DutchParl 1.0
A Corpus of Parliamentary Documents in Dutch

Maarten Marx and Anne Schuth
ISLA, University of Amsterdam
Kruislaan 403 1098 SJ Amsterdam, The Netherlands
maartenmarx@uva.nl aschuth@science.uva.nl

ABSTRACT

A corpus called DutchParl is created which aims to contain all digitally available parliamentary documents written in the Dutch language. The first version of DutchParl contains documents from the parliaments of The Netherlands, Flanders and Belgium. The corpus is divided along three dimensions: per parliament, scanned or digital documents, written recordings of spoken text and others. The digital collection contains more than 800 million tokens, the scanned collection more than 1 billion.

All documents are available as UTF-8 encoded XML files with extensive metadata in Dublin Core standard. The text itself is divided into pages which are divided into paragraphs. Every document, page and paragraph has a unique URN which resolves to a web page. Every page element in the XML files is connected to a facsimile image of that page in PDF or JPEG format. We created a viewer in which both versions can be inspected simultaneously. A search-engine for the complete collection is available online.

The corpus is available for download in several formats. The corpus can be used for corpus-linguistic and political science research, and is suitable for performing scalability tests for XML information systems.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

Keywords

Dutch, Text corpus, Politics, XML

1. INTRODUCTION

The aim of DutchParl is to create a corpus containing

> all digitally available parliamentary documents written in the Dutch language.

The main reason to create the corpus is to provide one portal from which these documents are accessible both in their original official version (in PDF format), and in a uniform XML format with extensive metadata [9]. The corpus was designed to be useful as a data set in all possible scientific disciplines. E.g., it can be used for (comparative) corpus-linguistic and political science research and as a test-set for information-theoretic experiments. This distinguishes DutchParl from EuroParl [6] which is developed for research in Statistical Machine Translation. The corpus was developed following the guidelines set out in [9].

One of the main difficulties with political data is the lack of permanent identifiers to documents. This simple fact hinders correctly referencing data-sources and to (re-)retrieve data, and thus makes it almost impossible to replicate or extend research. In the DutchParl corpus, every digital object has a unique permanent identifier in the form of a Uniform Resource Name (URN) [8] which resolves to a digital object and its associated metadata. This conforms to the recommendations of publishing eGovernment material as set out by the eGov working group of the W3C [2].

DutchParl distinguishes three types of digital objects: documents, pages and paragraphs. This facilitates fine grained referencing. More importantly, re-use and integration of the data with other datasets, as advocated in the LinkedData initiative [3, 1], becomes easy and reliable.

Besides making the data available for bulk download we created a search-engine from which the corpus can be queried with NEXI expressions [7]. The corpus is updated every night. Thus the search engine functions as a mediator over a number of data providers.

The paper is structured as follows. Section 2 describes the coverage and the size of the corpus and its partition into subcorpora. Section 3 describes the data format and the data collection process. Evaluation of the quality of the corpus is in Section 4. Section 5 describes some additional datasets which can be used directly for corpus-linguistic research. Section 6 lists other parliamentary corpora and related work. Section 7 concludes.

How to get the corpus?

The corpus is available for download at `http://politicalmashup.nl/DutchParl`. We are not aware of copyright restrictions on the material. If you use the corpus, please sent an email to `maartenmarx@uva.nl`.

2. COVERAGE AND SIZE OF DUTCHPARL

Spatial and temporal coverage.

Parliamentary documents in the Dutch language are pro-

DIR2010 Nijmegen

duced in the following locations:

Belgium Flemish parliament, and the Belgian federal parliament.

European Union Original texts by Dutch speaking members (Belgium and The Netherlands), and translations into Dutch.

The Netherlands Dutch parliament.

Suriname National parliament

The present version of DutchParl does not yet contain data from the EU nor from Suriname. We further exclude these two sources from the description.

The periods for which data is available differ per source. Table 1 lists the periods for which digital and scanned data is available on the web for each source (measured in September 2009). This is exactly the data available in DutchParl.

Subcorpora.

