
GELYKHEID, VRYHEID, BROEDERSCHAP.
N°. 100.
HAAGSCHE COURANT.
Maandag den 20 Augustus. A°. 1798.
HET VIERDE JAAR DER BATAAFSCHE VRYHEID.
FRANSCHE REPUBLIEK.
PARYS 14 Aug. Brieven van Genua verzekeren op eene stellige wyze, dat de Paus in het Carthuizer Klooster by Florence overleden zy.
Men wil, dat de Generaal en Chef des Legers van Engeland zich te Parys bevindt, om over het uitlopen van het Esader te Brest te raadplegen; anderen zyn van gevoelen
by het bevel over het leger van den Rhyn
Van Arau wordt gemeldt
Bern bevel gekomen
land der

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

Lotte Wilms | @lottewilms

www.kb.nl

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

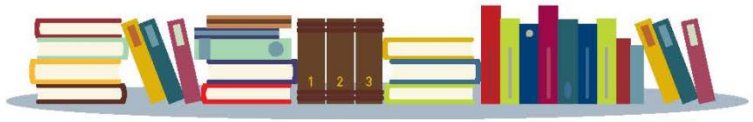
- The KB brings people and information together;
- **The KB makes the library collection of the Netherlands visible, preservable and usable;**
- The KB holds a central position in the library network;
- The KB helps people to become more skilled, smart and creative.

Board of Governors

Public Library Policy | Director General | Corporate Communication

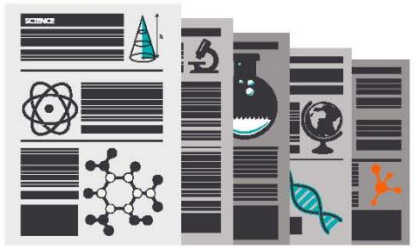
Finance & Corporate Services	Processing & Preservation	Marketing & Services	Innovation & IT
Internal Services	Digitization & Document Processing	Collections	Innovative Projects
Governance, Risk & Compliance	Digital Preservation	User Services	Information Management
Building & Facilities Management	Collection Care	Marketing & Education	Research
			IT Service Management
			Information Technology
			Metamorfoze Office

Digital collection KB

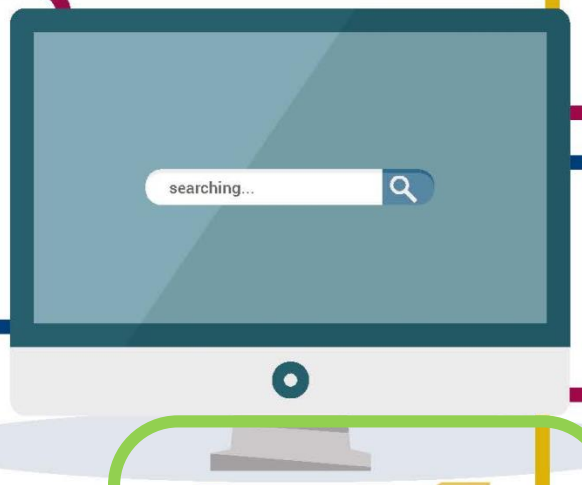


383.000 books (+6.000 new books every month)

