Linking the Kingdom: Enriched Access To A Historiographical Text

Ribbens, C.R.; de Boer, V.; van Doornik, J.; Buitinck, L.; Marx, M.; Veken, T.

published in
K-CAP 13 Proceedings of the seventh international conference on Knowledge capture
2013

document version
Publisher's PDF, also known as Version of record

Linking the Kingdom: Enriched Access To A Historiographical Text

Victor de Boer
Dept. of Computer Science,
VU University Amsterdam,
the Netherlands
v.de.boer@vu.nl

Johan van Doornik
Informatics Institute,
Universiteit van Amsterdam,
the Netherlands
vandoornik@gmail.com

Lars Buitinck
Informatics Institute,
Universiteit van Amsterdam,
the Netherlands
l.j.buitinck@uva.nl

Maarten Marx
Informatics Institute,
Universiteit van Amsterdam,
the Netherlands
maartenmarx@uva.nl

Tim Veken
NIOD Institute for War,
Holocaust and Genocide
Studies, Amsterdam,
the Netherlands
t.veken@niod.knaw.nl

Kees Ribbens
NIOD Institute for War,
Holocaust and Genocide
Studies, Amsterdam,
the Netherlands
k.ribbens@niod.knaw.nl

ABSTRACT

Digital history is a branch of digital humanities concerned using ICT to improve study of history. Linked Data provides a way of effective enriched digital access to scientific texts about history (historiographies). In this paper, we present a method for connecting a historiographical text to the Linked Data cloud. We present the method and tools that we use in each of the method's steps. We focus on one extensive case study: the enriched access of an important work of Dutch World War II historiography "Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog". We describe the digitization and present two sources of structured knowledge that link to individual text sources, retrievable on the Web of Data. The first is the manually constructed and highly curated "Back of the Book Index". The second is a list of extracted Named Entities. We compare both structured sources as stepping stones to the Web of Data and present a number of use cases relevant for both historical researchers as well as for the general public.

1. INTRODUCTION

Digital history (or e-history) is a branch of digital humanities which seeks to use modern ICTs to improve the historical research. A large part of current digital history research revolves around the digitization of historically relevant material and providing adequate access to these digitized assets for individual researchers. Representing and sharing data, information and knowledge is key to furthering the digital history research agenda[3]. Many archives, libraries and historical findings have already been digitized and have been made accessible in re-usable formats. The Web of (Semantic) Data provides a well-described and standardized framework for modeling and linking this type of knowledge. An example of a projects that uses linked data for historical research is the Civil War 150 project[1], where digitized datasets about the US civil war are connected as linked data and used for analysis and visualisations. In the related cultural heritage domain, publishing of metadata as linked data is gaining ground. Examples include Europeana [5] which uses the Linked Data architecture to provide access to Europe's cultural heritage metadata. These projects use methods and tools for digitizing, converting and publishing historical data in this format. However, these methods focus on either publishing existing structured data or metadata while a large part of historical research is concerned with the analysis of historical or historiographical texts (a 'historiography' refers to a body of work on a specialized historical topic).

We here present our work in the form of an extensive case study: the enriched publication of the important Dutch historiographical work *Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog* (The Kingdom of the Netherlands in WWII) by Dr. Loe de Jong. The *Koninkrijk* -as we will refer to this text from here on- remains the most appealing history of German occupied Dutch society (1940-1945). Published between 1969 and 1991, the 14 volumes, consisting of 30 parts and 18,000 pages combine the qualities of an authoritative work for a general audience, and an inevitable point of reference for scholars. In the *Verrijkt Koninkrijk* (Enriched Kingdom) project, we aim to provide enriched access to the original text to assist historians in their research.

We describe a method and tools to make a historiographical text available in structured form on the Web and to connect it to external sources on the Web of Data. We show how these explicit links between text fragments and external background knowledge can be used by historical researchers to investigate relevant hypotheses. Important in this respect is that this data-driven approach still connects to the historical methodology by providing explicit links to the original text from manually constructed as well as automatically gen-

[1] http://www.civilwardata150.net

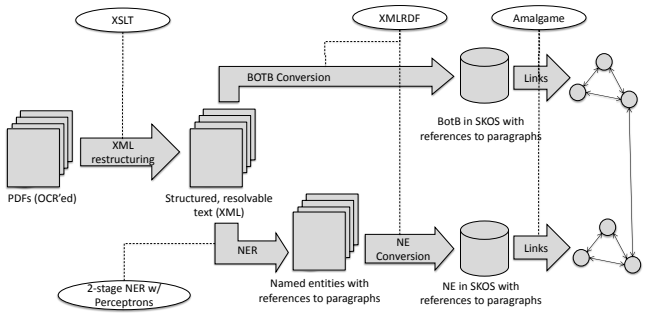

Figure 1: Overview of the conversion pipeline, including the tools used.

erated data. In Section 4 we show a number of use cases in the context of the case study exemplifying this. In Figure 1, we show the overall approach and the tools used in each of the steps. In the next sections, we will detail each of the steps, the tools used and the results in the *Koninkrijk* case study.

