Creating a Colonial Architecture Pipeline

Gossa Lô

Department of Computer Science, the Network Institute,
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
{a.g.lo}@vu.nl

Abstract. European colonialism has left its marks in many countries around the world. Traces of this heritage can still be found today in the infrastructure, planning and architecture in former colonies. Documents and images have since been collected and are stored in the online Colonial Architecture repository. This paper investigates a possible contribution of computational linguistic and Linked Data techniques on the annotation and formalization of these documents, by means of a Python pipeline. We finally validate its usefulness by testing the pipeline on a subset of the Colonial Architecture corpus.

1 Introduction

European colonialism is an important component of European history and refers to the exploration, conquering and exploitation of certain areas of the world. This expansion occurred in the 15th century, also known as the Age of Discovery, and can be divided into three waves, occurring in chronological order of colonization.

The main countries targeted are located in the Americas & the Caribbean (first wave), Asia (second wave) and Africa (third wave). The European countries that played a role in colonization were Portugal, Spain, the Netherlands, France and Britain. After World War II, one by one the areas were decolonized, starting with Asia and finally Africa in the 1950s [1]. The colonized countries were heavily impacted and some of the economic, political and social impairments are still noticeable today.

The infrastructure, planning and architecture were also affected under European ruling and are seen as important elements of the former colonies' heritage. Since the 1980s, the importance of preservation of documents and images of this heritage has been acknowledged and stimulated. The Colonial Architecture repository[1] contains a significant amount of data on European colonial architecture and town planning. The focus is on the creation of civic architecture in European style that can be found in former colonies. The documents are mostly unstructured and there is almost no meta data available.

This paper presents a proof of concept aimed at annotating relevant entities and terms that are used in the documents. This is accomplished by building a pipeline in Python that annotates and links these entities to knowledge existing

[1] colonialarchitecture.eu/, accessed 2017-12-18

on the Web, by means of Optical Character Recognition and Named Entity Recognition (NER). Finally, we validate the usefulness and functioning of the pipeline by testing it on a subset of the documents that are part of the collection.

2 Methods

This section describes a step-by-step walk-through of the development process. The aim of the project is to enable those working with the data to easily extract relevant entities and link them to Getty Vocabularies. These vocabularies are used because they include terminology for architecture and conservation, which is similar to the topics of the documents we investigate. Furthermore, the vocabularies are available as Linked Open Data, which can be used to publish and reuse the data on the Web. Fig. 1 gives an overview of the steps that were taken to create the pipeline. The section is divided according to the order shown in this figure.

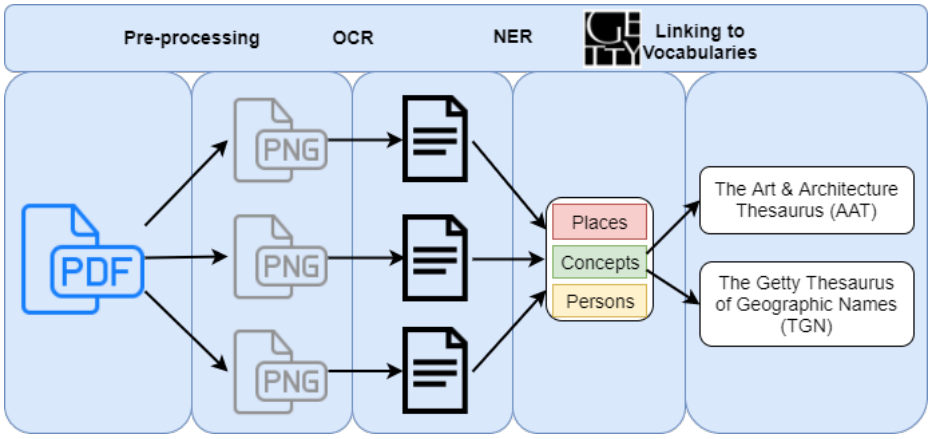

Fig. 1. Pipeline step-by-step

2.1 Step 0: Data selection

The documents used for this project were derived from a subset of the Colonial Architecture corpus. This corpus contains both journal issues and books. Besides the title of the author, the publisher and the abstract, there is not a lot of metadata available. For this project we have chosen to include some of the journal issues of the subset. These journal issues are small in size (ca. 20 pages) and are therefore easy to process. Moreover, the information on architecture and colonial artifacts that the documents contain forms a suitable first subset before extensive testing is done. The books are left out of consideration because of their large sizes. All the files used are in PDF format and are scanned. The pipeline exists of one Python script and can be found in appendix B.

2.2 Step 1: Data pre-processing

As shown in fig. 1, the first step is to pre-process the PDF file in order to enable easier extraction of the text in the next stadium. ImageMagick software offers several features to convert and edit files of different formats. The format conversion feature was used to convert the PDFs into multiple PNG files, each covering one page. To make the text more readable, the density, scaling and level are adjusted, resulting in a set of gray-scale PNG files.

