Building a national Infrastructure to enable research with Dutch Digital Heritage Collections

Dr. Martijn Kleppe
Board Member Research & Discovery
KB, national library of the Netherlands

martijn.kleppe@kb.nl
@martijnkleppe
https://www.kb.nl/over-ons/experts/martijn-kleppe
Today

- What is the KB and what do we do with data for research?
- Case 1: FAIR Cultural Heritage data for research
- Case 2: Secure Data Analysis Environment
- Wrap up & some lessons
7 million items
120 kilometres publications
Full text (OCR) access to:

- 1,000,000 books (1486 – 2013)
- 1,700,000 newspapers (1618 – 1995)
- 10,000,000 periodical pages (1840 – 1940)
- 1,500,000 million ANP radio items (1937 – 1984)

https://www.delpher.nl/
www.delpher.nl

www.kb.nl/dataservices
Dataservices

= result of
more than 200 years of collecting
over 30 years of digitisation
10+ years of collecting born-digital publications

= machine readable, mostly textual
= structured or semi-structured
= legally as open as possible
We can provide data in several ways:

1. There are two APIs: a metadata harvest API on the basis of OAI-PMH, and a search API on the basis of SRU. Manuals for these APIs can be supplied once legal access has been granted via dataservices@kb.nl. Please note: users must have some experience of programming.

2. The Delpher newspapers comprise the texts (OCR, ALTO, XML) from all newspapers dating from 1618 to 1879. The archive is 111 GB and split into 23 ZIP files.

We can sometimes provide customised support. Ask your question via dataservices@kb.nl.

http://data.bibliotheken.nl/

Dataset

Dutch Novels 1800-2000
Dataset that contains a corpus of 1346 novels from OBNL.

Accessible e-books and audiobooks
We asked specialists to convert several public domain publications into accessible versions.

https://www.kb.nl/en/research-find/for-researchers/dataservices-apis-and-downloads
**Mining Shifting Concepts Through Time (ShiCo)**

**DIVE+**
Dynamicity Linking Collections to the Basis of Events

**BEELD & GEWELD**

**AVResearcher XL**

**GOLDEN AGENTS**
CREATIVE INDUSTRIES AND THE MAKING OF THE DUTCH GOLDEN AGE

**Time Capsule**
Making Pharmaceutical and Botanical Digital Heritage Accessible and Usable

**HiTiME**
Historical Timeline Mining and Extraction

**Political Mashup**

**Translantis**
Digital Humanities Approaches to Reference Cultures: The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990

**PoliMedia**
http://ngramviewer.kbresearch.nl/
## Beperking van samenstelling
- Onzelfstandige titels
- Zelfstandige titels
- Koeptitels
- Herdrukken
- Uitslagen

## Beperking van genre(s)
- Fictie
- Non-fictie
- Periode
den
- Bloemlezing
- Hertaling
- Verzameld werk
- Verzamelhandboeken

## Beperking van collectie(s)
- Acta Zuid-Holland
- Beschrijvende Oostindische Compagnie
- Briefwisseling Heinsius
- Briefen Van Gogh
- CRM
- Correspondentie Clusius
- DBNL
- Dagboek De Clercq
- Dagboeken Aelbert
- Dagboeken De Beaufort
- EDBO
- Gekapte brieven

## Documenten gevonden
- Alle woorden: `belg`.
- Gezocht in tekst, metadata van bron: tekst aanwezig.
- 172.551 documenten gevonden.

