1. Project Title & Acronym and Abstract

Title: Verrijkt Koninkrijk (Enriched Kingdom)
Acronym: VK
Abstract: 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 29 volumes still combine the qualities of an authoritative work for a general audience, and an inevitable point of reference for scholars. The aim of this project is twofold; in the *demonstrator* part of the project advanced tools and techniques are applied to gather data on De Jong's much debated perception of pillarization and group identity. In the resource curation part of the project the corpus will be enriched and made available to the CLARIN-community for further research. The overall budget for the project is € 118,503 and the partners are: NIOD, UvA, VUA, Meertens and DANS.
Target Start Date: 01-01-2012
Target End Date: 20-12-2012
Type: The project is a combination of a Demonstrator Project and a Resource Curation Project
Call: Open Call

2. Coordinator

Name: Dr. Kees Ribbens
Function: Researcher
Organization: NIOD Institute for War, Holocaust and Genocide Studies
Address: Herengracht 380, Amsterdam
E-mail: k.ribbens@niod.knaw.nl
Tel: 020 52 33 730
Role(s): User, Data Provider

3. Composition of the Project Team

Name: Dr. Maarten Marx
Function: Assistant Professor
Organization: University of Amsterdam
Address: Science Park 904, Amsterdam
E-mail: maartenmarx@UvA.nl
Tel: 020-525 28888
Role(s): Technology Provider

Name: Dr. Victor de Boer
Function: Researcher
Organization: Vrije Universiteit Amsterdam (VUA)
Address: De Boelelaan 1081a
E-mail: v.de.boer@vu.nl
Tel: 0647418031
Role(s): Technology Provider

Name: Dr. Marjan Grootveld
Function: Information Expert
Organization: DANS-KNAW
Address: Anna van Saksenlaan 10, Den Haag
E-mail: marjan.grootveld@dans.knaw.nl
Tel: 070-3446484
Fax: 070-3446482
Role(s): Technology Provider, Infrastructure Specialist

Name: Marc Kemps-Snijders
Function: Information Expert
Organization: Meertens Instituut
Address: Joan Muyskenweg 25
E-mail: marc.kemps.snijders@meertens.knaw.nl
Tel: 020 462 85 50
Role(s): Technology Provider, Infrastructure Specialist

Name: Tim Veken, MA
Function: Information Specialist Collections
Organization: NIOD Institute for War, Holocaust and Genocide Studies
Address: Herengracht 380, Amsterdam
E-mail: t.veken @niod.knaw.nl
Tel: 020 52 33 830
Role(s): User, Data Provider

4. CLARIN centre

The curated resources will be made available by DANS-KNAW for the long-term (*archival environment*) and the web application by Meertens (*live environment*).

5. Requested Budget:

€ 118.503,00

6. Description of the Proposed Project

1.1 Research Question(s)

While Dutch society disposed itself from the pillarization (in Dutch: 'verzuiling') during the 1960s and 1970s, director Lou De Jong of the State Institute for War Documentation (currently NIOD) wrote the largest share of his monumental series *Koninkrijk der Nederlanden in de Tweede Wereldoorlog*. The project *Verrijkt Koninkrijk* revolves around the central research question: *how did De Jong, as sole author of Koninkrijk, portrays Dutch society as a collectivity of various groups?* The question is not only where and how frequent De Jong explicitly uses the concept of *verzuiling* or synonyms/related terms, but also which other terms are in its proximity. And with which nouns, adjectives, and adverbs that can be distinguished as 'positive' or 'negative' (such as e.g. *hokjesgeest* - parochialism) does De Jong associate pillarization terms in a polarizing way?

In the 20th century Dutch society was divided into *protestant-christian*, *catholic*, *social-democrat* and *liberal* pillars. A comparison with three other specific subnational communities representing essential positions in the Dutch WW2 history - *communists*, *national-socialists*, and *Jews* - will help to understand the dynamics of collective identity. The conditions in which these three groups were structured may have varied fundamentally, but their inclusion in the analysis will strongly illuminate the contrast and cohesion between all groups, as perceived and portrayed by De Jong - who has been accused of a too polarized depiction of certain positions in WW2.