The corpus can be divided into many subcorpora. This is facilitated by the uniform metadata using a controlled vocabulary. In the description below we partition the data along three dimensions. First by source: Belgium, Flanders and The Netherlands. Secondly, digitally produced documents are separated from scanned and OCR-ed documents. The latter contain noise in the form of wrongly recognized characters, mistakes in paragraph splitting, non UTF-8 characters, or simply no extractable text.

A special subset of the parliamentary documents are the verbatim notes of sessions of parliament. These can be plenary sessions or sessions of, usually smaller, committees. Even though the texts are edited and transcribed to be read, they are accounts of spoken language. For this reason, we present details both for the complete collections and for the verbatim notes separately.

Size of DutchParl.

Table 2 displays information about the size of the subcorpora. We list the following information: the size of the text of the documents in Megabytes; the number of documents; the number of pages in the original documents; and the number of tokens. Except for the OCR-ed text from the Netherlands, these numbers were obtained from the original PDF files using `pdfinfo` for the page counts and `pdftotext` followed by the Unix command `wc -w -c` for the token and byte counts.[1] The OCR-ed text from the Netherlands is available in the form of XML files and we obtained the figures directly from these XML files (The size in GB is the size of the raw text with XML tags).

We group these numbers for the three different parliaments, and separate the counts for the digital and the OCR-ed documents. The numbers for the verbatim notes are given separately.

We note that the documents from the Belgian parliament are bilingual, with text in Dutch and French interspersed in many different ways. The following paragraph contains counts for the Dutch tokens only.

The majority of Belgian and Flemish documents are verbatim notes. For the Netherlands the opposite holds. The Belgian and Flemish meeting notes come in one document for a day (with an average page length of 39 and 24, respectively). In the Netherlands the notes of one day are divided over a number of documents, corresponding to the number of topics discussed that day. This accounts for the much lower average page length of 6.3 per document.

Number of tokens.

Table 4 presents figures on the number of tokens occurring in the different subcorpora. Again we make a distinction between digital and scanned documents and present the numbers for the spoken texts separately. Tokenization was done as follows. We used the pure text files as described in the previous paragraph. On these files tokens were split on the regex `\W`, all tokens were lower-cased and leading and trailing whitespace was removed. The counts for the bilingual Belgian federal corpus consist of the words occurring in paragraphs that were detected as being in Dutch.

Official documents contain a large number of, mostly numeric, codes referring to other documents. From a corpus-linguistic point-of-view these are not very interesting. For that reason, we restricted the counts in Table 4 to tokens which contain at least three consecutive alphabetical characters. To get a feeling of the differences in counts, Table 3 presents the counts of all tokens and the adjusted counts for the digital meeting notes of the Flanders parliament.[2]

	All tokens	**Tokens with ≥ 3 consecutive letters**
total # tokens	50.549.284	38.629.223
unique tokens	267.005	258.304
occurring once	121.278	118.992
occurring ≥ 2	145.727	139.312
occurring ≥ 4	93.344	88.518
occurring ≥ 20	38.927	36.413

Table 3: Token counts for the digital meeting notes of the Flanders parliament.

3. TECHNICAL DESCRIPTION

3.1 Description of the data format

Every document in the DutchParl corpus is a UTF-8 encoded XML file which is valid with respect to the Relax NG schema in compact syntax[3] in Table 5. The description of the metadata is postponed to Table 6. We briefly describe the structure of the documents. The root element `root` of each document has three children:

meta this element contains meta-information of the document described using the 15 elements from the Dublin Core Metadata Element Set Version 1.1[4];

header this element contains textual data extracted from the source-text which may be used for displaying purposes;

[1]Both PDF commands are part of the Xpdf software, see `http://www.foolabs.com/xpdf`.

[2]The slight difference between the total number of tokens here and that in Table 2 is due to the different way of tokenization used.

[3]`http://relaxng.org/compact.html`

[4]`http://dublincore.org/documents/dces/`

Source	Digital	OCR-ed	Planned
Belgium	From 1999-07-01	-	1844–1999 is scanned
Flanders	From 1995-10-17	1971-12-07 to 1995-10-17	-
The Netherlands	From 1995-01-01	1917-01-01 to 1995-01-01	1814–1917 available in 2010

Table 1: Availability of parliamentary data in the Dutch language.

