Enticing Local Governments to Produce FAIR Freedom of Information Act Dossiers

Maarten Marx(✉), Maik Larooij, Filipp Perasedillo, and Jaap Kamps

University of Amsterdam, Amsterdam, The Netherlands
{maartenmarx,larooij,kamps}@uva.nl

Abstract. Government transparency is central in a democratic society, and increasingly governments at all levels are required to publish records and data either proactively, or upon so-called Freedom of Information (FIA) requests. However, public bodies who are required by law to publish many of their documents turn out to have great difficulty to do so. And what they publish often is in a format that still breaches the requirements of the law, stipulating principles comparable to the FAIR data principles. Hence, this demo is addressing a timely problem: the FAIR publication of FIA dossiers, which is obligatory in The Netherlands since May 1st 2022.

Keywords: IR data collection · FAIR data · Governments records · Transparency

1 Introduction

Freedom of Information Act (FIA), sometimes called *Access to Information* or *sunhine* laws are effective in over 100 countries. They give citizens the right to request previously unreleased documents on policy issues from governmental bodies. In the Netherlands, more than 1000 such bodies exist and they are obliged to also publish the requested dossiers to the general public, for instance through their own website. For a good functioning of democracy it is important that these documents are easy to retrieve and to discover relevant information in them [11]. Together with democracy watchdog *OpenState* and the platform for investigative journalism *Follow The Money*, we decided to create a vertical search engine for these FIA dossiers, bringing them together in one portal.

Finding and harvesting the published documents was not difficult and possible with standard crawling technology. Somewhat to our surprise, the main

This research was supported in part by the Netherlands Organization for Scientific Research (NWO) through the ACCESS project grant CISC.CC.016, and by the University of Amsterdam through Humane AI.

Demo material is available at https://ecir2023.wooverheid.nl.

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13982, pp. 269–274, 2023.
https://doi.org/10.1007/978-3-031-28241-6_25

problem was that the released data violated all 3 core assumptions behind Information Retrieval: that a collection consists of *documents* containing *words*, and having basic *metadata*. The predominant strategy to release FIA dossiers (containing the request, the argued decision, the list of relevant documents, and the released documents) is to print it out completely, scan the pile of pages and publish it as one (often huge) PDF file. Metadata is hidden inside the decision letter, individual documents are often published as one big PDF document without boundaries between the different documents and if the PDF contains characters, they should be obtained via OCR. This type of data is known in IR from the TREC legal and total recall tracks [3,4] and from IR for due diligence [6].

Of course, we can try to repair the data, using OCR, page stream segmentation [9] and knowledge extraction [7], and we have done so [5], but these processes are never error free and by nature frustrating, *as the process is most often reverse engineering what was originally in digital format and well structured available.* So we decided to repair this problem at the source and this demo describes our solution. Hence, our research question is

> *How can we entice, especially small local, governments to publish their FIA dossiers according to the FAIR data principles* [10].

We targeted municipal governments because they form the vast majority of FIA publishers, have little IT infrastructure and lack resources. So our solution had to be robust, simple and cheap with large, directly visible, concrete gains. We developed this FIA publication software in cooperation with the association of Dutch municipalities VNG.

At the demo we present our solution to the problem. Here, we first describe the model that best fits these FIA dossiers, then the requirements for us and for the users of the software, followed by the chosen implementation. We evaluate it by checking the desiderata and requirements. We use the extra page in the Proceedings to describe the state of our system in January 2023.

2 The Data: FIA Dossiers

A FIA dossier can be seen as an argued response to an uttered information need. It contains a request for information, and the response consisting of 1) a list of all relevant documents, 2) those documents from the list that the government wants to (partly) release, and 3) a decision letter, typically drafted by a lawyer, explaining and motivating the response. The resemblance to a TREC topic and its corresponding set of relevant documents is remarkable. The decision letter is drafted as text, but can better be seen as a set of attribute-value pairs, containing a number of required and optional attributes. The values can be of free type, like the text describing the information need, or constrained, like the dates of the request and decision or the applied articles of law to withheld information.

