Metadata publication and extraction from Dutch Convenants

submitted in partial fulfillment for the degree of master of science

Sander Oud

12288667

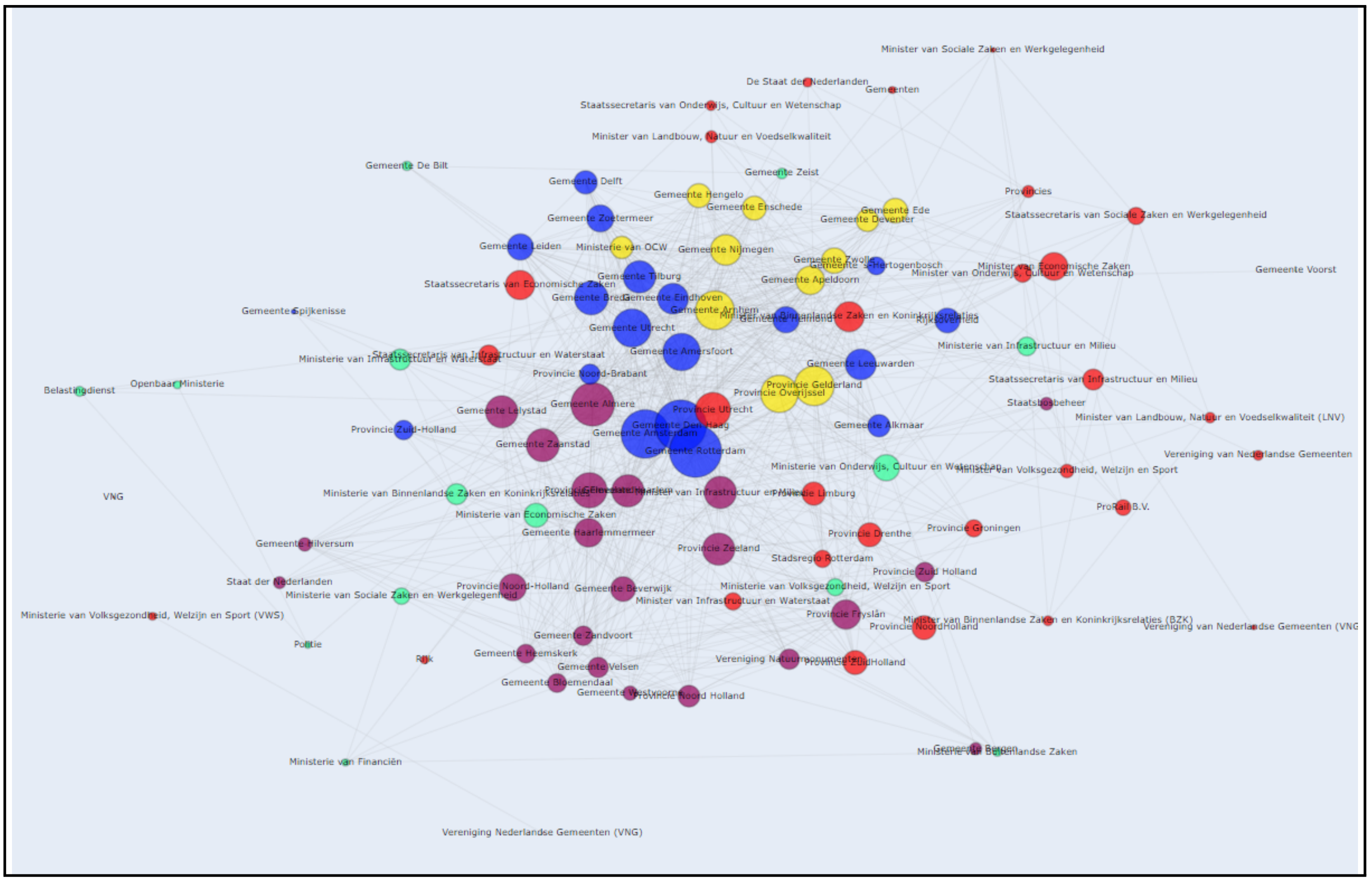
Master Information Studies

data science

Faculty of Science

University of Amsterdam

Submitted on 29.06.2024



	UvA Supervisor
Title, Name	Maarten Marx
Affiliation	UvA Supervisor
Email	m.j.marx@uva.nl

ABSTRACT

This research looks into how many convenants are published by Dutch governmental organizations in accordance with the Dutch openness act (WOO), and with how much metadata they come with. Next to that, the ability to gather the metadata afterwards using GPT-3.5 is tested. Convenants of 302 different organisations have been scraped, resulting in a dataset of 3011 documents. The publication of metadata with the convenants can be improved in many ways. Generally, larger governmental organizations perform better on this front than smaller ones. GPT-3.5's ability to classify convenants, extract dates and parties, model topics and descriptions is tested. The largest flaw in performing this task was concluded to be the inability of the model to take in more than 4096 characters. While the GPT-3.5 showed potential, the easiest way to improve findability for convenants is publishing them with metadata.

KEYWORDS

Government, WOO, GPT-3.5, Metadata extraction

GITHUB REPOSITORY

https://github.com/sanderoud/Thesis-project-covenants

1 INTRODUCTION

In 2022, the Dutch Government Openness Act (WOO) replaced the Open Government Law (WOB). Both laws require that all government agencies make information about their activities publicly available [20]. This transparency allows Dutch citizens to supervise the government, aiming to increase trust in democracy. Both the WOO and the WOB state that citizens are allowed to request information from local and national governmental organizations. One aspect introduced in the WOO that was not in the WOB, is the law's active disclosure requirement. Seventeen different types of government documents must be published on a web-accessible platform, so citizens are able to find the documents easily. Where possible, documents have to be published in electronic form, in a machine-readable open format, together with the metadata (article 2.4(3a)).

Convenants are one of the seventeen document types that hold the active disclosure requirements. A convenant is a written agreement between the government and one or more parties. The purpose of a convenant is to realize certain policies of the central government [22].

