**Schema**

## File format

A suitable file format should fit the data it is storing. In order to identify a compatible file format, the type of data that is dealt with needs to be researched. A few characteristics can be identified. Firstly, the documents and their overarching 'publication' can be thought of as objects. A publication contains document objects. Both the publication and documents need to have attributes, such as a title and a date, which do not have to be ordered. Secondly, not all objects have the same relevant attributes. An active disclosure of information does not have the need to store the same data as decisions on information requests, for example a decision date or if the request is adjourned. However, both active disclosures and decisions on information requests are grouped as publications. Similarly, a released document has more relevant attributes than a decision or inventory document. Metadata on released documents contain for example a originator and recipient, and sometimes grounds of refusal according to a relevant Wob- or Woo-article. Nevertheless, all types of documents are grouped together. A third characteristic is that there is no fixed number of documents belonging to a publication. Some publications contain tens or even hundreds of documents, while other publications are refused according to a relevant ground, resulting in the sole existence of a decision document. Likewise, some attributes, like grounds of refusal, can contain more values.

Also, a more user-centered approach can be taken in finding an optimal data model. For creators of the metadata, it is convenient that the data are human-readable. This will ideally improve archiving and reuse of the publications, as well as the communication between the government and it's inhabitants. For end-users of the data, like researchers, political figures and journalists, the data also must be human-readable, referring back to the experiment by van Oostveen & van Loenen (2014), which concluded that the responses to Dutch information requests were too mixed to make sense of. Also, the data must be computer-readable, as the current publications lack computer-readability. This should improve the researchability and transparency of the published information. Research trying to monitor the government on for example response time of the information requests, carried out by the Instituut Maatschappelijke Innovatie and Open State Foundation (2022), was seriously impeded by the lack of computer-readable data. About 1000 information request were analyzed by hand, producing about 12.000 records of data. Such time-consuming data gathering would be obsolete if there was suitable metadata on the publications available. More computer-readable data also provide better searchability.

From analyzing the publications and associated documents, their relations and attributes, the conclusion can be drawn that the needed data model is neither structured nor unstructured. It is not structured, because the model needs to deal with changing attributes and a variable number of documents attached to a publication. Also, it is not unstructured, because a number of tags or attributes can be identified to indicate and separate semantic elements of the data. Looking at the perspective of the data users, an unstructured approach would also not be a great fit, as this does not solve the problem of human- and computer-readability. There is a need for a flexible format. The answer is a *semi-structured* approach. The key idea is the representation of data as a tree-like structure (Buneman, 1997). Also, semi-structured data is said to be self-describing and suitable for object-oriented databases. This fits the challenge of storing publications and document objects, where documents belong to publications. However, the taken approach will incline towards a structured approach, as type and value constraints on the data can be introduced as well as predetermined, well-defined attributes.

**JSON**

**Met opmerkingen [ML1]:** Meer vergelijk? Bijna geen wetenschappelijke bronnen hierover…

Two prominent file formats satisfy the requirement of supporting the storage of semi-structured data. These are the Extensible Markup Language (XML) and JavaScript Object Notation (JSON). XML requires structuring the data into a document structure, JSON structures are based on arrays and

records, making it a simpler data exchange format (maybe 2 examples). XML and JSON have much in common, but JSON has a smaller grammar and is based on data structures used in modern programming languages, making it better compatible with languages like Python and Javascript when analyzing the data (Crockford, 2006). A case study by Nurseitov *et al.* (2009) indicates that JSON is faster and uses fewer resources than XML. Because of the simplicity of JSON, it fits the requirement of human-readability better than XML. Being better integrated into modern programming languages, JSON complies to the requirement of researchability, making it easier to analyze data directly before having to parse all the data first. Moreover, JSON is used in the Elasticsearch stack, tackling searchability. While using JSON for this exact purpose is a recommendation, the remainder of this research will contain examples and use-cases in JSON.

## Attribute selection

Relevant attributes for publications and documents can be found and extracted from different points of view. These include consulting existing metadata standards on files, exploring existing journalistic solutions, using earlier work and research, analyzing provided metadata for current publications and the more scientific approach of identifying technical metadata for research purposes. This chapter will treat all points of view, going from abstract metadata from existing standards to more case specific attributes. At the end of each source, the additional relevant attributes that will be used are listed.

