Digital Sustainable Publication of Legacy Parliamentary Proceedings

Maarten Marx
ISLA, University of Amsterdam
Science Park 107 1098 XG
Amsterdam, The Netherlands
maartenmarx@uva.nl

Nelleke Aders
Dienst Informatievoorziening
Tweede Kamer der
Staten-Generaal
Den Haag, The Netherlands
n.aders@tweedekamer.nl

Anne Schuth
ISLA, University of Amsterdam
Science Park 107 1098 XG
Amsterdam, The Netherlands

ABSTRACT

We address the problem of publishing parliamentary proceedings in a digital sustainable manner. We give an extensive requirements analysis, and based on that propose a uniform XML format. We evaluated our approach by collecting and automatically processing proceedings from six parliaments spanning almost 200 years in total. Most of this data is real legacy data consisting of scanned and OCRed documents. The approach scales very well and produces high quality data.

All documents are transformed into 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.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; E.2 [**Data Storage Representations**]: Linked Representations

Keywords

Knowledge Representation, XML, Parliamentary Proceedings, Linked Data

1. INTRODUCTION

Many democratic countries recently witness a rise in publishing governmental data on the web. Data is made available by different institutions: local and central governments, commercial publishing houses and non-commercial initiatives like theyworkforyou.com. Because many countries create very similar data, e.g. constitutions, tax laws and parliamentary proceedings, it is beneficial to try to standardize the format in which data is published. A standardized format has many advantages: comparative studies are facilitated, software can be exchanged and universally applied, emergence of best practises and a community of expert users. Several initiatives to standardization are or have been taken. An active group is the W3C working group on Egovernment which published two recommendations [**?**, **?**].

This paper focuses on one particular dataset which is produced in almost every democratic country and which has great appeal to both the general public, the media, and the scientific community: the proceedings of the plenary meetings of parliament. These are very well structured, verbatim notes of everything that is being said and that happens during plenary sessions.

An important aspect of parliamentary proceedings is their longitudinal character. In many states, the proceedings are collected for more than 100 years. Numerous states are currently digitizing their legacy using scan and OCR techniques. This trend gave rise to our research question:

> What is the best data format for publishing both legacy and current parliamentary proceedings in a digital sustainable manner?

Our main results are recommendations for representation schemas for the most important data format for publishing open government data, XML [**?**, **?**, **?**]. We evaluated the effectiveness of our representation by collecting almost 200 years of proceedings from five parliaments and transforming these into the common representation.

The paper is organized as follows. Section 2 contains an extensive analysis of the requirements on a good representation. Section 3 gives a detailed description of the XML format we have developed. We describe our data collection and processing in Section 4 and provide an extensive evaluation of the quality of the data. We end with conclusions.

1.0.0.1 Methodology.

We used the following methodology for arriving at the requirements on the representation of parliamentary documents. We surveyed existing comparative scientific research based on parliamentary proceedings and distilled desiderata. We investigated current representations and information systems in six states[1], and we took the recommendations for publishing governmental data as linked data from the W3C [**?**, **?**]. This resulted in a large wish list which no country in our survey could yet satisfy.

We then investigated which parts of the wish list could be fulfilled effectively with fully automated processes. The main criterion used was *scalability*: techniques tailor made for specific time periods in specific states were mostly discarded. We used techniques from information extraction

[1]Austria, Belgium, Germany, The Netherlands, Spain and the UK.

dg.0 2010, May 12-17, 2010, Puebla, Mexico
ACM [ISBN] 978-1-4503-0070-4/10/05

and retrieval [?, ?, ?] to *automatically* convert currently available data into the desired formats. Automatic conversion is an essential requirement because of the vast amounts of legacy data around. Usually this legacy data is only available as scanned and OCRed copies of printed versions.

The dataset reminds us of the youth of the digital age. The eldest proceedings in our survey which are available in an original digital format are from 1995.

We implemented the chosen techniques and tested them by converting proceedings into machine processable format for six parliaments covering almost 200 years. We then evaluated the accuracy of these techniques.

2. REQUIREMENTS ASSESSMENT

To answer our research question we collect requirements on publishing parliamentary data from four different sources. We first investigate the intrinsic qualities of the data itself. Then we survey typical scientific research done on parliamentary data and extract requirements from that. Thirdly, we look at W3C recommendations on publishing governmental data. We finish with a list of features collected from our survey websites publishing parliamentary data.