Considerable research has been conducted on the FAIRness of articles published because of the WOO. The FAIR principles introduced by Wilkinson et al. [13] aim to enhance the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. They emphasize that data should be findable, accessible, interoperable, and reusable. In the context of the WOO, following these principles in publishing the government documents means the data needed for further research is easily accessible. This can save a lot of time for social researchers in the data collection phase of their research.

However, despite the potential benefits of following the FAIR principles, earlier research into the extent to which WOO articles are published in accordance with these principles [16] concluded that most published documents score low on FAIRness. Specifically in relation to convenants, the research states that the FAIRness score can be improved upon by utilizing its metadata structure. The document type inherently holds a semi-structured form that can be leveraged to enhance findability and interoperability. The agreements published in the convenants include metadata about the involved parties, subjects, and the date and place of signing. When this metadata is properly provided, most of the information in a convenant can be deduced at a glance, together with significantly improving the findability and interoperability of the data.

In a report published by Marx and Kamps, the digital sustainability of all published documents was tested for ten provinces of the Netherlands [14]. While the report states that the active disclosure of documents of the WOO has improved, there is still a lot to be gained on the digital sustainability of the publications. The report reviewed four different aspects of digital sustainability in the WOO files.

The first aspect looked into is the existence of metadata with the given publication. The minimum metadata of a WOO file include a title and a brief description and the dates of request and decision. None of the 10 provinces provide all of this metadata.

In addition to the metadata, the machine readability was tested. Machine readability entails the structuring of the metadata and whether the released files are processable by computers. A WOO file consists of four elements: the request, the decision, the inventory, and the released documents. Frequently all elements are put into the same file, without real borders between the elements. This means that computers are not able to distinguish the different elements in a WOO file. This way of publishing is a lot more time-consuming because the entire document needs to be processed instead of just the relevant part of the document.

The files need to be scrapable, this means that they need to be accessible without human interaction. There are a number of aspects that hinder this accessibility: the list of WOO files is not uniform and the formatting changes annually; automatic downloads are deliberately hindered; files are located on Google Drive or other not easily accessible external providers; and finally, meaningless file names.

Lastly, there is the referability of the articles. None of the provinces use a persistent form of identifying their articles. The setup of an easy doi would facilitate a lot more clarity and digital sustainability.

To address these issues effectively, Marx et al. set up a website to centralize all publications [15]. At the moment, Woogle consists of 3,343,369 documents scraped from the internet. The website allows for the analysis of large-scale computer-assisted diachronic comparative research [15]. Collection of data at large scale often takes 80% of the research time and requires technical skills that social researchers often do not possess. The website provides the collection of data for further research. Convenats are mostly not yet implemented on Woogle and are still scattered across the internet.

This research will add to the Woogle project by trying to scrape all convenants from the internet, and publishing them in accordance with the WOO and the FAIR principles. The main obstacle in publishing the convenants in accordance with the WOO is the publication of metadata. As concluded in the research of Marx and Kamps, a lot of the required metadata is missing from the online publications on other websites [14]. This data can therefore not be collected by scraping it from the website with the documents. The only method of acquiring this missing metadata is by structurally getting it from the content of the documents.

For the task of extracting metadata, the GPT-3.5 model has been chosen for its versatility in performing language tasks. The ability of GPT-3.5 to structurally gather information from the governmental documents will be tested. For the gathering of subject, date, involved parties and description three different machine learning capabilities of the model will be tested. For the subjest of the convenant topic modeling will be utilized. This machine learning technique is used to identify patterns and themes within a collection of documents by grouping words into topics, allowing for an understanding of the underlying structure of the text [9]. For dates and involved parties Named Entity recognition will be tested. Named Entity Recognition is a fundamental task in natural language processing that involves identifying text spans associated with proper names and classifying them into predefined classes such as organization or dates [4]. For the description text generation will be tested. Text generation is a natural language processing task that involves generating coherent and contextually relevant text based on a given input, leveraging the model's ability to predict and produce human-like language [17].

In summary, the focus of this research will be on determining the extent all convenants can be scraped from the internet with relevant metadata. Gathering metadata from publications where possible, and from the content of the documents.

The research question central to this paper is as follows:

> *To what extent is it feasible to semi-automatically gather the convenants of all Dutch governmental agencies along with their accompanying metadata? And if the metadata is not available, how effectively can we extract it from the document?*

To answer the research question, the following subquestions are created:

(1) How many convenants are there and to what extent can they be structurally retrieved from the web to form a dataset?

(2) To what extent are convenants published with the necessary metadata?

(3) To what extent can missing metadata be extracted from the document text?

In the related work section of this research the previous counts of convenants published will be looked in to, together with the current research on metadata publication. The choice of using of the GPT-3.5 model over other machine learning models will be argued and other relevant research to the use of GPT-3.5 for data extraction tasks will be looked into. In the method section of this research the method to answering all subquestions will be described per question. In the result section the results of the executed method will be presented. In the discussion these results will be discussed, reflecting back on the theoretical framework. Finally, the conclusion will give answer to the research questions.

2 RELATED WORK

2.1 Convenant scraping

The latest count of the amount of published convenants was done in 1995 and resulted in a total of 154 convenants [19]. The count was executed by the Dutch Court of Audit. The publication of the court lead to the conclusion that the amount of convenants had increased due changing governmental culture where the central government tries to be less binding and regulating.

A current count of the amount of convenants does not exist [2]. However, when looking at a handful of websites where convenants are published there can be concluded that the increase of convenants has persisted, making them a more mainstream approach to government action. The website officielebekendmakingen.nl is a platform where all documents of the Dutch Staatscourant and other governmental magazines are published. It currently contains 1192 convenants on its own, over a thousand more than the total count of 1995.

2.2 Availability of metadata

In 2023, Marx and Kamps have already done research to the degree to which provinces provide metadata with their documents [14]. The digital sustainability of all provinces was tested by checking if the provinces provided a title, description, date of request, decision date and date of publication. None of the provinces provide all five metadata points. Every province did at least provide a title, and a date. In most of the cases it is not clear on what date it is about. The metadata publication of the provinces is compared to the publication of the ministries. Next to providing a title, a description, a document date and date of publication, they also provide a subject subject of the documents, the research shows that the ministries are overall more structured in the publication of metadata than the provinces. This research expands on the research of Kamps and Marx by also taking municipalities and independent governing bodies into account. Since municipalities are in many cases smaller than the provinces, it is hypothesised that the smaller governmental organisations will be less structured than the ministries.