1,5 million journal pages



1,5 million radio bulletins



1,2 million
pages of parliamentary papers

almost 20.000 images

11 million
newspaper pages

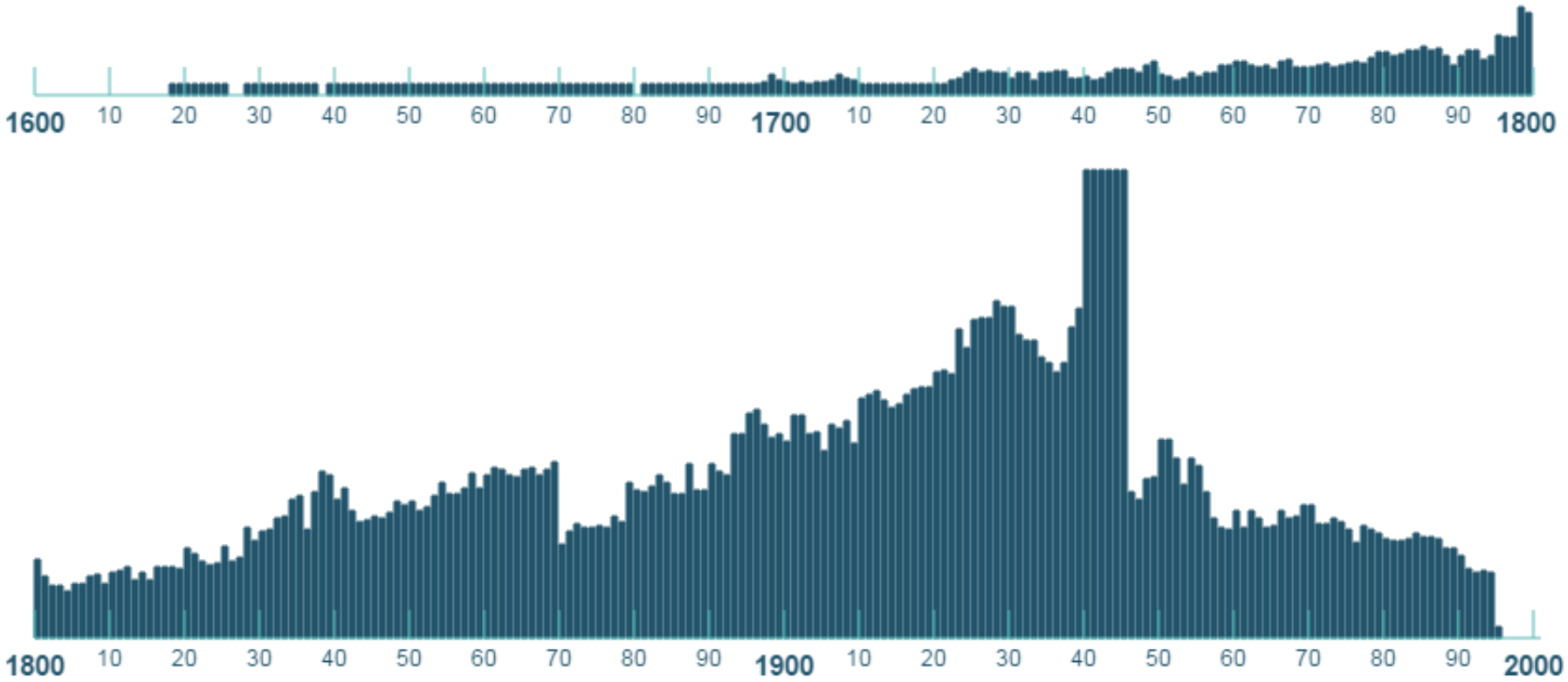
We also host **the Digital Library of Dutch Literature (DBNL)** which contains:

- **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


1600 10 20 30 40 50 60 70 80 90 1700 10 20 30 40 50 60 70 80 90 1800
1800 10 20 30 40 50 60 70 80 90 1900 10 20 30 40 50 60 70 80 90 2000

The data

type/format	level	comments
PDF	issue	Searchable text + scan
JPEG-2000	page	Access: JPEG 2000 lossy compression, colour (or greyscale in case of original from microfilm)
		Master: JPEG 2000 part 1, lossless compression, greyscale or colour
Dublin Core	iss./p./art.	Descriptive metadata
OCR	article	XML
ALTO	page	
mpeg21-didl	issue	Structural metadata