Subcorpus	Mbyte text	# Documents	# Pages	# Tokens
Belgian Federal	800	3.901	216.522	129.085.483
Flanders	454	5.470	161.881	72.958.408
Netherlands	4.331	198.433	1.594.845	684.932.669
Flanders OCR	146	1.018	34.867	23.924.567
Netherlands OCR	7.043	328.722	1.701.130	1.003.555.596

Subcorpus	Mbyte text	# Documents	# Pages	# Tokens
Belgian	502	3.462	137.366	81.086.575
Flanders	311	3.799	93.591	50.715.218
Netherlands	781	21.604	137.610	131.681.453
Flanders OCR	142	932	33.147	23.378.215
Netherlands OCR	2.644	12.796	383.863	402.657.396

Table 2: Number of documents, pages and tokens for the complete corpus (top) and only for verbatim notes of parliamentary and committee sessions (bottom).

text this element contains the complete text of the source document. Each `text` element has one or more `page` elements (corresponding to physical pages of the document), which in turn are divided in one or more `p` (for paragraph) elements.

Within the `text` element there is a strict separation between content and metadata. All metadata is stored in attributes. All text is contained in the `p` elements. The XPath expression `doc('file.xml')//text//text()` will return the complete text of the source document.

The attributes of the `page` and `p` elements contain provenance information [5]. The `root, page` and `p` elements have an obligatory `docno` attribute whose value is unique in the corpus. Each `page` also has an obligatory `imageref` attribute which points to a facsimile image of that particular page (these can be in PDF or JPEG format). All other attributes are optional. We briefly list them:

originalpagenr an integer denoting the page number of the page in the original document. This is extracted from the text using a special pattern. If the confidence in the extracted value is too low a '-' is given as a value.

class Its value is either "header" or "footer". Determined from the text using heuristics.

top and left Integers denoting the position of the upper left hand corner of the bounding box of the paragraph. The length of each page is normalized to 1000 units.

fulltextref and wordcoordinatesref These are two URLs referring to files which are specific for the Dutch OCR-ed part of the collection.

Dublin Core metadata.

Metadata is described in a uniform way for all sub-collections using the 15 Dublin Core properties. A number of elements obtained a fixed value for the complete DutchParl collection, see Table 6. We briefly discuss the others. `dc:coverage` indicates the country or region of the parliament. `dc:date` refers to the date of the document. This is often hard to determine, and in many cases not available. For documents of `dc:type` "Written Questions" the `dc:date` element is subdivided into the date of the question, the date of the answer and the difference between these two in number of days, whenever these could be obtained from the metadata.

`dc:description` and `dc:title` are free text describing the document.

`dc:publisher` contains the URL of the website from which the data is harvested. `dc:rights` contains the name of the parliament which produced the document. `dc:identifier` contains the URL of the present XML file. `dc:source` contains URLS to the text source and (if available) the source of the metadata.

`dc:type` indicates the kind of parliamentary documents. We distinguish two types: *Verbatim Proceedings* contain the meeting notes of plenary sessions of the parliament; *Written Questions* contain written question of members of parliament to members of the government and the answers. All other documents obtain type *Parliamentary Documents*.

The properties `dc:relation` and `dc:subject` contain semantic information which is usually not available and needs to be extracted from the text. These are not used yet.

We tried to restrict the fields as much as possible. With the data-type restrictions this may lead to validation errors due to typos or mistakes in the data. For instance, the string `2008-04-31` will not be accepted as being of type `xsd:date`, because that date does not exist.