2.3 Metadata gathering

Automatic metadata generation provides scalability and usability for digital libraries and their collections [7]. Previously, metadata extraction has been done using machine learning methods. Named Entity Recognition (NER) is the task of detecting mentions of real-world entities from text and classifying them into predefined types such as companies, government agencies or dates. Traditionally, NER systems relied on hand-crafted features and domain-specific knowledge due to limited supervised training data [10]. However, recent advancements have introduced novel neural network architectures that automate feature detection, reducing the need for extensive feature engineering [5].

Spacy's NER model is one of the leading machine learning methods used in named entity recognition [4]. SpaCy can recognize various types of named entities in a document. For example, Gemeente

Amsterdam can be classified as a geopolitical entity and Shell as an organization. SpaCy works by asking a model for a prediction. Because models are statistical and strongly depend on the examples they were trained on, this does not always work perfectly and might need some tuning later, depending on the use case [1].

In domain-specific applications, a significant challenge is the scarcity of annotated data, which limits model performance [18]. Generic NER tools remain limited in recognizing entities specific to a domain, such as drug use and public health. To improve the domain specific knowledge, pre-labeled data in the domain is required.

The topics covered in convenants of Dutch governmental organizations span a large number of domains. The topics cover all different fields that are discussed in politics. Labeling text for the model to train on in all different domains would be highly labor intensive. Therefore, it would be better to find a method that does not require any form of retraining.

GPT is a large language model developed by OpenAI that is capable of producing response text that is nearly indistinguishable from natural human language [6]. The GPT models are first trained without supervision on unlabeled data. This way the model learns naturally, same way as a person would. Afterwards the model is trained to improve on specific tasks with the goal of more guided and structured refinement by the creators [12]. The large benefit of GPT over other large language models like BERT, RoBERTa and XLNe is that GPT has the ability to generate high-quality text responses [12]. The other models focus on understanding and analyzing the text.

The use of GPT to extract metadata from documents has two large benefits. First of all, with 175B parameters and 96 layers trained on a corpus of 499B tokens of web content, It is far the largest language model constructed to date [6]. This training means that for almost all domains the model already has knowledge, meaning additional training is not required. Next to the existing domain knowledge, the model also has the ability to generate text. For generating metadata this is useful for returning words that are lost within enumeration. For example, when the involved parties are "gemeenten Amsterdam and Amstelveen", a NER model will recognise gemeente Amsterdam as an organisation and Amstelveen as a location. A GPT model can see the relation between the word gemeente and Amstelveen and therefore see *gemeente Amstelveen* as an organisation as well.

To date, few studies have examined the potential of LLMs in reading and interpreting clinical notes, turning unstructured texts into structured, analyzable data [8].

Huang et al. have examined the potential of LLMs in reading and interpreting clinical notes, turning unstructured texts into structured, analyzable data [8]. The study concluded that ChatGPT-3.5 has the ability to extract pathological classifications with an overall accuracy of 89%, outperforming NER and keyword search algorithms. The added benefit is that it does not require extra annotated data. The research by Huang et al suggests that the use of large language models is the best method to gather metadata from the Dutch convenants. The largest difference between the lung cancer pathology notes and this research is the structuredness of the data. A key finding of this research will be how ChatGPT-3.5 handles poorly scanned text and different textual structures in the documents.

3 METHODOLOGY

3.1 Obtaining all convenants

The central government has guidelines for the location of publication of convenants. Most important in these guidelines, everything has to be linked to the WOO-index, a list of all governmental organisations [3].

A distinction of four different types of governmental organizations is made [2]. First of all, there are the organizations of the state. These include the ministries of the Netherlands and all their subdivisions. These organizations are required to publish their convenants in the Staatscourant, which in turn will be uploaded to Officielebekendmakingen.nl. The link to the Officielebekendmakingen must be published in the WOO-index. All municipalities and governments can choose their own location of publishing as long as it is listed in the WOO-index. Since, by definition, at least one government organization is part of the convenant, going through the entire WOO-index should result in finding all convenants. What is left are independent governing bodies, which have the choice to publish their documents in the Staatscourant or a location of choice, with the registration of this location in the WOO-index.

In practice the link to the location is often not provided in the WOO-index. For organisations of state the WOO-index always links to the page of the Staatscourant, yet there are also convenants on the website of Rijksoverheid.

Since the ministries are officially required to publish in the Staatscourant, the website of Officielebekendmakingen.nl was scraped first. On this website all publications of the Staatscourant and other official notification magazines are published. The website makes it possible to filter on convenants. All results from the website were scraped by building a scraper using python library Beautifulsoup [11]. First, all the results were loaded into a textual overview of the HTML of the page. Then all individual items were separated by finding all 'li' tags in the HTML. From this 'li' tag the link to the publication is scraped. The link is used to create a new textual overview of the publication page. From this page the file is scraped. This is the standard scraping procedure used in this research. Most of the publications are from the ministries. Next to that there are some provinces, municipalities and independent governing bodies that publish here.

The website Rijksoverheid.nl contains a list of convenants by ministries as well. On this website and advanced search option makes it possible to filter on convenants. The same standard scraping process as Officielebekendmakingen.nl was used. To ensure the completeness of the list of convenants available on the advanced search bar, a second search is conducted on the Rijksoverheid website. The main page of the Rijksoverheid website contains a search bar. This keyword search was used with the search term 'convenant'. Every item that contains convenant in its title is scraped using the standard scraping method mentioned above. If new convenants are found in this search that would suggest that the advanced search of the Rijksoverheid is incomplete in its publication of the articles. There is a separate website for the Belastingdienst, which is part of the Ministry of Finances. For this website another standard scraper was build to scrape all results of a keyword search.

