# Enriched Access to a Large War Historical Text using the Back of the Book Index

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

[1]Dept. of Computer Science, VU University Amsterdam
[2]Informatics Institute, Universiteit van Amsterdam
[3]NIOD Instituut voor Oorlogs-, Holocaust- en Genocidestudies

## 1 Introduction

Dr. Loe de Jong's *Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog* remains the most appealing history of German occupied Dutch society (1940-1945). Published between 1969 and 1991, the 30 volumes combine the qualities of an authoritative work for a general audience, and an inevitable point of reference for scholars. In the *Verrijkt Koninkrijk* project, we provide enriched access to a digitized version of this seminal text for historians researching Dutch WWII history in general and de Jong's view of it in particular. It will be used to gather data on De Jong's perception of the much debated issue of pillarization (Dutch: 'verzuiling') and group identity. To achieve this enriched access, the original text has been digitized using OCR technology and converted to an XML format[1]. Named Entity Recognition software identifies persons, places etc. in the text. The individual chapters, sections and paragraphs are assigned persistent identifiers. Combined with a resolver basename, these identifiers are realized as URIs[2].

[1]More specifically, the FoLiA format `http://ilk.uvt.nl/folia/`
[2]A sample paragraph URI: `http://resolver.loedejongdigitaal.nl/nl.vk.d.3.2.17.1`

# 2 Using the Back-of-the-Book Index

The back-of-the-book index contains 15.234 lemmas refering to pages in the text. It was carefully constructed and curated by domain experts as part of the publication de Jong's work. In our setup it serves as a structured vocabulary that provides a gateway from external data sources into the text itself. To achieve this, the index (in XML) was converted to RDF using the SKOS schema, where each entry yields a SKOS concept. The data source was loaded in a ClioPatria triple store and exposed as Linked Data. We then aligned this data with other Linked Data sources including GeoNames and a national war historic thesaurus (NIOD Thesaurus). All datasources are available through the Verrijkt Koninkrijk semantic server at `http://eculture.cs.vu.nl:1914`.

## 2.1 Exploiting the external Links

The back-of-the-book index provides an excellent gateway from other structured sources into the text. For example, through the link with GeoNames, we can construct a simple SPARQL query that generates all paragraphs that mention a city or village in a given region. Another example is a query which links paragraphs through the index and the local NIOD thesaurus to WW2 images[3] that are annotated using this thesaurus. For the sake of brevity, these queries are presented on a separate website at `http://tinyurl.com/vk-sparql`.

## 2.2 Historical Information Extraction.

We are currently setting up an experiment where a historian can use the infrastructure to gather information related to the pillarization research question. In a first step, the historian has defined 7 of such 'pillars' (Catholicism, Socialism etc.) and selected 10 back-of-the-book concepts for each pillar concept. Through the index, the historian is able to retrieve all paragraphs related to a pillar. In a next step, the contents of these paragraphs will be analyzed in multiple ways. These include comparing word frequencies and co-occurrence across the pillars and combining this with the NER results. Through this, we expect that the historian will be able to identify patterns in the text which will allow him to more efficiently investigate the nature of pillarization in the work of de Jong.

[3]From BeeldbankWO2 `http://beeldbankwo2.nl`