data

MDPI

Article

Enhancing Access Across Europe for Documents Published According to Freedom of Information Act: Applying Woogle Design and Technique to Estonian Public Information Act Document

Gerda Viira * and Maarten Marx *

Faculty of Science, Informatics Institute, University of Amsterdam, 1098 XH Amsterdam, The Netherlands
* Correspondence: gerda.viira@student.uva.nl (G.V.); maartenmarx@uva.nl (M.M.)

Abstract: In the Netherlands, the Open Government Act (Wet openbare overheid or Woo/Wob in Dutch) is in effect, with the primary objective of ensuring a more transparent government. In line with the legislation, a search engine named Woogle has been designed and developed to centralize documents published under the Open Government Act. The Estonian Public Information Act serves a similar purpose and requires all public institutions to publish information generated during official duties, fostering transparency and public oversight. Currently, Estonia's document repositories are decentralized, and content search is not supported, which hinders people's ability to efficiently locate information. This study aims to assess public information accessibility in Estonia and to apply Woogle's design and techniques to Estonia's document repositories, thereby evaluating its potential for broader European implementation. The methodology involved web scraping data and documents from 57 Estonian public institutions' document repositories. The results indicate that Woogle's design and techniques can be implemented in Estonia. From a technical perspective, the alignment of the fields was successful, while it was found that content-wise, the Estonian data present challenges due to inconsistencies and lack of comprehensive categorization. The findings suggest potential scalability across European countries, pointing to a broader applicability of the Woogle model for creating a corpus of Freedom of Information Act documents in Europe. The collected data are available as a dataset.

Keywords: Freedom of Information Act; FAIR data; open government

Citation: Viira, G.; Marx, M. Enhancing Access Across Europe for Documents Published According to Freedom of Information Act: Applying Woogle Design and Technique to Estonian Public Information Act Document. *Data* **2024**, *9*, 125. https://doi.org/10.3390/data9110125

Received: 17 September 2024
Revised: 16 October 2024
Accepted: 25 October 2024
Published: 29 October 2024

1. Introduction

Freedom of Information Act (FOIA) refers to the public's right to access information that the government, local municipalities, and other public institutions have created when performing public duties. Access to this information is considered one of the pillars of democracy. FOIA laws are crucial for information distribution in countries, and the EU in general, from a political accountability and media scrutiny perspective. Recent decades have seen significant legislative efforts aimed at enhancing transparency, particularly within the European Union [1]. In the Netherlands, since 2022, the Open Government Act (known as *Wet Open Overheid* (Woo) in Dutch) has been in effect, with the primary objective of ensuring a more transparent government. A search engine called Woogle [2] was developed to align with this legislative framework. It provides a centralized, machine-readable repository of documents released under the Act [3]. A similar legal framework exists in Estonia where the Public Information Act mandates that public institutions disclose any information obtained or generated while performing public duties [4]. This law serves the purpose of creating opportunities for society to monitor the activities of public institutions. It ensures that information is published systematically and in a straightforward manner for public access. Currently, Estonia's document repositories are decentralized, and searches are mainly conducted on the title of the documents as content searches are not supported.

In this paper, we aim to, firstly, assess public information accessibility in Estonia, and secondly, apply Woogle's design and techniques to Estonia's document repositories, assessing its suitability for other European countries. This study seeks to identify best practices, enhance transparency, and lay a foundation for a harmonized approach to FOIA document accessibility in Europe. The integration of Estonia into Woogle is seen as a pilot study for a larger initiative of adding FOIA documents from various European countries. This case study draws parallels to the ParlaMint project, which standardized parliamentary proceedings from 32 countries into a unified format [5,6]. In summary, we focus on the following three tasks:

- Compliance and Accessibility: The extent to which Estonian public institutions adhere to the Public Information Act is assessed, determining the accessibility and availability of documents within the current repository system. This serves as a measure of Estonia's transparency efforts. The adherence is measured based on the requirements for entries in the database.
- Suitability of Woogle's Data Model: The feasibility of adapting Woogle's data model and architecture to manage Estonia's FOIA documents and linked data are analyzed, regarding specific local data and document management practices.
- Potential for Wider Application: The adaptability of the Woogle design to other European legal and administrative contexts is explored, aiming to foster a unified approach to FOIA document accessibility that enhances governmental transparency across Europe.

This paper is structured as follows: Section 2 explores background and related work. Section 3 details the methodology, including data collection via web scraping, validation, and the transformation of Estonia's FOIA data to fit the Woogle model. Section 4 presents the analysis results, and Section 5 evaluates the adaptability and effectiveness of Woogle in Estonia and discusses findings in the context of broader European applications. Section 6 concludes the work by summarizing key insights and suggesting future research directions.

Main Findings

Our results indicate that Woogle's design and techniques can be implemented for Estonian FOIA data and documents. We examined the feasibility of alignment from three perspectives: technical specifications, content perspective, and file perspective. From a technical perspective, the alignment of the fields was successful. However, content-wise, the Estonian data present challenges due to inconsistencies and lack of comprehensive categorization. From a file perspective, we encountered challenges with processing the digital signature file types (ASiC-E and BDOC), but these were resolved within Woogle. The collected data will be made available as a dataset and will be partially accessible through Woogle. The findings suggest potential scalability across European countries, pointing to a broader applicability of the Woogle model for creating a corpus of FOIA documents in Europe.

A subset of almost 10 K documents was uploaded to Woogle and can be found by restricting to the 'Estonian Woogle' on https://woogle.wooverheid.nl/search?country=ee (accessed on 24 October 2024). The complete dataset is available from the authors.

2. Background and Related Work

The significance of FOIA is increasingly recognized at both international and national levels, as it forms a cornerstone of democratic transparency and governance. This research aims to bridge the current gap in the accessibility of FOIA documents across Europe by examining existing practices and exploring the development of a potential uniform corpus. A similar initiative called ParlaMint has successfully combined parliamentary proceedings into a unified format [5]. In the FOIA domain, Woogle has been launched in the Netherlands, with various improvement activities being undertaken. Additionally, initiatives have begun in countries like Belgium and Spain to publish FOIA documents in a digital document repository accessible to the public.

2.1. Unified Approach to Parliamentary Proceedings

Aggregating datasets from diverse countries into a unified corpus has facilitated large-scale comparative research endeavors. The ParlaMint corpus exemplifies this approach by consolidating parliamentary proceedings data from 26 parliaments over 10 years [5]. The corpus employs a decentralized approach where countries who manage their data contribute to the integrated project. Parliamentary proceedings represent just one instance of a resource possessing the requisite characteristics for a large-scale data collection and alignment project. These characteristics include the following:

- Adherence to a shared data model;
- Consistent interpretation across different countries;
- Significant contextual overlap among diverse resources.

The availability of translations into English enabled large-scale analysis of various global topics [7], making this corpus a valuable resource for corpus linguists and social or political scientists seeking insights into various socio-political phenomena. The success of initiatives like the ParlaMint corpus underscores the potential for similar efforts across other resources with comparable properties [5].

Our research aligns with such efforts and advocates for the consolidation of publicly available resources following local Freedom of Information Acts, building upon the demonstrated methods of the ParlaMint project.

2.2. Fair Research Data Principles

The FAIR research data principles are designed to enhance the re-usability and sharing of data, focusing on improving the ability of machines to automatically find and use data [8]. FAIR stands for Findability, Accessibility, Interoperability, and Reusability. Findability ensures that data are clearly identified, described, and indexed. The abstract principles are fleshed out in [9] and we briefly recall these here. The data should include unique identifiers which are stored in a public resource. Accessibility requires data to be accessible through a well-defined process, ideally, the process would be automated and including authentication or authorization procedures. Interoperability means that metadata are conceptualized and structured using common published standards, which involves standard technical and semantic formats, variables, and ontologies. Reusability ensures that data characteristics are described in detail and are aligned with relevant standards.