When looking at the link provided by the WOO-index for the municipalities and provinces, most websites do not provide as many convenants as expected. Instead the municipalities and provinces publish almost all their documents on an external digital platform. This is either with 'Notubiz' or 'Bestuurlijke Informatie'.

The URL of Notubiz or Bestuurlijke Informatie always has the same format for each municipality. For Bestuurlijke informatie this is https://body.bestuurlijkeinformatie.nl, where body is the name of the governmental body. For Notubiz this is https://body.notubiz.nl. To find out what decentralized government uses what platform a list of all 343 municipalities and 12 provinces was composed. All items are tried in both URLs. If the status code of the url is 200 it is scraped.

For the provinces and municipalities that use Notubiz a scraper is build that is similar to the standard scraping process. The biggest difference being that clicking the results would not link to a separate page, but download the results immediately on selecting. The scraper was rebuild to be able to gather all results from the search page. A problem with the website of Notubiz for scraping is that no results exist within the HTML of the page without interacting with it. They are only loaded after a user is actually on the page. This problem is solved by first opening a webdriver using python library Selenium Webdriver [21]. This driver opens the URL of the Notubiz website. Then the driver scrolls to the bottom of the page automatically to load in all results. All information form results that contain a document are scraped. Only results that are classified as document are scraped. This way the results that are agenda items in conferences are ignored. Once again only results that have convenant in the title of the file are put in the dataframe. Results that are likely not convenants based on title and filename are cleaned out. This is done by only selecting documents that have the word convenant within the first four words of the title.

The website of Bestuurlijke informatie is similar to Notubiz in the sense that results are only loaded into the HTML upon interaction with the page. Therefore, a webdriver was once again used to load in all results. The difference being that this time a new page had to be clicked instead of scrolling down to the bottom. Next to that, there is no download link to the files. The website makes use of a download button within an iframe. To solve this, first the driver is switched to the second iframe. Then the file is downloaded. Instead of saving the download link, the directory path to the right file is saved. Any files that seem like they are not convenants based on title and filename are filtered out of the dataframe in the same method as Notubiz.

Next to the municipalities and provinces, other potential parties in a convenant can be the independent administrative bodies. Overheid.nl provided a list of all individual governing bodies. In the Netherlands there are 431 independent idministrative bodies. All websites were checked for convenants using keyword search or the sitemap of the page.

The scraped convenants are then all downloaded and set into the same directory. The file path is saved into the dataframe. This file path is used to load in the textual data from the PDF file using python library PyPDF2. This library can take in a pdf document and return all text that was in the pdf file. All files that cannot be read in by PyPDF2 are taken out of the dataset. These are files that are scanned in and published, and therefore not processable by a machine. Next to the non readable files, duplicate files are also filtered out. Anything that has reasonable suspicion of not being a convenant is also taken out of the dataset. This involves examining filenames and removing any files that lead to suspicion of not being a convenant. This is done through a regex list with abbreviations for entries like 'rv', which stands for council meeting.

3.2 Metadata scraping

From the page where a convenant was publised the publication date, description and the title are scraped in the same method as the download link of the page was scraped. This data is saved into a dataset together with the path to the relevant document. Figure 1 gives an example of how the metadata is published in the HTML of the Staatscourant.

Figure 1: Example of metadata publication of Officielebekendmakingen.nl

3.3 Metadata extraction

After the dataset is obtained the metadata can be gathered from inside the documents themselves using ChatGPT-3.5. The large language model will be used to gather the following information.

- Description - small description of the convenant
- Topic - the topic of the project
- Signing date - the date the document was signed
- Starting date - the date the decisions made in the document start
- Parties - the involved parties in the document.
- Convenant - Whether the document is a convenant or not

The Python script utilizes the OpenAI API to generate descriptions for each row of a dataframe. For each metadata item a different prompt is used. The data is stored into a json file. After execution of the prompt the json file is extracted into the dataset with a new column for all gathered data.

You are a model that performs six tasks:

(1) *Provide a small two to three sentence description in Dutch about the full convenant.*
(2) *Classify the topic of the convenant into one of the following categories: Environment, Housing, Transportation, Healthcare, Education, Economic development, Public Safety, Energy, Social Services, Agriculture, Technology and Innovation.*

(3) *Extract the signing date of the convenant format dd-mm-yyyy. If it is unfindable, return None.*
(4) *Extract the starting date of the convenant in the format dd-mm-yyyy. If it is unfindable, return None.*
(5) *Extract all the involved organisations that have signed the convenant.*
(6) *Classify whether this is a convenant or not (True or False). A convenant is an agreement by the government with one or more parties, aimed at achieving certain (policy) goals. A convenant includes written agreements on (the delivery of performance).*

Your output should be a JSON object with six keys: 'description' (the short description), 'topic' (the classified topic), 'signingdate' (the signing date), 'startingdate' (the starting date), 'parties' (a list of involved parties), and 'convenant' (classification).

For the validation of the method there the generated information will be tested for Precision, recall and F1Score. For the involved parties a list will be composed manually for a subset of the data to check whether the information generated is correct. Next to that the Spacy NER algorithm is used to compare the effectiveness of the large language model. Only exact matches of parties are counted as correct classifications.

The publish date and description can be evaluated against existing descriptions and dates that are scraped from the web. A subset of the descriptions will be evaluated after creation. The topic of the model is hand labeled and evaluated against the chatgpt outcome.

4 RESULTS

4.1 Convenant scraping

A total of 3011 documents were scraped from 302 organisations. Some organisations published in multiple locations. The total amount of distinct organisations lies around 250. Table 1 displays the distribution of convenants per source.

Source	Organisations	Amount
Officiele bekenmakingen	49	1119
Rijksoverheid	12	184
Belastingdienst	1	42
Notubiz	115	1011
Bestuurlijke Informatie	112	617
Manual	13	38
Total	**302**	**3011**

Table 1: Total Number of Convenants Scraped per Source

The website of Officiele bekendmakingen resulted in 1137 documents when filtering on convenants. A total of 49 different governmental organisations published on this website, mostly consisting of ministries and some municipalities and independant governing bodies. After scraping and cleaning all non-readable documents 1119 convenants were added to the dataset.