Such a dossier is best modeled as semi-structured data combining metadata with raw text, with many optional attributes and unbounded cardinality constraints. In addition, the released documents are currently all in PDF format,

but obviously this is most often not their original technical format. Requested documents are often about *communication* (mail, social media like WhatsApp messages) and *structured data* (maps, spreadsheets), which are far more valuable in their original technical format than as a PDF produced to print.

To summarize, the data model of a FIA dossier consists of metadata on the level of the dossier, metadata for each released document, and the documents themselves. We chose to store a dossier as a zip file containing the released files, together with all metadata, including the text extracted from the documents, in XML.

The small municipality Waalwijk used our system to publish their data, resulting in this publication page with complete dossiers. Our datamodel is visible in the XML metadata file added to the zip file created for downloading a dossier.

3 Requirements

Our final aim thus is being able to harvest FIA dossiers of as high as possible technical (FAIR data) quality, in order to maintain a well functioning FIA search engine. This translated into 3 requirements: an API to harvest all data in validated XML format, OCR and full text extraction at the source, and checks and services to increase the availability and quality of the metadata.

To entice municipalities to use our system, data entry had to be almost effortless and the system had to have direct, visible and easy implementable advantages *for them*. This translated to an easy data entry interface with many prefilled values; automatic attractive publishing on a website of the municipality; automatic pushing to the obliged central government API; an internal search system for the FIA lawyers, and a data dashboard for managers/annual reports.

The FAIR data principles of course also had to apply to our system, so we required only open data standards and free open source software, both preferably top quality and with a large user base and community.

4 Implementation

Given these requirements, and the nature of the data, XML or JSON seemed the best option for data representation. We choose XML because we wanted extensive validation with good error messages, which is available through the XML constraint language Relax NG and the Jing validation software [8].

MySQL was well suited as our database backend because it is open source, has good full-text search mainly based on TF-IDF and BM25, solid security, and a wide active user base. We stored all metadata, the full text of all documents, and the original PDF files in the database. The relational schema is a straightforward implementation of the XML model of a dossier, witnessed by obvious translations back and forth. We used Python as the scripting language tying the different components together. The different components were implemented as follows:

- the translations from XML to the relational DB schema and back were done in Python using the etree module;
- publishing a list of dossiers as a webpage was done through an SQL query generating an answer which was processed by Python into an HTML page;
- publishing to the government API using Python, Flask, and SQL;
- search engine using MySQL full-text `MATCH AGAINST` queries on several text fields, with ranking by internal BM25 and TF-IDF relevancy ranking;
- data dashboard using MySQL, Pandas, Seaborn and Plotly.

The application runs on a server. Municipalities can choose to use the server or run it on their own platform by using a Docker image.

5 FAIRiscore and Full Text score

To encourage municipalities to create data of high FAIRness quality, we created a 5 point A–E *FAIRiscore* scale, reminiscent of the Nutriscore [2]. Using imported and overruled RelaxNG schemas it is easy to define and maintain 5 schemas of monotonically increasing tightness, one for each fairiscore value. After data entry, the data is validated immediately and the user receives not only the score but also suggestions for improving the score. As the score is monotonic it can be explained and visualized easily by coloring the XML tree using the 5 Nutriscore colors, as in this figure on the web. It must be read as follows: in order to reach Nutriscore say B, one needs to have score C, and fill in all fields coloured light green (the color of score B).

Similarly we provide a 5-point A-E score indicating the quality of the full text. Many documents are released as PDFs without underlying text, but the civil servants doing this are often not aware of this. We estimate the amount of real text per page using OCR and redacted text recognition, compare that to the text inside the PDF and give a score based on the overlap.

6 Evaluation

As the system is very new, we have no user studies yet, but expect to have them at the time of the conference. The system works well as a filter for data entry into our large Freedom of Information Act search engine https://woogle.wooverheid.nl. Even if the dossiers have the (lowest) FAIRiscore E, we have the full text (either the original or through OCR), and a tiny bit of metadata. This was our main goal, and thus reached. How well our data quality encouragment methods worked need to be seen, and will be reported on at the time of the conference.