**Metadata standard (Dublin Core)**

Numerous of metadata standards and ontologies have been proposed to consistently store metadata on subjects. The Dublin Core Metadata Initiative (DCMI) created a set of core elements to describe resources, such as video, images and files. The element set has been formally standardized as ISO 15836 (International Organization for Standardization, 2019). The original version consists of fifteen metadata elements, however not all are applicable on this case. The standard aids in formulating some basic metadata on publications and documents. For publications, we can identify the following attributes: identifier, title, subject, description, date and type. The type attribute will help in determining if the publication is based on a request or an active disclosure. For documents, the following set of attributes is extracted from the Dublin Core standard: identifier, title, description, format, date, type, language. The format will be the mime type of the files. The type attribute can be used twice. Firstly to denote the document type (request, decision, inventory list or released document) and secondly, in the case of a released document, the type of released document, such as a memo, an agenda or an email. Attributes like contributor, creator, publisher and source are considered too generalized for the publications and documents, as these attributes call for more case specific metadata. This will be further elaborated in the next sections. The remaining Dublin Core attributes, coverage, relation and rights, seem not applicable for publications and documents.

**Publication:** identifier, title, subject, description, date, type

**Document:** identifier, title, description, format, date, type, language

**Journalistic standard (Follow the Money)**

Next to considering the somewhat technical Dublin Core standard, a more end-user centered approach can be taken. This approach is more focused on extracting meaningful information on the data. Dutch research journalism platform ‘Follow the Money’ (FTM) created their own Follow the Money data model to organize concepts in a for journalists useful way to investigate things (Follow the Money, 2021). The hierarchical data model contains two relevant objects with attached metadata. On top of the hierarchy is a ‘thing’, which describes a real-world object. A ‘document’ is a subtype of a ‘thing’, directly relating to this case’s released publication documents. A thing contains some of the same attributes as Dublin Core, like a name (title), a description and topics. Two new useful and relevant attributes can be added to that list: a source url (more specific than the Dublin Core ‘source’) and the

‘retrievedAt’, which can be interpreted as a download date. These attributes introduce a new kind of attributes. This new kind includes attributes that are obtained after publication and can be filled during research of the metadata, while the source url and a download date are not available when publishing the documents. Inspection of the defined ‘document’ object also yields new attributes: file size, file name and body text all contain useful information on documents. A document also contains the mime type and extension attributes, replacing Dublin Core’s ‘format’ attribute.

**Publication:** sourceUrl, retrievedAt

**Document:** sourceUrl, fileSize, fileName, bodyText, mimeType, fileExtension

**Earlier work (Open State, VNG, North-Holland)**

From research perspective, The Open State Foundation in cooperation with the province of North-Holland and the association of Dutch municipalities (VNG) already created a concept version of a guideline for standardized attributes to attach to information request, to improve publishing and archiving the information requests (Open State, VNG, Provincie Noord-Holland, 2021). The guideline, named OpenWob, contains recommendations on request level and is composed in cooperation with a province and municipalities. As these bodies will handle the information requests, the guideline contains useful attributes for handlers and creators of the information requests, as well as some request specific attributes. Firstly, some attributes are already covered by Dublin Core or FTM: identifier, title, subject (called a theme in the guideline) and a description. The guideline splits the date out in a file date and a decision date. Furthermore, the guideline provides some new attributes: the responsible body for handling the request, the requester, an internal id, a boolean attribute indicating adjournment of response, the first recipient of the request and some geographical attributes. The handler of the request of publication seems a better describing attribute than Dublin Core’s ‘author’, ‘publisher’ or ‘contributor’.

**Publication:** fileDate, decisionDate, handledBy, requester, internalId, adjourned, firstRecipient

**Analysis of current publicly available publications**

Webpages from the Dutch government, municipalities and provinces often already contain a number of publications with documents. By analyzing these publications, together with their decision documents and inventory lists, already used attributes and potential new attributes can be listed. This section will discuss three ways of identifying attributes: by analyzing the publication webpages, the decision documents and inventory lists.

Publication pages

The Dutch government usually provides a title, description and responsible body on the publication page. For documents, the title, format, number of pages and file size are provided. Also, the date is provided. However, this date is always the decision date, and not the date of the creation of the individual documents. By inspecting and downloading the individual documents, a file name and extension can be retrieved. The title or filename of a file often describes the type of document, for example a decision document or an annex. All found attributes but the number of pages of a document are already defined by Dublin Core, FTM or the guideline by OpenState.