Searching for convenant on the advanced search of Rijksoverheid.nl generates 166 results. After the cleaning process, 151 documents were saved as convenants in the dataset. The keyword search generated 547 results. Scraping all items that have convenant in the file title left over 33 documents. When these files are compared to the files in the advanced search, seventeen files were in both datasets, meaning that another sixteen files were not yet found using the advanced search. Upon further inspection the missing files in the advanced search are due to misclassification of the documents. The documents are in the advanced search, but classified as reports or publications and not as convenants. The scraper for the website of the Belastingdienst resulted into another 47 convenants.

A total of 115 municipalities and provinces had an existing Notubiz with convenants on them. From these websites a total of 968 convenants were scraped. 43 of these convenants came from the provinces. The biggest provider being Zuid-Holland with 31 convenants. Then Gelderland with 8, Flevoland with 3 and Friesland with 1. Zeeland did not provide any convenants on its website after cleaning. 925 convenants came from the municipalities. The biggest provider of the these is Amsterdam with 98 convenanten. There are a total of 29 municipalities that only contibute one convenant.

The organisations that use Bestuurlijke informatie yielded a total of 612 convenants divided over 112 different organisations. The biggest provider being Hilversum with 38 convenants.

82% of provinces and municipalities had a findable Notubiz/Bestuurlijke. The only provinces that miss a platform are Drenthe and Friesland. For both provinces there is no link to convenanten in the WOO-index. This would suggest that there are no convenants for the provinces. In most cases of municipalities the same situation as the provinces is true.

For the independant administrative bodies all locations in the WOO-index were checked. All organisations did not provide enough results to build a seperate scraper. Of the 431 websites searched only 12 resulted in actual convenants, leading to a total 38 extra convenants.

A sample set of 156 documents was taken to evaluate how many of the documents were actually convenants. Seventeen of these documents were not actually convenants, resulting in an accuracy of 89%. Further analysis of the non convenants scraped leads to the conclusion that in most cases, there was no reasonable suspicion that these files were not convenants. There is nothing in title or filename that might suggest that these files are not convenants. Examples of these files are titled “Convenant Centrale Toegang”, “Samenwerkingsconvenant lokale alliantie voorkomen en aanpak financieel misbruik”, which are diplomatic notes. Or “Convenant milieuzone lichte bedrijfsauto’s” and “Convenant verzekering” which are both letters from the board of directors of a municipality.

4.2 Metadata extraction

Table 2 gives an overview of how much metadata is published with the documents depending on the different platform. The different platforms very in the amount of metadata published with ducuments. The smaller organisations publish metadata overall less structured than the bigger organisations

On the website of Officielebekendmakingen it is possible to filter on convenants. The documents come with a title, publishdate and

	Convenant classification	Title	Topic	Description	Publish date	Involved parties
Officielebekendmakingen	●	●			●	❍
Rijksoverheid	●	●	●	●	●	❍
Belastingdienst		●		●		❍
Notubiz		❍		❍	●	
Bestuurlijke informatie		❍		❍	●	
Manual		❍		❍	●	

Table 2: Metadata Publication per source

Note: In this table, ● denotes true for all cases, while ❍ indicates the item exists in some cases but not universally.

the responsible party for the convenant. Other involved parties in the document cannot be seen in the metadata. A description and topic is completely missing from the publication.

Rijksoverheid.nl is the best performing website on metadata. The documents of the Rijksoverheid are often neatly published with large amounts of metadata. First of all, there are no convenants that come without a small description of the content. All convenants have a meaningful title assigned to them. 100% documents come with a publication date and at least one responsible organ. This organ does however not go beyond the ministries, so it is missing other external parties. Finally, all documents contain one or more subjects, making it possible to find other documents related to this convenant.

Convenants published on the website of the Belastingdienst always have a small description. The junction of the Ministry of finance is the only scraped website that does not provide a date of publication. The title of the document is always "convenant" and then the involved party in the document. This gives little exlpanation on what the convenant is about, but does give information on the involved parties.

The results from Notubiz are published with a small amount of metadata. All documents do get a publish date with them. Next to the date there is a link to the conference the convenant was discussed. 90% of convenants do come with a small amount of text, but this is not a description of the content. The text given with the document is a small snippet of the text in the document.

The documents from Bestuurlijk Informatie come with a title, publishdate and source. None of the documents have a description or a subject of the document. Documents published on Bestuurlijke informatie do come with the option for classifying documents. However the problem here is that there is no option to classify as convenant.

The convenants that were manually scraped from websites differ in the amount of metadata given. The publication for smaller organisations was less structured, but often had a lot of information for each document. In no cases of the manually scraped documents was there a classification of convenants

4.2.1 Convenant classification. One of the objectives of this research is to assess the capabilities of the GPT-3.5-turbo model to extract missing metadata from convenants. The first function of the model is to classify whether the document is a convenant or not. 156 documents were manually evaluated on whether they are convenants or not. Of the 156 documents evaluated 17 were not actually convenants. The model was able to correctly identify one of these seventeen. Leading to a recall score of 0.06. Since the model did not predict any convenants as non-convenants the precision is 1. The combined F1 score for the classification of convenants by gpt-3.5-turbo is 0.11.

	Precision	Recall	F1 Score
Classifying convenants	1.00	0.06	0.11

Table 3: Performance Metrics of the GPT Model for Classifying Convenants

4.2.2 Party extraction. To validate the extraction of parties from convenants, the same validation set was used, excluding all non-convenants. This left a dataset of 139 actual convenants. Within these convenants there were a total 828 parties to extract. Gpt-3.5-turbo was able to correctly extract 750 of these, making 78 false classifications. This results in a precision score of 0.91, a recall of 0.77 and a F1 score of 0.83. When looking at the results at convenants level, the model made no mistakes or misses in 64% of convenants. Meaning more than two in three of convenants are extracted fully correct. After doing an in depth analysis of the mistakes made by the model, two main error causes are identified. First of all, Gpt-3.5-turbo has a token intake limit of 4096, meaning it can only take in the first 4096 characters. In larger convenants this means that the section of involved parties is missing, meaning all involved parties in the document are able to be read in by the model. The second common mistake is merging multiple parties into the same instance. When a convenant has a large number of similar parties the model will merge them into one entity. For example, merging all twelve provinces into 'Deputies of different provinces'. Appendix 1 displays an example of what can be done with having convenants that structurally contain all involved metadata with the convenants.