We will analyse where and how frequently De Jong refers to the 7 identified pillars/communities and to the institutions related to specific pillars (political parties, churches, media, schools, unions and other interest groups) and to persons that can be identified as representatives of a certain pillar or community. With what possible stereotypes does De Jong associate these groups, their institutions and representatives? And where and with which frequency does De Jong present examples of consultation, cooperation, exchange, and other forms of the concept of rapprochement between pillars/communities, their institutions and representatives?

Finally, the elements of time and space will be addressed. The distribution of findings in the various volumes may show a development during De Jong's writing process. Named entity recognition will identify topographical names and their proximity to pillarization and rapprochement terms, helping to clarify the geographical distribution and regional importance of both phenomena.

Pillarization is a concept subtly but distinctly woven into the fabric of *Koninkrijk*. De Jong's lack of explicit attention for it, makes it a highly intriguing topic for a computerized analysis. In a more general sense the project *Verrijkt Koninkrijk* will explore a method to increase our understanding of collective identities, stereotyping and group cohesion, as well as our understanding of the personal element in history writing.

1.2 Research Data

The project curates the series *Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog*. The data will be made available by NIOD as scanned images and OCRed text, in three different formats: JPEG and RTF per page, searchable PDF per volume. The series consists of 30 volumes (apr.19,000 pages). All participants are familiar with the research data.

1.3 Technology

6.3.1. Pre-processing and curating the data

Before enrichment the data will be pre-processed and turned into XML by using the XML transformer developed at UvA for the PoliticalMashup project [ref LREC 2010]. This transformer preserves reading order, chapter, page and paragraph information, extracts page numbers and detects and tags non-running text like headers and footers. The transformer also adds unique identifiers to each document, page and paragraph. Using the urn:nbn namespace of the UvA, these identifiers are worldwide unique and persistent.

6.3.2.Enriching the data

The UvA will enrich the corpus with the following techniques:
* *Tokenization*, *sentence splitting*, *part-of-speech tagging* and *lemmatization* is done with the FROG software from Tilburg University.
* *Named entity recognition* will be done using UvA's NE tagger (specially trained for Dutch within the Stevin DuoMan project).
* *Polarity tagging* (positive/negative connotation of words) will be done using UvA's FietsTas software (developed for Dutch within the Stevin DuoMan project).
* *Named entity reconciliation* by linking to Wikipedia will be done using software developed by Edgar Meij (UvA).

The corpus will be published in the Sonar XML format. Unlike Sonar, we guarantee that all encodings are UTF-8.

Further enrichment of the corpus is done by the VUA, that will use its experience in semantic technologies to provide advanced curation of and metadata-based access to the corpus. This is achieved by constructing authority lists, structured vocabularies and thesauri. These will be processed from a number of already existing sources:
a) The NIOD thesaurus, which is based on the existing 'trefwoordenlijst' and currently used for cataloguing library and image objects. This vocabulary will be re-structured and represented using semantic technology standards such as SKOS.
b) A source-specific vocabulary, extracted from indices accompanying the corpus ('Register op de wetenschappelijke editie' and back-of-book registers) and Named Entity Recognition results. The thesauri will have information on 1) persons 2) places 3) time intervals and 4) relevant contexts and will be mapped to the NIOD thesaurus.

The result of the enrichment and mapping of the vocabularies is a semantic network of well-curated data and metadata. VUA will facilitate the exposition of the data to the Linked Data Cloud.

6.3.3. Search engine

UvA will create a best-entry-point faceted search engine for the *Koninkrijk* corpus with advanced search capabilities. This will be comparable to the search engine build for the Dutch Parliamentary Proceedings in Clarin 2.

6.3.4. Documentation

The user documentation of the search engine, API documentation and technical documentation for developers will all be part of the web-based demonstrator and publicly accessible as such.