7 Conclusion

Our system is simple but it contains for municipalties very desirable functionality. If the uploaded data is already slightly better than the worst FAIRiscore

E, a FIA search engine can both rank better and present much richer search snippets, making it easier for users to judge relevance already before they have to open a document.

At the time of writing we cannot yet answer the implicit question in the title of this demo. We expect to be able to do that at the time of the conference.

8 Update January 2023

We use the extra page in the proceedings to describe the state of our system in January 2023. On December 28, 2023, the Dutch central government announced that they stopped developing a platform having similar functionality as the system described here, after very negative advice from an independent ICT in government review committee [1].

After publishing our initiative we received over 30 reactions from interested municipalities, from which only 2 went further and 1 succeeded, Waalwijk. We decided to collect the large bulk of available data by dedicated crawling and around Christmas 2022 our search engine was complete for the published FIA dossiers of all Dutch ministeries, all provinces (10 of the 12 publish), and the top 80 largest municipalities (only 6 of them publish), see the overview on https://woogle.wooverheid.nl/overview.

The quality of the data turned out to be as bad as described in this paper. To summarize:

Publish seperate documents None of the ministeries, only 2 of the 10 provinces, and only 1 of the 6 municipalities.

Documents containing text From all 41K documents in our system, 46% contained not a single machine readable character (30% of all 1M pages). The amount of non machine readable pages ranges from 2% with the municipality Waalwijk and one ministery (publishing 40K pages) to 85%. It seems reasonable to assume that with 2% all documents are born digital, sometimes also called native PDFs.

Metadata Not a single publisher provided even these 4 basic properties as metadata: the dates of the request and the decision, the text of the request (the information need), and the decision.

Positive News. The single publisher Waalwijk using our system showed that it is possible to deliver what we asked for: release native PDF documents separately in a zip file instead of one concatenated PDF; each page machine readable and all basic metadata nicely in order; see https://doi.wooverheid.nl/?doi=nl.gm0867.

Negative News. Obviously with only one participant, we did not entice the municipalities. This seems partly due tp the unclear situation regarding the platform of the central government, but there is also another reason. From interviews with several publishers we learned that we simply asked too much, especially with

regard to metadata. So, we changed strategy, focusing on just getting the raw data, preferably machine readable, using an API that can directly connect to the IT infrastructure of the data publisher.

Next Steps. We hope to fill our search engine using this API. We will highlight data publishers which produce FAIR data and hope that seeing good examples in action (besides their own not so good example) entices municipalities to change their publishing habits. On https://woogle.wooverheid.nl the reader can track our progress.

References

1. Adviescollege ICT-toetsing: BIT-advies Plooi. https://www.adviescollegeicttoetsing.nl/onderzoeken/documenten/publicaties/2022/11/28/bit-advies-plooi Accessed Dec 28 2022
2. Chantal, J., Hercberg, S., Organization, W.H., et al.: Development of a new front-of-pack nutrition label in france: the five-colour nutri-score. Public Health Panorama **3**(04), 712–725 (2017)
3. Grossman, M.R., Cormack, G.V.: Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. Rich. JL Tech. **17**, 1 (2010)
4. Grossman, M.R., Cormack, G.V., Roegiest, A.: Trec 2016 total recall track overview. In: TREC (2016)
5. van Heusden, R., Kamps, J., Marx, M.: Wooir: A new open page stream segmentation dataset. In: Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 24–33 (2022)
6. Roegiest, A., Hudek, A.K., McNulty, A.: A dataset and an examination of identifying passages for due diligence. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 465–474 (2018)
7. Shi, P., Lin, J.: Simple bert models for relation extraction and semantic role labeling. arXiv preprint arXiv:1904.05255 (2019)
8. Van der Vlist, E.: Relax ng: A simpler schema language for xml. " O'Reilly Media, Inc." (2003)
9. Wiedemann, G., Heyer, G.: Multi-modal page stream segmentation with convolutional neural networks. Lang. Resour. Eval. **55**(1), 127–150 (2021)
10. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016)
11. Worthy, B.: More open but not more trusted? the effect of the Freedom of Information Act 2000 on the United Kingdom Central Government. Governance **23**(4), 561–582 (2010). https://doi.org/10.1111/j.1468-0491.2010.01498.x