The model is compared to Spacy's named entity recognition model (nl_core_news_lg). This model was able to identify 384 parties correctly, with 4047 false classifications. Resulting into a precision of 0.09, a recall of 0.44 and an F1 of 0.14. The Spacy model made a lot of mistakes due to the combination of lack of domain

	Precision	Recall	F1
GPT	0.91	0.77	0.83
SPC	0.09	0.44	0.14

Table 4: Performance metrics for GPT and SPC on party extraction

knowledge and unstructured text. While reading in the text using PyPDF2, a lot of spaces, whitelines and structure got lost. For the model lead to a lot of mistakes like: “deArbeidsomstandighedenwet”, “deelconvenant” and “geheleconvenantperiode”. These spacing mistakes were seen by the model as organisations.

4.2.3 *Date Extraction.*

Gpt-3.5-turbo was able to generate 1281 values for signing dates. Of these 1281 values, only ten were actual dates containing a day, month and year. The rest of the values were just months or just a day and month. This does not give much information about the possible duration or relevance of the convenant. The main cause of the model not generating dates based on the documents is because they are not within the first 4096 characters of the document. In many convenants the signing date is at the end of the document or not in the document at all. To test this a sample set was created and tested for the date giving the model the last 4096 characters of the convenant. Two hundred convenants were tested on giving the last 4096 characters of the document. In this case the model performed better, but was still only able to generate a signing date in 18.5% of convenants in the format dd-mm-yyyy. When validating the correctness of these generated dates only 21 of these were correct, meaning that in just 10.5% of cases the model was able to generate the correct signing date.

When analyzing the mistakes the model made it becomes clear that the model has trouble with the signing date because it is often not there. For many scraped documents the signing date is left to be hand-written like in the image below.

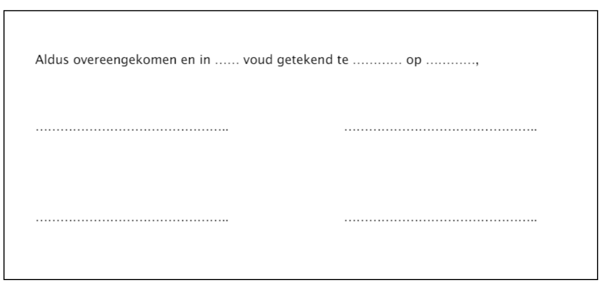

Aldus overeengekomen en in ...... voud getekend te .... op ....,

Figure 2: **Example of how the signing date is left open to be filled in by hand.**

In other cases the model had trouble with deciding with what the signing date is, since in many cases no context is given to the signing date. The model then chooses another date in the file that has context. The starting date extraction of the convenant generated 1991 results. Of these results, 1550 were actual dates in the dd-mm-yyyy format. When doing the validation on the documents, almost none of them returned the correct starting date. The overall accuracy of the model came out to about 1%.

4.2.4 *Description & topic modeling.*

The description created by the model resulted in an accuracy score of 86%. The descriptions given by the model were often based on the first sentences of the document. The most common mistake of the model is being too short in its description, forming more of a title than an actual description. The descriptions were most clear when there was a small section dedicated to what the convenant was about in the beginning of the document. Even when documents contain a long list of involved parties, and therefore much of the input data is taken up by the parties, the model is able to provide an adequate description on what the document is about.

Metric	Accuracy
Description	0.86
Topic	0.91

Table 5: Accuracy of Description and Topic Classification

For topic modeling, out of 139 topics classified, 126 were correctly identified, resulting in an accuracy of 91%. Most errors occurred with documents whose subjects did not clearly fit into any predefined categories. In some cases the model would make up a new category. All topics have a large amount of convenants classified to them. The most common topic classified by is the model is education with 361 documents assigned to it. The least common category is housing with 192 convenants.

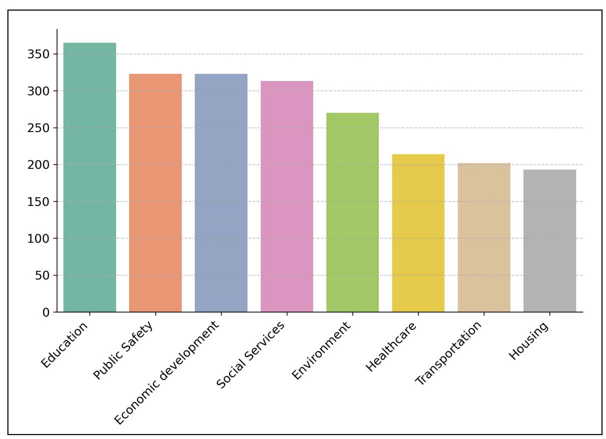


Figure 3: Amount of topics classified per category

5 DISCUSSION

Scraping convenants from the internet is challenging because many websites do not classify these documents by type. For the website of Rijksoverheid and officiëlebekendmakeingen there is such classification. Still, for these website there are errors being made marking documents as convenants that are not and vice versa. By classifying the document type, convenants and such become a lot more findable.

The websites that contain a large number of convenants are structurally scraped in this research. But there are still a lot of websites that contain just one or two convenants. Building a scraper for each of these websites is not sustainable.

Publishing documents with metadata also makes them more findable and interoperable. For the website of the Rijksoverheid this goes well. The documents are all published on their own page and have a topic and responsible party for the document. The topic and party is clickable, which links to more information on the entity. The other websites can still improve a lot on this.

The websites that contain a large number of convenants are structurally scraped in this research. But there are still a lot of websites that contain just one or two convenants. building a scraper for each of these websites is not sustainable.