These guidelines aim to support knowledge discovery and innovation rather than merely managing data [8]. In addition to the previously mentioned guidelines, FAIR data must be machine-readable; in large-scale data-intensive research projects, computers play a crucial role in tasks such as indexation, retrieval, and analysis. Various EU data principle directives further reinforce these guidelines and promote a standardized approach across Europe.

2.3. Freedom of Information in Europe

The EU Directive 2019/1024, known as the 'Open Data and the Reuse of Public Sector Information Directive', regulates the availability and reuse of information produced or held by public sector bodies across EU member states [10]. This directive aims to enhance the efficiency and extent of public sector information reuse, fostering innovation, transparency, and economic growth. Freedom of Information laws fundamentally aim to establish transparency in the public sector, with almost all European countries having enacted some form of FOIA law by 2018, except Luxembourg [1]. These laws typically mandate that government agencies provide access to information pertinent to their functions unless classified otherwise. Differences among national FOIA laws include the scope of accessible information, the methods, and the timelines for information requests, and specific exceptions to disclosure [11]. For instance, while countries like Sweden, Denmark, and Finland emphasize the proactive publication of information, others focus on responding to individual requests.

To promote transparency and ease of access, several European countries have developed digital portals dedicated to publishing information proactively or publishing FOIA requests and responses. One example is Belgium, which offers a region-specific portal in Brussels known as *Transparencia* [12]. This portal allows citizens to access government-held information online and facilitates the requesting and online publication of responses, accessible to all [13]. It stands out by permitting requests under certain pseudonyms, a feature that is not commonly used in other systems. Spain has a similar government-developed portal that provides information about governmental activities [12]. Users must identify themselves through electronic authentication procedures to request information, ensuring security and authenticity in interactions. In Estonia, the law mandates that all public institutions upload new documents in a digital repository [4]. In the Netherlands, the Open Government Act (Wet open overheid, abbreviated as Woo) of 2022 led to the development of the platform *Woogle*.

2.3.1. Woogle—Corpus for Dutch Freedom of Information Requests

Woogle serves as a centralized search engine and is designed to convert materials published under the jurisdiction of the act into a machine-readable format. A data model and corpus have been established for Dutch FOIA Requests. This continuously updated corpus consists of more than 3 million FOIA dossiers obtained from over 1.000 distinct governing bodies, totaling more than 11 million pages, all presented in a standardized format and accessible via the Woogle search engine https://woogle.wooverheid.nl, (accessed on 24 October 2024). All data within the Woogle search engine are freely available for scientific research in FAIR open formats[1]. It is important to acknowledge that the inherent raw nature of FOIA documents diverges from the FAIR research data principles, requiring work to align with standards of findability, accessibility, interoperability, and reusability as mandated by both Dutch FOIA law and broader European guidelines [8,10].

Turning released FOIA documents into FAIR data poses several challenges. One of them is the poor quality of the optical character recognition (OCR) applied to documents containing sensitive information after the redaction process [14]. Additionally, the Dutch government's habit of merging documents into large, undifferentiated PDFs necessitated Page Stream Segmentation techniques to delineate original document boundaries [15]. Several procedures were implemented to address metadata scarcity, extensive document classification, and information extraction [14,16,17].

2.3.2. Public Information Act in Estonia

In the Estonian Freedom of Information law, public information has been defined as information that is recorded or documented in any matter upon the performance of public duties provided by law or legislation [4]. In that law, a document registry is defined as a digital database where state or local government agencies register public information documents. These agencies are required to ensure access to that document registry for the public. New documents need to be uploaded to the document registry the morning of the day after receiving them. Article §12 of the law specifies requirements for the document registries, including the scope, data, and access methods for the repository. It details the requirements for documents that necessitate action from the institution and mandates institutions to publish certain data points for all incoming and outgoing documents. The following metadata must be published regarding received and released documents and entered into the document registry: (1) From whom the document(s) were received or to whom they were sent; (2) The date when they were received or sent; (3) How documents were received or sent; (4) Specific information regarding the documents; (5) The type of the documents; (6) Access restrictions for the documents [4].

3. Materials and Methods

This study follows a case study methodology. It is seen as a pilot study for a larger initiative of adding FOIA documents from various European countries into one repository.

The methodology for incorporating Estonian documents into Woogle consists of multiple sequential stages.

As a first step, the scope is defined, and the current situation is analyzed. Subsequently, data collection involves web scraping existing document repositories to extract publisher, record, and document details, followed by data validation and cleaning. Additionally, a daily scraping job is established. The collected data are then transformed into the Woogle JSON schema to evaluate compatibility and explore the feasibility of extending the model. Finally, an exploratory analysis assesses compatibility with Woogle. The methodology process is visualized in Figure 1, with further details provided in this section. This methodology aligns with the dual objectives of this research: firstly, to explore the current situation, and secondly, to implement improvements aimed at assessing the compatibility with Woogle's structure [18].

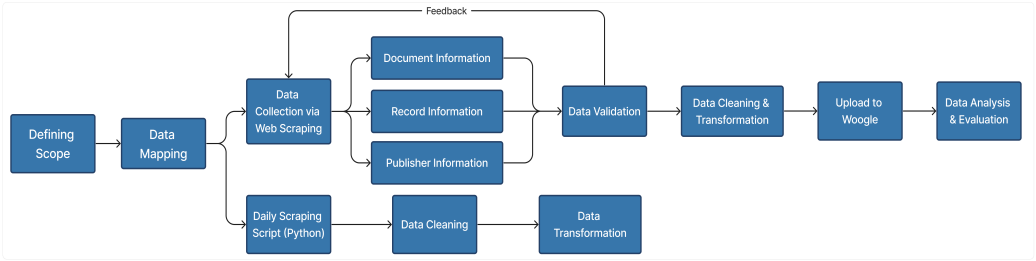

Figure 1. Visualization of the method. The figure illustrates the sequential methodology stages for incorporating Estonian FOIA documents into Woogle, from scope definition and web scraping to data validation, transformation, and compatibility assessment.

3.1. Scope

The document repositories of government agencies in Estonia are decentralized, with each public institution maintaining its own repository.

Searches can only be conducted within the documents published by a specific institution. According to the Estonian National Information Systems database, there are a total of 209 actively used document repositories nationwide [19]. These registries are mainly hosted on two platforms. This paper will focus on the newer and more extensive document repository system, Delta ADR. The Information Systems database indicates that Delta ADR hosts a total of 63 repositories [19]. All sites were assessed to determine if they were active and followed the same architecture and user interface. The sites meeting these criteria were included in the scope of the study. From a total of 63 repositories, four were excluded: three had been discontinued, and one was inaccessible. Thus, the final list comprises 59 document repositories.

These repositories are distributed across various types of institutions, as summarized in Table 1. The full list of institutions included in the study can be seen in Table A1 in Appendix A. The data are collected in bulk for the years 2021 to 2023.

Table 1. Number of institutions for each category present in the Delta ADR repository system and included in the present study.

Type of Institution	Count
Government Agency	23
Local Government	16
Constitutional Institution	10
Other State Agencies	8
Educational Institution	1
State Held Companies	1
Total	59

3.2. Data Collection

We briefly explain each step in the web scraping process: website analysis, website crawling, and data organization [20].

The analysis began with an exploration of the document repositories' structural framework, delving into the user journey, navigation, and the user interface (UI). These elements are crucial for effective web scraping, as they dictate the accessibility and usability of the repository's data. It was found that the document repository websites' UI enables searches based on a limited set of fields, such as document title, upload date, and document type. To see the search results, the user must enter at least one search criteria. In addition, the uniform UI across all websites within the scope of this study facilitates a web scraping job. Furthermore, the analysis entailed the creation of an entity–relationship (ER) model to map the entities, their attributes, and the relationships between them (Appendix B). Following the website analysis and data modeling, the collected data were mapped to the existing Woogle schema, facilitating compatibility assessment and further analysis (Appendix C). This was performed by comparing the data fields, their values, and formatting between the two models.