Decision documents

Decision documents contain the reaction of the handling body to an information request. In the case of an active disclosure, there is also a decision document explaining the released documents. A decision document often contains the file date of the request and the response date. Furthermore, the decision contains information about the possible adjournment of the response. An overall valuation of the response is given, for example public, partially public or not public. The decisions mentions the

requester, however this information is mostly hidden based on privacy reasons. Again, mostly already identified attributes are found. The valuation of the publication is a newly added attribute.

Inventory lists

Not all publications contain an inventory list. This type of document lists all the released documents, together with some useful information about the documents. From 392 readable inventory lists the table inside the pdf document is extracted using the Python module 'pdfplumber'. Table 1 shows a frequency table of the ten most occurring column headers. The names of the headers are generalized, as for example 'nummer, 'nr' and 'documentnr' all mean the same. This also indicates that the reported count is the minimum count, as not all the tables were perfectly readable. The column headers are mostly in Dutch, but are translated in the table. Also, an attribute name is given to the column. By analyzing the inventory lists, five new document attributes are identified.

*Table 1. Frequency table of the ten most occurring column headers of 392 inventory list tables*

| **Column name (Dutch)** | **Column name (English)** | **Attribute name** | **Count** |
|---|---|---|---|
| Nummer | Number | identifier | 295 |
| Documentnaam | Document name | title | 278 |
| Weigeringsgrond | Ground of refusal | groundsOfRefusal | 268 |
| Beoordeling | Valuation | valuation | 260 |
| Afzender | Originator | originator | 204 |
| Datum | Date | date | 203 |
| Ontvanger | Recipient | recipient | 198 |
| Soort | Type | annexType | 82 |
| Tijd | Time | date (generalized) | 42 |
| Vindplaats reeds openbaar | Location of already public document | alreadyPublicLocation | 28 |

**Publication:** numberPages, valuation

**Document:** groundsOfRefusal, valuation, originator, recipient, alreadyPublicLocation

**Extractable metadata**

To improve monitoring research on information publication, a few easily extracted attributes can be added. From the document collection, an attribute on number of documents can be added to quickly determine the size of the publications. Also, the documents itself can be enriched with useful attributes, such as the number of pages, the number of pages with text on it, the total number of words and characters, if the documents is a scan and requires Optical Character Recognition (OCR) in order to read the document. These attributes stimulate researchability and transparency on the quality of the publications. While the information gain would not outweigh the amount of work to extract these attributes per document or publication, these values can be filled by analyzers of the data. Next to that, an easy software tool may be able to automatically extract and fill these fields by providing the document itself. The automatic production of desirable JSON documents, with the extraction of useful additional metadata, will be covered extensively in the next chapter.

**Publication:** numberDocuments

**Document:** numberTextPages, numberWords, numberCharacters, isScan

**Uniform publications**

Using the same attributes on a national, or even bigger scale is the first step in improving the coordination and cooperability of governmental bodies. However, uniform agreements about the attribute's values have to be in place to prevent ambiguity. Three types of restrictions on the values can help in achieving this.

To avoid ambiguity, some attributes should only contain values in accordance with predefined standards. Dates can be represented in many ways. ISO 8601 is a unambiguous calendar format that is internationally understood (International Organization for Standardization, 2019). All dates should be conforming to this standard, representing dates in the format 'YYYY-MM-DD'. Similarly, ISO 639-3 (International Organization for Standardization, 2007) provided three-letter language code elements. Language values should follow this standard, writing languages in the form 'eng' or 'nld'. The mime type attribute already indicates the use of mime types, or media types as specified in IETF RFC 6838 (Internet Engineering Task Force, 2013). As most documents are PDF documents, mostly the mime type 'application/pdf' will be used. Each attribute should also be restricted to one or more valid JSON data types. These consist of strings, numbers, objects, arrays, booleans and null values. The validation of the use of the correct standards and data types will be covered in the next chapter on the automatic production of JSON documents.

To further constrain attribute values, some of the attributes should be a constrained choice. For publications, the type of publication should be a choice of two possible values, 'request' or 'active disclosure'. Both publications and documents have the 'valuation' attribute. This is a constrained choice of 'public', 'partially public', 'not public' or 'already public'. The document type is a choice of either 'request', 'decision', 'inventory list' or 'released document'.