2.1 Intrinsic qualities

The most valuable characteristic of a collection of parliamentary proceedings is its longitudinal nature. The collection consists of periodic measurement points conducted in a uniform and consistent manner over a (possibly very long) period of time. The data is thus well suited for temporal comparisons. Also, measurements are rather similar across states which facilitates cross-national comparative studies, common in the political sciences.

The collection is a record of spoken language with very rich metadata. For every word spoken in parliament, the following facts are known, and can be extracted from the written proceedings:

1. when it was said,
2. who said it,
3. in what function,
4. speaking on behalf of which party,
5. in which context, and
6. who was actively present during the speech act.

These features enable all kind of groupings and comparisons. Findings in different states may also be compared. It is desirable that a representation makes these six features machine processable.

2.2 Scientific research

We distinguish qualitative and quantitative research, as each comes with their own requirements.

Because of their longitudinal nature, parliamentary proceedings are important data for historical research. It is a goldmine for historic-linguistic and etymological research looking for first (spoken) occurrences of terms. This qualitative research requires powerful search capabilities (e.g. using wildcards for characters to allow for OCR-errors), fast access to processed and raw data (in this case usually the OCRed text and the scanned images, conveniently linked), and the ability to make precise references into the source material (comparable to the very fine-grained reference system of the Bible).

Fields as political science, sociology, communication science and content analysis additionally use quantitative methods to study large amounts of textual data [?]. Modern text analytics techniques from the fields of information retrieval [?] and web data mining [?, ?] are applied here. Examples include agenda-setting research [?], research correlating parts of the political spectrum with specific (e.g. populist) language [?], and trend detection in media and parliament [?]. This research uses exactly the six features from the previous section.

The Text Encoding Initiative (TEI `http://www.tei-c.org`) publishes XML schemas for various kind of publications, but not for parliamentary proceedings.

2.3 W3C recommendations

The W3C created three notes on publishing government data [?, ?, ?]. The main points are:

- make data both machine and human readable;
- link data, make data linkable, provide permanent identifiers for each government object and data item;
- provide metadata using common standards (e.g. Dublin Core);
- make the data as easy to reuse (e.g. in mashups) as possible.

Tim Berners Lee [?, ?] emphasizes the fact that government data should be published as *linked data*. This means that it is open (expressed in non-proprietary formats; XML and RDF are preferred), modular (data can be combined with other pieces of data), and scalable.

According to [?], "much public sector information was and is stil being published using proprietary formats or in ways that create barriers of use for various interested parties". Potential benefits of open and linkable data include multiple views (e.g., list everything being said by MP X), reuse of information, improved web search and data integration.

[?] also mentions provenance and trust explicitly, here in connection with mashups. The data-format should make it very easy to refer and return to the original data source, both for machines and humans. Making data linkable using permanent identifiers is also recognized by the OECD who use Digital Object Identifiers (DOI's) for permanent links [?].

The eight principles published by *The Open Government Group* (`http://www.opengovdata.org`) and reproduced in Table 1, neatly summarize the W3C recommendations.

2.4 Best practices

We analysed the websites of six parliaments and two independent foundations providing access to parliamentary information. They are listed in Table 2. Here we provide a list of best practices that we found and that are relatively easy to implement. Nonetheless, none of these points was present at the majority of the sites.

- Publish extensive metadata to each file, appropriately linked and in a common format.

1. **Complete** All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. **Primary** Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. **Timely** Data is made available as quickly as necessary to preserve the value of the data.
4. **Accessible** Data is available to the widest range of users for the widest range of purposes.
5. **Machine processable** Data is reasonably structured to allow automated processing.
6. **Non-discriminatory** Data is available to anyone, with no requirement of registration.
7. **Non-proprietary** Data is available in a format over which no entity has exclusive control.
8. **License-free** Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Compliance must be reviewable.

Table 1: Open Government data principles, C. Malamud et. al. 8 December 2007, `http://resource.org/8_principles.html`.

Austria	`http://www.parlinkom.gv.at`
Belgium	`http://www.dekamer.be`
European Union	`http://www.europarl.europa.eu/activities/plenary/home.do`
Flanders	`http://www.vlaamsparlement.be`
Germany	`http://www.bundestag.de`
The Netherlands	`http://parlando.sdu.nl/cgi/login/anonymous`
Spain	`http://www.congreso.es`
MySociety.org (UK)	`http://theyworkforyou.com`
PoliticalMashup (NL)	`http://polidocs.nl`

Table 2: Parliamentary information websites that were surveyed.