No real estimate of how many convenants exist on the internet can be made after this research yet. Manually scraping the internet a lot of websites with just a handful of convenants can be found. As an example, the website of the police has a url where each unit can publish their convenants (https://www.politie.nl/wet-open-overheid/convenanten). However on this URL just one convenant is published by the National Expertise and Operations Unit. When looking at the dataset collected there are a lot more convenants published by the police that are not on this website. The convenants are scattered through the web too unstructured to make a real estimate of how many there actually are.

Some aspects of metadata can be successfully extracted by GPT-3.5, but can not give a guarantee of structurally extracting all data. The biggest limitation of the model is its inability to read over 4096 characters. Since convenants are in many cases far over this amount of characters. The model cannot return the correct data when the input does not provide the complete picture. Since the convenants are in many cases structured in such a way where the parties are the first thing mentioned in the document these are often extracted successfully. The beginning of the document gives enough context for a description and topic to be successfully generated. Yet, when it comes to extraction of a specific date the documents are often too large to find the right date. The input maximum of 4096 characters is fixed and cannot be increased. Next to that, even if the correct date is given within the input characters, the convenants often lack context regarding the dates. The model will often take a random date that is found in the document since there is no clear context on what is the correct date. Regardless of the problems with the maximum input characters and lack of context, the model does show a lot of potential. No extra training data is required for the model to run successfully. As long as the input data is structured and contains all required information it can retain information very effectively. The model outperforms Spacy's most comprehensive Dutch language model in party extraction F1 score by 492.86%. This difference is mainly due to Spacy making a lot of mistakes. Likely, these mistakes are because of the model lacking domain knowledge on specific smaller organizations.

Similar research by Huang et al., analyzed the potential of GPT-3.5 on extracting structured data on clinical notes. The research concluded that the model is very capable of performing the task, scoring an overall accuracy of 89%. With the party extraction and topic- and description modeling this research showed similar statistics. The flaws of the model in this research are overlapping with the flaws in this research. Huang et. al concluded that most mistakes made by the model are due its ability to infer from logical reasoning. In this research this can be seen in the date extraction. The model is not able to infer that a standalone date on top or bottom of the page is the date of signing and will therefore choose another date.

This research has mostly focussed on the scraping and extraction of convenants published by the central government and provinces and municipalities. Some independent administrative bodies have been manually scrapped, but in many cases the convenants of these bodies were not on the website provided by Overheid.nl. Future research may look further into how the smaller governmental bodies are publishing their documents and if there are actual patterns in their publishing that were not found in this research.

6 CONCLUSION

The Dutch Government Openness Act (WOO) requires governmental bodies to publish their documents in a findable manner. A great way to improve findability of documents is publishing them with metadata. This research has looked into how many convenants are published by Dutch governmental organizations in accordance with the WOO, and with how much metadata they come with. Next to that the ability to gather the metadata afterwards using GPT-3.5 is tested.

Convenants of 302 different organisations have been scraped, resulting in a dataset of 3011 documents. A large problem in the publication of convenants is not classifying them, in some cases making everything seem like it is a convenant. Next to that, the scatteredness of the convenants on the internet makes it impossible to find them all. There are too many websites that contain just one or two convenants to be able to find them all. These two factors combined make it impossible to get an actual number on the amount of convenants that are published.

The publication of metadata with the convenants can be improved on many fronts. The ministries of the central government publish their documents with a lot of metadata already. Smaller governmental bodies like local governments and independent administrative bodies can still improve in the publication of descriptions, topics and in some cases even the correct title of the document.

GPT-3.5's ability to classify convenants was tested on the dataset, resulting in a F1 score 0.11. In most cases the model is not able to classify non-convenants as such The GPT-3.5 model generally performs well on extracting parties from the document. The model produced an F1 score of 0.81 outperforming without any extra domain training required. The model far outperformed the most advanced Dutch Spacy model, which only scored an F1 of 0.14. Extraction of dates resulted in an accuracy of .11. This huge difference is mainly due to the lack of context that convenants often give to dates, and the models inability to take in more than 4096 characters. This makes GPT-3.5 less viable for extraction of dates in convenants. Finally, when modeling a topic or description to a convenant the model resulted in an accuracy of 0.91 and 0.86 respectively. In conclusion, the Dutch government needs to improve the findability of convenants published by governmental organizations. While GPT-3.5 showed potential in extracting parties, topics, and descriptions from convenants, its inability to handle dates and lack of accuracy in convenant classification limit its usefulness in this specific task.

REFERENCES

[1] [n. d.]. spaCy 101: Everything you need to know. https://spacy.io/usage/spacy-101. Accessed: 2024-02-23.

[2] [n. d.]. Wijze van openbaar maken Woo-informatiecategorie 'Convenanten'. https://www.open-overheid.nl/instrumenten-en-diensten/richtlijnen/2024/2/16/hulpmiddel-convenanten. Accessed: 2024-06-26.

[3] [n. d.]. Woo-index. https://organisaties.overheid.nl/woo. Accessed: 2024-06-26.

[4] 2022. A framework for domain-specific distant supervised named entity recognition. *Proceedings of WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications* (2022). https://doi.org/10.18178/wcse.2022.04.003

[5] J. Chiu and E. Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics* 4 (2016), 357–370. https://doi.org/10.1162/tacl_a_00104

[6] Robert Dale. 2021. GPT-3: What's it good for? *Natural Language Engineering* 27, 1 (2021), 113–118. https://doi.org/10.1017/S1351324920000601

[7] Hui Han, C.L. Giles, E. Manavoglu, Hongyuan Zha, Zhenyue Zhang, and E.A. Fox. 2003. Automatic document metadata extraction using support vector machines. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings.* 37–48. https://doi.org/10.1109/JCDL.2003.1204842

[8] Jingwei Huang, Donghan M. Yang, Ruichen Rong, and et al. 2024. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine* 7 (2024). https://doi.org/10.1038/s41746-024-01079-8

[9] Pooja Kherwa and Poonam Bansal. 2019. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems* 7, 24 (2019).

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua* (2016). https://doi.org/10.18653/v1/n16-1030

[11] Leonard Richardson and Matthew L. Ward. [n. d.]. Beautiful Soup Documentation. https://pypi.org/project/beautifulsoup4/. Accessed: 2024-06-07.