2. PREPROCESSING

In 2011, the entire *Koninkrijk* was scanned and Optical Character Recognition (OCR) was performed using proprietary software. The digitized documents are available online as PDF downloads at `http://www.niod.knaw.nl/koninkrijk/`. The fact that this server went offline shortly after publication due to the enormous popularity of the website speaks to the appeal not only to professional users but also the general public. An example page is shown in Figure 2(a).

We then transformed the pdf collection into XML with the open-source tool pdf2xml[2]. An example of the resulting XML is shown in Figure 2(b). In the following we will discuss the XML restructuring steps applied to the resulting document collection.

2.1 XML restructuring

We first performed a pre-processing clean-up step for all documents filtering out combinations of non-legible characters (the result of dirt on the paper or on the scannerbed) and characters outside the margins of the text. The documents were then transformed into a structured book format with an XSLT script[3]. All resulting XML files validate against a RelaxNG schema[4].

The root element root of each document separates Dublin Core metadata about the origin and formatting of the document from the actual book content. This *book* element is created based on an automatic detection of a number of visual and textual cues that can be found throughout the different pages. Thanks to a relatively consistent layout used throughout the different parts (with an unfortunate exception of 'deel 14', see Section 3.1) it was possible to use the same feature detector for all books. Additionally, each of

[2] `http://sourceforge.net/projects/pdf2xml/`

[3] `http://transformer.loudejongdigitaal.nl/d/vk/loudejong.xsl`

[4] `http://schema.loudejongdigitaal.nl/book.rnc` see *.html* page for a human-readable presentation of the schema

Figure 2: Results of the postprocessing steps for one page of the "Koninkrijk". The figure shows a) the original page in the OCR'ed PDF document, b) the result of the OCR in XML and c) the resulting structured XML (ellipses used for brevity).

XML tag	Description	Occ.
<book>	a book	30
<chapter>	a chapter	226
<section>	a section	1,885
<p>	a paragraph (smallest resolvable unit)	86,257
<quote>	indicates "reported speech"	56,547
<page>	indicates a page number	16,922
<lemma>	a back of the book lemma	16,186
<lemma-ref>	back of the book lemma's reference	148,370

Table 1: Statistics for the most important elements in the restructured XML.

the elements is assigned a unique hierarchical identifier. For example, the paragraph with the identifier *nl.vk.d.1.6.1.43* is the 43rd paragraph in the first subsection of the sixth chapter of the first volume of the Koninkrijk. In Table 1, we list statistics for the most important elements detected in the converted XML text. Figure 2(c) shows an example of the resulting XML.

2.2 Access to the structured text

The digitized text can be accessed in a number of ways. First of all, a resolver server was installed which responds by presenting the structure (in XML) when presented with an identifier. For example, `http://resolver.loedejongdigitaal.nl/nl.vk.d.1.6.1.43` is resolved to the XML fragment of the identified paragraph. Removing the last number of the identifier (43) results in its broader section, etcetera. This resolver essentially makes URIs out of the identifiers, which in turn are used to link to existing web sources as we will describe in Section 3. A full text search engine was installed at `http://search.loedejongdigitaal.nl`.

3. LINKING TO THE WEB OF DATA

To enrich the digitized text with links to linked data, we need "stepping stones" of structured data that can be used to link the work-specific resources to external resources. These structured datasets should contain links into the text itself

(ie. to the URIs created through the resolver). We model the datasets as SKOS vocabularies[5]. In sections 3.1 and 3.2, we present two separate methods of producing such structured datasets and compare them in Section 3.3.

3.1 Using the Back-of-the-Book Index

The first structured data source is constructed out of the Back of the Book index (the *BotB index*). The original index is a separate volume of the *Koninkrijk* which was finalized in 1994 and created by employees of the NIOD by combining the indexes for the individual books. In this process errors and inconsistencies were removed. The indexes of the individual books were also created by hand, under supervision of the author of said books. The BotB index represents a highly curated source of information linking lemmas, identified as important by experts to representative pages in the text. There are a total of 15,234 lemmas in the BotB index. These include "Named Entities" such as persons, places, organizations and named events but also more general concepts such as "Hospitals", "Unemployment", etc. The references in the index point to one or more occurrences of these lemmas in het *Koninkrijk*. This reference takes the form of a single page number or a page range, preceded by a volume indicator. Furthermore, a lemma might have a *see:* or *see also* reference to another lemma, pointing to a preferred or related lemma respectively. An example is "Volksziekten. Zie: Epidemieën" (English "Public diseases, see Epidemics").