This study employs two distinct methods for data collection. Firstly, archival data are gathered through bulk data collection using an online web scraping tool to gather document repository entries from 2021 to 2023. Secondly, a daily scripting job developed in Python facilitated the collection of daily newly added entries in the repositories. These methods aim to ensure data collection for historical analysis and compatibility testing with Woogle, as well as ongoing data collection for process improvement.

Both methods adopt a similar scraping process described in Appendix D. In both scraping methods, the data extraction occurs on three levels:

- Publisher level: publisher name, page title, source link;
- Record level: reference, record title, source URL, document type, published date, other attributes (such as responsible person, sent/received date);
- Document level: document URL, document title, and the document itself.

Due to differences between the Estonian and Dutch models of publishing FOIA documents, adjustments are necessary to ensure the accurate grouping of records. In Estonia, each new file or communication is published individually if there are no access restrictions on the specific entry. In contrast, the Dutch model consolidates documents within a dossier, comprising the original request, decision, and requested documents. Individual entries in the Estonian document repository do not constitute complete dossiers; instead, they are interlinked, collectively forming a dossier. The data descriptions and mapping are visible in Appendix C.

3.3. Data Description

Websites from 57 institutions were scraped for the publications of three years, collecting 1,159,165 entries linked to 666,638 documents. Summary statistics for the dossier and documents datasets are provided in Appendix E (Tables A3 and A4).

The distribution of entries collected per publisher category shows that data from Government Agencies constitute almost half of the dataset, whereas data from educational institutions and state-held companies are the least represented (Table 2). The distribution per publisher is detailed in Table A5 in Appendix E.

The Estonian document repository includes entries with documents for public access ('Avalik') and documents with restricted access ('AK'). Out of the total entries, 338,040 entries are classified as public and 821,125 as not public. Within the public scope, the majority of the entries have a small number of documents linked to them, with a median of 0 and a maximum of 165. Table 3 shows the frequency of the number of documents per entry. For entries with documents, the dataset shows that 162,515 entries contain two to four documents, while 150,681 entries have only one document. Entries with more than five documents are less frequent.

Table 2. Dataset Distribution by Publisher Category. The table presents the number of entries collected and the number of documents for each publisher category.

Publisher Category	Number of Entries	Number of Documents
Government Agency	585,359	250,859
Local Government	394,440	390,311
Other State Agency	129,858	14,949
Constitutional Institution	36,319	5880
Educational Institution	10,962	3536
State Held Company	2227	1103
Total	1,159,165	666,638

Table 3. Frequency table of the number of documents within each entry in the document repositories.

Number of Documents	Frequency
0	828,600
1	150,681
2–4	162,515
5–9	14,491
10–19	2110
20+	1180

When analyzing the classification by document type, it was found that 548,187 were classified as incoming letters and 306,371 as outgoing letters. Overall, 121 different labels are used inconsistently across institutions (see Table A3).

Cross-tabulation and heatmaps were utilized to explore the relationships between categorical variables, excluding variables with a direct relationship such as 'publisher name', 'publisher category', and 'FOIA page title'. The analysis investigated the relationship between 'access restriction' and 'publisher category', as well as 'publisher name', utilizing normalization for heatmaps to address data imbalances across categories. The normalized heatmap for access restriction and publisher category (Figure A3) revealed that the category 'Other State Agency' has the lowest concentration of public documents (9% of all published documents), while the 'State-Held Company' category published documents for 37% of the entries. The heatmap analyzing documents per publisher category indicated that institutions such as the Military, the Southern District Prosecutor's Office, and the Center of Registers and Information Systems had more than 91% entries restricted for public access. In contrast, the Ministry of Finance, Ministry of Economic Affairs and Communications, and the Rescue Service published the most entries to the public, with over 60% of entries being public.

Additionally, a time-series analysis was conducted and was divided into four components: observed, trend, seasonal, and residual. Firstly, looking at the observed component provides an overview of the changes in the data over time. The entry publications per month vary from 27,500 to the peak of 36,877, with the lowest number of publications occurring in July in all years (Figure 2). Document publications follow a similar pattern, with 14,000 to 22,000 documents published monthly. Notably, July is the only month with fewer than 15,000 documents published.

Secondly, the trend for entries shows a slight upward trajectory from July 2021 to January 2023, followed by a downward trend after January 2023. On the contrary, the documents show a slight downward trend.

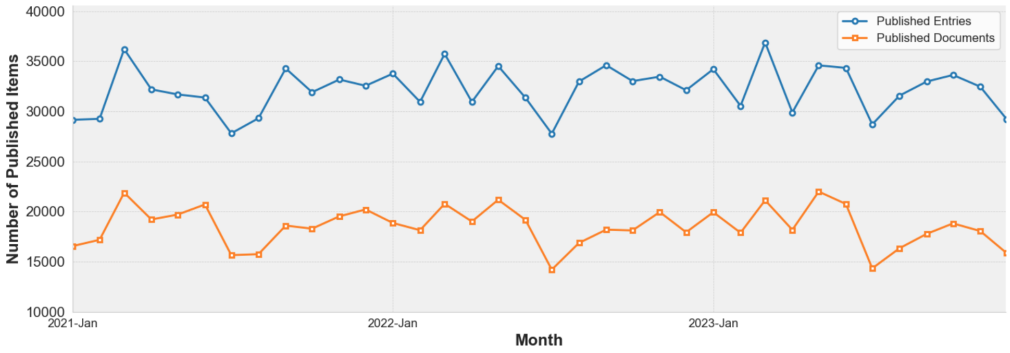

Figure 2. Monthly Document Publication Counts (2021–2023). Represents the number of documents published per month between 2021 and 2023 on the public institutions' document repositories included in the scope of this study.

4. Results

4.1. Compliance with Public Information Act

The analysis first focused on the adherence of public institutions to Estonia's Public Information Act, which mandates specific criteria for each document entry, such as sender or recipient details, transmission dates and methods, document properties, and access restrictions. In this study, the compliance of the documents with the law was assessed based on the Estonian Public Information Act §12(3). It specifies the information that has to be entered into the document repository, including the following:

- Sender or recipient details (name or reference);
- Date of receipt or release;
- Transmission method;
- Requisite information on the documents;
- Type of documents;
- Access restrictions.

Compliance with these criteria was assessed with the population rate metric, except for the requisite information on the documents, which cannot be assessed due to ambiguity in the law regarding what constitutes the requisite information. Requirements in §12(1), §12(2) and §12(4) were not assessed. §12(1) and §12(2) indicate the types of documents that should be and should not be registered in the repository. This was not assessed as there is no golden source available to compare with. In addition, §12(4) requires that for documents that need a response or a resolution, the responsible person and a deadline have to be mentioned—this is not assessed as no tag determines which documents need a response. §12(5) requires full-text search to be available on the data but does not mention a search capability of the documents' content. This was confirmed to be present on every document repository site in the scope of this study.

The compliance rate was assessed in three categories: All entries in the document repository, outgoing documents, and incoming documents. The next subsections will highlight the findings in these categories. We emphasize that the reason that documents do not comply is because the data in the source is truly not available. It is not due to a technical fault in our information extraction process.

4.1.1. All Documents

Access restrictions and document types are required to be published for all entries in the repository. A 100% population rate was found for these fields.

4.1.2. Outgoing Documents

For outgoing documents, the required fields are as follows: (1) Sender or recipient reference or name; (2) Arrival/sending method; (3) Date sent. The population rates are presented in Table 4. Two fields can contain sender or recipient details: 'recipient' and

'sender or recipient reference'. Therefore, the occurrence of either field was also assessed, resulting in a combined population rate of 95.9%. The lowest population rate was found for the date sent field (78.6%).