- Make the status of the Intellectual Property Rights of the data clear and easy to find.
- Publish fast, thus also publish non-definitive versions (UK).
- Link named entities occurring in the proceedings. In Austria, members of parliament and government who are speaking are linked to their biographical page. Also numbers referring to laws or dossiers are linked to their pages.
- Put timestamps in the proceedings in order to allow alignment with audiovisual material. Austria does this at the start of every new speaker.
- Publish in well formed XML (XHTML) in order to allow for machine processing.
- Publish in coherent wholes. E.g. publish the proceedings of one day together in one file.
- Publish multiple views which link into the original and to other documents. For instance, publish on the biographical page of each member of parliament a list of all her speeches. Each speech should be linked to its place in the debate to provide context (Austria, EU).

All of these best practices are instances of the principles from Table 1.

2.5 In summary

In summary we can distill four main points:

- add metadata in a uniform standard format;
- publish in XML with links to the original sources;
- make each object linkable by unique permanent identifiers;
- make entities explicit in the representation and link them appropriately.

The next section formalizes these points as constraints on the XML representation.

3. REPRESENTATION IN XML

This section contains the XML schema in which we represented parliamentary proceedings. [**?**] contains the following high level schema for documents describing the proceedings of one day:

```
meeting          ⟶  (topic)+
topic            ⟶  (speech | stage-direction)+
speech           ⟶  (p | stage-direction)+
p                ⟶  (#PCDATA | stage-direction)*
stage-direction  ⟶  (#PCDATA).
```

All elements contain metadata stored in attributes. The speech elements contain attributes specifying the name, the

party and the function in parliament of the speaker. This purely semantical representation satisfies the requirements of Section 2.1.

Because of its semantical nature, special purpose extraction scripts need to be created for each parliament. Moreover, for each change in layout or organization of the proceedings, adaptations of the extraction scripts are called for. Thus this approach does not scale well.

An alternative layout oriented representation does not have this scalability problem, but still allows further processing which extracts the semantical information. We now describe that in detail.

Every document is a UTF-8 encoded XML file which is valid with respect to the Relax NG schema, available from the authors. 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[2];

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

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 [**?**]. 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.

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

3.0.0.2 *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 collection, see Table **??**. 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.

4. IMPLEMENTATION AND EVALUATION

We collected proceedings data from Belgium, Flanders, Germany, The Netherlands and Spain. Tables 3 and 4 gives an indication of the amount of data for the proceedings in Dutch.

We now describe the data collection and processing approach and evaluate the quality of the data.

4.1 Description of the data collection and processing

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

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 3: Availability of parliamentary data in the Dutch language.

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

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.

4.2 Data quality (Evaluation)

We evaluate completeness and correctness of the Dutch part of the corpus. We have also performed these evaluations for the Spanish and German proceedings with similar results. Completeness means that every parliamentary document that is published on the official web-pages of the respective parliaments is contained in our corpus 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?.

4.2.1 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[3]. 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.

4.2.2 Correctness

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

We now evaluate the transformation and the storage steps described in Section 4.1. Some of these procedures use heuristics and some do not. We start with an evaluation of the latter.

Table 5 describes the quality of the transformation process in terms of well-formed and valid XML output. Validity is measured with respect to the Relax NG schema from Table **??**.

Some of the data in the corpus are extracted from the text using heuristic methods. We list these here and evaluate the performance of the used methods. Table 6 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 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*.

Subcorpus	# Documents	Well formed XML		Valid XML	
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 5: Percentage of the total number of documents that are well formed and valid. Validity is measured with respect to the Relax NG schema described in Section 3.

	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 6: 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.

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 [**?**], trained on parts of the publicly available EuroParl corpus [**?**], which has in-domain data in the languages we are interested in. [4]

Table 7 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.

	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 7: 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. CONCLUSIONS

We addressed the problem of publishing parliamentary proceedings in a digital sustainable manner. We gave an extensive requirements analysis, and based on that proposed a uniform XML format. An extensive evaluation shows that the approach scales very well and produces high quality data.

Although the paper only discussed parliamentary proceedings we believe that both the findings and the used methodology are applicable to other governmental datasets. We have successfully applied our techniques to political speeches and written questions and answers. The thus obtained datasets were used for several applications ranging from political search systems [**?**] and electoral advice systems [**?**] to debate summarization systems [**?**, **?**].

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

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 Robert–Jan de Groot, Tim Gielissen, Steven Grijzenhout, Carlos Martin, Marina Lacroix, Arjan Nusselder, Hendrike Peetz, Anne Schuth, and Cor Veninga, for their help in the data collection and transformation processes.