UNIVERSITEIT VAN AMSTERDAM

Research proposal

Maik Larooij – 12761915

Thesis supervisor: Dr. M.J. (Maarten) Marx

18 EC

Introduction. According to article 3:1 of the Dutch 'Wet Openbaarheid van Bestuur' (Wob), Dutch inhabitants have the right to file a request, a so called 'Wob-request', for information about administrative matters. A request can be sent to any governing body, such as ministries of the government, provinces and municipalities. The body then typically sends a reaction, containing a decision and providing documents of information where possible. While the Wob will be replaced by the 'Wet open overheid' (Woo), the right to receive information on request persists in the new law in article 4:1. Decisions on these requests are publicly published, accessible for everyone. There seems to be, however, no clear rule or standard on how to provide the requested information. With the introduction of the new law, the Dutch Ministry of the Interior and Kingdom Relations published a plan of action to create a more open government. This plan acknowledges the need for a standardized way of providing public information (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2020).

As a result of this plan, in September 2021, the Open State Foundation in cooperation with the province of North-Holland and the association of Dutch municipalities (VNG) created a concept version of a guideline for a national standard for providing and archiving the information requests (Open State, VNG, Provincie Noord-Holland, 2021). The guideline gives a first recommendation on what metadata to attach to the requests. However, the guideline does not speak of a standard file format in which to save the metadata and it fails to provide more technical metadata for further processing purposes. Furthermore, the guideline speaks of metadata on request-level, while it can be very interesting to also store metadata on document-level, enabling users of the data to search for specific documents instead of only being able to filter the requests on their main themes.

Research goal. The main goal of this research is to further enhance the guideline by detecting both useful semantic- and technical metadata on the request- and document-level of an information request. The aim is that this metadata standard needs to be computer-readable as well as human-readable. To make this possible, existing information request publications from different kinds of sources will be inspected to create a clear view of the current state of providing information request documents. Next step is to identify a standard file format considering both the intentions of creators (governing bodies) and end users of the data. The last goal is the creation of software that can be helpful in the process of working with the proposed standard. This software should automatically create and validate information requests from user input, according to the created metadata standard. It is targeted to be used by employees of the governmental bodies to both increase the reuse of data and the level of communication with the citizens.

Relevance. With the Dutch government aiming on becoming more transparent, it is more important than ever to collaborate and communicate clearly. In the first place, this means communicating and working together as governmental bodies. A comparison of Dutch public organizations and their open data policies showed that there is a lack of systemic collaboration, which impedes mutual exchange and reuse of information (Zuiderwijk & Janssen, 2014). Brummelkamp *et al.* (2016) state that the impact of openness grows when the different parts of the government coordinate and work together. Fortunately, a metadata standard can help in developing interoperable tools and services (Duval, 2001). Attaching metadata to published documents relating to an information request will help in working together and creating an open government. A national metadata standard will make it easier to reuse documents, even if the information is provided by an other government body.

On the other hand, the communication between the governmental bodies and their citizens is also important. Research at Delft University of Technology concluded that the substantive response to an information request (Wob-verzoek) was very mixed (van Oostveen & van Loenen, 2014). Every governing body

seemed to have their own way of responding to information requests, which makes it hard to structure the information for use and reuse. Article 2:5 of the Wet open overheid, taking effect in May 2022, states that "the application of the law is based on the general public interest of public access to public information for a democratic society". This depicts a contradiction, with the government aiming for public access for a more democratic society, while the responses to the information requests seem too mixed to make sense of them.

Both the vertical and horizontal communication problems denote the need for research on how to improve this communication of documents. This paper will serve as a recommendation on the first steps of achieving this, by providing a metadata standard. Other datasets, like the registry of government bodies, already use frameworks like the Resource Description Framework (RDF) and Extensible Markup Language (XML). RDF proved to perform really well when working with related entities, like in the DBpedia project (Auer *et al.*, 2007), where different Wikipedia pages (entities) can be represented with relationships. However, for the information requests there is no necessity to link the requests with each other. Also, it is important to keep the possible use cases of the data in mind, like facilitating a search engine. This research will build on the file formats already used to try to find other file formats that have the potential to be a fit for the purpose of storing metadata of information requests.

Question. This research will formulate an answer to the question: "How can an information request be digitally represented and standardized to empower the concept of an open government?". Five sub-questions have been formulated to aid in finding an answer to the stated question.

1. What 'open data' factors make a government 'open'?
2. What information do current publicly available information requests contain?
3. What does an optimal metadata standard look like, storing at least all the found information and being usable for production and search engine purposes?
4. To what extent do the current publicly available information requests comply with the metadata standard?
5. In what way can information requests be automatically turned into the valid and desirable metadata format?

Method. This research consists of five sub-questions, where every sub-question has its own method of finding an answer. To answer the first question on what factors make an open government, a literature study will be carried out. There is already quite some literature on the topic of open governments. This research will have a focus on using open data in an open government. Brummelkamp *et al.* (2016) provide a literature study on the impact of an open government, combining academic literature and interviews with Dutch researchers and experts. Meijer *et al.* (2012) tried defining an open government. Another interesting Dutch research is a thesis by Zuiderwijk-Van Eijk (2015) on an interface to coordinate open data usage. It discusses factors and conditions for open data. Dawes & Helbig (2010) discuss information strategies for an open government. The question will be answered by identifying what open data factors and practices positively influence an open government.