2.3 Step 2: Optical Character Recognition

The next step is to use the PYOCR tool wrapper[2] to extract the text from the PNG files. This tool uses Tesseract[3], an Open Source OCR Engine. Despite the pre-processing, the OCR does not function perfectly. Some of the mistakes made by the tool are caused by rotation and the old condition of some of the documents. Taking this into account, the tool works properly and we were able to extract most of the relevant words.

2.4 Step 3: Named Entity Recognition

Now that the text is available, the entities that are deemed relevant can be extracted. Several open source NLP tools are available online. Unfortunately, most of them focus on English, Spanish and French, and only a few are available for Dutch.

Frog[4] is an NLP suite for Dutch, based on Timbl, the Tilburg memory-based learning software package. Python-frog[5] is a Python binding to Frog, and was used to parse the text. The installation required a number of dependencies, which were difficult to integrate in Windows. To solve this, Linux running on a virtual machine was used instead.

The text is tokenized and each token is processed as follows:

```
{'index': u'14', 'morph': u'[Borneo]', 'dep': u'obj1', 'text': u'Borneo',
'pos': u'SPEC(deeleigen)', 'chunker': u'I-NP', 'lemma': u'Borneo',
'posprob': 1.0, 'depindex': u'11', 'ner': u'B-LOC'}
```

The pieces of information relevant for this project are the part-of-speech tag ('pos'), the confidence in the POS tag ('posprob') and, most importantly, the named entity type ('ner'). Since we are only interested in the named entities, we filter out the rest of the words in the text. Frog distinguishes six types of entities, namely:

[2] github.com/openpaperwork/pyocr, accessed 2017-12-22
[3] github.com/tesseract-ocr/tesseract, accessed 2017-12-22
[4] languagemachines.github.io/frog/, accessed 2017-12-22
[5] github.com/proycon/python-frog, accessed 2017-12-22

1. person (PER)
2. organization (ORG)
3. location (LOC)
4. product (PRO)
5. event (EVE)
6. miscellaneous (MISC)

Our focus is solely on persons, locations and architectural terms. These are terms of which many appear in the Getty Vocabularies, which enables us to link them to the Web. The latter group consist solely of nouns and we adjusted the POS tag confidence probability to a minimum of 0.95. The confidence probability is a numeric value that indicates the extent to which the recognizer is confident that a word is correctly tagged. Some values were tested, and 0.95 seemed to be the best trade-off between precision and recall, as this filters out most of the irrelevant nouns. Subsequently, the most frequent occurring words, with the exception of stop words, are counted and used.

2.5 Step 4: Linking to Getty Vocabularies

The final step is to link these concepts to entities from the Getty Vocabularies. The two vocabularies we link to, as shown in fig. 1, are the Art & Architecture Thesaurus (AAT) and the Getty Thesaurus of Geographic Names (TGN). The former refers to architectural terms, the latter refers to the location entities.

With SPARQL queries we exported two relevant subsets of these vocabularies in CSV format for offline use. The TGN and AAT entities that correspond to entities in the text are put in two different dictionaries, with the values being the Getty URIs.

The query below refers to TGN and contains 628074 names and URIs referring to locations all over the world, such as "Java" and "China". The US based locations were filtered out, since they contain a lot of duplicates with locations in Asia and Europe, which are deemed more relevant given the context of this project.

```
SELECT ?s ?literal WHERE {
  ?s rdf:type ?o.
  ?s xl:prefLabel ?label.
  ?label xl:literalForm ?literal.
  ?s gvp:parentString ?parent.
  FILTER(regex(str(?o),'PlaceConcept')).
  FILTER NOT EXISTS {
    FILTER(regex(str(?parent),'United States')).
    }
}
```

The following query gives an overview of the AAT concepts and their URIs. This file contains 18086 concepts such as "landvoertuigen" and "drukkerijen".

```
SELECT * WHERE {
  ?concept xl:prefLabel ?prefl.
  ?prefl dcterms:language aat:300388256.
  ?prefl xl:literalForm ?litf}
```

The Union List of Artist Names vocabulary, which is a part of the Getty vocabularies, contains meta data about historical figures such as artists, architects and patrons. However, we decided not to refer to this vocabulary, since the NER tool we used does not yet sufficiently extract the whole names of the persons mentioned in our data. Instead, the names that were extracted by Frog were often location names. The output of the pipeline script consists of a list of names that were obtained by Frog. This list was then compared with the list of location names to filter out the duplicate names. In addition to the person names and the TGN and AAT entities, the output contains a list of most frequent occurring words. This is made clear in the example used in the next section.

