KB Lab: Exploring the National Library of the Netherlands’ digital treasure trove

Lotte Wilms | @lottewilms
1. KB, National Library of the Netherlands
KB in a nutshell

**Founded**
1798

7 million items = 115 km of library materials

10,800 current periodicals;

500 licensed databases and e-journals;

In 2015 the collection grew by:

49,000 books;

42,000 issues of periodicals;

6.5 million digital items;

2,700 e-books;

3,800 websites.
Building

Net floor space of the building: **80,000 m²**

Library: **37,000 m²**, including **28,000 m²** storage

Other institutions: **15,000 m²**

Capacity

500 study seats (including 125 with a work station)

WiFi available
Mission

- The KB brings people and information together;
- The KB makes the library collection of the Netherlands visible, preservable and usable;
- The KB holds a central position in the library network;
- The KB helps people to become more skilled, smart and creative.
1,5 million radio bulletins

383,000 books (+6,000 new books every month)

1,5 million journal pages

1,2 million pages of parliamentary papers

almost 20,000 images

11 million newspaper pages

We also host the Digital Library of Dutch Literature (DBNL) which contains:

- 13,000 titles with corrected OCR
- 14,000 scanned titles
- 2,500 eBooks
2. Digitised historical newspapers
Delpher newspaper corpus

• Digitised Dutch newspapers
• 1618-1995
• Images + metadata + text
• now: 11 million pages (in 1,351,123 issues)
• prognosis 2020: 20 million pages
• Full text searchable on: www.delpher.nl
## The data

<table>
<thead>
<tr>
<th>type/format</th>
<th>level</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF</td>
<td>issue</td>
<td>Searchable text + scan</td>
</tr>
<tr>
<td>JPEG-2000</td>
<td>page</td>
<td>Access: JPEG 2000 lossy compression, colour (or greyscale in case of original from microfilm)</td>
</tr>
<tr>
<td>Dublin Core</td>
<td>iss./p./art.</td>
<td>Descriptive metadata</td>
</tr>
<tr>
<td>OCR</td>
<td>article</td>
<td>XML</td>
</tr>
<tr>
<td>ALTO</td>
<td>page</td>
<td></td>
</tr>
<tr>
<td>mpeg21-didl</td>
<td>issue</td>
<td>Structural metadata</td>
</tr>
</tbody>
</table>
Copyright

→ 1618 – 1876 = Public Domain

→ 1877 – 1995 = (Potentially) in copyright

**However:**
- All content online at delpher.nl
- Available for research
3. Access
Delpher

- > 60 million digitised text pages
- Newspapers, books, magazines, radio bulletins
- 1618 – 1995
- Full text search
- Updated regularly

www.delpher.nl
Delpher data

- 22 downloadable zip files
- All newspapers of 1618 - 1876
- Metadata + OCR + ALTO
- CC-BY
- Unzipped: 622 GB

www.delpher.nl/data
Data services

- Access to data via APIs
- Open data
  - Set descriptions
  - Manual to access via API
- Data in copyright
- Contact dataservices@kb.nl
- Access for research

www.kb.nl/dataservices
4. OCR
OCR correction

• Crowdsourced correction

• 17th century newspapers
  Partner: Meertens Instituut

• WWII newspapers
  Partner: NIOD
Automated OCR correction

- Research project
- 2000 pages
- Evaluation of OCR
- ReOCR existing material
- Automated correction using machine learning (LSTM)

- Partner: Netherlands eScience Center
5. Use of Delpher newspapers: external projects
Polimedia

- Links Dutch parliamentary papers to newspapers and ANP in one search interface
- CLARIN-NL Project
- TU Delft, VU University, Sound & Vision, Erasmus University

- http://www.polimedia.nl/
Translantis: Texcavator

- “Digital Humanities Approaches to Reference Cultures: The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990.”
- Utrecht University: 2013 – 2018
- Funded by Dutch government
- Uses analysis tool Texcavator with Delpher newspapers

- [http://translantis.wp.hum.uu.nl/](http://translantis.wp.hum.uu.nl/)
6. Use of Delpher newspapers: internal projects
KB Lab

= online and offline sandbox
  2 advisors
  1 curator Digital Collections
  2 research software engineers

= experimental data sets

= tools and software

= workshops
Cooperation via KB initiated programs

1. Fellowship Program
   
   = 4 months, 1 fte | invited | successful academic

2. Researcher-in-Residence Program
   
   = 6 months, 0.5 fte | open call | early career

→ Technical support from KB Lab
→ Tools and/or data always available for others
Tool: Frame generator

- In collaboration with fellow Prof. Dr. Joris van Eijnatten
- Automatic extraction of keywords from a text
- Identification of co-occurrence patterns
- “How is the concept Europe described in newspapers?”
The KBK-1M Dataset is a collection of 1,603,316 images and accompanying captions of the period 1922–1994. We extracted the images from digitised newspapers that are stored in the National Library (KB) Newspaper Archive and that are publicly accessible via www.delpher.nl. Via Delpher visitors can search and browse through several collections, including Dutch newspapers. One way to narrow down retrieved results is by clicking on facets. One of these is ‘Illustrations met oorsprong’, illustrations with captions that contain photographs (black & white and colour), comic strips, political cartoons and weather forecasts. This KBK-1M dataset contains these illustrations with captions of all newspapers in the period 1922–1994 which were on Delpher when we crawled the illustrations in August 2015.

**Creation of the dataset**

In the newspaper archive of the KB, each issue is stored as a set of scanned pages with one JPEG per newspaper page. Each page is associated with a set of metadata files which describe the locations of each image, caption and article on that page. During the digitisation process of the newspapers, these locations were manually annotated by trained workers. The article and caption texts are available through automatic OCR-processed output. We took these data as starting point when we built the harvester to create the KBK-1M dataset. The data harvester was built using the Python programming language which prepared and extracted the images and captions using KB-internal RESTful APIs. Figure 1 below, shows how we transformed the raw source material into the dataset that contains JPEG files for the images and JSON files for the metadata.

### Data: KBK-1M

- 1,6 million images extracted from digitised newspapers
- 1922-1994
- Combination of image & byline
- Available for research purposes
- Computer vision
Computer vision applications

• Researchers-in-residence:
  • Thomas Smits (RU): Shift from illustration to photograph in newspapers & recognition of subjects
  • Melvin Wevers (UU): Nearest neighbours of advertisements
Open: Github

- Online code repository
- Open source
- Free for re-use
Future plans

• Be more transparent about collections

• Research own collections & publish about findings

• Make access to open sets easier

• Find new user groups such as social sciences
Any questions?