1.4 Description

The research questions on identity of groups in *Koninkrijk* can be addressed by:
- frequency counts of specific terms, taking into account the subject-and time-related context in which they are used.
- proximity searches by combining specific terms and analyzing the context.

Opening up the corpus by applying SKOS and ISOcat and publishing it under Creative Commons license (CC BY-NC-SA 3.0) will ensure that *Koninkrijk* can be used be by a large community of developers, researchers and the general public.

By linking related resources and other open information resources like e.g. Wikipedia, will enable a large user community to re-use De Jong's work and perform further research.

Verrijkt Koninkrijk will benefit from the experience gained in the CLARIN-project War in Parliament. There are however a number of essential characteristics that make it a fundamentally different project:

- *Koninkrijk* is a corpus on one specific knowledge domain (WW2), produced by one person, throughout a period of 22 years. It is in many ways a unique information hub for many different types of research, with the added value that it allows analyses throughout a consistent period in time.
- *Koninkrijk* is less structured than for instance the Parliamentary Papers, which sometimes have already been pre-defined in specific classes.

De Jong's *Koninkrijk* is a series with a large public appeal; it is bound to attract a lot of attention from researchers, as well as the general public.

1.5 Plan

Type: Combination a Demonstrator Project and a Resource Curation project

Demonstrator project:
The core component of the demonstrator is built on an XML database system with full text search capability. Interaction with the database is through a REST API which returns data in XML. The API offers all the functionality of the website-demonstrator plus a limited form of XQuery interaction. Of course all documents and fragments can be retrieved by their permanent identifier. The software behind the core component is already used at UvA.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
XML Database	UvA	0,5PM	No	No	D4
IR module	UvA	0,5PM	No	No	D4
API	UvA	0,5PM	No	No	D4

The web-based application: UvA will create a best-entry-point faceted search engine for the de Jong corpus with advanced search capabilities. This will be comparable to the search engine build for the Dutch Parliamentary Proceedings during Clarin 2. Interaction with the core component is through the REST API. The whole application is build using XQuery, XSLT and standard web technology (HTML, CSS, Javascript, AJAX).

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
Basic search	UvA	0,5PM	Yes	No	D6
Advanced search	UvA	0,5PM	Yes	No	D6
User interface	UvA	0,5PM	No	No	D6

Classifier: For the NE and Timex taggers we use ISO standard categories.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
NE recognizer and reconciliation	UvA	1PM	No	No	D5

Demonstration scenario: Specialized knowledge of the subject is imperative to make the project successful because an effective search requires a sophisticated use of metadata**.** Expertise regarding these metadata is present at NIOD. The research knowhow and experience of Ribbens is indispensable for securing appropriate research questions to make the results of this project effective.

The demonstration scenario will result in: (1) requirement specification which reflects the needs of users in the humanities (D8). (2) an enhanced publication demonstrating the research potential of using language resources in the historical sciences (D9).

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
Requirements specification	NIOD	1PM	No	No	D8
Enhanced publication	NIOD	3PM	No	No	D9

Documentation: The user documentation of the search engine will be part of the website. The API is described using standard attribute-value pairing. The clean interaction between front-and-backend through the API and the creation of the website using

declarative programming languages XQuery and XSLT ensures an easily maintainable system. Developed documentation will be web-based and fully text searchable.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
User documentation	NIOD	1PM	No	No	D6
API documentation	UvA	0,25PM	No	No	D4
Developer documentation	UvA	0,25PM	No	No	D4

Clarin centre and curation plan: All components are built using open source technology which runs on every LinuX system. The complete system will be developed on a dedicated server at the UvA. For future access we distinguish between a live environment (hosted by Meertens Institute) and an archival environment (DANS); the plan for the latter is described under "Resource Curation plan" hereafter. UvA will transfer the data to DANS and all developed software for hosting the demonstrator website to Meertens.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
Data and metadata migration	Meertens	0,5PM	No	No	D3
Frontend installation	Meertens	0,25PM	No	No	D7
Backend installation	Meertens	0,25PM	No	No	D7