3.2 Description of the data collection and processing

Each part of the corpus needed its own specialized data-collection, extraction and transformation scripts. We de-

	NL-DIGITAL	NL-SCAN	Flanders DIGITAL	Flanders SCAN	BE-federal
Total number of words	514087570	782029017	9081282	13668172	61579706
Unique words	2583035	3601829	142705	195416	445100
Words occurring just once	1219262	2228857	61429	93690	174395
Words occurring more than once	1363773	1372972	81276	101726	270705
Words occurring at least 4 times	852703	762736	49564	60390	165563
Words occurring at least 20 times	334891	264612	17784	22229	65228

	NL-DIGITAL	NL-SCAN	Flanders DIGITAL	Flanders SCAN	BE-federal
Total number of words	102870201	329540359	38629223	17120704	41152224
Unique words	353677	1963712	258304	184945	245447
Words occurring just once	149719	1311243	118992	91889	102093
Words occurring more than once	203958	652469	139312	93056	143354
Words occurring at least 4 times	130008	370932	88518	57277	90911
Words occurring at least 20 times	55054	134735	36413	22945	37250

Table 4: Token counts; all data (top) and verbatim notes of parliamentary sessions (bottom).

scribe here the main steps common to all subcorpora. The next section contains an evaluation of these steps.

Analysis: determine where on the web a corpus is located; determine its scope and see what kind of metadata are available for each document.

Harvest: collect the sources of the texts and the corresponding metadata.

Transform: turn the metadata into the uniform Dublin Core format. Extract the text from PDF files and store in UTF-8 format. Create PDF files for each page. Use text-analytics to determine headers and footers, to extract page-numbers, and to partition each page into paragraphs. Perform language detection on the level of paragraphs, for the bilingual documents from Belgian, and on the document-level for all documents.

Compose, validate and store: collect all information together into one XML document; add values for the `docno` attributes, validate against the Relax NG schema; store the XML document on disc and import it into a DBMS. Create pure text and word list files for subcorpora.

4. DATA QUALITY (EVALUATION)

We evaluate completeness and correctness of the DutchParl corpus. Completeness means that every parliamentary document that is published on the official web-pages of the respective parliaments is contained in DutchParl and nothing more. Correctness has a number of dimensions: is the content of the documents faithfully represented in the XML format?, are the metadata correct?, are the XML files themselves well-formed and valid?.

4.1 Completeness

Establishing completeness is difficult for a number of reasons. Most importantly because listings of documents are not available. On top of that, the parliamentary websites do not offer support for harvesting their collection. Instead sites have to be scraped using specially crafted scripts.

The Dutch National Library, which provides access to the Dutch parliamentary data from before 1995 provides a harvesting service according to the Protocol for Metadata Harvesting of the Open Archives Initiative[5]. This protocol uses a two-step process: first harvest a list of permanent identifiers, and then download the documents named by these identifiers. This system works very well. We collected a list of over 1.7 million of XML files. All were downloaded correctly. Only 2 of them were not valid XML after our transformation, both due to non UTF-8 characters in the originals. After consulting with the Dutch National Library these mistakes were repaired and the correct files added.

4.2 Correctness

We now evaluate the transformation and the storage steps described in Section 3.2. Some of these procedures use heuristics and some do not. We start with an evaluation of the latter.

Some of the data in the DutchParl corpus are extracted from the text using heuristic methods. We list these here and evaluate the performance of the used methods. Table 8 contains the figures of the evaluation.

Header and Footer detection Most documents we consider have either a header, a footer or both. These, in a sense, disturb the normal text-flow of the document and should thus be detected as such before we proceed. Furthermore, headers or footers often contain interesting meta data such as page numbers. We detect headers and footers by searching for repeating patterns on the left or right page, allowing for minor discrepancies, such as incrementing page numbers. Once detected, we label these paragraphs elements with attributes `class='header'` and `class='footer'`.

Page number detection From the found headers and footers we collect those tokens that differ from page to page, given that the token is a number. If we can find these numbers for more than half the pages, and if