Table 4. Population rates of mandatory fields for incoming (N = 548,187) and outgoing (N = 306,371) documents. The mandatory fields are defined by the Public Information Act §12(3).

Field Name	Population Rate (%)	
	Incoming Documents	Outgoing Documents
Recipient	96.01	95.79
Sender/Recipient Reference	37.32	14.62
Sender/Recipient Reference OR Recipient	96.15	95.89
Arrival/Sending Method	96.32	91.26
Date Sent	na	78.58

4.1.3. Incoming Documents

For incoming documents, the required fields are as follows: (1) Sender or recipient reference or name; (2) Arrival/sending method; (3) Date sent; again, see Table 4. The receiving date is not mentioned as a separate data item. Similarly to the outgoing documents category the sender and recipient details can be in two fields. A combined population rate of 96.1% was found.

4.2. Suitability of Woogle's Data Model

To assess the compatibility of Woogle's data model and architecture with Estonian FOIA documents, the key fields required for uploading information were examined for alignment with Woogle's system specifications. Compatibility was assessed from three perspectives:

- Data model perspective: Evaluating technical requirements for the data;
- Content perspective: Assessing the alignment of the content within the fields to ensure similar and comparable data are used;
- File and File Type Compatibility: Analyzing compatibility of different file formats and types used in Estonian repositories with Woogle's system requirements.

An overview of the results can be found in Table 5.

Table 5. Field compatibility between Woogle schema and Estonian data.

Category	Field Names
Compatible	dc_identifier, dc_title, dc_source, dc_publisher, FOIA_page_title, dc_publisher_name, FOIA_nrDocuments, FOIA_retrievedDate, FOIA_publishedDate, dc_date_year, FOIA_fileName
Not Compatible	dc_type
Added	function, series

From a technical requirements perspective, all fields were found to be compatible with Woogle's data schema. From a content perspective, the dc_type field was found to be not compatible with Woogle's categorization of document types. Compared to the Netherlands, the usage of document types (dc_type) in Estonia is less comprehensive. In Woogle, the document types include specific categories such as drafts of laws, annual plans, reports, investigative reports, etc. In Estonia, entries are primarily categorized as incoming or outgoing letters without specifying the document type. In addition, 548,187 were classified as incoming letters and 306,371 as outgoing letters. Overall, 121 different labels are used inconsistently across institutions, which complicates categorization and can lead

to duplication. As the document type field was incompatible, it was substituted with the function and series fields used in the Estonian repository system. These fields describe the institutional function served by the document and allow for a more refined categorization of entries and dossiers by type. There are 4993 unique function and series combinations, making labeling into Woogle categories not feasible. See Table 6 for three examples.

Table 6. Examples of unique combinations of the Function and Series Fields. Count represents the number of occurrences within the dataset.

Function	Series	Count
8 Administration of public services [1]	8-8 Criminal record documents [2]	27,714
PRP-10 Memos, clarification requests, statements, complaints, requests and requests for information [3]	PRP-10 Memos, clarification requests, statements, complaints, requests and requests for information [3]	9473
RP-6 Main activities of the Prosecutor's Office [4]	RP-6-11 Requests for foreign legal aid [5]	11,434

[1] 8 Avalike teenuste haldus, [2] 8-8 Karistusregistri dokumendid, [3] PRP-10 Märgukirjad, selgitustaotlused, avaldused, kaebused, taotlused ja teabenõuded, [4] RP-6 Prokuratuuri põhitegevus, [5] RP-6-11 Välisriikide õigusabitaotlused.

In addition, the entry title (dc_title) in Estonian repositories is highly repetitive, with the top 15 titles occurring in the dataset more than 10,000 times each. The most frequent title appeared 46,247 times. The word cloud in Figure 3 represents the repetitiveness of the titles in the dataset. For Estonian entries in Woogle, the document type, title, and original unique identifier were combined to form a more descriptive title.

Figure 3. Word cloud of entry titles. Displaying titles with over 100 occurrences in the dataset. Titles were translated with Google Translate.

The analysis of file formats, displayed in Figure 4, revealed that approximately 41.15% of the files scraped from the Estonian document repositories use the .bdoc or .asice file format, 42.93% are PDF files, and 4.08% are Word (docx) documents. Other file types such as .png, .msg, .jpg, and .txt each make up less than 2% of all the documents.

The accessibility of the files is a crucial aspect of the upload to Woogle. Documents must be accessible for download. This is essential for the pipeline in Woogle to download and OCR the documents. In analyzing the data, specific file extensions to Estonia were identified. These .bdoc and .asice file formats were previously unrecognized by Woogle. ASiC-E format serves as the current standard in Estonia for digitally signed files, whereas historically the .bdoc format was predominantly used [21]. The ASiC-E signature is a BDOC signature with a timestamp that corresponds to the RFC 3161 standard [21]. Therefore, this observation indicates the necessity of expanding Woogle's capabilities to accommodate the file formats found in the Estonian document repositories. Ultimately, we were able

to open the ASiC-E and BDOC containers and convert the contents to PDF files, enabling us to run the Woogle pipeline with the OCR process and indexing to make the content machine-readable and searchable.

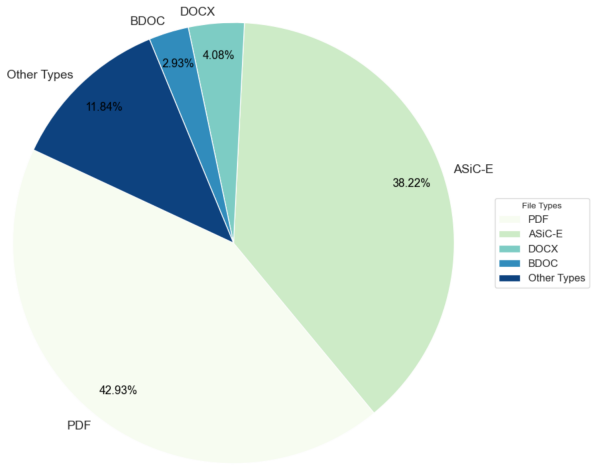

Figure 4. File format distribution in the Estonian document repositories. PDF files are the most common in the dataset, followed by ASiC-E. The 'Other' category includes formats such as PNG, MSG, and TXT.

5. Discussion

This study investigated public information accessibility in Estonia and applied Woogle's design and techniques to the country's document repositories, evaluating their potential suitability for other European nations. Taking inspiration from other large-scale data collection initiatives such as ParlaMint, the research involved collecting data from the document repositories of 57 Estonian public institutions for analysis.

We showed that compliance with the Public Information Act in Estonia is high. The compliance for the mandatory fields for all documents was 100% for both document type and access restrictions fields. For outgoing documents, the lowest population rate was found at 78.5%, while for incoming documents, all requirements were fulfilled on average 96% of the time. Despite the high compliance, content analysis revealed that several fields are not used for their purpose. Institutions frequently use the document type as the entry title, resulting in many repetitions and not unique titles. Additionally, over 73% of the entries are labeled as 'incoming letter' or 'outgoing letter', without referencing to the actual document or inquiry type.

The current user interface of Estonia's document repositories facilitates searches based on a limited set of fields, including the entry title, upload date, and document type. Due to the numerous repetitions and the misuse of the fields, it is difficult to conduct targeted searches. The absence of content-based search functionality limits the system's effectiveness and prevents users from retrieving documents based on specific content within the published documents. Despite the repository's easy access and the policy of publishing all documents by the next day, technological and data shortcomings hinder usability. Consequently, the current state of the document repository system poses significant barriers to the public's access to information, which is contrary to the objectives of the Public Information Act, aimed at fostering transparency and facilitating public oversight.