3.1.1 Conversion to SKOS

We first produced an XML version of the OCR'ed index. For the XML restructuring, different rules were needed, since the indentation and layout of this volume are significantly different from the rest of the work (eg. the index uses a two-column instead of a one-column layout). In this process, the original reference to a page number is made explicit by retrieving the identifiers of the paragraphs that occur on the referred page. Since a paragraph can span more than one page, this results in a list of consecutive paragraphs that span the whole page plus any part of the paragraph that occurs on a previous or consecutive page. The XML restructuring resulted in a XML document with the 15,234 lemmas as individual XML records.

We then converted this XML document to a SKOS document using the XMLRDF tool of the ClioPatria semantic framework [10]. This tool allows us to convert XML documents to arbitrary RDF structures in an interactive way, using simple RDF graph rewriting rules. The XMLRDF tool has been previously described in [4]. We manually created six of such rewriting rules to create a SKOS representation of the BotB index. We list the rules below to illustrate the conversion procedure:

Rule 1 creates RDF resources of type `skos:Concept` from XML lemma records, with `skos:prefLabel` being the original lemma.
Rule 2 assigns URIs to these concepts by concatenating the namespace `http://purl.org/collections/nl/niod/` (shorthand "niod:'), the prefix `botb-` and the lemma itself.
Rule 3 converts the paragraph reference into a `niod:pageRef` triple whose object is a blank node, which in turn has `niod:PRef` triples linking it to individual paragraph URIs.
Rule 4 converts the *see:* construct to constructs. The lemma is removed and added as a `skos:altLabel` for the concept to which it refers.
Rule 5 converts the *see also:* construct to a *skos:related* triple, linking to the refered concept.
Rule 6 the cleaning rule links the original lemma entry and the reference to the concept as `rdfs:literal`s. Any empty references are removed.

[5] `www.w3.org/2004/02/skos`

```
niod:botb-Bagehor_W
a skos:Concept ;
skos:prefLabel "Bagehor, W." ;
niod:original "Bagehor, W., 2: 30; 9: 103" ;
niod:pageRef [ niod:parRef
 <http://resolver.loedejongdigitaal.nl/nl.vk.d.2.2.3.1.4> ,
   <http://resolver.loedejongdigitaal.nl/nl.vk.d.2.2.3.1.5> ,
   <http://resolver.loedejongdigitaal.nl/nl.vk.d.2.2.3.1.6>
     ] ,
     [ niod:parRef
       <http://resolver.loedejongdigitaal.nl/nl.vk.d.9-1.3.1.2.1> ,
   <http://resolver.loedejongdigitaal.nl/nl.vk.d.9-1.3.1.2.2>
     ] ;
niod:vkId "nl.vk.d.reg.4.767" ;
skos:inScheme niod:BotBScheme.
```

Figure 3: Example of a BotB index concept in RDF Turtle syntax. This concept has two references into the work, which are represented as `niod:pageRef` triples to blank nodes, which in turn link to consecutive paragraph URIs using `niod:parRef` triples.

After conversion, the BotB index is a `skos:ConceptScheme` with 15,234 Concepts. Table 2 lists the statistics for a number of constructs in the generated SKOS graph. Figure 3 shows an example of a BotB concept in RDF turtle syntax.

3.1.2 Links

To enrich the *Koninkrijk*, we align the BotB index with a number of existing thesauri. For this, we use the Amalgame alignment tool for finding, evaluating and managing vocabulary alignments [6]. Amalgame supports a semi-automatic interactive approach to vocabulary alignment where the user can manually combine existing automatic matching techniques into an alignment workflow targeted to the data set at hand using a workflow setup [12].

Currently, the BotB SKOS thesaurus has been aligned with four datasets. Below, we describe each of these alignments. In each case, we use very simple label matching algorithms to find potential mappings between concepts. Ambiguous mappings (one source concept matching several target concepts or vice versa) are discarded. This is a very simple procedure that produces fairly high-precision mappings while recall remains relatively low. This was done to reduce the effort for experts in evaluating the produced More sophisticated combinations of matching algorithms can boost precision and more likely recall. We here present the numbers merely as an indication of the possibilities of the described method. The produced links were added as separate RDF graphs in the form of skos:exactMatch triples between the BotB concepts and external URIs.

[6] Amalgame is freely available and documented at `http://semanticweb.cs.vu.nl/amalgame/`

- The index was aligned with the **NIOD thesaurus**. This in-house thesaurus consists of 1241 concepts that are used to annotate various objects in the NIOD archives. This includes the BeeldbankWO2, an image archive with over 175,000 images gathered from various Dutch national war- and resistance-museums [7]. We were able to map 171 concepts (14% of the NIOD thesaurus) to BOTB concepts. The fact that the majority of the NIOD thesaurus concepts are not mapped is due to a number of reasons. Manual inspection of a number of unmatched concepts showed that they mostly are indeed not mentioned in the back-of-the-book index but cover topics related to the classification of images or library items. The NIOD thesaurus itself is aligned with a number of external sources, including the Dutch audiovisual GTAA thesaurus [8] as well as the aforementioned Cornetto.