3 Validation

The pipeline is tested in two stages. First, the functioning of the OCR is evaluated by comparing the original text with the OCR-text. The second stage will consist of testing whole pipeline by comparing whether the script generates a functioning output.

3.1 Validating the OCR

The OCR is validated by comparing the processed text to its original variant. It should be mentioned that some of the parts were very badly processed and have therefore been left out of the OCR validation. These parts for instance include tables, or italicized and rotated text surrounding figures. For instance, the following is a snipped of what the text around an image in the document "De Ingenieur in Indonesië" (see footnote 8) looks like.

```
_ . 7
y_/emlx
dania:
          d='y* _
 El /" (uw dav_
:z*/'EI

Z l
```

This document has 22 of these difficult to interpret parts, which is a small number, considering the high amount of mathematical functions and images and the relatively large size of the document. The amount of parts that are difficult to interpret depends on the state of the document, its size, and the amount of figures/functions.

Instead our validation focuses on paragraphs that are readable but may still contain mistakes. These texts were randomly selected over three different documents that are also used to validate the NER. The following table shows the percentage of error for each of the five paragraphs evaluated, for three different PDF inputs. By dividing the number of misinterpreted words by the number of words in the paragraph, and multiplying that number by one hundred, the value which we call the percentage of error is calculated.

Table 1. Validation of the OCR in percentage of error per paragraph for five documents.

	De Priokweg	Algemeene Koloniale en Internationale Tentoonstelling Semarang	De Ingenieur in Indonesië	Average
I	0%	1.27%	2.90%	1.39%
II	1.75%	4.35%	2.35%	2.82%
III	1.23%	5.36%	1.52%	2.70%
IV	6.82%	2.17%	0.5%	3.16%
V	18.92%	2.29%	0%	7.07%
Average	5.74%	3.09%	1.45%	3.42%

As we can see, there is quite a difference in error rate per paragraph and per text. The highest average of error percentage in a text is 5.74% and the overall average is 3.42%, which is considered a low number given the condition of some of the document. However, as mentioned before, the parts that are difficult to read were left out of the equation, and would have otherwise increased this number significantly.

3.2 Validating the NER

The functioning of the overall pipeline is validated by assessing whether the pipeline generates an output and if this can be considered relevant given the input. The time limit of the project unfortunately prevented us from being able to let the pipeline be validated by experts.

The NER is validated by examining the output of a subset of five different documents that were derived from the Colonial Architecture repository. The documents are quite diverse with sizes ranging from 16 to 32 pages, differing in subject, years of publication and conditions. Of the following journal issues, only the first is shown in this section. The other four outputs can be found in appendix A.

- De Priokweg[6] (22 pages)
- Algemeene Koloniale en Internationale Tentoonstelling Semarang[7] (22 pages)
- De ingenieur in Indonesië (year 1948, volume 001, issue 002)[8] (26 pages)
- Nederlandsch-Indië, oud en nieuw (year 1931, volume 016, issue 004)[9] (32 pages)
- 7de Internationale Wegencongres No. 2[10] (16 pages)

[6] colonialarchitecture.eu/obj?sq=id%3Auuid%3Afccda65c-bb24-4923-a3cd-cd6c7bfa2236, accessed 2017-12-27

[7] colonialarchitecture.eu/obj?sq=id%3Auuid%3Ab31bb870-71fb-4ff2-8805-e307c9865013, accessed 2017-12-27

[8] colonialarchitecture.eu/obj?sq=id%3Auuid%3Acf4442b4-a25f-4e98-abab-81be18ba6c97, accessed 2017-12-27

[9] colonialarchitecture.eu/obj?sq=id%3Auuid%3A957b0f81-96ee-43db-81e7-ee1096eedc8b, accessed 2017-12-27

[10] colonialarchitecture.eu/obj?sq=id%3Auuid%3A76bde3c7-d576-4226-8799-eb74719f1332, accessed 2017-12-27