In exploring the adaptability of Woogle's design to the Estonian context, the results reveal that most fields were successfully integrated into Woogle's schema. However, a misuse of the fields resulted in content misalignment with Woogle in two instances. The requirements for unique entry titles were not met, so we generated unique titles by combining several fields. The document type field was populated as type 00, with different types being used compared to the Netherlands, and the entries were inconsistent across various Estonian institutions.

Furthermore, the use of the ASiC-E and BDOC file formats in the Estonian document repositories presents an accessibility issue. The content of these files is not indexed by search engines such as Google [22], preventing finding and retrieving the documents. As mentioned previously, the document titles are often generic and repetitive, making it difficult to locate the information without a robust search functionality. The inability to index the content of the documents amplifies the problem, as users cannot rely on the search functionality to find the documents based on the content. At Woogle, we had to extract and store the underlying PDF files from the ASiC-E and BDOC formats. This required extraction process raises the question of the necessity of using these file formats. Considering the drawbacks in accessibility of using these file formats, it is worth considering more accessible file formats such as PDFs. PDF files are universally supported and can be indexed by search engines, thereby improving accessibility and the utility of publishing FOIA documents. Alternatively, the ASiC-E and BDOC containers could be opened when the documents are published to the document repository to provide easy access and indexing for search engines. In Estonia, there is a document viewer available for opening the ASiC-E and BDOC containers, DigiDoc4, which is the same application used for digital signatures. However, in other European countries, these file types are not a standard and, therefore, the general public does not have the necessary tooling to open the containers.

The potential scalability of Woogle is dependent on the prerequisites for a large-scale data collection and alignment project, such as Parlamint. The characteristics mentioned are contextual overlap, consistent interpretation and a shared data model. Our study confirms that contextual overlap and interpretation criteria are satisfied in the Estonian context. At the same time, the shared data model presents challenges in Estonia and potentially in other countries due to national differences in document handling and publication. Further investigation is required to test the Woogle model with more countries. Each country needs to have an online source for accessing FOIA documents. These sources can vary significantly—countries like Estonia and the Netherlands proactively publish documents or do so upon request. This study found portals in Belgium and Spain that publish national or regional FOIA documents on a website that is accessible to citizens or everyone. However, the adoption of broader, uniform initiatives to publish and standardize the format of FOIA documents across various countries is still notably lacking. Continued research into the scalability of Woogle across other European countries could further make a way toward a unified European corpus of FOIA documents.

Limitations

This paper acknowledges several limitations potentially affecting generalizability, reproducibility, and scalability. This research was fixed to three years and included 57 public institutions using the ADR Delta repository systems, limiting the results to that context. Additionally, the method of web scraping used in this research is not a future-proof method of data collection compared to API-based methods. Changes in the UI or data placement would disrupt data collection. Regarding compliance with the Estonian Public Information Act, the study only verifies the presence of data points, not the completeness of the entries in the repository, due to the lack of a golden source for comparison. Despite these limitations, the study provides insights into public information accessibility in Estonia and lays a foundation for future research in other European countries.

6. Conclusions

This research aimed to assess public information accessibility in Estonia and apply Woogle's design and techniques to Estonia's document repositories, evaluating its suitability for other European countries. The study found high compliance with the Estonian Public Information Act in aspects quantifiable within the dataset's context. It demonstrated the feasibility of applying the Woogle design and technique to the Estonian Freedom of Information (FOIA) documents. It provides a blueprint for enhancing the accessibility of public information across European countries. Our analysis highlights that, despite

the high compliance rates with the Public Information Act in Estonia, significant barriers remain in document accessibility due to technical limitations in the current storage method. This highlights a need for advancements in search capabilities and document management practices. Our findings reveal that Woogle facilitates the aggregation of data into a uniform format for more comprehensive research. Based on the individual differences in data structuring and modeling, adaptions may be needed to cater to the country-specific frameworks. Future research should thus focus on adapting the Woogle model to accommodate FOIA documentation from various other European countries. Testing the model in different contexts will contribute to the development of a unified European corpus of FOIA documents.

Author Contributions: Conceptualization, G.V. and M.M.; methodology, G.V. and M.M.; software, G.V.; validation, G.V.; resources, G.V. and M.M.; data curation, G.V.; writing—original draft preparation, G.V.; writing—review and editing, G.V. and M.M.; supervision, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Netherlands Organization for Scientific Research (NWO) through the ACCESS project grant CISC.CC.016 and an Open Science Fund grant no. 01607400. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-5788.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A subset of almost 10 K documents is searchable and available in the Woogle datadump when restricting to the 'Estonian Woogle' on https://woogle.wooverheid.nl/search?country=ee, (accessed on 24 October 2024). The complete dataset is available from the authors.

Acknowledgments: We thank Maik Larooij for extensive help in preparing the Estonian documents for uploading to Woogle.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Institutional Scope

Table A1. List of institutions included in this study, with the institution names translated to English, the types of organizations, and the URLs to the document repositories for each specific institution.

No	Institution Name	Type	URL
1	Harju County Court	Constitutional Institutions	https://adr.rik.ee/harjumk/, (accessed on 24 October 2024)
2	Chancellor of justice	Constitutional Institutions	https://adr.rik.ee/okk/, (accessed on 24 October 2024)
3	Pärnu County Court	Constitutional Institutions	https://adr.rik.ee/parnumk/, (accessed on 24 October 2024)
4	Supreme Court	Constitutional Institutions	https://adr.rik.ee/riigikohus/, (accessed on 24 October 2024)
5	Tallinn Administrative Court	Constitutional Institutions	https://adr.rik.ee/tallinnahk/, (accessed on 24 October 2024)
6	Tallinn District Court	Constitutional Institutions	https://adr.rik.ee/tallinnark/, (accessed on 24 October 2024)
7	Tartu Administrative Court	Constitutional Institutions	https://adr.rik.ee/tartuhk/, (accessed on 24 October 2024)
8	Tartu County Court	Constitutional Institutions	https://adr.rik.ee/tartumk/, (accessed on 24 October 2024)

Table A1. *Cont.*

No	Institution Name	Type	URL
9	Tartu District Court	Constitutional Institutions	https://adr.rik.ee/tarturk/, (accessed on 24 October 2024)
10	Viru County Court	Constitutional Institutions	https://adr.rik.ee/virumk/, (accessed on 24 October 2024)
11	Estonian Academy of Security Sciences	Educational Institutions	https://adr.sisekaitse.ee/ska/, (accessed on 24 October 2024)
12	Data Protection Inspectorate	Government Agencies	https://adr.rik.ee/aki/, (accessed on 24 October 2024)
13	Emergency Center	Government Agencies	https://adr.112.ee/hk/, (accessed on 24 October 2024)
14	Ministry of Justice	Government Agencies	https://adr.rik.ee/jm/, (accessed on 24 October 2024)
15	Ministry of Defence	Government Agencies	https://adr.rik.ee/kmin/, (accessed on 24 October 2024)
16	Defense Resources Board	Government Agencies	https://adr.rik.ee/kra/, (accessed on 24 October 2024)
17	Military	Government Agencies	https://adr.rik.ee/mil/, (accessed on 24 October 2024)
18	Estonian Competition Authority	Government Agencies	https://adr.rik.ee/ka/, (accessed on 24 October 2024)
19	Ministry of Culture	Government Agencies	https://adr.rik.ee/kum/, (accessed on 24 October 2024)
20	Western District Prosecutor's Office	Government Agencies	https://adr.rik.ee/laanerp/, (accessed on 24 October 2024)
21	Southern District Prosecutor's Office	Government Agencies	https://adr.rik.ee/lounarp/, (accessed on 24 October 2024)
22	Ministry of Economic Affairs and Communications	Government Agencies	https://adr.rik.ee/mkm/, (accessed on 24 October 2024)
23	Estonian Rescue Board	Government Agencies	https://adr.rescue.ee/paa/, (accessed on 24 October 2024)
24	Estonian Patent Office	Government Agencies	https://adr.rik.ee/pa/, (accessed on 24 October 2024)
25	North District Prosecutor's Office	Government Agencies	https://adr.rik.ee/pohjarp/, (accessed on 24 October 2024)
26	Police and Border Guard Board	Government Agencies	https://adr.politsei.ee/ppa/, (accessed on 24 October 2024)
27	Ministry of Finance	Government Agencies	https://adr.rik.ee/ram/, (accessed on 24 October 2024)
28	Estonian Information System Authority	Government Agencies	https://adr.rik.ee/rik/, (accessed on 24 October 2024)
29	State Prosecutor's Office	Government Agencies	https://adr.rik.ee/riigiprokuratuur/, (accessed on 24 October 2024)
30	Ministry of Interior	Government Agencies	https://adr.siseministeerium.ee/sisemin/, (accessed on 24 October 2024)
31	Ministry of Social Affairs	Government Agencies	https://adr.rik.ee/som/, (accessed on 24 October 2024)