Plan to make a mapping between any resource-specific categories and ISOcat categories and extend ISOcat with new categories if unavoidable: ISOcat does not have categories yet on which we can map the NIOD thesaurus. Thus ISOcat will be extended with the part of the thesaurus used by our classifier.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
NIOD thesaurus	NIOD	1,5PM	No	No	D10
SKOS	VUA	3,5PM	No	No	D10

Hackathon for third-party developers: emphasis will be on combining external data sources with the LOD version of the Verrijkt Koninkrijk

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
Hackathon	VUA	1PM	No	No	D11

Resource Curation Project
Format resources
The ocred text provided by NIOD will be delivered as PDF files consisting of images with aligned underlying text. Before we can enrich the text we need to preprocess it and turn it into XML. We use the PDF to XML transformer developed at UvA for the PoliticalMashup project [ref LREC 2010]. This transformer preserves reading order, chapter, page and paragraph information, extracts page numbers and detects and tags non-running text like headers and footers. The transformer also adds unique identifiers to each document, page and paragraph.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
Data cleaning and transformation	UvA	2PM	No	No	D1

Metadata
Metadata of the documents for harvesting will be in DIDL format. Semantic metadata of the documents uses CMDI categories and data-values as described in ISO/TC 37/SC 4. Transformation of metadata is done with a hand-crafted, rule-based system implemented in XSLT. No risks.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
Document metadata	VUA	1PM	No	No	D1

Persistent identifiers: UvA assigns persistent identifiers according to the URN:NBN system. These will be harvested by the Dutch URN:NBN resolver (http://www.persistent-identifier.nl/).

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
PIDs for data and fragments	UvA	0,25PM	Yes	DANS	D2
PID's test	UvA	0,25PM	Yes	DANS	D2

Clarin centre and curation plan: Data and metadata will be stored at DANS. The metadata will be made harvestable via the OAI-PMH protocol.

Task	Participant	Effort	Compliancy test	External IS required	Related Deliverable
Long-term storage and access	DANS	1PM	No	No	D3

A plan for making a mapping between the resource-specific linguistic categories and ISOcat categories: See part under 'demonstrator'

7. Deliverables and Milestones

Deliverable	Title	Type	Responsibility	Month of completion
D1a	Data and Metadata transformation	DATA	UvA	M3
D1b	Document metadata	DATA	VUA	M3
D2	PIDs for documents and fragments	M	UvA, DANS	M4
D3	Permanent storage of data and metadata	M	Meertens, DANS	M11
D4a	Demonstrator backend and API	SW	UvA	M6
D4b	Demonstrator API documentation	SW	UvA	M6
D5	Classifier	SW	UvA, VUA	M8
D6a	Demonstrator frontend	SW	UvA	M9
D6b	User documentation	DOC	NIOD	M10
D6c	Developer documentation	DOC	?	M10
D7	Demonstrator installed at Clarin centre	M	Meertens	M10
D8	Requirements and desiderata for CLARIN infrastructure, in particular for historical research	DOC	NIOD	M6
D9	Demonstrator scenerio enhanced publication	DOC	NIOD, UvA	M12
D10	Mapping to and extension of ISOCAT	DATA	NIOD, VUA	M12
D11	Hackathon	M	VUA	M12

8. IPR and Ethical Issues: Risks

NIOD owns the copyright to the corpus, with the exception of the illustrations, which will be removed in advance. The resulting data will be placed on CLARIN servers and can freely accessed and used by everyone who has access to these servers.

9. Expertise of the applicant(s)

DANS

- Dr. Marjan Grootveld is a computational linguist. She is project manager and is managing the DANS efforts for CLARIN.

Meertens

- Marc Kemps-Snijders has been involved in the CLARIN project on both the European and Dutch level. He is currently Head of Development.