The first journal issue is about "De Priokweg" and is a publication from the "Nederlandsch-Indische Wegenvereeniging". The content is a presentation of the condition of the road, technical construction, costs and materials used. In addition, it contains some pictures that show the process of building the road. We would want our pipeline to tell us more about the people involved, the locations discussed, and the entities that relate to road construction. The fragment below is the output of the pipeline for this document. No persons were detected, and most of the TGN sites are located in Indonesia, Asia. This is of course correct, since the Priok road connects two parts of Indonesia, namely Batavia and Tandjong Priok. Most of the AAT entities refer to the techniques and materials used, as do the frequently occurring words.

```
Persons:  []
TGN locations:  {'Kali': 'http://vocab.getty.edu/tgn/8476952',
'Borneo': 'http://vocab.getty.edu/tgn/7015963',
'Philadelphia': 'http://vocab.getty.edu/tgn/7119338',
'Batavia': 'http://vocab.getty.edu/tgn/1019208'}
AAT entities:  {'treinen': 'http://vocab.getty.edu/aat/300212738',
'administratie': 'http://vocab.getty.edu/aat/300027425',
'naden': 'http://vocab.getty.edu/aat/300228472',
'duikers': 'http://vocab.getty.edu/aat/300006116',
'cijfers': 'http://vocab.getty.edu/aat/300194307',
'ketels': 'http://vocab.getty.edu/aat/300195349',
'zijkanten': 'http://vocab.getty.edu/aat/300190706',
'strijkijzers': 'http://vocab.getty.edu/aat/300185311',
u'steenslag': 'http://vocab.getty.edu/aat/300011681
'olie': 'http://vocab.getty.edu/aat/300014254'}
Frequently occurring words::  {'grade': 3, 'zeef-analyse': 3,
'materiaal': 5, 'Priokweg': 4, 'deel': 3, 'bezwaar': 4,
'wals': 3, 'breedte': 5, 'materialen': 4, 'uitvoering': 4,
'vereischte': 3, 'samenstelling': 3, 'asphaltlagen': 5,
'duikers': 3, 'zand': 12, 'steenslaglaag': 3, 'machine': 4,
'onderlaag': 6, 'kwaliteit': 3, 'asphaltmortel': 4,
'mengsel': 8, 'Tandjong_Priok': 7, 'minerale': 3, 'goede':
3, 'slijtlaag': 13, 'droogtrommel': 4, 'uitgevoerde': 3,
'gevaar': 3, 'verharding': 5, 'hoeveelheid': 3,
'temperatuur': 5, 'weghelft': 3, 'halve': 3, 'bindlaag': 13,
'Augustus': 3, 'asphaltmortellaag': 3, 'zijkanten': 3,
'verkeer': 7, 'gedeelte': 6}
```

The frequently occurring words are those with more than three letters occurring at least three times in the document. These numbers were used to filter out stop words. Text mining techniques are generally better at recognizing and eliminating these words, but a Python text mining library for Dutch words is yet difficult to find. The fairly simple technique we used for this project seems to work properly. However, there is a positive correlation between the amount of pages and the the number of frequent occurring terms. Table 2 shows that the number of concepts is too high in the largest documents, with the highest number of concepts (i.e. AAT conceps and frequently occurring words) being 205. Since we are not interested in words such as "aantal" and "hoge", the parameters should be tuned better. An option would be to divide the highest number of occurrence by two and to remove the ones below this number. In the example of Appendix A.2. this would result in a list of words that occur at least 14 times and would reduce this amount from 155 to three. However, this would be at the cost of relevant words such as "civiel-ingenieur" and "delfstoffen".

The NER works properly for the location names and the AAT entities. Most of the locations in the five examples are in Asia, which corresponds to our expectations and limited knowledge of the content of the documents. Table 2 shows that out of the five documents, only three contain person names. One is a location instead of a person name and the others only consist of first names. These location names were not filtered out by the TGN location recognition. This problem could be solved by extracting the whole names with a more advanced entity recognizer.

As table 2 shows, the number of places is highest in the document on the Colonial and International exhibition. This corresponds to our understanding of the content, which is an overview of artifacts discovered in different regions and countries in Asia. The document about the Priokweg contains only four location names. This was expected, as its content is on a local area in Indonesia, and thus solely contains local names.

Linking to the Getty vocabularies works as intended. The TGN and AAT URIs refer to the right pages, which gives us more information on the entity in question. By including both the AAT referrals as context and the isolated words in the "frequently occurring words", we believe that the output represents a more or less complete overview of the outline of the document. For instance, the document used in Appendix A.3 is on ornaments and artifacts from the Dutch East Indies. The AAT entities and frequently occurring words in the output show many entities related to this, such as "beeldjes", "galerijen", "sieraden" and "lotusbloem". This indicates that, despite the limitations of the tools used, the pipeline seems to be working properly.

Table 2. Number of entities detected per document

Documents	Nr. of pages	Nr. of persons	Nr. of places	Nr. of concepts	Total nr. of concepts
De Priokweg	22	0	4	48	52
Algemeene Koloniale en Internationale Tentoonstelling Semarang	22	1	18	59	78
De Ingenieur in Indonesië	26	3	12	205	220
Nederlandsch-Indië, oud en nieuw	32	5	11	118	134
7de Internationale Wegencongres No. 2	16	0	4	46	50

4 Conclusion

European colonialism has heavily affected the infrastructure, planning and architecture in many former colonies. A collection of the documentation on these topics is available in the Colonial Architecture repository online. Numerous journals and books have been preserved, but contain little to no meta data or further annotation regarding the contents. This paper introduced a pipeline aimed to

extract and link relevant entities (e.g. persons, locations AAT concepts) to vocabularies online, by using computational linguistic and Linked Data techniques.