Table A1. *Cont.*

No	Institution Name	Type	URL
32	Health Board	Government Agencies	https://adr.rik.ee/ta/
33	Viru District Prosecutor's Office	Government Agencies	https://adr.rik.ee/virurp/, (accessed on 24 October 2024)
34	Alutaguse Municipality Government	Local Governments	https://adr.novian.ee/alutaguse_vald/, (accessed on 24 October 2024)
35	Kanepi Municipality Government	Local Governments	https://adr.novian.ee/kanepi_vald/, (accessed on 24 October 2024)
36	Kohila Municipality Government	Local Governments	https://adr.novian.ee/kohila_vald, (accessed on 24 October 2024)
37	Lääne-Nigula Municipality Government	Local Governments	https://adr.novian.ee/laane-nigula_vald, (accessed on 24 October 2024)
38	Lääneranna Municipality Government	Local Governments	https://adr.novian.ee/laaneranna_vald/, (accessed on 24 October 2024)
39	Paide Town Government	Local Governments	https://adr.novian.ee/paide_linn/, (accessed on 24 October 2024)
40	Põltsamaa Municipality Government	Local Governments	https://adr.novian.ee/poltsamaa_vald, (accessed on 24 October 2024)
41	Räpina Municipality Government	Local Governments	https://adr.novian.ee/rapina_vald/, (accessed on 24 October 2024)
42	Saaremaa Municipality Government	Local Governments	https://adr.novian.ee/saaremaa_vald, (accessed on 24 October 2024)
43	Saku Municipality Government	Local Governments	https://adr.novian.ee/saku_vald, (accessed on 24 October 2024)
44	Sillamäe Town Government	Local Governments	https://adr.novian.ee/sillamae_linn/, (accessed on 24 October 2024)
45	Tapa Municipality Government	Local Governments	https://adr.novian.ee/tapa_vald/, (accessed on 24 October 2024)
46	Tartu Municipality Government	Local Governments	https://adr.novian.ee/tartu_vald, (accessed on 24 October 2024)
47	Tõrva Municipality Government	Local Governments	https://adr.novian.ee/torva_linn, (accessed on 24 October 2024)
48	Luunja Municipality Government	Local Governments	https://adr.novian.ee/luunja_vald, (accessed on 24 October 2024)
49	The Geological Survey of Estonia	Other State Agencies	https://adr.rik.ee/egt/, (accessed on 24 October 2024)
50	Estonian Forensic Science Institute	Other State Agencies	https://adr.rik.ee/ekei/, (accessed on 24 October 2024)
51	Financial Intelligence Unit	Other State Agencies	https://adr.fiu.ee/smit/, (accessed on 24 October 2024)
52	Center of Registers and Information Systems	Other State Agencies	https://adr.rik.ee/rik/, (accessed on 24 October 2024)

Table A1. *Cont.*

No	Institution Name	Type	URL
53	IT and Development Center, Ministry of the Interior	Other State Agencies	https://adr.smit.ee/smit/, (accessed on 24 October 2024)
54	Tallinn prison	Other State Agencies	https://adr.rik.ee/tallinnav/, (accessed on 24 October 2024)
55	Tartu prison	Other State Agencies	https://adr.rik.ee/tartuv/, (accessed on 24 October 2024)
56	Viru prison	Other State Agencies	https://adr.rik.ee/viruv/, (accessed on 24 October 2024)
57	State Infocommunication Foundation	State Hold Companies	https://adr.rik.ee/rit/, (accessed on 24 October 2024)

Appendix B. Data Model

The model, represented in Figure A1, serves as a visual representation of the data architecture. Each entry in the document repository was found to be associated with several key attributes, including title, registration date, and document type. These entries exhibit a one-to-zero or many relationships with properties, which includes fields such as sent and received dates —these fields are populated only when applicable. Furthermore, each entry may be linked to zero or many documents, contingent upon the presence of access restrictions. Each document entry comprises a title and a URL. It is important to understand the data structure to scrape and organize data from the current repository, ensuring it is easy to analyze in the further stages.

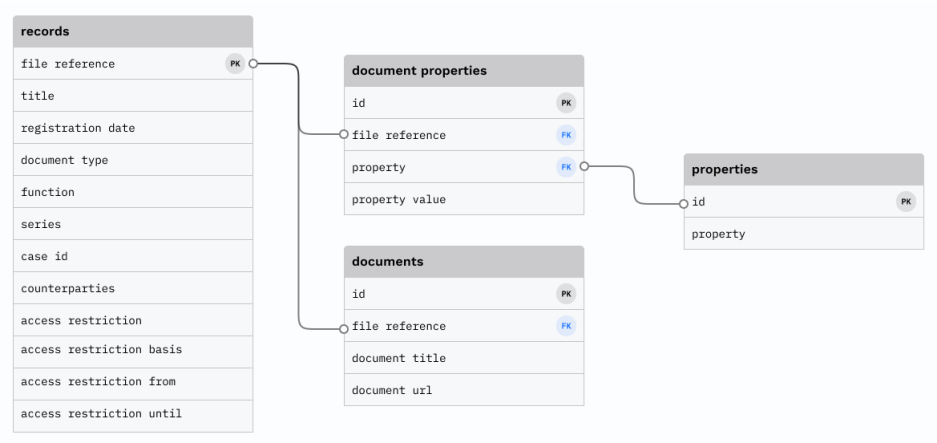

Figure A1. Entity–relationship model.

Appendix C. Data Mapping

Table A2. Data mapping between Estonian document repository fields and Woogle schema.

Field Name in Woogle	Field Explanation According to the Mapping to Estonian Model
dc_identifier	Unique identifier in the following format: country_code.dc_publisher.dc_type.dc_date_year.count.
dc_title	Title of the entry is combined from the type of document, title of the document and original unique identifier.
dc_type	Type of document. Type 00 is used as a default value.

Table A2. *Cont.*

Field Name in Woogle	Field Explanation According to the Mapping to Estonian Model
dc_source	Entry URL.
dc_publisher	Abbreviation of the publishing institution.
FOIA_nrDocuments	Total number of documents in the dossier.
FOIA_retrievedDate	Date of scraping.
FOIA_publishedDate	Publication date of the entry.
dc_date_year	Year of publishing the entry, derived from the publication date.
dc_publisher_name	Publishing institution's full name.
FOIA_page_title	Website title from which the information was scraped.
function	Estonia-specific field containing the categorization of the topic (main category).
series	Estonia-specific field containing the categorization of topic (subcategory).
FOIA_files: FOIA_fileName	File name.
FOIA_files: dc_source	File URL.