- We aligned it with the **GeoNames** [9] geographical dataset. We made a selection of only Dutch locations, resulting in a dataset of 21,405 locations. The alignment yielded 487 BotB entries matched (3%).

- A mapping was created to **Cornetto**, a Dutch lexical dataset similar WordNet[10]. The SKOS version of Cornetto has 70,370 concepts. Using the abovementioned simple matching techniques, we were able to align 250 of these concepts with BotB concepts. This corresponds to 2% of our index. A majority of the BotB terms is not found in this lexical dataset as they concern Named Entities or context-specific compound terms, not present in Cornetto.

- The BotB index was also aligned with the Dutch Art and Architecture Thesaurus **AATNed** [11], which consists of 33,689 concepts. Here, we found mappings from 159 source concepts to AATNed concepts (1% of the BotB index).

3.1.3 Quality

There are a number of issues with the quality of the BotB index. First of all, many errors stem from the OCR process in the same way as the digitization of the main text is also affected by this. For the BotB index, little or no context information is available that can be used to improve the OCR as it concerns only individual lemmas. This results in a number of erroneous labels for the BotB concepts as well as for the page references of these lemmas. The XML conversion also introduces a number of errors even though great care has been taken to use rules that cover most cases. For example, in some cases, see/see also references are not correctly recognized because they occur halfway through the reference list. Also, some page numbers could not be transformed to paragraph references because of errors in the conversion of those pages in the text. In the conversion to SKOS, no additional errors are produced since the SKOS is mostly a representation of the structured XML. In the alignment, however, erroneous matches can be produced since simple label matching algorithms are used. However, the fact that these label matchers are fairly strict, results in the errors being mostly in terms of recall rather than precision. We manually evaluated random subsets of each of the mappings and found that the precision of the matches is very high: >95%. Usage of more fuzzy string matching algorithms can boost recall, but this comes at the expense of precision. To increase the chance of acceptance of these matches data for the historical research, we here chose to focus on high precision matches.

Even though the quality is not optimal, the BotB index is a valuable method of linking paragraphs in the text to the Web of Data. The quality can be further increased through manual inspection and evaluation of the concepts. In Section 4, we will present a number of concrete use cases that use the current version of the index.

3.2 Using Named Entities

The second structured data source is a SKOS vocabulary constructed out of the Named Entities extracted from the text in all of the volumes (excluding the index). We refer to this vocabulary as the NE index.

3.2.1 Named Entity Extraction

For the extraction, we employed the the perceptron-based named entity recognizer of [2]. The resulting set of named entities was consolidated to a single XML document which contains the 88,243 extracted Named Entities. Each entity has one or more references to the paragraph identifier where it was found as well as a type. Table 2 lists the number of extracted references per type. Additionally, for each lemma in the back of the book, an attempt was made to match it to a concept from the Dutch or English wikipedia. For this, an algorithm based on machine learning with features related to the N-gram (term frequency, number of terms, etc.) and concept (number of wikipedia articles linking to it, number of redirect pages linking to it, etc.) was used [7]. The concept mapping was only retained if the computed confidence was more than 95% resulting in 16,480 wikilinks.

3.2.2 Conversion to SKOS

In this case, minimal transformation was necessary to convert this XML document to a SKOS vocabulary using the XMLRDF tool. We again used one rule to construct the SKOS Concepts and one for URI assignment. Another rule converted the paragraph reference to an RDF triple that refers to that paragraph's resolver URI. The wikilinks were converted to DBPedia links by a simple URI replacement rule (replacing "http://nl.wikipedia.org/page/" by "http://nl.dbpedia.org/resource/"). The NE class was also maintained. Here, no `skos:altLabel` or `skos:related` relations were found. In Table 2, we compare the statistics to those of the BotB index.

3.2.3 Links

For the NE index, we also produced a number of alignments with existing sources using the Amalgame tool alongside the alignment that was derived from the wikilinks. We here list the results.

[7] `http://www.beeldbankwo2.nl/`
[8] `http://www.den.nl/standaard/262/`
[9] `http://www.geonames.org`
[10] `http://datahub.io/dataset/cornetto`
[11] `http://www.aat-ned.nl`

construct	BotB index	NE index
skos:Concepts	15,234	88,243
person		42,379
location		24,135
organization		20,717
miscellaneous		8,437
product		4,178
event		919
skos:altLabels	377	0
skos:related	486	0
no. references	100,681	364,928
average no. references	6.6	4.1
outgoing links	896	18,699

Table 2: Statistics for Back of the Book index and the Named Entity index in SKOS. The number of references refers to pages for the BotB index and to paragraphs for the NE index.