This has resulted in a Python script and is validated by testing on a subset of the corpus. The output is an overview of the content of the documents in the form of entities. Aside from the person names, the results give an interesting insight into the content of the documents. We have shown that the pipeline functions to an extent that it extracts the most relevant terms and concepts and is indeed able to link them to the Web. This facilitates the process of retrieving necessary background information on the documents in question, which is needed for deeper understanding of the content.

Further steps include the use of more advanced OCR and NER tools, adjusted to the condition and content of the documents. Furthermore, the extraction of entities should be narrowed down to present only the most relevant ones. Additionally, extensive testing by experts would give us a more in-depth view of the validity of the pipeline.

Despite these points of improvement, this proof of concept has demonstrated that computational linguistics can certainly contribute to enriching the documents on European Colonialism.

References

1. Gallaher, C., Dahlman, C.T., Gilmartin, M., Mountz, A., Shirlow, P.: Key concepts in political geography. Sage (2009)

A
Output pipeline

A.1 Algemeene Koloniale en Internationale Tentoonstelling Semarang

```
Persons: ['Probolinggo']
TGN Locations: {'Ceram': 'http://vocab.getty.edu/tgn/1009102', 'Java':
'http://vocab.getty.edu/tgn/7003695', 'Nederland':
'http://vocab.getty.edu/tgn/7016845', 'Bali':
'http://vocab.getty.edu/tgn/8083869', 'Borneo':
'http://vocab.getty.edu/tgn/7015963', 'Kediri':
'http://vocab.getty.edu/tgn/7016704', 'Madama':
'http://vocab.getty.edu/tgn/8476044', 'Semarang':
'http://vocab.getty.edu/tgn/1078528', 'Batavia':
'http://vocab.getty.edu/tgn/1019208', 'Levant':
'http://vocab.getty.edu/tgn/7001519', 'Makassar':
'http://vocab.getty.edu/tgn/1078833', 'Sumatra':
'http://vocab.getty.edu/tgn/7016484', 'China':
'http://vocab.getty.edu/tgn/7994975', 'Australi\xc3\xab':
'http://vocab.getty.edu/tgn/7000490', 'Pekalongan':
'http://vocab.getty.edu/tgn/1078336', 'Palembang':
'http://vocab.getty.edu/tgn/7016481', 'Japan':
'http://vocab.getty.edu/tgn/7030366', 'Medan':
'http://vocab.getty.edu/tgn/1078112'}
AAT entities: {'gebouwen': 'http://vocab.getty.edu/aat/300004792',
'pijlen': 'http://vocab.getty.edu/aat/300036976',
'wapens': 'http://vocab.getty.edu/aat/300036926',
'hoeken': 'http://vocab.getty.edu/aat/300266471',
'cijfers': 'http://vocab.getty.edu/aat/300194307',
'eilanden': 'http://vocab.getty.edu/aat/300008791',
'heuvels': 'http://vocab.getty.edu/aat/300008777',
'inzendingen': 'http://vocab.getty.edu/aat/300265018',
'uniformen': 'http://vocab.getty.edu/aat/300212393',
'rijst': 'http://vocab.getty.edu/aat/300343827',
'poorten': 'http://vocab.getty.edu/aat/300069189',
'suiker': 'http://vocab.getty.edu/aat/300183931',
'grafieken': 'http://vocab.getty.edu/aat/300027020',
'monsters': 'http://vocab.getty.edu/aat/300028875',
'drukkerijen': 'http://vocab.getty.edu/aat/300006247'}
Frequently occurring words:: [('tentoonstelling', 23), ('paviljoen', 13),
('groep', 10), ('inlandsche', 7), ('nijverheid', 7), ('gebied', 6),
('inheemsche', 6), ('terrein', 6), ('verkeer', 6), ('gebouw', 5),
('inzending', 5), ('kolonie', 5), ('aantal', 5), ('afdeeling', 5),
('gebouwen', 5), ('tuin', 4), ('verzameling', 4), ('inlandsch', 4),
('uitvoer', 3), ('groepen', 3), ('stijl', 3), ('overzicht', 3),
('eerste', 3), ('perceelen', 3), ('jaren', 3), ('Gouvernement', 3),
('oppervlakte', 3), ('1901', 3), ('meest', 3), ('handel', 3),
('beide', 3), ('inzendingen', 3), ('handwerk', 3), ('indruk', 3),
('voorstellingen', 3), ('Archipel', 3), ('onderwijs', 3),
('afdeelingen', 3), ('plan', 3), ('stelling', 3), ('invoer', 3),
('restaurant', 3), ('paviljoens', 3), ('comite, 3)]
```