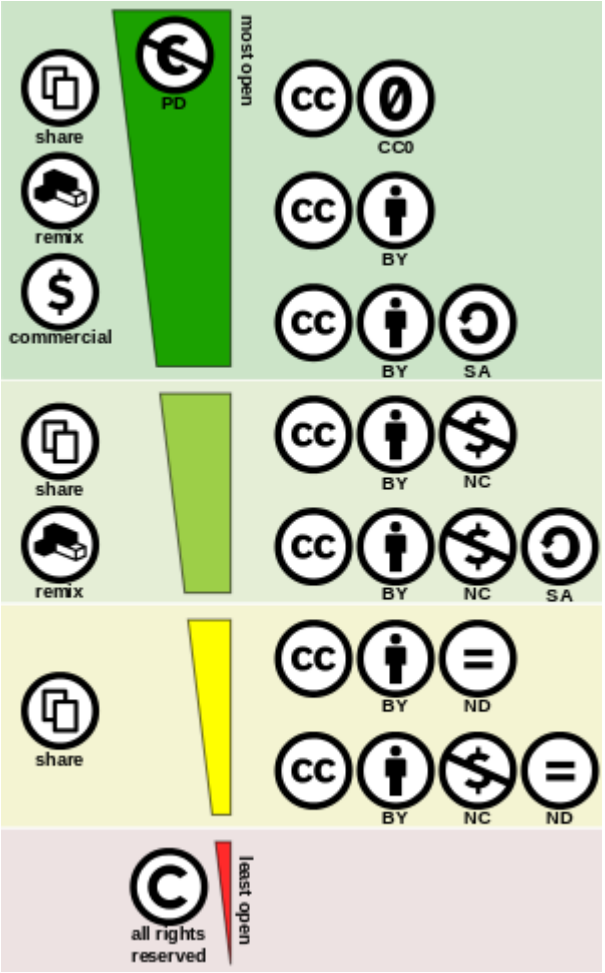
Copyright

→ 1618 – 1876 = Public Domain

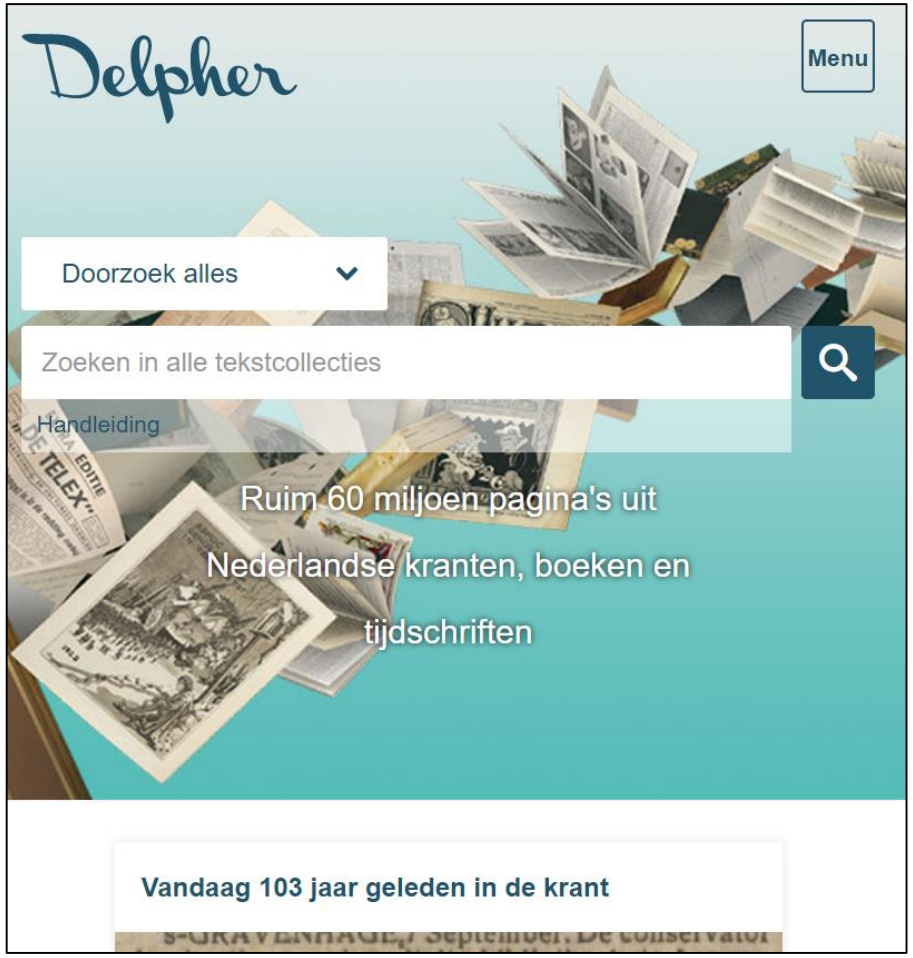
→ 1877 – 1995 = (Potentially) in copyright