- The NE index was also aligned with the **NIOD thesaurus**. In this case, we were able to match 420 NE index concepts to NIOD Thesaurus concepts using the same alignment procedure. This is a surprisingly high number of matches, given that the NIOD thesaurus contains mainly concepts and not Named Entities. However, the Named Entity extractor extracted a large number of entities that it considers Named Entities but formally are not. These are spread over the various NE classes and include resources such as "Artillery", "Taxes" etcetera.
- We aligned the NE index concepts of type "location" to the Dutch subset of **GeoNames** using the Amalgame tool (again using simple label matching algorithms). This resulted in 814 links (1%)
- We aligned the NE index with **GTAA**, which resulted in 1,589 links (2%).
- As explained above, we converted the wikilinks to **DBPedia** links. This resulted in 13,160 relations to Dutch DBPedia resources. An additional 3,320 links to English DBPedia are present. In total, 19% of the concepts are matched with a DBPedia concept.

3.2.4 Quality

Like the BotB index, the NE index also has quality issues stemming from the OCR procedure of the text. Secondly, the Named Entity Recognition process is probabilistic and errors do occur. [VIC: can we say something about the precision and recall of the NER]. Another issue are errors in Named Entity reconciliation, where one entity that occurs in multiple forms in the text is not recognized as such. For example, in our case "H. Colijn" and "Hendrik Colijn" are not consolidated to one entity. Also in the alignment process errors can occur -as they do in the BotB index. Manual inspection of a random subset showed that here, the precision is also very high (>90%). Post-processing or manual evaluation of the extracted entities could improve the quality of these entities and as with the BotB index, using more fuzzy matching string matching algorithms can boost recall for the external links.

3.3 Comparing NE and BotB indexes

Table 2 lists the main characteristics of the two structured vocabularies we use to link the

One way in which we compared the BotB and NE indexes was by aligning the two SKOS vocabularies. For this, we again used the Amalgame alignment tool. We used simple label matching algorithms and excluded ambiguous matches. This resulted in 2,853 matches, meaning that 19% of the BotB concepts are matched to NE concepts.

Both sources have their own strengths and weaknesses. The original BotB index is the result of careful curation and involves manual identification of the occurrence of important concepts in the text. These occurrences might not be in the form of the actual term used in the index. For example, the BotB term "Spertijd" (Curfew) refers to a number of pages -and in our SKOS version to paragraphs. One of these is to paragraph `http://resolver.loedejongdigitaal.nl/nl.vk.d.6-2.3.8.6.4` where the following Dutch statement occurs "[...] en na tien uur zou zich niemand meer op straat mogen bevinden". In this paragraph, the word "Spertijd" occurs nowhere, rather this sentence describes an occurrence of the curfew. This is an occurrence of a reference that Named Entity Recognition algorithms would never be able to detect.

The BotB concepts that have a matching NE concept have a total of 21,857 references to pages. Those NE concepts have a total of 37,466 references to paragraphs. Out of these, 11,383 have a referenced paragraph in common with its matching NE concept. This means that 48% of the BotB references is not present in the matching NE concepts and that 70% of the NE concept references is not present for the matching BotB concept. This indicates that the two sources are not redundant but rather complement each other a great deal.

On the other hand, the NE index covers a wider range of concepts, which is reflected in its larger size. No manual selection process has determined which concepts are important enough to be included. The BotB index is designed in such a way that it concerns a limited amount of terms and allows researchers to identify the pages relevant to their interest. It proved to be more accepted by the researchers and is easier to browse and relate to existing sources, (cf. Section 4.1.2). The NE index does not does not have a limited size and is constructed without any specific goal in mind. It does however list a great deal of less important concepts. In the use case section below, we present specific use case that use either one of the indexes.

Since the two indexes stem from different sources and have different sources of errors, this subset of matched BotB and NE index concepts could be consolidated as a high-quality separate vocabulary. In fact, 127 out of the 171 BotB concepts that were matched to the NIOD thesaurus are also matched to the NE index, indicating that this is indeed a relevant source.

3.4 Linked Data access

We loaded the two vocabularies, as well as the described mapping sets into an instance of the ClioPatria semantic server. This Verrijkt Koninkrijk semantic server can be browsed at `http://semanticweb.cs.vu.nl/verrijktkoninkrijk/`. This server also hosts the NIOD Thesaurus in SKOS format. The PURL URIs redirect to the specific resources on this server which will respond by returning the RDF triples con-

cerning the resource (in the case of an RDF request header) or by showing a human-readable local view page (in the case of a web browser), conforming to Linked Data principles [1]. A SPARQL endpoint is also available at `http://semanticweb.cs.vu.nl/verrijktkoninkrijk/sparql/` with an interactive SPARQL editor available at `http://semanticweb.cs.vu.nl/verrijktkoninkrijk/flint/`. Only the structured sources are present in the semantic server. The text itself is hosted on a separate server, accessible through the `http://resolver.loedejong.nl` URL. The SPARQL endpoint therefore does not allow for more complex queries that analyze the actual textual content beyond the indexes.