[5] `http://www.openarchives.org/OAI/openarchivesprotocol.html`

```
# Dublin Core namespaces and our own local addition

namespace dc = "http://purl.org/dc/elements/1.1/"
namespace dcterms = "http://purl.org/dc/terms/"
namespace pm = "http://www.politicalmashup.nl"

# The Elements
start =
    element root {
        Meta,
        Header?,
        Text,
        attribute docno {
            xsd:string { pattern = "[\-\w]+\.[\-\w]+" }
        },
        attribute imageref { xsd:anyURI }?,
        # (the same as source)
        attribute source { xsd:anyURI }?,
        # (the same as imageref)
        attribute metadata { xsd:anyURI }?,
        attribute didlurl { xsd:anyURI }?
    }
Header =
    element header {
        mixed { Paragraph? }
    }
# Mixed content
Text = element text { Page+ }
Page =
    element page {
        Paragraph*,
        attribute docno {
            xsd:string { pattern = "[\-\w]+\.[\-\w]+.\d{4}" }
        },
        attribute imageref { xsd:anyURI },
        # URL to facsimile of the page(PDF/JPEG)
        attribute originalpagenr { "-" | xsd:integer }?,
        attribute fulltextref { xsd:anyURI }?,
        attribute woordcoordinatesref { xsd:anyURI }?
    }
Paragraph =
    element p {
        text,
        attribute language { language }?,
        # ISO 639-1 language codes
        attribute docno {
            xsd:string { pattern = "[\-\w]+\.[\-\w]+.\d{4}\.\d{3}" }
        },
        #
        attribute class { "h" | "header" | "footer" }?,
        attribute top { xsd:integer },
        attribute left { xsd:integer }
    }
```

Table 5: Relax NG schema for the XML documents in DutchParl.

```
Meta =
    element meta {
        (element dc:contributor { "http://www.politicalmashup.nl" },
         element dc:coverage {
             mixed { external "country.rnc"? } #list of ISO31663166 country codes
         }+,
         element dc:creator { "http://www.politicalmashup.nl" },
         element dc:date {
             xsd:date
             | # yyyy-mm-dd
               element pm:parliamentary-year {
                   xsd:string { pattern = "\d{4}/\d{4}" }
               }
             | (element pm:dateQuestion { xsd:date },
                element pm:dateResponse { xsd:date },
                element pm:ResponseDuration { xsd:integer })
             |
               empty
         },
         element dc:description { text },
         element dc:format { "text/xml" },
         element dc:identifier { text },
         # URL of this XML file
         element dc:language { language }+,
         element dc:publisher {
             "http://statengeneraaldigitaal.nl"
             | "http://parlando.sdu.nl"
             | "http://www.vlaamsparlement.be"
             | "http://www.dekamer.be"
         },
         element dc:relation {
             element dcterms:media  {
                (text | xsd:anyURI  ),
                attribute mediatype {"audio" |"video"|"other"}
                }*,
             element pm:dossiers {
                 text
                 | element item { text }*
             },
             element pm:person {
                 text
                 | element item { text }*
             }
         },
         element dc:rights { text },
         element dc:source {
             element pm:textsource { xsd:anyURI }?,
             element pm:metasource { xsd:anyURI }?
         },
         element dc:subject {
             element pm:legislative_period { text }?,
             element pm:session_number { text }?,
             element pm:keywords {
                 element item { text }*
             }?,
             element pm:categories {
                 element item { text }*
             }?,
             element dcterms:abstract {
                 element item { text }*
             }?
         },
         element dc:title { text },
         element dc:type {
             "Verbatim Proceedings" | "Parliamentary Documents" | "Written Questions"
         }) # 3 fixed types
    }

language = xsd:language { pattern = "nl|fr|es|en|de" }
```

Table 6: Definition of the Dublin Core metadata elements.

Subcorpus	# Documents	xmllint		relaxng	
Belgian Federal	3.462	3462	100.0%	3456	99.83%
Flanders	2.284	2114	92.56%	2038	89.23%
Netherlands	198.433	198,421	99.99%	184,274	92.86%

Table 7: The number and percentage of correctly validated xml files. Xmllint just checks for valid xml, relaxng also uses a schema to validate against.

	Correct		Incorrect						N/A	
Pagenumber	87	58.00%			27	18.00%			36	24.00%
Reading order	102	68.00%			48	32.00%				
			too large		**too small**		**other**			
Header detection	120	80.00%	4	2.67%	0	0.00%	26	17.33%		
Footer detection	91	60.67%	5	3.33%	14	9.33%	40	26.67%		

Table 8: Evaluation result for a stratified random sample of 150 pages (50 from each subcorpus; for each subcorpus we choose documents from all three document types). We evaluated whether the correct pagenumber was detected, whether the detected paragraphs where in the right order and how we did with respect to detecting headers and footers.

these numbers are incrementing as expected for page numbers, we assume these are the original page numbers, and tag *all* pages in the document accordingly.