To pressure the creators of the metadata, a few attributes are marked as required. These are attributes that should always be present and are always attributes that already are present, but hard to extract for a computer. For publications, the following attributes are required: identifier, title, type, handledBy, decisionDate and valuation. If the publication is based on a request, the fileDate and adjourned attributes are also required. For documents, the required attributes are: identifier, title, date, fileName, fileExtension, mimeType and documentType. If the document is a released document, also annexType and valuation are required. If not completely public, the groundsOfRefusal are required and when the valuation is 'already public', the alreadyPublicLocation shall be provided.

*Table 2. Value constraints on the defined attributes*

| **Object – attribute** | **Constraint** |
|---|---|
| Publication – fileDate | ISO 8601 date format |
| Publication – decisionDate | ISO 8601 date format |
| Publication – type | Constrained choice ('request', 'active disclosure') |
| Publication – valuation | Constrained choice ('public', 'partially public', 'not public', 'already public') |
| Document - date | ISO 8601 date format |
| Document - downloadDate | ISO 8601 date format |
| Document – language | ISO 639-3 language format |
| Document – mimeType | IETF RFC 6838 media type format |
| Document – documentType | Constrained choice ('request', 'decision', 'inventory list', 'released document') |
| Document - valuation | Constrained choice ('public', 'partially public', 'not public', 'already public') |

Buneman, P. (1997, May). Semistructured data. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (pp. 117-121).

Crockford, D. (2006). JSON: The fat-free alternative to XML. *http://www.json.org/xml.html*.

Follow the Money. (2021). *followthemoney documentation*. Geraadpleegd op 13 april 2022, van https://followthemoney.readthedocs.io/en/latest/entity.html

Instituut Maatschappelijke Innovatie & Open State Foundation. (2022). *Ondraaglijk traag – Analyse afhandeling Wob-verzoeken*.

International Organization for Standardization. (2019). DATE AND TIME FORMAT. (ISO Standard No. 8601). Retrieved from https://www.iso.org/iso-8601-date-and-time-format.html

International Organization for Standardization. (2019). Information and documentation — The Dublin Core metadata element set — Part 2: DCMI Properties and classes. (ISO Standard No. 15836-2). Retrieved from https://www.iso.org/standard/71341.html

Nurseitov, N., Paulson, M., Reynolds, R., & Izurieta, C. (2009). Comparison of JSON and XML data interchange formats: a case study. *Caine*, *9*, 157-162.

| Attribute | Dublin Core (DC) | FTM | OpenWob | Woo | Research |
|---|---|---|---|---|---|
| identifier | x | | x | | |
| title | x | x | x | x | |
| subject (topic, theme) | x | x | x | x | |
| subjectId | | | x (based on) | | |
| description | x | x | x | x | |
| requestDate | x (date) | x (date) | x | x | |
| decisionDate | x (date) | x (date) | x | x | |
| type | x | | | x | |
| sourceUrl | | x | | | |
| retrievedAt | | x | | | |
| handledBy | | | x | x | |
| requester | | | x | x | |
| internalId | | | x | | |
| geographical attributes | | | x | | |
| adjourned | | | x | x | |
| firstRecipient | | | x | | |
| valuation | | | | x | |
| numberDocuments | | | | x | x |

| Attribute | Dublin Core (DC) | FTM | OpenWob | Woo | Research |
|---|---|---|---|---|---|
| identifier | x | | | x | |
| title | x | x | | x | |
| description | x | x | | | |
| date | x | x | | x | |
| fileExtension | x (format) | x | | x | |
| mimeType | | x | | | |
| documentType | x (type) | | | x | |
| language | x | x | | | |
| sourceUrl | | x | | | |
| fileSize | | x | | x | |
| fileName | | x | | x | |
| bodyText | | x | | | |
| annexType | x (type) | | | x | |
| valuation | | | | x | |
| groundsOfRefusal | | | | x | |
| alreadyPublicLocation | | | | x | |
| originator | | | | x | |
| recipient | | | | x | |
| numberPages | | | | | x |
| numberWords | | | | | x |
| isScan | | | | | x |
| numberTextPages | | | | | x |
| numberCharacters | | | | | x |