4. USE CASES

The two indexes provide excellent gateways from other structured sources into the text and from text fragments to external sources. In this section, we present a number of use cases relevant to historical researchers and the general public. For the sake of brevity, we omit the complete SPARQL queries used in these use cases here, but they are reproduced on a separate web site that accompanies this paper at `http://few.vu.nl/~vbr240/verrijktkoninkrijk/`.

4.1 Historical research

The following use cases are based on historical research questions that were explored in collaboration with a Dutch historian (co-author of this paper). One important finding in this collaboration is about the role that this type of data can play in *narrative* historical science methodology. In our view, the linked data analysis is not used within this context to directly provide numerical answers to historical research questions but rather helps in two ways. First, it allows a researcher to analyze the a text using semantic enrichment. He or she can make broad numerical analyses of texts, providing opportunities for preliminary quantitative evaluation of historical hypotheses. The data can be used to find anomalies in the text generating interesting hypotheses that can be further researched by studying the text in detail. Secondly, the fact that the structured vocabularies through which this happens have explicit links to specific paragraphs, the historian can easily identify exactly these paragraphs. Rather than having to read entire volumes or chapters, concepts are related to smaller fragments, allowing for more efficient analysis of the text. In the next subsections, we provide a number of examples of this.

4.1.1 Geographic analysis

The link of the BotB index as well as the NE index allows us to use the geographical hierarchy of GeoNames for analyzing the text. For example, we can construct a simple SPARQL query that generates all paragraphs that mention a city or village in a given Dutch province region. Of course, this uses the current geographical reality as represented in GeoNames and mixes it with the historical context of WWII. In this case, the differences are not significant since most WWII locations still are part of the same provinces.

For this use case, we can use both the BotB index and the NE index as both are aligned with GeoNames. We used the SPARQL package for the statistical analysis tool R to provide a quantitative analysis and visualizations of the results[12]. In total 11,388 place references that are part of one of the Dutch provinces were found through the BotB index and 58,452 were found using the NE index. Analysis per province shows that the two sources correlate significantly with a Pearson correlation coefficient of 0.85. Figure 4 shows the distribution of location occurrences per Dutch province. The results of the BotB and NE indexes are combined in each bar to show the sum of the occurrences. This is an indication that the author of the work describes mostly events in the provinces Gelderland and Zuid-Holland. Further steps could be to normalize these bars by population size (also available in the GeoNames linked data) or surface area. Note also that the province Flevoland almost has no mentions, which can be explained by the fact that this province did not exist until after WWII and was at the time mostly still under water.

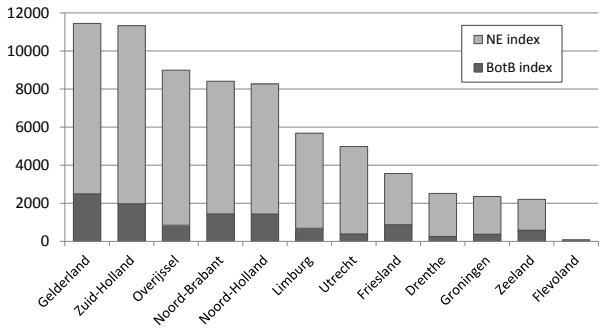

Figure 4: Analysis of location occurrences in the text, categorized by province. The BotB and NE indexes make up the total number.

Not only can these quantitative results be used as a starting point to formulate hypotheses. The individual mentions can be traced back as the SPARQL queries that lead to this data produce URIs that resolve to individual text fragments. More complicated queries can show correlations between these geographical features and others such as number of persons found (from the NE index concepts) or through concepts from external vocabularies such as DBPedia.

4.1.2 Pillarization

One of the specific historiographical research questions that informed this project concerns the quintessentially Dutch notion of "Pillarization". In Dutch cultural history, pillars refer to religious or political group identities (Catholicism, Socialism, etc.) that permeated Dutch daily life in the 20th Century [13]. The amount to which de Jong used these concepts in his description of WWII is an open question and our enrichment of the text can shed light on this. First of all, the NIOD thesaurus contains concepts for the seven main pillars that were identified by the historian (these are shown in table 3). Rather than identifying mentions of the pillars themselves, the historian was interested in persons and organizations associated with that pillar. To this end, he manually expanded the vocabulary with a total of 60 *pillarLinks*, linking pillar concepts from the NIOD thesaurus

[12] `http://cran.r-project.org/web/packages/SPARQL/`
[13] `http://en.wikipedia.org/wiki/Pillarisation`

pillar	occurrence	with `niod:Pillarization`
National-Socialist	885	9
Social-Democrat	645	22
Protestant	417	40
Liberal	378	10
Roman-Catholic	365	58
Communist	259	9
Jewish	150	12

Table 3: Pillarization concepts and the number of occurrences of linked entities

to BotB concepts.