Sort to reading order The text extraction method we use, gives per page a number blocks of text with its original coordinates. Since we want to be able to detect paragraphs in the right order and across columns, it is helpful to detect the number of columns and assign each text block, excluding the previously detected headers and footers, to a column. Once we have done that, sorting the text blocks to reading order comes down to sorting on *column*, then on *top location* and finally on *left location*.

Paragraph detection Now that the text blocks are in reading order we can merge the blocks, that were together in the original document, into paragraphs. This is done using some simple heuristics: we always merge the next text block with the current one, unless one of the following conditions occurs: a) there is no next block, b) the font size of the next block is different, c) the start of the next block is indented, d) the horizontally separating whitespace with the next block is higher than average.

Language detection The Belgian Federal documents are bilingual, in both Dutch and French. Written questions and answers are available in both languages in an aligned translation. In the verbatim proceedings, the spoken text is given in the original language, and a translated summary is provided. There is no systematic way in which one can distinguish the two languages. Thus we used a language-recognizer on the paragraphs. This recognizer uses a simple Bayesian classifier [4], trained on parts of the publicly available EuroParl corpus [6], which has in-domain data in the languages we are interested in. [6]

Table 9 contains an evaluation of the precision. For both languages, we randomly picked 200 paragraphs tagged as being in that language, and containing at least 5 tokens with 3 consecutive letters. We obtain precision scores of .95 and .85 for Dutch and French, respectively. Most mistakes (83%) were in paragraphs with mixed language. In our sample these were all either a mistake of the paragraph splitter or a header or footer which has mixed language by design.

[6]Our implementation uses `http://divmod.org/trac/wiki/DivmodReverend`

	Dutch	French
p's solely in the language	190	170
mixed language	6	27
p's not in the language	4	3
total	200	200

Table 9: Evaluation of the language recognition for the Belgian Federal documents. For both Dutch and French, 200 paragraphs where randomly picked and scored (for both languages: 100 from written questions, and 100 from verbatim notes).

5. CORPUS LINGUISTICS

As an illustration of the wealth of the corpus we present a small investigation into political compound-creation.

Politicians are renown for creating new words. Try to make sense of for instance *antibioticasensibiliseringscampagne*. The Dutch language allows complex compounds , so we may expect rather long words. Here we list some results from the Belgian corpus.

Word length	# Words
≥ 25	3793
≥ 30	453
≥ 35	53
≥ 40	10

We found 55 tokens of at least 35 characters. Manual inspection showed that 2 of these were errors.[7] Here are the 10 (lower-cased) tokens of at least 40 characters, together with the number of occurrence:

[7]Both were repetitions of a short token, e.g., *huishuishuishuishuishuishuishuishuisbezoek*.

```
4 verpleegkundigenverzekeringsinstellingen
2 brigadecommissarissenverbindingsambtenaren
1 verzekeringstegemoetkomingonderhoudsgeld
1 verplegingsinrichtingenverzekeringsinstellingen
1 standaardluchtvaartbeveiligingsmaatregelen
1 rechtsbijstandsverzekeringsovereenkomsten
1 kwaliteitswarmtekrachtkoppelingsinstallaties
1 kwaliteitswarmtekrachtkoppelingseenheden
1 burgerlijkeaansprakelijkheidsverzekering
1 brigadecommissarissenverbintenisambtenaren
```

It seems likely that the second and the last item in the list refer to the same concept.

6. OTHER PARLIAMENTARY CORPORA AND RELATED WORK

The best known parliamentary corpus is probably EuroParl[8] [6]. Its latest version (V3) contains the verbatim notes of the plenary sessions of the European Parliament from April 1996 until October 2006, in UTF-8 but not in XML format. For Dutch, it has almost 40 million words. Its main purpose is to train statistical machine translators. Probably for that reason, it has hardly any metadata nor any provenance information. Thus for political science research, EuroParl seems insufficient. Unfortunately one cannot consistently extract all text originally spoken in Dutch. This is indicated in the `LANGUAGE` attribute of the `SPEAKER` element, but this attribute is often missing.