NIOD

- Dr. Kees Ribbens is a cultural historian. He obtained his PhD at Utrecht University in 2001 and is specialized in collective memory and the dynamic relations between history (writing) and identity. He has published widely in the field of 20th century Dutch history, in particular concerning World War II and its aftermath.
- Tim Veken is information specialist.

UvA

- Dr. Maarten Marx has a Master in Political Science and a PhD in mathematical logic. Currently working at a computer science department, his main research interests are in semi-structured data. Within his PoliticalMashup project he integrates textual political data from a variety of sources and countries within one data warehouse system. Dr. Marx is involved in the CLARIN-project War in Parliament.

VUA

- Dr. Victor de Boer has extensive research experience in using and developing semantic technologies in the Cultural Heritage domain and has worked on the MultimediaN E-Culture and EuropeanaConnect projects. He is currently employed as a post-doctoral researcher at the Web & Media group at the VU University Amsterdam.

10. Project budget details

Participant	Organization	Effort (PM)	Salary Costs/PM (Euro)	Salary Costs (Euro)	Travel & subsistence (Euro)	Total (Euro)	
Kees Ribbens	NIOD	5	€ 5.330,00	€ 26.650,00	€ 1.250,00	€ 27.900,00	
Tim Veken	NIOD	1,5	€ 5.708,00	€ 8.562,00	€ 375,00	€ 8.937,00	
Victor de Boer	VU	5,5	€ 5.330,00	€ 29.315,00	€ 1.375,00	€ 30.690,00	
Maarten Marx	Uva	7	€ 5.330,00	€ 37.310,00	€ 1.750,00	€ 39.060,00	
Marjan Grootveld	DANS-KNAW	1	€ 5.708,00	€ 5.708,00	€ 250,00	€ 5.958,00	
Marc Kemps-Snijders	Meertens	1	€ 5.708,00	€ 5.708,00	€ 250,00	€ 5.958,00	
Total				**€ 113.253,00**	**€ 5.250,00**	**€ 118.503,00**	

Literature

J.Th.M. Bank and P. Romijn, eds., *Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog. Deel 14: reacties*, Sdu Uitgeverij Den Haag, 1991

J.C.H. Blom and J. Talsma (red.), *De verzuiling voorbij: godsdienst, stand en natie in de lange negentiende eeuw,* Amsterdam: Spinhuis 2000

P. van Dam, *Staat van verzuiling. Over een Nederlandse mythe*, Amsterdam: Wereldbibliotheek 2011

L. de Jong, *Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog*, Staatsuitgeverij, Den Haag, 1969-1991

L. de Jong, *De bezetting: tekst en beeldmateriaal van de uitzendingen van de Nederlandse Televisie-Stichting over het Koninkrijk der Nederlanden in de Tweede Wereldoorlog, 1940-1945*, Querido, Amsterdam, 1966

M. de Keizer, red., *Een dure verplichting en een kostelijk voorrecht. Dr. L. de Jong en zijn geschiedwerk*, Sdu Uitgeverij Den Haag, 1995

H. Knippenberg & H. van der Wusten, 'De zuilen, hun lokale manifestaties en hun restanten in vergelijkend perspectief' in: C. van Eijl, L. Heerma van Voss & P. de Rooy

(red.), *Sociaal Nederland, contouren van de twintigste eeuw*, Amsterdam: Het Spinhuis IISG 2001, 129-150.
Dr. C. Kristel, *Geschiedschrijving als opdracht. Abel Herzberg, Jacques Presser en Loe de Jong over de Jodenvervolging*, Meulenhoff, Amsterdam, 1998
A. Lijphart, *Verzuiling, pacificatie en kentering in de Nederlandse politiek*, Amsterdam: De Bussy 1968 [http://www.dbnl.org/tekst/lijp001verz01_01/lijp001verz01_01.pdf]
F.P.I.M. van Vree, *In de schaduw van Auschwitz. Herinneringen, beelden, geschiedenis*, Historische uitgeverij Groningen, 1995