A.2 De ingenieur in Indonesië

```
Persons: ['Reith', 'Medan', 'Brandon']
TGN Locations: {'Sumatra': 'http://vocab.getty.edu/tgn/7016484',
'Balei': 'http://vocab.getty.edu/tgn/8345029',
'Nederland': 'http://vocab.getty.edu/tgn/7016845',
'Banka': 'http://vocab.getty.edu/tgn/8469680',
'Belawan': 'http://vocab.getty.edu/tgn/1077482',
'Batavia': 'http://vocab.getty.edu/tgn/1019208',
'Rijn': 'http://vocab.getty.edu/tgn/7032606',
'Second': 'http://vocab.getty.edu/tgn/7787554',
'Wonogiri': 'http://vocab.getty.edu/tgn/1078891',
'Japan': 'http://vocab.getty.edu/tgn/7030366',
'Palembang': 'http://vocab.getty.edu/tgn/7016481',
'Bergen': 'http://vocab.getty.edu/tgn/7211714'}
AAT entities: {'regelafstand': 'http://vocab.getty.edu/aat/300216347',
u'transis': 'http://vocab.getty.edu/aat/300252108',
'zones': 'http://vocab.getty.edu/aat/300000853',
'meters': 'http://vocab.getty.edu/aat/300198989',
'woonhuizen': 'http://vocab.getty.edu/aat/300005433',
'takken': 'http://vocab.getty.edu/aat/300379798',
'publicaties': 'http://vocab.getty.edu/aat/300111999',
'gangen': 'http://vocab.getty.edu/aat/300004294',
'putten': 'http://vocab.getty.edu/aat/300006207',
'kust': 'http://vocab.getty.edu/aat/300008733',
'universiteiten': 'http://vocab.getty.edu/aat/300006617',
'wijzigingen': 'http://vocab.getty.edu/aat/300055457',
'kool': 'http://vocab.getty.edu/aat/300015149',
'bouwsteen': 'http://vocab.getty.edu/aat/300011700',
'verdiepingen': 'http://vocab.getty.edu/aat/300002667',
'tenten': 'http://vocab.getty.edu/aat/300005694',
'artikelen': 'http://vocab.getty.edu/aat/300048715',
'bouwstenen': 'http://vocab.getty.edu/aat/300211303',
u'he': 'http://vocab.getty.edu/aat/300265824',
'gebouwen': 'http://vocab.getty.edu/aat/300004792',
'richtlijnen': 'http://vocab.getty.edu/aat/300027029',
'banden': 'http://vocab.getty.edu/aat/300266412',
'goudmijnen': 'http://vocab.getty.edu/aat/300132656',
'kamers': 'http://vocab.getty.edu/aat/300004044',
'eilanden': 'http://vocab.getty.edu/aat/300008791',
'stellen': 'http://vocab.getty.edu/aat/300133146',
'zijkanten': 'http://vocab.getty.edu/aat/300190706',
'vrucht': 'http://vocab.getty.edu/aat/300011868',
'eisen': 'http://vocab.getty.edu/aat/300027622',
'atomen': 'http://vocab.getty.edu/aat/300264242',
u'olie': 'http://vocab.getty.edu/aat/300014254'}
Frequently occurring words:: [('water', 28), ('grote', 19), ('tijd', 16),
('aantal', 13), ('jaar', 13), ('gedeelte', 13), ('lijn', 12),
('materiaal', 12), ('taak', 12), ('ligging', 12), ('nieuwe', 12),
('geval', 11), ('wijze', 10), ('oorlog', 9), ('1948', 9),
('methode', 9), ('plaats', 9), ('eerste', 8), ('examen', 8),
('hoogleraar', 8), ('stroom', 8), ('diameter', 8), ('weerstand', 8),
('debiet', 8), ('gaten', 7), ('oppervlak', 7), ('hoge', 7),
('Regering', 7), ('opleiding', 7), ('deel', 7), ('afstand', 7),
('spanning', 7), ('hoogte', 7), ('medewerking', 7), ('reservoir', 6),
('gebied', 6), ('jaren', 6), ('parabool', 6), ('brandstof', 6),
('ingenieur', 6), ('negatieve', 6), ('temperatuur', 6), ('Faculteit', 6),
('waterspiegel', 6), ('gewone', 6), ('leden', 6), ('positieve', 6),
('gevolg', 6), ('wereld', 6), ('inrichting', 5), ('twee', 5),
('waarde', 5), ('kracht', 5), ('snelheid', 5), ('bouw', 5),
('toekomst', 5), ('Ingenieur', 5), ('stijghoogte', 5), ('vraag', 5),
('kans', 5), ('doel', 5), ('gevolgen', 5), ('formule', 5),
('gehele', 5), ('ingenieurs', 5), ('grond', 5), ('basis', 5), ('druk', 5),
('Technische_Hogeschool', 5), ('ontwikkeling', 5),
('belasting', 5), ('huisvesting', 5), ('toestand', 5), ('plan', 5),
('land', 5), ('banden', 5), ('medewerkers', 5), ('belangrijk', 4),
('tweede', 4), ('afdelingen', 4), ('onderwijs', 4), ('overzicht', 4),
```

```
('stijgkolom', 4), ('artesische', 4), ('vele', 4), ('afdeling', 4),
('stoffen', 4), ('electroden', 4), ('collector', 4), ('ter_Oostkust', 4),
('scheikunde', 4), ('putten', 4), ('totale', 4), ('voordelen', 4),
('jaarverslag', 4), ('September', 4), ('technische', 4), ('lengte', 4),
('aanvang', 4), ('beeld', 4), ('arbeid', 4), ('belang', 4),
('type', 4), ('afleiding', 4), ('vloeibare', 4), ('fouten', 4),
('landen', 4), ('vraagstuk', 4), ('halfgeleider', 4), ('inzicht', 4),
('gelegenheid', 4), ('studenten', 4), ('behoefte', 4), ('buis', 4),
('transistor', 4), ('propaedeutisch', 4), ('staf', 4), ('resultaten', 4),
('groot', 4), ('beschouwing', 4), ('andere', 3), ('Ter_Oostkust', 3),
('geringe', 3), ('nadeel', 3), ('veranderingen', 3), ('keuze', 3),
('onderzoek', 3), ('maanden', 3), ('vrije', 3), ('impedantie', 3),
('energie', 3), ('bezwaren', 3), ('productie', 3), ('speciale', 3),
('mogelijk', 3), ('omstandigheden', 3), ('behoeften', 3), ('laag', 3),
('aard', 3), ('hoogleraren', 3), ('vraagstukken', 3), ('delfstoffen', 3),
('oude', 3), ('personeel', 3), ('mechanisme', 3), ('uitbreiding', 3),
('werking', 3), ('instelling', 3), ('maatregelen', 3), ('golflengte', 3),
('korte', 3), ('stand', 3), ('ding', 3), ('ring', 3),
('stijgkolorn', 3), ('omvang', 3), ('mogelijkheden', 3), ('zich', 3),
('laatste', 3), ('mededeling', 3), ('tekst', 3), ('deling', 3),
('reserves', 3), ('blad', 3), ('variatie', 3), ('prijs', 3),
('huidige', 3), ('gegevens', 3), ('middelen', 3), ('ruimte', 3),
('civiel-ingenieur', 3), ('beroep', 3), ('nummers', 3), ('plannen', 3)]
```