[12] Brady D Lund and Ting Wang. 2023. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library hi tech news* 40, 3 (2023), 26–29.

[13] M. Aalbersberg I. J. Appleton G. Axton M. Baak A. … Mons B. M. D., Dumontier. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data* (2016).

[14] Jaap Kamps Maarten Marx. 2023. Digitaal duurzaam publiceren van Woo-dossiers. *Archievenblad* (2023).

[15] Jaap Kamps Maarten Marx, Joep Meindertsma. 2023. Maak hergebruik van Woo-informatie nu echt mogelijk. *od-online* (2023).

[16] Maarten Marx. 2024. Onderzoeksrapport naar de digitale toegankelijkheid van documenten die de overheid publiceert onder de Woo. *Adviescollege Openbaarheid en informatiehuishouding* (2024).

[17] Yuhong Mo, Hao Qin, Yushan Dong, Ziyi Zhu, and Zhenglin Li. 2024. Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *arXiv preprint arXiv:2405.06652* (2024).

[18] J. Qin, A. Leathers, and V. T. Tompkins. 2020. Linking mechanisms in data repositories: a case study of biosample database. *Proceedings of the Association for Information Science and Technology* 57 (2020). Issue 1. https://doi.org/10.1002/pra2.365

[19] Algemene Rekenkamer. 1995. Convenanten van het Rijk met bedrijven en instellingen. *Tweede Kamer* 1996, 24 (1995), 480.

[20] Rijksoverheid. 2024. Hoofdlijnen Wet open overheid. https://www.rijksoverheid.nl/onderwerpen/wet-open-overheid-woo/hoofdlijnen-woo.

[21] Selenium Project. 2024. WebDriver. https://www.selenium.dev/documentation/webdriver Last accessed: June 10, 2024.

[22] Kenniscentrum voor beleid en regelgeving. 2022. Aanwijzingen voor convenanten.

Appendix A FIRST APPENDIX

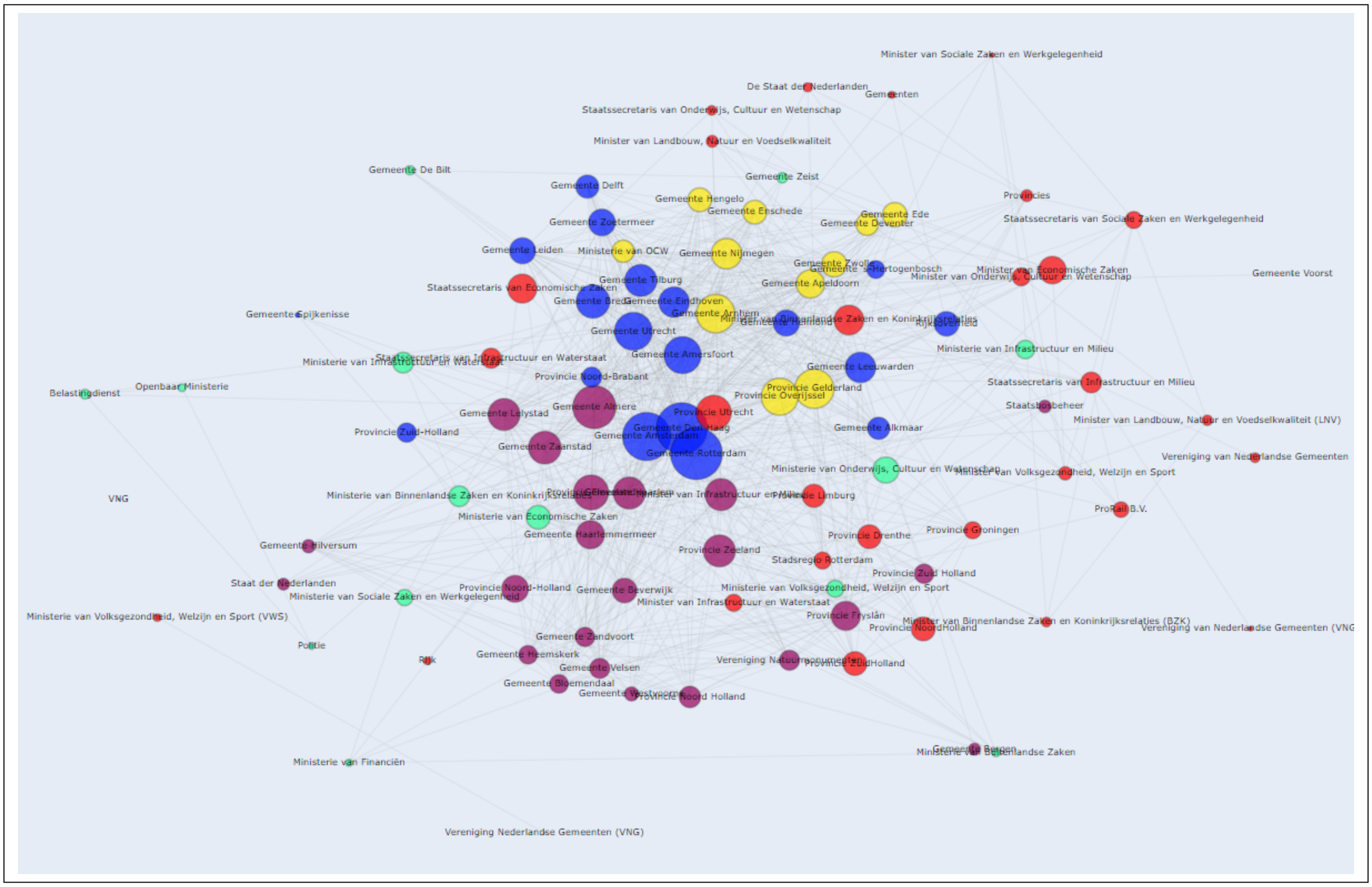


Figure 4: Clustering graph of top 100 involved parties in convenants. Each node is an organisation. The size of each node is increased with each convenant they are a part of. Each edge is a cooperation in a convenant. The colors are clusters of nodes that often work together.