### Bronnen
- Visual overzicht
- Tijdslijn
- Statistieken
- Frequentielijsten
- Groeperingen

### Type woorden
- Top 100
- Min. lengte 4
- Begint met:
- Eindigt op:
- Regel: `[z-z]` (X)

<table>
<thead>
<tr>
<th>Token</th>
<th>Som</th>
<th>Documenten</th>
<th>Gem. per document</th>
<th>Maximum</th>
<th>Mediaan</th>
</tr>
</thead>
<tbody>
<tr>
<td>heeft</td>
<td>1.364.321 (0.026%)</td>
<td>68.859</td>
<td>20</td>
<td>9.376</td>
<td>3</td>
</tr>
<tr>
<td>worden</td>
<td>1.073.270 (0.020%)</td>
<td>63.947</td>
<td>17</td>
<td>3.336</td>
<td>2</td>
</tr>
<tr>
<td>hebben</td>
<td>996.326 (0.019%)</td>
<td>61.629</td>
<td>16</td>
<td>2.941</td>
<td>2</td>
</tr>
<tr>
<td>wordt</td>
<td>777.953 (0.015%)</td>
<td>55.584</td>
<td>14</td>
<td>3.877</td>
<td>2</td>
</tr>
<tr>
<td>onder</td>
<td>726.796 (0.015%)</td>
<td>49.589</td>
<td>15</td>
<td>3.563</td>
<td>2</td>
</tr>
<tr>
<td>waren</td>
<td>563.174 (0.011%)</td>
<td>64.714</td>
<td>9</td>
<td>2.464</td>
<td>2</td>
</tr>
<tr>
<td>andere</td>
<td>555.497 (0.010%)</td>
<td>43.910</td>
<td>13</td>
<td>1.630</td>
<td>1</td>
</tr>
<tr>
<td>tegen</td>
<td>556.410 (0.010%)</td>
<td>52.309</td>
<td>11</td>
<td>1.987</td>
<td>2</td>
</tr>
<tr>
<td>eerste</td>
<td>492.789 (0.009%)</td>
<td>51.391</td>
<td>10</td>
<td>1.667</td>
<td>2</td>
</tr>
<tr>
<td>kunnen</td>
<td>476.868 (0.009%)</td>
<td>40.769</td>
<td>12</td>
<td>1.662</td>
<td>2</td>
</tr>
<tr>
<td>alleen</td>
<td>456.525 (0.009%)</td>
<td>37.864</td>
<td>12</td>
<td>1.245</td>
<td>1</td>
</tr>
<tr>
<td>welke</td>
<td>441.054 (0.009%)</td>
<td>39.071</td>
<td>11</td>
<td>2.775</td>
<td>2</td>
</tr>
<tr>
<td>zijne</td>
<td>405.858 (0.008%)</td>
<td>12.663</td>
<td>32</td>
<td>4.523</td>
<td>2</td>
</tr>
<tr>
<td>Amsterdam</td>
<td>401.525 (0.007%)</td>
<td>83.533</td>
<td>5</td>
<td>8.564</td>
<td>2</td>
</tr>
<tr>
<td>zonder</td>
<td>377.746 (0.007%)</td>
<td>39.690</td>
<td>10</td>
<td>1.572</td>
<td>1</td>
</tr>
<tr>
<td>nieuwe</td>
<td>342.521 (0.006%)</td>
<td>48.209</td>
<td>7</td>
<td>1.017</td>
<td>1</td>
</tr>
<tr>
<td>leven</td>
<td>339.737 (0.006%)</td>
<td>17.067</td>
<td>20</td>
<td>1.096</td>
<td>1</td>
</tr>
<tr>
<td>reeds</td>
<td>331.078 (0.006%)</td>
<td>38.041</td>
<td>9</td>
<td>1.552</td>
<td>2</td>
</tr>
<tr>
<td>werden</td>
<td>326.547 (0.006%)</td>
<td>42.227</td>
<td>8</td>
<td>1.496</td>
<td>2</td>
</tr>
</tbody>
</table>

https://www.nederlab.nl/
Search "Public Dutch Newspapers"

Filters
- Date
  - 01-01-1600
  - 31-12-1876
- OCR confidence
  - 0 - 100
- Newspaper title
  - Choose
- Publication frequency
  - Choose

Display as Percentage

Graph showing activity over time with a peak in around 1700.

https://ianalyzer.hum.uu.nl/
Why don’t we build an advanced research environment ourselves?
How can we better fit the needs of Digital Humanities scholars by co-designing the solutions together with the researchers?
Recommendations

We conclude there are sufficient opportunities for developing a text suite. For this development, we provide the following recommendations:

1. Position a text suite as a corpus selection tool and support the discovery and selection research phases. A text suite hereby functions as a user-friendly front end toDataservices with more advanced features that do not fit within Delpher and DBNL. Users can then make their own selection of sources and export them (possibly after approval by a KB employee).

https://zenodo.org/record/6591572
Why don’t we build an advanced research environment ourselves?

Just give me access and allow me to select the data I want and allow me to do everything I want.