Many parliaments make their proceedings (often called Hansards) available online. The Inter-Parliamentary Union (IPU) is the international organization of Parliaments. IPU maintains a useful list of parliamentary websites at `http://www.ipu.org/english/parlweb.htm`. A useful list of both official and alternative websites offering access to parliamentary information is avaliable at `http://en.wikipedia.org/wiki/Parliamentary_informatics` The list is ordered by country.

The debates from the British Hansard collection are available in XML format with high OCR quality. `http://www.hansard-archive.parliament.uk/` contains the digitised debates from 1803 until 2004. Unfortunately the XML contains no directly resolvable links to the images of the scans. From 2004 the debates are available in XML (unfortunately using a different schema) from `http://www.theyworkforyou.com`.

The website `www.ikregeer.nl` provides an API to collect parliamentary documents from The Netherlands published after 1995 in PDF-format.

A large collection of political writings in many languages is available at `http://www.marxists.org`.

Making governmental and/or political data easily accessible through the internet is a major research area with a lot of ongoing activity. The W3C has a special interest group on eGovernment (`http://www.w3.org/2007/eGov/`) which encourages governments to publish their data in reusable, linkable, human- and machine-readable formats using open standards such as XML, RDF and Dublin Core [1, 2]. Independent non-profit organisations scrape governmental websites and create vertical search engines, mashups or appealing visualizations, e.g. `http://theyworkforyou.com` and `http://capitolwords.org`.

[8] `http://www.statmt.org/europarl/`

7. CONCLUSIONS AND FUTURE WORK

This work started out as a challenging data integration project: Can we collect and bring together under one uniform schema parliamentary data from different countries, produced in different periods of time, and available in different formats? DutchParl showed that we partly succeeded. We created a rich metadata schema based on Dublin Core standards. However, it is not always easy or possible to collect meaningful data for all fields (we did not manage for Belgium Federal). Also, even after many tries and promises, we did not receive the data from Suriname. A hard problem is checking completeness. Even if we are confident that we downloaded all material available on the web, we cannot be sure that we have all material. It is difficult to find reliable independent listings of material.

We paid extra care to providing provenance information [5]. Because we assigned corpus unique ID's to every paragraph, page and document, specific referencing of material (common in the social sciences) is possible using hyperlinks. The connection of the data in XML with the original official publications is quite specific and convenient because we provide a facsimile image of every page.

Future challenges include 1) keeping the corpus daily up to date, 2) managing the data in an XML database managment system, 3) scaling to other countries, 4) linking the data with other datasets, e.g. bibliographies of MP's, 5) performing text analytics on noisy data.

Acknowledgements

Maarten Marx acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Open grant agreement FOX, number FP7-ICT-233599.

Thanks are due to Tim Gielissen, Fons Laan, Marina Lacroix, Arjan Nusselder and Martin Reynaert.

8. REFERENCES

[1] J. Alonso et al. Improving access to government through better use of the web. W3C Interest Group Note 12 May 2009 `http://www.w3.org/TR/egov-improving/`, May 2009.

[2] D. Bennet and A. Harvey. Publishing open government data (W3C Working Draft 8 September 2009). `http://www.w3.org/TR/gov-data/`, 2009.

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

[4] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *International Conference on Machine Learning*, pages 105–112, 1996.

[5] O. Hartig. Provenance Information in the Web of Data. In *Proc. of the Linked Data on the Web Workshop at WWW*, 2009.

[6] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proc. MT summit*, volume 5, 2005.

[7] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In *Advances in XML Information Retrieval*, pages 16–40, 2005.

[8] W3C/IETF URI Planning Interest Group. URIs, URLs, and URNs: Clarifications and Recommendations 1.0. W3C Note 21 September 2001, 2001. `http://www.w3.org/TR/uri-clarification`.

[9] M. Wynne. Archiving, distribution and preservation. In M. Wynne, editor, *Developing Linguistic Corpora: a Guide to Good Practice*, pages 71–78. Oxford: Oxbow Books, 2005. Available online from `http://ahds.ac.uk/linguistic-corpora` [Accessed 2009-07-01].