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

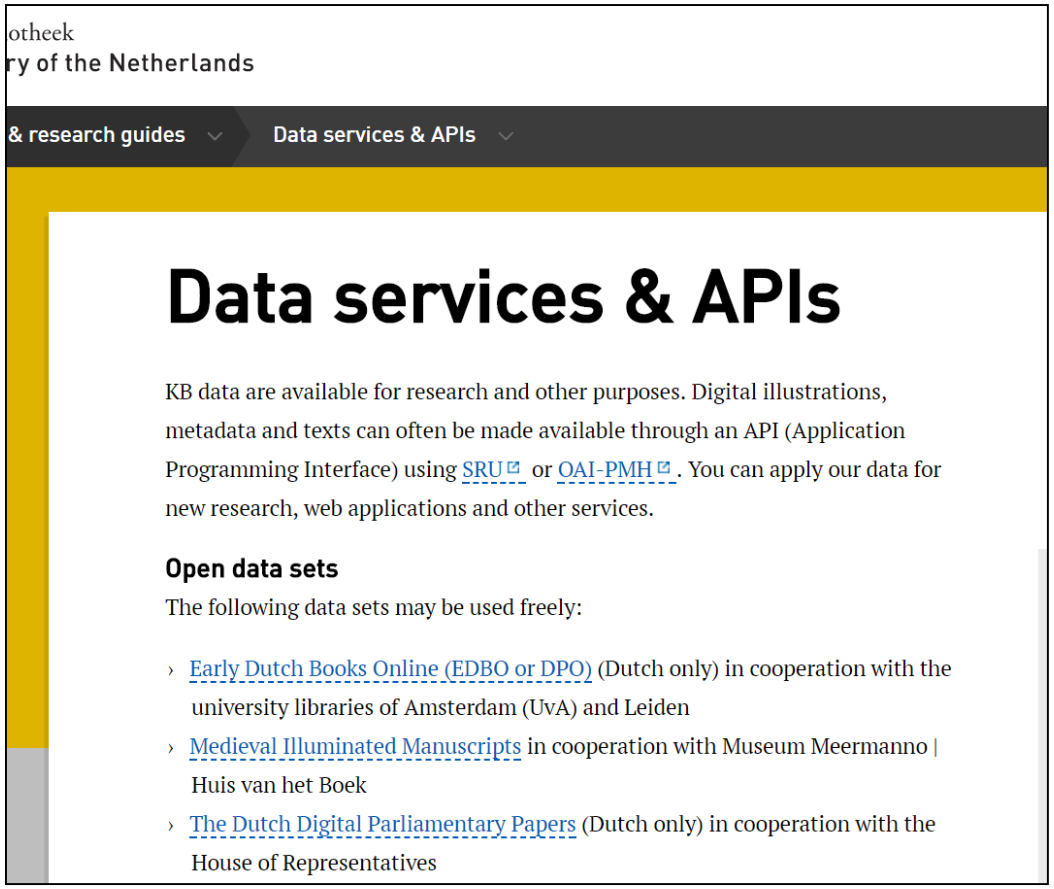
- 22 downloadable zip files
- All newspapers of 1618 - 1876
- Metadata + OCR + ALTO
- CC-BY
- Unzipped: 622 GB

