Verrijkt Koninkrijk: Linking a Historiographical Text to the Web of Data

Victor de Boer*[1], Johan van Doornik[2], Lars Buitinck[2], Maarten Marx[2], Tim Veken[3], and Kees Ribbens[3]

[1]The Network Institute, VU University Amsterdam, the Netherlands
[2]Informatics Institute, Universiteit van Amsterdam, the Netherlands
[3]NIOD Institute for War, Holocaust and Genocide Studies, Amsterdam, the Netherlands

March 29, 2013

1 Introduction

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 generated data. In Figure 1, we show the overall approach and the tools used in each

*corresponding author. email: v.de.boer@vu.nl

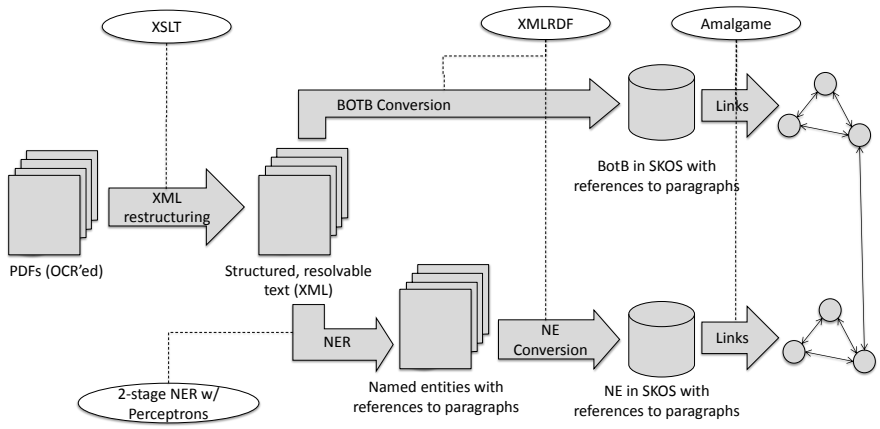

Figure 1: Overview of the conversion pipeline, including the tools used.

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. We then transformed the pdf collection into XML with the open-source tool pdf2xml[1]. After clean-up. the documents were then transformed into a structured book format with an XSLT script[2].

Each of the elements (book, section, paragraph, ...) 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.

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`.

[1] `http://sourceforge.net/projects/pdf2xml/`
[2] `http://transformer.loudejongdigitaal.nl/d/vk/loudejong.xsl`

Figure 2: 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).

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[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* and represents a highly curated source of information linking 15,234 lemmas , identified as important by experts to representative pages in the text. References 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: Epidemien" (English "Public diseases, see Epidemics").

We first produced an XML version of the OCR'ed index and converted this XML document to a SKOS document using the XMLRDF tool of the ClioPatria semantic framework[3]. After conversion, the BotB index is a `skos:ConceptScheme` with 15,234 Concepts. Table 1 lists the statistics for a number of constructs in the generated SKOS graph.

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

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 1: 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.

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. 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. Table 1 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. The concept mapping was only retained if the computed confidence was more than 95% resulting in 16,480 wikilinks. 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.

Table 1 lists the main characteristics of the two structured vocabularies. 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.

3.3 Links

To enrich the *Koninkrijk*, we align the BotB and NE index with a number of existing thesauri. For this, we use the Amalgame alignment tool for finding, evaluating and managing vocabulary alignments [4]. Currently, the BotB SKOS thesaurus has been aligned with a number datasets of which we describe a subset:

- **NIOD thesaurus**. This in-house thesaurus consists of 1241 concepts that are used to annotate various objects in the NIOD archives. We were

able to map 171 concepts (14% of the NIOD thesaurus) to BOTB concepts and 420 concepts to NE concepts.

- The **GeoNames** [4] geographical dataset. We made a selection of only Dutch locations, resulting in a dataset of 21,405 locations. The alignment yielded 487 BotB concepts and 814 NE concepts matched.
- The BotB index was aligned with the Dutch Art and Architecture Thesaurus **AATNed** [5]. Here, we found mappings from 159 source concepts to AATNed concepts (1% of the BotB index).
- Finally, 16,480 links between NE concepts and DBPedia were consolidated from the wikilinks found in the NER process.

3.4 Linked Data access

This Verrijkt Koninkrijk semantic server can be browsed at `http://semanticweb.cs.vu.nl/verrijktkoninkrijk/`. The PURL URIs redirect to the specific resources on this server which will respond by returning the RDF triples concerning 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/`.

4 Use Cases

4.1 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[6] that generates all paragraphs that mention a city or village in a given Dutch province region. We used the SPARQL package for the statistical analysis tool R to provide a quantitative analysis and visualizations of the results[7]. Figure 3 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. Not only can these quantitative results be used as a starting point to formulate hypotheses. The individual mentions can be traced back to

[4] `http://www.geonames.org`
[5] `http://www.aat-ned.nl`
[6] We list the complete SPARQL queries at `http://few.vu.nl/~vbr240/verrijktkoninkrijk/`.
[7] `http://cran.r-project.org/web/packages/SPARQL/`

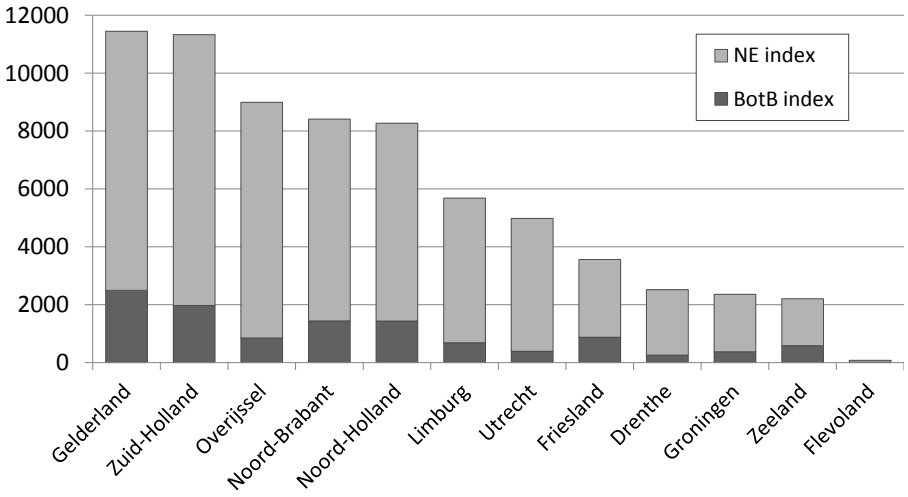

Figure 3: Analysis of location occurrences in the text, categorized by province. The BotB and NE indexes make up the total number.

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.

4.2 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. To investigate how de jong discusses pillarization, we manually expanded the vocabulary with a total of 60 *pillarLinks*, linking pillar concepts from the NIOD thesaurus 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 2 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.

5 Discussion and Future work

We presented a method of linking a digitized historiographical text to the Web of Data in the form of a case study using the *Koninkrijk*. An important issue with the presented work is the quality of the digitized text and linked data. From our

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 2: Pillarization concepts and the number of occurrences of linked entities

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.

Acknowledgements

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

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] 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.

[4] 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.