Appendix D. Web Scraping Process and Tooling

Web Scraping Process Flow

The web scraping process is illustrated in Figure A2. The process starts with navigation to the first repository website. To access the entries in the repository, at least one search criteria must be entered. In this case, the date-of-upload field is utilized for the search queries. The tool enters the specified date range in the begin and end date fields and initiates the search. The results are presented in a paginated table format. The tool extracts data from this table and then iterates through each result on the page, accessing the respective sub-page to gather additional details and documents. Upon completion, the tool navigates to the next page of search results, repeating the process for all websites within the specified scope and date range.

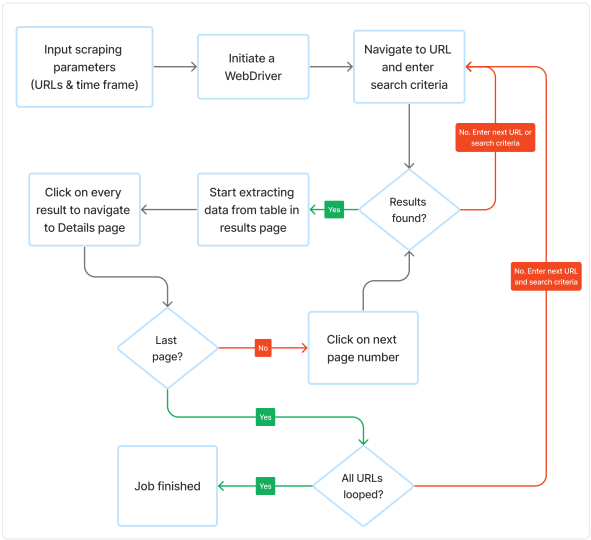

Figure A2. Web scraping workflow.

Appendix E. Exploratory Data Analysis Results

Appendix E.1. Summary Statistics

Table A3. Dossier dataset summary statistics. This summary includes the count of data points collected for each field, the number of unique values for each field, the most frequent (top) value, and the frequency of the most frequent value.

Variable	Count	Unique	Top	Freq
dc_identifier	1,159,165	930,962	2-2	206
FOIA_publishedDate	1,159,165	1106	2023-05-02	1870
dc_title	1,159,165	237,377	Taotlus	46,247
dc_source	1,159,165	1,159,165	https://adr.rik.ee/mkm/dokument/13799251, (accessed on 24 October 2024)	1
dc_type_description	1,159,165	121	Sissetulev kiri	548,187
dc_publisher_name	1,159,165	57	Saaremaa Vallavalitsus	51,991
FOIA_page_title	1,159,165	57	Saaremaa Vallavalitsuse avalik dokumendiregister	51,991
function	1,159,142	738	8 Sotsiaalse kaitse ja tervishoiu korraldamine	59,543
series	1,159,160	4869	- -	47,981
case_id	1,111,480	35,333	-	46,955
access_restriction	1,159,165	2	AK	821,125
access_restriction_basis	830,487	6231	AvTS § 35 lg 1 p 12	373,942
access_restriction_from	821,596	1854	12-9-2022	1416
access_restriction_until	818,386	6631	1-11-2096	1174
FOIA_retrievedDate	1,159,165	15	14-2-2024	450,710
sender/recipient_reference	257,787	178,159	e-mail	762
recipient	864,160	98,745	Eraisik	48,699
arrival/sending_method	856,763	689	e-post	471,261
responsible_person	1,013,352	11,946	Jelena Viilas (Registrite ja Infosüsteemide Keskus, Riikliku sunni registrite osakond, Karistusregistri talitus)	13,050
resolution_date	238,641	1180	17-10-2022	516
resolution_deadline	279,413	1360	30-4-2021	838
other_parties	44,644	16,608	Eraisik	4874
end_date	7711	1531	31-12-2023	450
sending_date	384,377	NaN	NaN	NaN
additions	4674	2533	sõidupäevik	126
publisher_category	1,159,165	6	Government Agency	585,359
FOIA_nrDocuments	1,159,165	NaN	NaN	NaN

Table A4. Documents dataset summary statistics. This summary includes the count of data points collected for each field, the number of unique values for each field, the most frequent (top) value, and the frequency of the most frequent value.

Variable	Count	Unique	Top	Freq
FOIA_fileName	666,638	439,959	E-kiri.pdf	62,749
dc_source	666,638	666,638	not applicable	1
page_url	666,638	330,565	https://adr.novian.ee/tartu_vald/dokument/5097917, (accessed on 24 October 2024)	165

Table A5. Number of entries and total number of documents collected from each institution's document repository over the period of three years (2021–2023).

Publisher Name	Frequency	Document Count
Saaremaa Municipality Government	51,991	64,203
Ministry of Justice	49,462	16,248
Health Board	48,528	25,864
Military	47,981	6737

Table A5. *Cont.*

Publisher Name	Frequency	Document Count
Estonian Rescue Board	41,278	46,744
Police and Border Guard Board	39,517	20,800
Paide Town Government	39,325	36,055
Ministry of Finance	38,962	38,030
Center of Registers and Information Systems	37,393	4968
North District Prosecutor's Office	34,893	1805
Tapa Municipality Government	34,659	23,151
Sillamäe Town Government	34,384	27,887
Põltsamaa Municipality Government	34,026	28,490
Data Protection Inspectorate	32,014	4986
Saku Municipality Government	31,821	27,575
Ministry of Interior	29,818	12,921
Tartu Municipality Government	29,772	37,359
Defense Resources Board	29,723	965
Ministry of Economic Affairs and Communications	29,430	28,523
Ministry of Defence	29,204	10,414
Lääne-Nigula Municipality Government	28,971	30,875
State Prosecutor's Office	26,609	1214
Tartu prison	26,340	2331
Tõrva Municipality Government	26,044	25,400
Viru prison	24,392	2418
Ministry of Social Affairs	23,652	16,155
Räpina Municipality Government	22,931	16,495
Lääneranna Municipality Government	21,330	22,555
Chancellor of justice	20,163	1433
Tallinn prison	19,449	498
Estonian Competition Authority	17,587	1581
Viru District Prosecutor's Office	16,846	1520
Southern District Prosecutor's Office	14,550	1261
Kohila Municipality Government	13,346	13,437
Ministry of Culture	11,721	9825
Estonian Academy of Security Sciences	10,962	3536
Alutaguse Municipality Government	10,561	15,070
Kanepi Municipality Government	9808	11,290
Western District Prosecutor's Office	9520	796
Estonian Forensic Science Institute	8939	1155
Estonian Information System Authority	7355	1699
IT and Development Center, Ministry of the Interior	6913	1789
Supreme Court	6692	1658
Luunja Municipality Government	5471	10,469
Emergency Center	5437	2136
Financial Intelligence Unit	5073	756
Tartu County Court	4142	477
State Infocommunication Foundation	2227	1103
Pärnu County Court	1542	1000
The Geological Survey of Estonia	1359	1034
Harju County Court	1346	308
Estonian Patent Office	1272	635
Viru County Court	771	119
Tartu Administrative Court	662	340
Tallinn District Court	482	250
Tartu District Court	381	256
Tallinn Administrative Court	138	39
Total	1,159,165	666,638

Appendix E.2. Categorical Variable Analysis

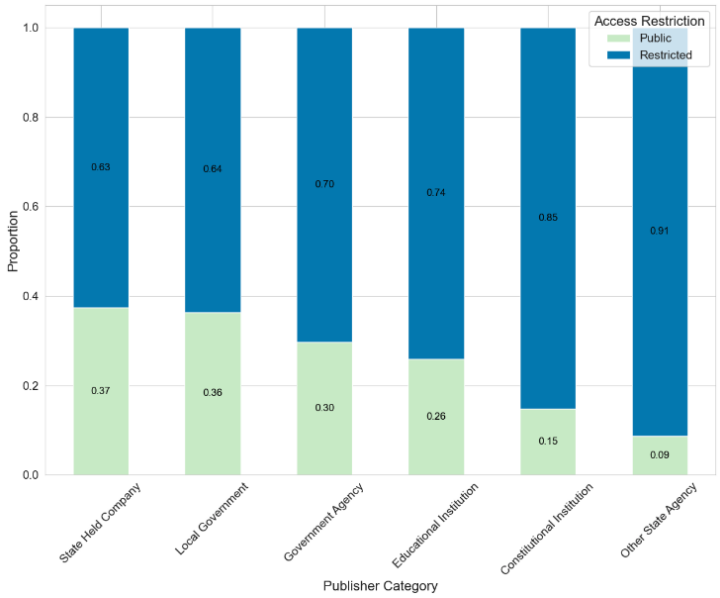

Figure A3. Proportion of Public and Restricted Entries per Publisher Category.