- www.delpher.nl/data

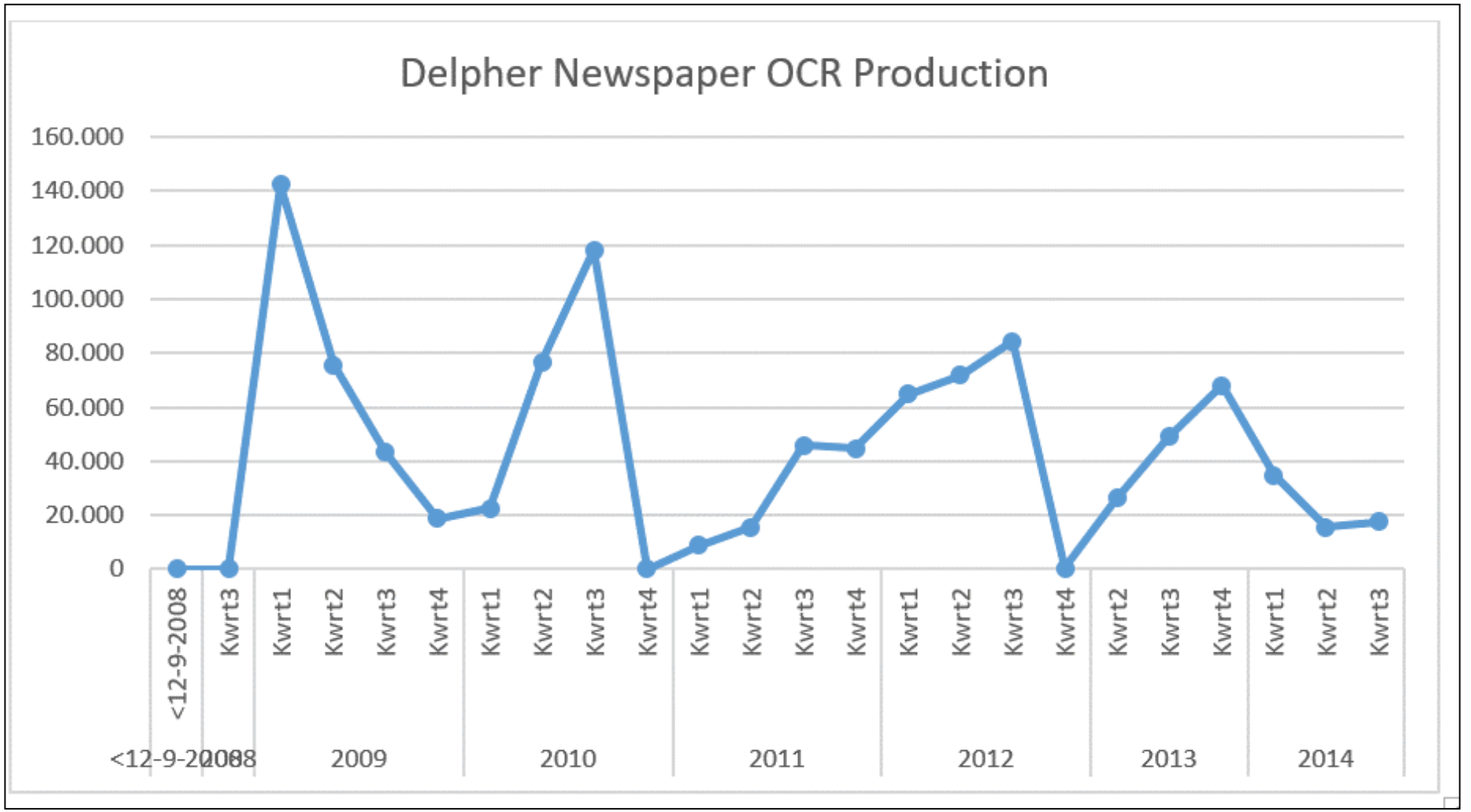
Data services

- 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

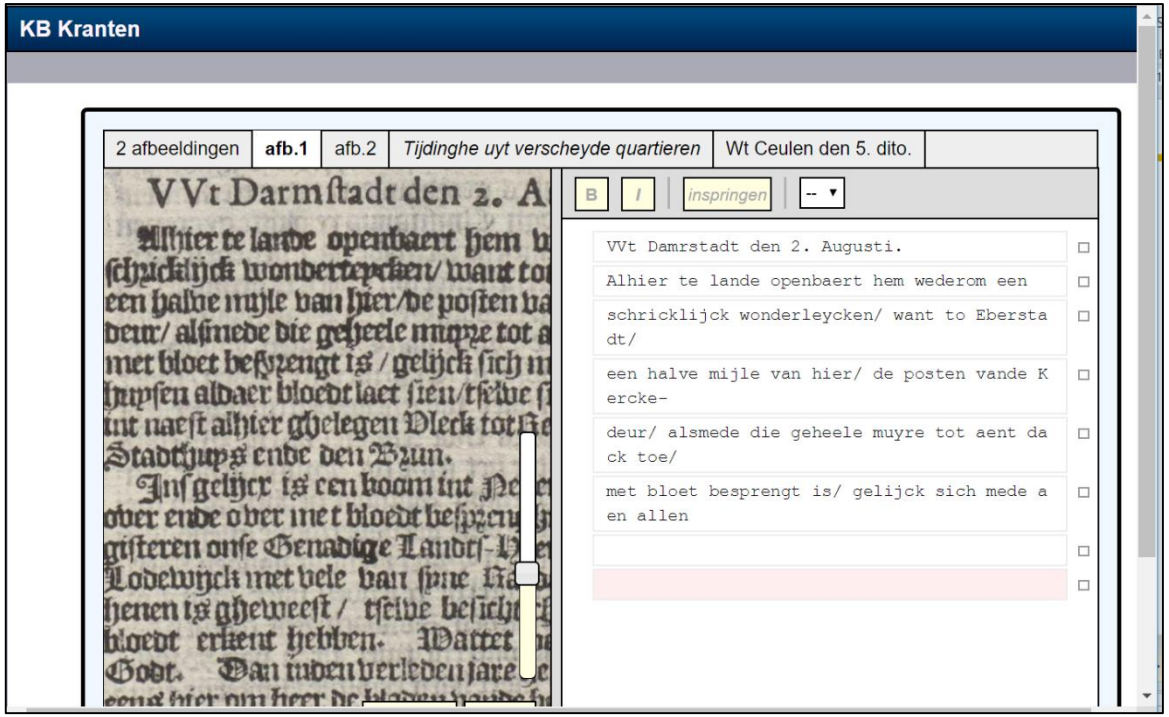

Delpher Newspaper OCR Production
160.000
140.000
120.000
100.000
80.000
60.000
40.000
20.000
0
<12-9-2008
Kwrt3
Kwrt1
Kwrt2
Kwrt3
Kwrt4
Kwrt1
Kwrt2
Kwrt3
Kwrt4
Kwrt1
Kwrt2
Kwrt3
Kwrt4
Kwrt1
Kwrt2
Kwrt3
Kwrt4
Kwrt2
Kwrt3
Kwrt4
Kwrt1
Kwrt2
Kwrt3
<12-9-2008
2008
2009
2010
2011
2012
2013
2014