To answer the second question on what information can be extracted from current public information requests, two different categories of data can be identified; data available on the website and data available in the documents itself. Firstly, an inventory has to be made on what metadata is currently available for the information requests on the website. https://wobcovid19.rijksoverheid.nl/ contains data on information requests, handled by Dutch ministries, on all COVID-19 related cases. Next to wobcovid, the municipality of Utrecht and the province of Noord-Holland archives will also be used to create a clear view of the different governmental bodies. A web scraper will be used to extract as much

metadata from the webpages as possible on request level. Also, there is information stored inside the provided documents. Some information requests contain inventory lists, providing metadata on a document level. 491 documents labeled as containing an inventory list will be examined by extracting the table from the pdf file with a Python module called 'pdfplumber'. The program will try to extract all rows containing column names to learn about possible metadata fields. Lastly, some metadata are incorporated in the documents. With Optical Character Recognition (OCR), text from PDF files can be made readable. This is done by another student, this research tries to build on his research and measure the information gain after OCR.

The outcomes of the second question will be used to create a metadata standard to uniformly store the information requests. First, literature will be consulted to find a file format that fits the purposes of the end user, like a search engine. Then, based on the characteristics of an information request and the found metadata, all the necessary fields can be identified and labeled as required or optional. Also, some more technical metadata fields can be identified.

The fourth question will focus on comparing the publicly available information requests to the proposed metadata standard. To answer the question, information request metadata will be scraped from their source. On request level, conclusions can be made about how many information requests already comply with the proposed required attributes and listing attributes that are not available at the moment. On document level, the same can be researched, using inventory lists and filenames of the documents to extract information. Also, an inventory can be made up of what type of documents the requests contain, like decision document, request documents and inventory lists. Lastly, a comparison between governmental bodies can be made, comparing the municipality, the province and ministries.

The fifth question will build on all the previous questions. It is an experimental question, where tools to ease working with the standard are suggested. These are tools like an automatic metadata creator, in the proposed and valid file format, where the user can fill in all the required fields, just as they do now for an inventory list. JSON Schema will be used to create a validation schema for the proposed standard (Pezoa *et al.*, 2016). Other tools can include a search engine, which may need to be researched on its own.

Data. The wobcovid dataset used contains 367 documents, consisting of decisions, released documents and inventory lists. The documents combined consist of 30.881 pages. This dataset will be used to measure the information gain after OCR.

To extract column headers from the inventory list files, 491 inventory lists have been retrieved. These inventory lists are extracted from the wobcovid dataset and extracted from other information requests publicly available on the government website.

2751 information requests, derived from the government website (rijksoverheid.nl/documenten) are going to be used to extract metadata from the filename and the information shared by the government. An additional 775 information requests are extracted from the municipality of Utrecht and all information requests from 2016 until 2022 from the province of Noord-Holland.

Literature.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In The semantic web (pp. 722-735). Springer, Berlin, Heidelberg.

Brummelkamp, G., van den Berg, J., & Meijer, A. (2016). Impact van Open Overheid.

Dawes, S., & Helbig, N. (2010). Information Strategies in Government. Challenges and prospects for deriving public value from government transparency initiatives. International Conference on Electronic Government.

Duval, E. (2001). Metadata Standards: What, Who & Why. *J. Univers. Comput. Sci.*, *7*(7), 591-601.

Meijer, A. J., Curtin, D., & Hillebrandt, M. (2012). Open government: connecting vision and voice. *International review of administrative sciences*, *78*(1), 10-29.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2020). *Actieplan Open Overheid 2020–2022*.

Oostveen, van, J., & Loenen, van, B. (2014). De praktijk van de Wet openbaarheid van bestuur: een gebruikersperspectief. *B&G*, *2014*(mei/juni).

Open State, VNG, & Provincie Noord-Holland. (2021, september). *Handreiking Open Wob*.

Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M., & Vrgoč, D. (2016, April). Foundations of JSON schema. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 263-273).

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government information quarterly*, *31*(1), 17-29.

Zuiderwijk-Van Eijk, A. (2015). Open Data Infrastructures, The design of an infrastructure to enhance the coordination of open data use (proefschrift). Delft: TU Delft.

Planning.

Date(s)	Plan
Already done	Wob-scraping, Wob-schema version 1.0, Wob-validator version 1.0, meet-up with students working on this project.
April, 1, 2022	Hand in Research proposal
April, 8, 2022	Complete measuring information gain from OCR
April, 15, 2022	Create a complete version of context and relevance for final version thesis
April, 22, 2022	Create a Wob-creator and validator (tool, research question 4)
April, 22, 2022	Create a complete version of method for final version thesis
May, 6, 2022	First research question writing (literature study on open government)
May, 20, 2022	Second research question writing (information extraction, technical part mostly already done)
June, 3, 2022	Third research question writing (metadata standard)
June, 10, 2022	Fourth research question writing
June, 17, 2022	Conclusion, discussion
June, 24, 2022	Final deadline hand-in thesis

Supervisor. Via GitHub I regularly contact my thesis supervisor Dr. Maarten Marx. Also, we already had a meetup with other students working on the same project. We plan on meeting on a regular basis (every 2-3 weeks) with at least the Information Sciences students working on this project.