krextor–An Extensible XML→ RDF Extraction Framework." In: *Workshop on Scripting and Development for the Semantic Web, co-located with 6th European Semantic Web Conference* 449, p. 38.

Latif, Shahid, Diao Liu, and Fabian Beck (2018). "Exploring Interactive Linking Between Text and Visualization." In: *EUROVIS*.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören

Auer, et al. (2015). "DBpedia–A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia." In: *Semantic Web* 6.2, pp. 167–195.

Lemke, Matthias and Gregor Wiedemann (2015). *Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Springer-Verlag.

Lorenz, Astrid (2005). "How to Measure Constitutional Rigidity: Four Concepts and Two Alternatives." In: *Journal of Theoretical Politics* 17.3, pp. 339–361.

Luhmann, Niklas (1984). *Soziale Systeme: Grundrisseiner allgemeine Theorie*. Suhrkamp Verlag.

— (1993). *Gesellschaftsstruktur und Semantik*. Vol. 3. Suhrkamp Frankfurt am Main.

Mayring, Philipp (2015). *Qualitative Inhaltsanalyse*. 12th ed. Beltz.

McCusker, James P, Michel Dumontier, Rui Yan, Sylvia He, Jonathan S Dordick, and Deborah L McGuinness (2017). "Finding Melanoma Drugs Through a Probabilistic Knowledge Graph." In: *PeerJ Computer Science* 3.

Meinel, Christoph and Harald Sack (2011). *Internetworking: Technische Grundlagen und Anwendungen*. Springer-Verlag.

Moore, Niamh (2007). "(Re) using Qualitative Data?" In: *Sociological Research Online* 12.3, pp. 1–13.

Morbidoni, Christian and Alessio Piccioli (2015). "Curating a Document Collection via Crowdsourcing with Pundit 2.0." In: *The Semantic Web: ESWC 2015 Satellite Events*. Springer, pp. 102–106.

Musen, Mark A (2015). "The protégé Project: A Look Back and a Look Forward." In: *AI matters* 1.4, pp. 4–12.

Nadeau, David and Satoshi Sekine (2007). "A Survey of Named Entity Recognition and Classification." In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Noy, Natalya F, Deborah L McGuinness, et al. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Tech. rep. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics, Stanford, CA.

Oeberst, Aileen, Ulrike Cress, Mitja Back, and Steffen Nestler (2016). "Individual Versus Collaborative Information Processing: The Case of Biases in Wikipedia." In: *Mass Collaboration and Education*. Springer International Publishing, pp. 165–185.

Oren, Eyal, Knud Möller, Simon Scerri, Siegfried Handschuh, and Michael Sintek (2006). "What are Semantic Annotations?" In: *Relatório técnico. DERI Galway* 9, p. 62.

Osgood, Charles E (1959). "The Representational Model and Relevant Research Materials." In: *Trends in content analysis*. University of Illinois Press, pp. 33–88.

Osterhoff, Johannes, Jörg Waitelonis, and Harald Sack (2012). "Widen the Peepholes! Entity-Based Auto-Suggestion as a rich and yet immediate Starting Point for Exploratory Search." In: *Proceedings of 2nd Workshop Interaction and Visualization in the Web of Data (IVDW)*. Gesellschaft für Informatik.

Pellegrini, Tassilo, Harald Sack, and Sören Auer (2014). *Linked Enterprise Data: Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien*. Springer-Verlag.

Popping, Roel (2000). *Computer-assisted Text Analysis*. Sage.

Prud'hommeaux, Eric and Carlos Buil-Aranda (2013). *SPARQL 1.1 Federated Query*. W3C Recommendation. W3C, https://www.w3.org/TR/sparql11-federated-query/.

Rahman, Muhammad Mahbubur and Tim Finin (2018). "Understanding and Representing the Semantics of Large Structured Documents." In: *Workshop on Semantic Deep Learning, co-located with the 17th International Semantic Web Conference*.

Sanderson, Robert, Paolo Ciccarese, and Benjamin Young (2016). *Web Annotation Ontology*. https://www.w3.org/ns/oa#, last visited: July 19, 2018.

Schandl, Thomas and Andreas Blumauer (2010). "PoolParty: SKOS Thesaurus Management Utilizing Linked Data." In: *Extended Semantic Web Conference*. Springer, pp. 421–425.