OCR correction

- Crowdsourced correction

- 17th century newspapers
- Partner: Meertens Instituut

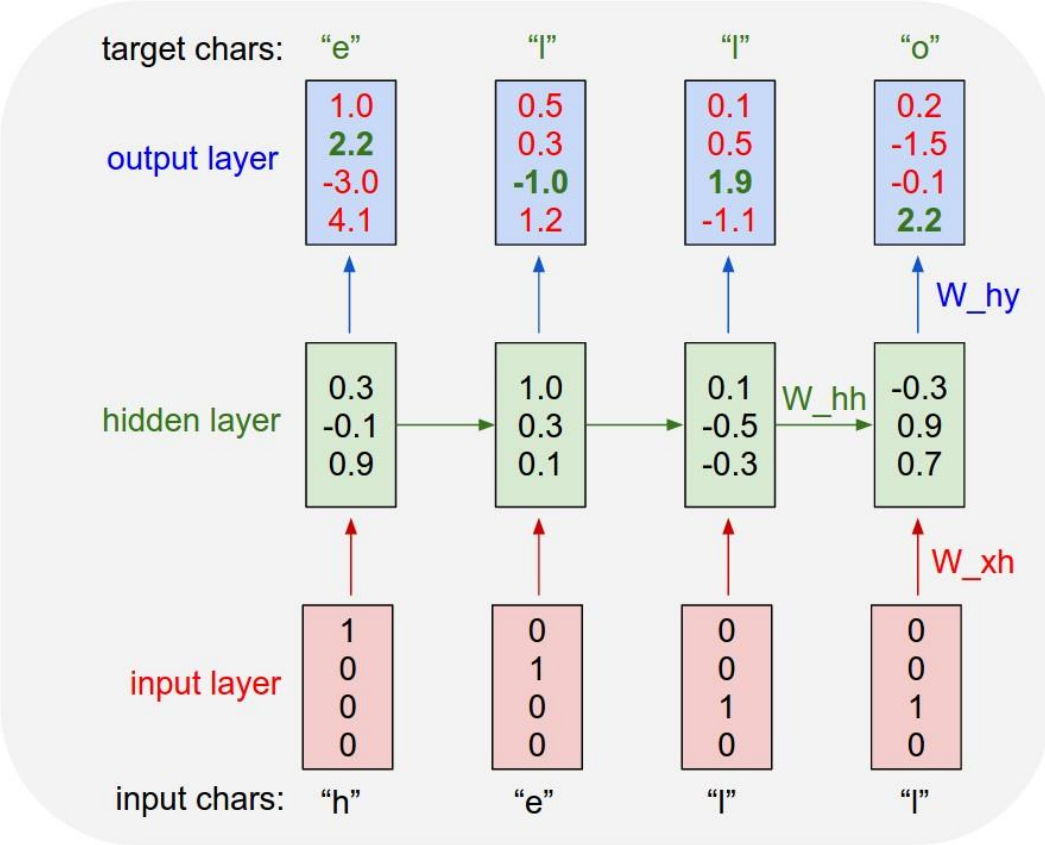
- WWII newspapers
- Partner: NIOD



Automated OCR correction

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

- Partner: Netherlands eScience Center