A.3 Nederlandsch-Indië, oud en nieuw

```
Persons: ['Gabriel', 'Tegel', 'Pawan', 'Meru', 'Kema']
TGN Locations: {'Java': 'http://vocab.getty.edu/tgn/7003695',
'Bali': 'http://vocab.getty.edu/tgn/8083869',
'Tegal': 'http://vocab.getty.edu/tgn/1078745',
'India': 'http://vocab.getty.edu/tgn/7000208',
'Minahasa': 'http://vocab.getty.edu/tgn/1012871',
'Tual': 'http://vocab.getty.edu/tgn/1078819',
'China': 'http://vocab.getty.edu/tgn/7994975',
'Kunde': 'http://vocab.getty.edu/tgn/7955958',
'Buddha': 'http://vocab.getty.edu/tgn/7903442',
'Lombok': 'http://vocab.getty.edu/tgn/1078022',
'Tabanan': 'http://vocab.getty.edu/tgn/1078669'}
AAT entities: {'rivieren': 'http://vocab.getty.edu/aat/300008707',
'vrucht': 'http://vocab.getty.edu/aat/300011868',
'beeldjes': 'http://vocab.getty.edu/aat/300047455',
'banden': 'http://vocab.getty.edu/aat/300266412',
'heuveltjes': 'http://vocab.getty.edu/aat/300008781',
'galerijen': 'http://vocab.getty.edu/aat/300004031',
"stoepa's": 'http://vocab.getty.edu/aat/300007576',
'eilanden': 'http://vocab.getty.edu/aat/300008791',
'hoofdstukken': 'http://vocab.getty.edu/aat/300311699',
'prototypes': 'http://vocab.getty.edu/aat/300166391',
'noot': 'http://vocab.getty.edu/aat/300011897',
'voetstukken': 'http://vocab.getty.edu/aat/300001656',
'bladzijden': 'http://vocab.getty.edu/aat/300194222',
'details': 'http://vocab.getty.edu/aat/300133138',
'sieraden': 'http://vocab.getty.edu/aat/300209286',
'heften': 'http://vocab.getty.edu/aat/300024926',
'voorschriften': 'http://vocab.getty.edu/aat/300027842',
'tempels': 'http://vocab.getty.edu/aat/300007595',
'artikelen': 'http://vocab.getty.edu/aat/300048715',
'ornamenten': 'http://vocab.getty.edu/aat/300266794'}
Frequently occurring words:: [('kris', 20), ('kwast', 13), ('water', 10),
('lotus', 9), ('vorm', 9), ('Boeddha', 8), ('huis', 8),('eilanden', 8),
('plaats', 8), ('ornament', 7), ('Lotus', 7), ('lotusbloem', 7),
('that', 7), ('eiland', 7), ('andere', 6), ('gebied', 6),('bloem', 6),
('from', 6), ('zaden', 6), ('jaren', 5), ('tijd', 5), ('bouwwerk', 5),
('naam', 5), ('nokversiering', 5), ('oude', 5), ('beteekenis', 5),
('type', 5), ('krissen', 5), ('nieuwe', 5), ('Boroboedoer', 5),
('dagobs', 4), ('voorstellingen', 4), ('The_Lotus', 4), ('vlakte', 4),
('Nieuwenkamp', 4), ('with', 4), ('monument', 4), ('aantal', 4),
('also', 4), ('hand', 4), ('heuvel', 4), ('hoofdeiland', 3),
('materiaal', 3),('wijze', 3), ('heften', 3), ('gouden', 3), ('kruik', 3),
('being', 3), ('avanen', 3), ('geheel', 3), ('bloei', 3), ('geen', 3),
('they', 3), ('overgang', 3), ('become', 3), ('opvatting', 3),
('vele', 3), ('idea', 3), ('Type', 3), ('verhoogingen', 3), ('what', 3),
('flower', 3), ('kegel', 3), ('Water', 3), ('Indische', 3),
('woord', 3), ('artikel', 3), ('studie', 3), ('both', 3), ('seeds', 3),
('beeld', 3), ('Maleisch', 3), ('geval', 3), ('eeuwen', 3),
('gelijkenis', 3), ('emblem', 3), ('symbool', 3), ('verschijnsel', 3),
('eilandje', 3), ('bestemming', 3), ('geheele', 3), ('nokbalk', 3),
('zeer', 3), ('kleine', 3), ('zaaddoos', 3), ('streken', 3),
('Lotus-flower', 3), ('oorspronkelijke', 3), ('beide', 3), ('same', 3),
('boom', 3), ('land', 3), ('soorten', 3), ('stolpen', 3), ('groot', 3),
('form', 3), ('hokjes', 3), ('munten', 3)]
```