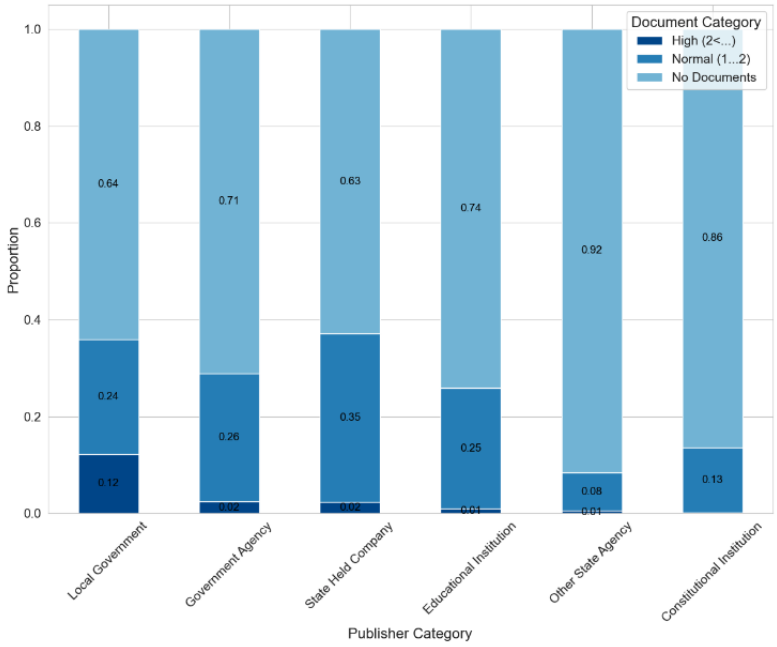

Figure A4. The proportion of Document Counts by Publisher Category. The categorization of the number of documents is as follows: No documents (0), Normal (1–2), and High (2+).

Notes

[1] Available at doi.org/10.17026/dans-zau-e3rk. See also https://wooverheid.nl/2023/05/12/woogle-data-nu-vrij-beschikbaar/ (accessed on 24 October 2024).

References

1. Mokrosinska, D. *Transparency and Secrecy in European Democracies: Contested Trade-Offs*; Routledge: Abingdon, UK, 2021.
2. Woogle. Home. Available online: https://woogle.wooverheid.nl/ (accessed on 11 December 2023).
3. Woogle. About. Available online: https://woogle.wooverheid.nl/about (accessed on 11 December 2023).
4. Estonian Government. Public Information Act. Available online: https://www.riigiteataja.ee/en/eli/514112013001/consolide (accessed on 3 March 2024).

5. Erjavec, T.; Ogrodniczuk, M.; Osenova, P.; Ljubesic, N.; Simov, K.; Pancur, A.; Rudolf, M.; Kopp, M.; Barkarson, S.; Steingr'ımsson, S.; et al. The ParlaMint corpora of parliamentary proceedings. *Lang. Resour. Eval.* **2023**, *57*, 415–448. [CrossRef] [PubMed]
6. Viira, G.; Marx, M.; Larooij, M. ParlaMint Widened: A European Dataset of Freedom of Information Act Documents (Position Paper). In Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024; 2024; pp. 171–172. Available online: https://aclanthology.org/2024.parlaclarin-1.0/ (accessed on 24 October 2024).
7. Kuzman, T.; Ljubešić, N.; Erjavec, T.; Kopp, M.; Ogrodniczuk, M.; Osenova, P.; Fišer, D.; Pirker, H.; Wissik, T.; Schopper, D.; et al. Linguistically Annotated Multilingual Comparable Corpora of Parliamentary Debates in English ParlaMint-en.ana 3.0, 2023. Slovenian Language Resource Repository CLARIN.SI. Dataset. Available online: http://hdl.handle.net/11356/1810 (accessed on 28 March 2024).
8. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
9. Boeckhout, M.; Zielhuis, G.A.; Bredenoord, A.L. The FAIR guiding principles for data stewardship: Fair enough? *Eur. J. Hum. Genet. EJHG* **2018**, *26*, 931–936. [CrossRef] [PubMed]
10. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on Open Data and the Re-Use of Public Sector Information (Recast). Available online: https://eur-lex.europa.eu/eli/dir/2019/1024/oj (accessed on 3 March 2024).
11. Stein, L.L.; Camaj, L. Freedom of Information. In *Oxford Research Encyclopedia of Communication*; Oxford University Press: Oxford, UK, 2018. [CrossRef]
12. Portal de la Transparencia. Available online: https://transparencia.gob.es/transparencia/transparencia_Home/index.html (accessed on 10 March 2024).
13. Lambrecht, I. New in Belgium: Transparencia—An Online Portal to Access Government-Held Information. Available online: https://www.law.kuleuven.be/citip/blog/new-in-belgium-transparencia-an-online-portal-to-access-government-held-information/ (accessed on 28 March 2024).
14. van Heusden, R.; Ling, H.; Nelissen, L.; Marx, M. Making PDFs Accessible for Visually Impaired Users (and Findable for Everybody Else). In Proceedings of the TPDL 2023, Zadar, Croatia, 26–29 September 2023.
15. Wiedemann, G.; Heyer, G. Multi-Modal Page Stream Segmentation with Convolutional Neural Networks. *Lang. Resour. Eval.* **2021**, *55*, 127–150. [CrossRef]
16. van Heusden, R.; de Ruijter, A.; Majoor, R.; Marx, M. Detection of Redacted Text in Legal Documents. In Proceedings of the TPDL 2023, Zadar, Croatia, 26–29 September 2023.
17. Bakker, F.; van Heusden, R.; Marx, M. Timeline Extraction from Decision Letters Using ChatGPT. In Proceedings of the CASE 2024 Colocated with EACL, St. Julians, Malta, 17–22 March 2024.
18. Runeson, P.; Höst, M. Guidelines for Conducting and Reporting Case Study Research in Software Engineering. *Empir. Softw. Eng. Int. J.* **2009**, *14*, 131–164. [CrossRef]
19. Riigi Infosüsteemi Amet. Infosüsteemid. Available online: https://www.riha.ee/Infos%C3%BCsteemid?sort=meta.update_timestamp&dir=DESC (accessed on 22 December 2023).
20. Krotov, V.; Johnson, L.; Silva, L. Legality and Ethics of Web Scraping. *Commun. Assoc. Inf. Syst.* **2020**, *47*, 539–563. [CrossRef]
21. Information System Authority. What is the Difference Between Digitally Signed Documents with .bdoc and .asice Extensions? Available online: https://www.id.ee/en/article/what-is-the-difference-between-digitally-signed-documents-with-bdoc-and-asice-extensions/ (accessed on 13 February 2024).
22. File Types Indexable by Google. Available online: https://developers.google.com/search/docs/crawling-indexing/indexable-file-types (accessed on 10 June 2024).