And not only KB data but data from different heritage institutes.
Case 1
FAIR Cultural Heritage Data
Network Digital Heritage

The Dutch Digital Heritage Network (NDE) aims at increasing the social value of the cultural heritage information maintained by libraries, archives, museums and other cultural institutions.

The NDE strategy starts from the end user perspective and encourages institutions to provide digital heritage information that is more visible, usable and sustainable.

The NDE program is about building strong cross sector networks on the level of expertise and information. Linked Data is regarded as one of the enabling technologies.

https://netwerkdigitaalerfgoed.nl/en/
Roadmap for the NDE discovery infrastructure

Services
- Web Portals (Powered by KGs)
- Dataset Register
- Knowledge Graph (data summaries)
- Network of Terms

Network
- "discover relevant datasets"
- "smarter, more dynamic services"
- "improve the visibility on the web"

Sources
- "advertise your data"
- "use things not strings"
Dataset Register
For all heritage datasets!

The dataset register provides insight into the availability of datasets in the heritage field and thus encourages the use of these datasets.

For heritage institutions with datasets: add a dataset description
For users of heritage data: search 754 dataset descriptions

https://datasetregister.netwerkdigitaaldergoed.nl/?lang=en
CLARIAH develops, facilitates and stimulates the use of Digital Humanities resources and infrastructures. We offer these resources to researchers and other professionals in an insightful and user-friendly way.

https://www.clariah.nl/
Start using digital humanities resources

Ineo lets you search, browse, find and select digital resources for your research in humanities and social sciences. The platform is already fully functional, but is still being filled with resource content. At the end of 2023, it will offer access to thousands of tools, datasets, workflows, standards and educational material.

https://www.ineo.tools/
Case 2
Secure Data Analysis Environment
Data

= result of more than 200 years of collecting over 30 years of digitisation 10+ years of collecting born-digital publications

= machine readable, mostly textual

= structured or semi-structured

= legally as open as possible
Data

= result of more than 200 years of collecting over 30 years of digitisation 10+ years of collecting born-digital publications

= machine readable, mostly textual

= structured or semi-structured

= legally as open as possible
Current workflow

Challenges for both researcher as well as KB:

- More data
- Better & richer data
- More recent data (legal issues)
Tools-to-data: a data-safe research environment

Tools-to-data allows researchers to work with data in a safe environment that prevents data leaks. Instead of sending a copy of the data to the researchers to work with, the researchers send their tool to the environment and receive the results.

Joris van Zundert
WP6 Co-leader, CLARIAH NL

Melvin Wevers
Assistant Professor in Urban History and Digital Methods, University of Amsterdam

Roel Smeets
Assistant Professor of Modern Literature and Digital Culture, Radboud University

https://www.clariah.nl/projects/tools-to-data
1. Data provider authorises researcher

2. Researcher selects data

3. Researcher develops/chooses algorithm

4. Researcher uploads algorithm

5. Data provider approves access

6. Algorithm runs in secure environment

7. Results checked by data provider

8. Results available to researcher

Future workflow

https://zenodo.org/record/7254517
SANE: secure data environment for social sciences and humanities

Privacy, copyright and competition barriers limit the sharing of sensitive data for scientific purposes. Together with several partners, SURF is working on a Secure ANalysis Environment (SANE). A virtual container in which the researcher can analyse sensitive data, but the data owner retains full control. This makes new research possible.

Sensitive data remain unused

Although non-academic parties have an increasing number of interesting datasets available, there is currently no
Wrap up

How can we better fit the needs of Digital Humanities scholars by co-designing the solutions together with the researchers?
Some lessons

COLLABORATE

- Collaborate with researchers
- Collaborate with other heritage institutes
- Collaborate with research IT partners
- Build generic solutions. There is no 1 size fits all
- Make access to data as simple and secure as possible
Questions?

Building a national Infrastructure to enable research with Dutch Digital Heritage Collections

Dr. Martijn Kleppe
Board Member Research & Discovery
KB, national library of the Netherlands

martijn.kleppe@kb.nl
@martijnkleppe
https://www.kb.nl/over-ons/experts/martijn-kleppe

Thanks fantastic colleagues

Marian Hellema
Enno Meijers
Steven Claeysens

KB nationale bibliotheek
National Library of the Netherlands