A.4 7de Internationale Wegencongres No. 2

```
Persons:  []
TGN Locations:  {'Java': 'http://vocab.getty.edu/tgn/7003695',
'Semarang': 'http://vocab.getty.edu/tgn/1078528',
'Batavia': 'http://vocab.getty.edu/tgn/1019208',
'Bali': 'http://vocab.getty.edu/tgn/8083869'}
AAT entities:  {'hoofdwegen': 'http://vocab.getty.edu/aat/300008283',
'oppervlakken': 'http://vocab.getty.edu/aat/300190708',
u'steenslag': 'http://vocab.getty.edu/aat/300011681',
'publicaties': 'http://vocab.getty.edu/aat/300111999',
'banden': 'http://vocab.getty.edu/aat/300266412',
'cijfers': 'http://vocab.getty.edu/aat/300194307',
'eilanden': 'http://vocab.getty.edu/aat/300008791',
'bruggen': 'http://vocab.getty.edu/aat/300007836',
'wielen': 'http://vocab.getty.edu/aat/300024976',
'rijtuigen': 'http://vocab.getty.edu/aat/300185335',
'rivieren': 'http://vocab.getty.edu/aat/300008707',
'wegen': 'http://vocab.getty.edu/aat/300008217',
'proefstukken': 'http://vocab.getty.edu/aat/300178236',
'gemeenten': 'http://vocab.getty.edu/aat/300265612'}
Frequently occurring words::  [('wegen', 8), ('wijze', 6),
('oppervlakte', 6), ('stroefheid', 6), ('split', 5), ('auto', 5),
('praktijk', 5), ('invloed', 5), ('voegen', 5), ('verkeer', 5),
('bevolking', 4), ('afdekking', 4), ('bindmiddel', 4), ('wegennet', 4),
('steenen', 4), ('achterwielen', 4), ('millioen', 4), ('lengte', 4),
('oppervlak', 3), ('asphalt', 3), ('onderzoekingen', 3), ('meest', 3),
('vlakte', 3), ('hoofdwegen', 3), ('toestand', 3), ('behandeling', 3),
('verharding', 3), ('hoeveelheid', 3), ('banden', 3), ('wegdek', 3),
('eeuw', 3), ('motorverkeer', 3)]
```