Schmachtenberg, Max, Christian Bizer, and Heiko Paulheim (2014). "Adoption of the Linked Data Best Practices in Different Topical Domains." In: *International Semantic Web Conference*. Springer, pp. 245–260.

Schreiber, Guus and Yves Raimond (2014). *RDF 1.1 Primer*. W3C Recommendation. W3C, https://www.w3.org/TR/rdf11-primer/.

Shneiderman, Ben, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson.

Silberman, Neil Asher (2012). *The Oxford Companion to Archaeology. 1. Ache-Hoho*. Vol. 1. Oxford University Press.

Stone, Philip J, Dexter C Dunphy, and Marshall S Smith (1966). "The General Inquirer: A Computer Approach to Content Analysis." In: *MIT press*.

Tietz, Tabea, Joscha Jäger, Jörg Waitelonis, and Harald Sack (2016). "Semantic Annotation and Information Visualization for Blogposts with refer." In: *Workshop on Visualization and Interaction for Ontologies and Linked Data, co-located with the 15th International Semantic Web Concernce*, pp. 28–40.

Usbeck, Ricardo et al. (2015). "GERBIL: General Entity Annotator Benchmarking Framework." In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1133–1143.

Van Rijsbergen, Cornelis Joost (1979). "Information Retrieval." In: *Dept. of Computer Science, University of Glasgow*.

Verborgh, Ruben, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht, Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, and Rik Van de Walle (2014). "Querying Datasets on the Web with High Availability." In: *International Semantic Web Conference*. Springer, pp. 180–196.

Verborgh, Ruben, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, and Pieter Colpaert (2016). "Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web." In: *Journal of Web Semantics* 37–38, pp. 184–206.

Villazón-Terrazas, Boris, Luis M Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez (2011). "Methodological Guidelines for Publishing Government Linked Data." In: *Linking Government Data*. Springer, pp. 27–49.

Waitelonis, Jörg, Henrik Jürges, and Harald Sack (2016). "Don't compare Apples to Oranges: Extending GERBIL for a fine grained NEL evaluation." In: *Proceedings of the 12th International Conference on Semantic Systems*. ACM, pp. 65–72.

Waitelonis, Jörg and Harald Sack (2012). "Towards Exploratory Video Search using Linked Data." In: *Multimedia Tools and Applications* 59.2, pp. 645–672.

— (2016). "Named Entity Linking in #Tweets with KEA." In: *Proceedings of 6th workshop on 'Making Sense of Microposts', Named Entity Recognition and Linking (NEEL) Challenge in conjunction with 25th International World Wide Web Conference*. CEUR-WS.

Waitelonis, Jörg (2018). "Linked Data Supported Information Retrieval." PhD thesis. Karlsruher Institut für Technologie (KIT). 256 pp. DOI: 10.5445/IR/1000084458.

Welty, Christopher and Deborah McGuinness (2004). *OWL Web Ontology Language Guide*. W3C Recommendation. W3C, http://www.w3.org/TR/2004/REC-owl-guide-20040210/.

Wilde, Erik and Martin Dürst (2008). *RFC5147: URI Fragment Identifiers for the text/-plain Media Type*. https://www.ietf.org/rfc/rfc5147.txt.

Zenk-Möltgen, Wolfgang and Greta Lepthien (2014). "Data Sharing in Sociology Journals." In: *Online Information Review* 38.6, pp. 709–722.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Ich versichere, dass ich die von mir vorgelegte schriftliche Arbeit einschließlich evtl. beigefügter Zeichnungen, Kartenskizzen, Darstellungen u.a.m. selbständig angefertigt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle Ausführungen, die dem Wortlaut oder dem Sinn nach anderen Texten entnommen sind, habe ich in jedem Fall unter genauer Angabe der Quelle deutlich als Entlehnung kenntlich gemacht. Das gilt auch für Daten oder Textteile aus dem Internet. Die Richtlinie zur Sicherung guter wissenschaftlicher Praxis für Studierende an der Universität Potsdam (Plagiatsrichtlinie) - Vom 20.Oktober 2010", im Internet unter https://www.uni-potsdam.de/am-up/2011/ambek-2011-01-037-039.pdf, ist mir bekannt.

Potsdam, den 25. August 2018

Tabea Tietz