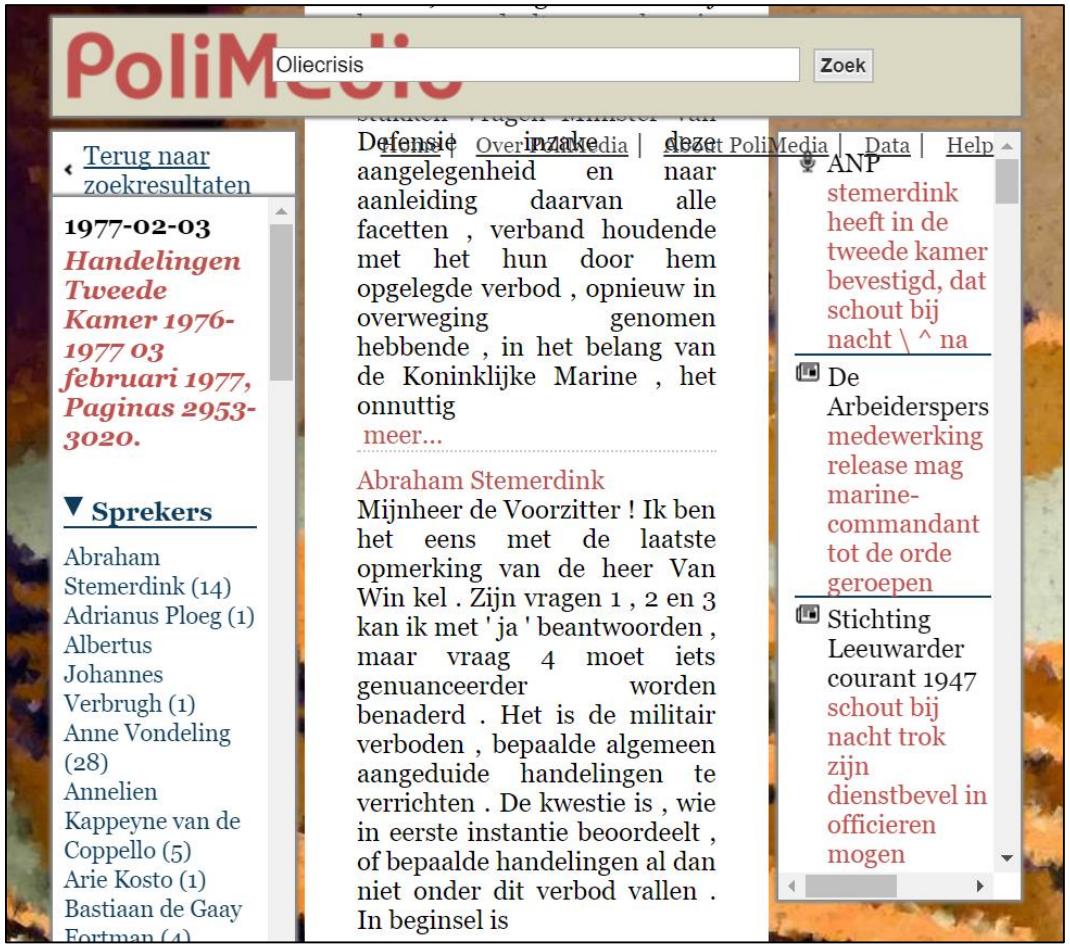


5. Use of Delpher newspapers: external projects

Polimedia

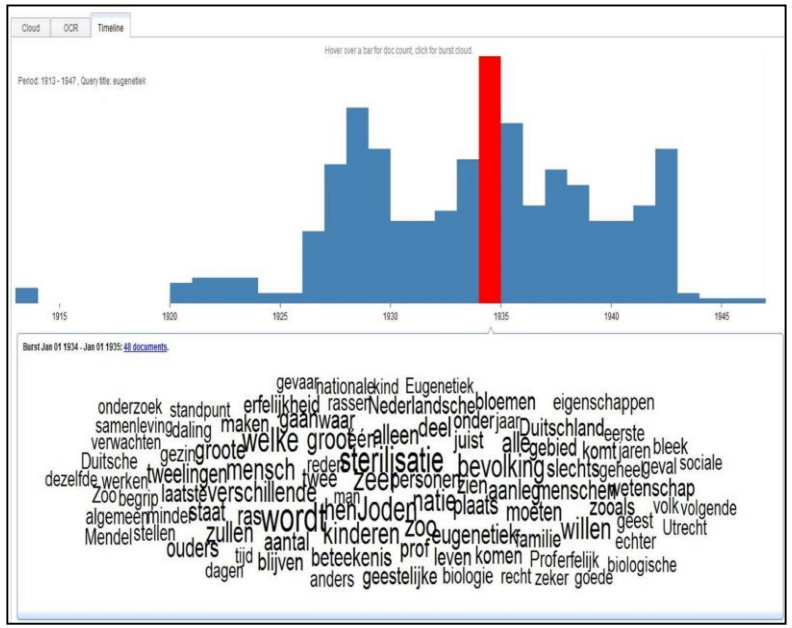
- Links Dutch parliamentary papers to newspapers and ANP in one search interface
- CLARIN-NL Project
- TU Delft, VU University, Sound & Vision, Erasmus University