This allowed us to produce a SPARQL query that retrieves the paragraphs that talk about one or more persons or organizations associated with a pillar. Table 3 shows the quantitative analysis. Moreover, the historian can use the specific links to identify *how* the author discusses these persons and organizations including the textual context. A follow-up question was to identify on which of the these paragraphs also a label for the BotB concept `niod:Pillarization` itself occurs. This is also shown in Table 3. These are examples of a number of queries that lead to a better understanding of de Jong's views on pillarization.

4.1.3 General background knowledge

In their research, historians combine analysis of historical and historiographical texts with their expert knowledge as well as common-sense knowledge. Through the link with DBPedia, we can use the formalized common sense and expert knowledge to automatically analyse the *Koninkrijk*. DBPedia covers a lot of knowledge about occupations of historical persons, which are partially matched to NE index concepts. This allowed us for example to use a simple SPARQL query to retrieve paragraphs referencing a person who before, during or after WWII was the prime minister of the Netherlands. We can also retrieve mentions of persons that were members of specific political parties. This allows for political analysis that relates to the pillarization question in the previous section.

4.2 Access for the general public

Other than for historical research, linking the *Koninkrijk* to the Web of Data allows for new types of access to the text as well as re-use. The NIOD thesaurus to which both indexes are linked can be used provide enriched access through the text by using query expansion based on SKOS properties such as the hierarchy. That thesaurus is also used to annotate over 175,000 images in BeeldbankWO2, as described in Section 3.1.2. Through the BotB and NE index, individual paragraphs are explicitly related to individual images. This link can be used in both directions: 1) the images can be used to automatically illustrate the paragraphs in a Web application and 2) the paragraphs can be used as example descriptions of the concepts shown in the images. We are currently working on development of such an application.

Another example is the use of the link to Cornetto, also described in Section 3.1.2. This data source is aligned with Wordnet and therefore we can use the links to provide partial English-language access to the text. Through links with other Wordnets a truly multilingual access to the text is possible. In the same fashion, the AATNed vocabulary is aligned with Getty's Art and Architecture thesaurus, allowing for access using those (English) terms. We are currently setting up a workshop in which we invite developers to link this data to their datasets and develop mashup applications using these linked datasets.

5. RELATED WORK

A number of approaches exists for extracting information from textual content and linking it to the Web of Data. In [9], the authors describe the NERD system that uses multiple information extraction tools to identify named entities in texts and link them to DBPedia or other Linked Data sources, including DBPedia Spotlight[14], OpenCalais[15] or Zemanta[16]. These tools could be used to replace the part in our approach where we make the NE index and link it to the Web of data. However, these tools work only on English texts and cannot be used for text in Dutch or other unsupported languages. A Dutch version of DBPedia spotlight is under construction [17] and we plan to compare our links to those extracted by this tool. However, the method we present in this paper does not extract entities from a specific data source (such as DBPedia), but is more general-purpose, allowing for linking to unspecified data sources.

The two indexes we present here represent rather shallow extracted information. In [8], the authors model the content of philosophical works such as that of Wittgenstein to allow students and researchers to investigate very complex questions. Although such a deep model of an historical text would be very useful to answer the type of questions we discuss in Section 4.1, constructing such a representation of the text requires even more manual labor. Using the back of the book index leverages work that was already done by experts. Deep NLP techniques could be used to extract even more information from the text. However, this is a very error-prone process and will require significant further research.

This work is related to that of linking library information to the Web of Data, as proposed by W3C's Library Linked Data Incubator Group[13]. Current linking of library data focuses on descriptive metadata of books and other media curated by libraries, rather than exposing the actual content of those texts to the Web of Data. Our work has a similar relation to other efforts that attempt to link historical data to the Web of data [6, 11]. Here also, pre-existing metadata rather than (references to) textual content are concerned.

6. DISCUSSION AND FUTURE WORK

In this paper, we describe a method of linking a digitized historiographical text to the Web of Data in the form of a case study using the *Koninkrijk*. The core idea is construct two SKOS vocabularies. One is derived from the manually constructed back of the book index and one is derived from automatically extracted named entities. Both vocabularies on the one hand refer to individual text fragments and on the other hand are aligned with external datasets. This allows us to provide enriched access to the text. We have shown how this enriched access can be used to support digital historical

[14] `http://dbpedia.org/spotlight`
[15] `http://www.opencalais.com`
[16] `http://www.zemanta.com`
[17] `http://nl.dbpedia.org/spotlight`

research as well as general-purpose access to the text. We also showed the possibilities of re-use of the data by others.

