ParlaMint Widened: a European Dataset of Freedom of Information Act Documents (Position Paper)

Maarten Marx, Maik Larooij, Gerda Viira
Information Retrieval Lab, Informatics Institute, University of Amsterdam.
{maartenmarx|larooij}@uva.nl, gerda.viira@student.uva.nl

Abstract

This position paper makes an argument for creating a corpus similar to that of ParlaMint, not consisting of parliamentary proceedings, but of documents released under Freedom of Information Acts. Over 100 countries have such an act, and almost all European countries. Bringing these now dispersed document collections together in a uniform format into one portal will result in a valuable language resource. Besides that, our Dutch experience shows that such new larger exposure of these documents leads to efforts to improve their quality at the sources.

Keywords: Freedom of Information Act, ParlaMint, Government Data

1. ParlaMint

The ParlaMint corpus of Parliamentary proceedings in 27 languages from 26 European parliaments covering at least 10 years of data for each parliament enables diachronic comparative research done by corpus linguists but also by social and political scientists (Erjavec et al., 2023). With the recently released translations into English (Kuzman et al., 2023), it is easy to conduct large scale comparative research on European and global topics like immigration, climate change, the pandemic, the War in Ukraine, or European integration.

The ParlaMint corpus shows that such a huge corpus in a tightly controlled format can be created with a decentralized approach with independent groups taking care of "their own data", and together creating an archive which derives its strength from the fact that it is an integrated data warehouse covering so many nations and languages.

The parliamentary proceedings are just one example of a resource which has the needed properties for such a huge socio-linguistic data collection and harmonization project. The key properties are:

- resources are built on top of a shared data model (for the parliamentary proceedings this is the Hansard model);
- the resources mean more or less the same in each country (what they represent is very similar: speech acts in parliament);
- there is enough overlap in context among the different resources.

There are other types of resources with these properties for which it is useful and desirable to collect and harmonize them. For instance, notes of cabinet meetings, Supreme Court rulings, and Addresses to the Nation (e.g., State of the Unions).

2. Freedom of Information Act

We are advocating in this position paper to bring together resources which are made public after a request based on the local Freedom of Information Act (FOIA). According to Wikipedia 102 nations have FOIA legislation by which citizens can request the public release of government documents on a certain topic. By 2018, every European country, except Luxembourg, has implemented some form of Freedom of Information law (Mokrosinska, 2021).

Also these FOIA resources share the desired properties needed to bring them together into a ParlaMint-like corpus.

We have created a data model and a corpus for Dutch Freedom of Information Requests, and extensively tested it with examples from very different sources: ministries, provinces, municipalities, the police, some universities, and regulating bodies like the gambling, the financial and the consumer authorities. This yielded a daily updated corpus of over 10K requests coming from 50 different governing bodies, consisting of 87K documents and 1.6M pages, all in a uniform format, accessible via a search engine called Woogle (the Dutch FOIA is abbreviated as Woo), and via datadumps in csv format (Marx, 2023).

We tested whether our data model could also fit FOIA documents from another country, and proved that it did with a corpus of 720K linked FOIA documents originating from 57 different Estonian governing bodies: the Estonian Woogle

3. Building the Corpus

Creating the corpus came with new challenges that we did not encounter when creating the Dutch ParlaMint corpus. As documents released under FOIA may contain sensitive information they often contain text redaction (pieces of the text made unreadable). This redaction process is often done by scanning

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 171–172

Type of Institution	Count
Government Agency	22
Local Government	15
Constitutional Institution	10
Other State Agencies	8
Educational Institution	1
State Held Companies	1
Total	57

Table 1: Institutions in the Estonian FOIA corpus.

the documents, thereby effectively removing (destroying is a better term) all the textual and structural content of the documents. Afterwards, often no OCR is applied, and if it is, it is usually of poor quality. Thus we had to OCR all documents ourselves (van Heusden et al., 2023). Besides this, the Dutch government has the habit of concatenating all released documents into one huge PDF document, *without clearly indicating the borders between the original documents*. To recover the original separate documents we had to use *Page Stream Segmentation* techniques (Wiedemann and Heyer, 2021). Thus much more low level document analysis was needed than we expected beforehand. Besides this, as the provided metadata was hardly existing, we needed to do document classification and information extraction (Bakker et al., 2024).

4. FAIR Data

As indicated above, "raw" FOIA documents are far from being FAIR research data, as defined in (Wilkinson et al., 2016). In fact the Dutch FOIA law stipulates that all documents released under this law have to be machine readable, contain all relevant metadata, and have to comply to European accessibility and re-use guidelines, covering exactly the four FAIR principles: data should be findable, accessible, interoperable and reusable.

Being rather frustrated that we had to use documents of such poor quality, we widely published about this in Dutch media directed to civil servants, and information professionals. The fact that these documents were being collected for scientific purposes and brought together in a convenient search platform like our Woogle, and thus could also be compared to documents from other publishers had a positive effect on the awareness by stakeholders of this problem. We already see the first changes and improved quality of data released under the Dutch FOIA.

By reusing data and exposing it, data will become more FAIR. We have seen this in the early 2000's with TheyWorkForYou.com and PoliticalMashup, two precursors of ParliaMint, and now we see the same happening with Woogle. The same process is known from self-organizing systems like Wikipedia: infoboxes have become so much more standardized after the advent of large knowledge graphs like DBPedia and Yago based on them.

5. Call for Action

Our goal with this position paper is to start an incentive to collect FOIA documents on a European scale, using a similar setup as ParlaMint. We believe that with Woogle we have already a strong foundation, in terms of a proven well fitting data model, a proven data processing methodology with reliable software, and a stable initial infrastructure for collecting and storing the data.

If you have FOIA data that you want to add to our collection, make contact with us, and we are happy to help.

6. Bibliographical References

Femke Bakker, Ruben van Heusden, and Maarten Marx. 2024. Timeline extraction from decision letters using ChatGPT. In *Proc. CASE 2024 Colocated with EACL*, St. Julians, Malta.

Tomaz Erjavec, Maciej Ogrodniczuk, Petya Osenova ..., and Darja Fiser. 2023. The ParlaMint corpora of parliamentary proceedings. *Lang. Resour. Evaluation*, 57(1):415–448.

Maarten Marx. 2023. Woogle dump. Technical report, DANS. doi.org/10.17026/dans-zau-e3rk.

Dorota Mokrosinska. 2021. *Transparency and secrecy in European democracies: contested trade-offs*. Routledge.

Ruben van Heusden, Hazel Ling, Lars Nelissen, and Maarten Marx. 2023. Making PDFs accessible for visually impaired users (and findable for everybody else). In *Proc. TPDL 2023*.

Gregor Wiedemann and Gerhard Heyer. 2021. Multi-modal page stream segmentation with convolutional neural networks. *Lang. Resour. and Evaluation*, 55(1):127–150.

Mark D Wilkinson et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific data*, 3(1):1–9.

7. Language Resource References

Kuzman, Taja and Ljubešić, Nikola and Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej et al. 2023. *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 3.0*. Slovenian language resource repository CLARIN.SI.