Although we here present a single case study, the method can be applied to other digitized texts as well. Although the general method will be the same, for specific texts specific XML restructuring rules will have to be constructed. How paragraphs, chapters, page numbers are denoted on printed pages varies from print to print, from time period to time period. The main work effort in applying the XML restructuring to another book or collection of books is in the required adjustments of the XSL transformer to accommodate the particulars of the layout. The detection of chapters, sections, etc., is based on a combination of layout cues (font size, location on the page, white spaces) and textual information (chapter titles may for instance always start with the word 'chapter' followed by a number). The difficulty is in finding the right balance between precision (no false positives) and recall (no false negatives), which may be caused by OCR mistakes. When a significant back-of-the-book index is available, this can be used to convert to a structured vocabulary as a stepping stone in the same way as presented here. Named Entity extraction results can also be used similarly.

An important issue with the presented work is the quality of the digitized text and linked data. From our own experience, a number of errors are still present in the current version of the Verrijkt Koninkrijk, due to errors in the OCR, the XML restructuring and the alignment procedures. Errors early on in the process are propagated through the process. This is one reason that we do not claim that our Linked Data enrichment can provide definitive answers to quantitative historical research questions. We do however show that the enrichment can be used to provide researchers with efficient access to the text for specific research questions. The fact that at every point, partial "answers" to these questions are explicitly linked to individual paragraphs allows for researchers to verify and contextualize these answers.

Future work includes linking to other linked data sets in the e-humanities domain. We are currently linking the Verrijkt Koninkrijk data to that of other digital history projects in the Netherlands. For example, in the Agora project, cultural heritage collections are enriched with event metadata [11]. We currently are in the process of linking our data to the Agora data to on the one hand provide more information about overlapping events, persons and locations. The recently started BiographyNed project aims at providing a linked data set of Dutch historical persons and their biographical information. This will be a high-quality, curated dataset that allows for even more effective biographical analysis of de Jong's work, in the use case described in Section 4.1.3.

Acknowledgements

This work was supported by CLARIN-NL (`http://www.clarin.nl`) under project number 11-014.

7. REFERENCES

[1] T. Berners-Lee. Linked data - design issues. http://www.w3.org/DesignIssues/ LinkedData.html, 2006.

[2] L. Buitinck and M. Marx. Two-stage named-entity recognition using averaged perceptrons. In *Proceedings of NLDB*, pages 171–176. Springer, 2012.

[3] Daniel J. Cohen et al. Interchange: The promise of digital history. *Special issue, Journal of American History*, 95, no.2 September 2008.

[4] V. de Boer, J. Wielemaker, J. van Gent, M. Hildebrand, A. Isaac, J. van Ossenbruggen, and G. Schreiber. Supporting linked data production for cultural heritage institutes: the amsterdam museum case study. In *Proceedings of the 9th international conference on The Semantic Web: research and applications*, ESWC'12, pages 733–747, Berlin, Heidelberg, 2012. Springer-Verlag.

[5] B. Haslhofer and A. Isaac. data.europeana.eu - the europeana linked open data pilot. In *DCMI International Conference on Dublin Core and Metadata Applications*, The Hague, The Netherlands, July 2011.

[6] E. Hyvönen, T. Lindquist, J. Törnroos, and E. Mäkelä. History on the semantic web as linked data – an event gazetteer and timeline for the world war i. In *Proceedings of CIDOC 2012 - Enriching Cultural Heritage, Helsinki, Finland*. CIDOC, June 2012.

[7] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Mapping queries to the linking open data cloud: A case study using dbpedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):418 – 433, 2011.

[8] M. Pasin and E. Motta. Philosurfical: an ontological approach to support philosophy learning. *The Future of Learning*, page 197, 2009.

[9] G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. Nerd meets nif: Lifting nlp extraction results to the linked data cloud. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *LDOW*, volume 937 of *CEUR Workshop Proceedings*, 2012.

[10] G. Schreiber, A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. van Kersen, M. de Niet, B. Omelayenko, J. van Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, and B. Wielinga. Multimedian e-culture demonstrator. In *Proceedings of the 5th international conference on The Semantic Web*, ISWC'06, pages 951–958, Berlin, Heidelberg, 2006. Springer-Verlag.

[11] M. van Erp, J. Oomen, R. Segers, C. van de Akker, L. Aroyo, G. Jacobs, S. Legêne, L. van der Meij, J. R. van Ossenbruggen, and G. Schreiber. Automatic Heritage Metadata Enrichment With Historic Events. In *Proceedings of International Conference for Culture and Heritage On-line-Museums and the Web 2011.* Archimuse, April 2011.

[12] J. van Ossenbruggen, M. Hildebrand, and V. de Boer. Interactive vocabulary alignment. In S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, editors, *TPDL*, volume 6966 of *Lecture Notes in Computer Science*, pages 296–307. Springer, 2011.

[13] W3C Library Linked Data Incubator Group. Library linked data incubator group: Datasets, value vocabularies, and metadata element sets. Group Report, 25 October 2011, 2011.