- 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/

6. Use of Delpher newspapers: internal projects

KB Lab

= online **and** offline sandbox

2 advisors

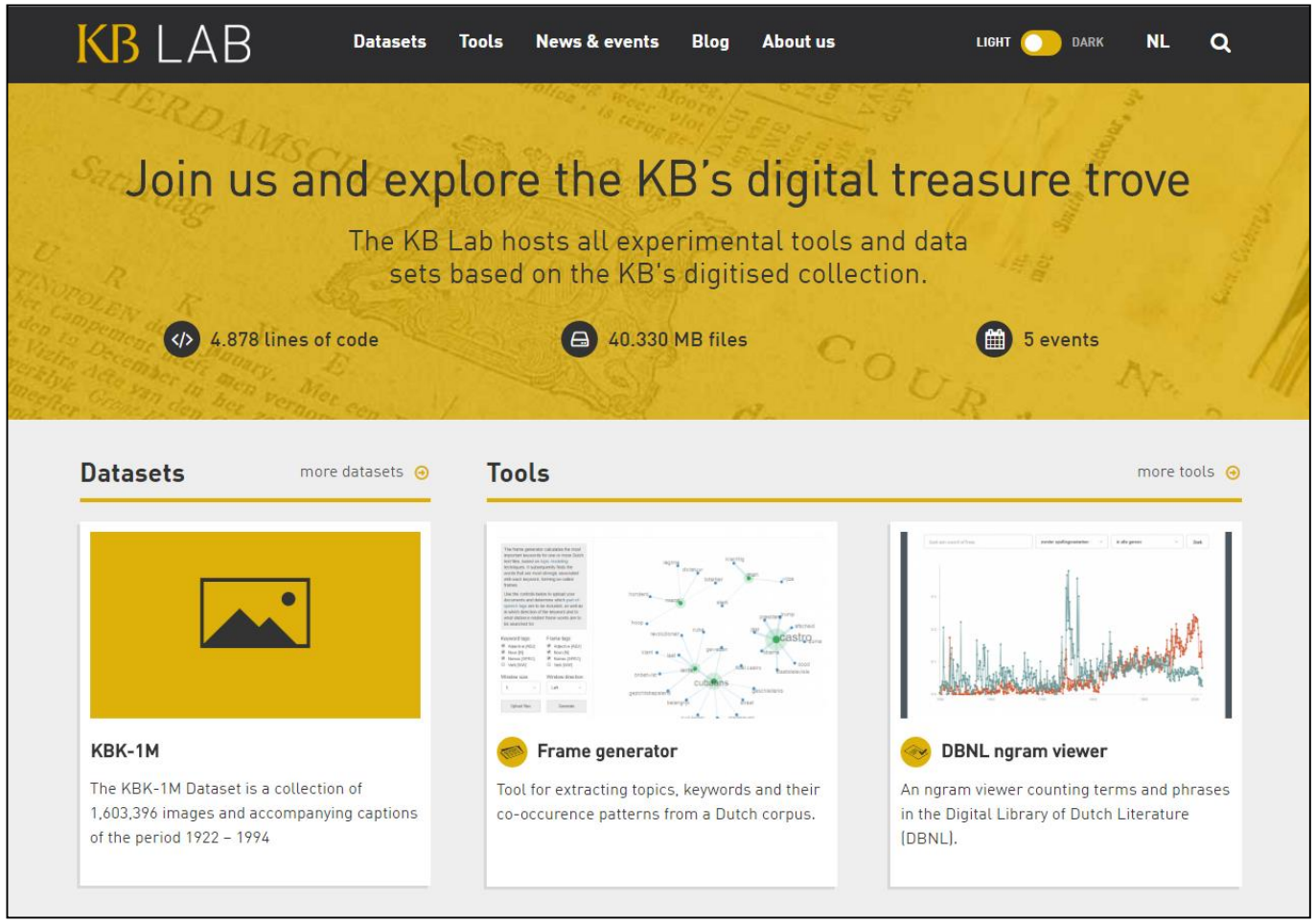
1 curator Digital Collections

2 research software engineers

= experimental data sets

= tools and software

= workshops



www.lab.kb.nl

Cooperation via KB initiated programs

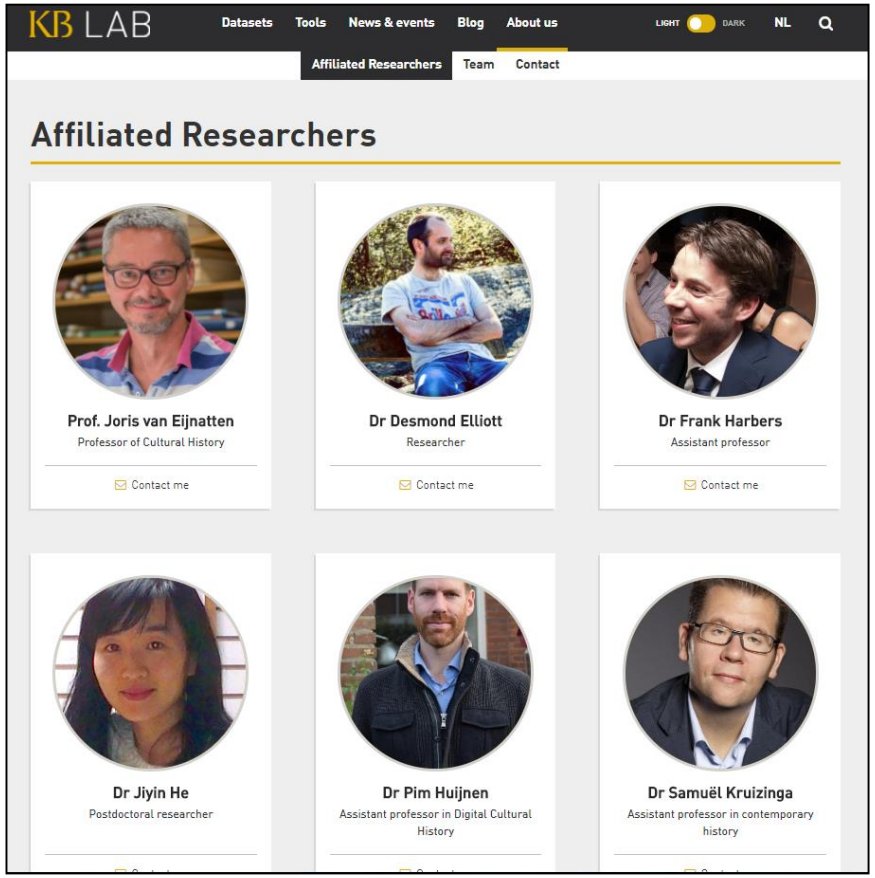
1. Fellowship Program

   = *4 months, 1 fte | invited | successful academic*

2. Researcher-in-Residence Program

   = *6 months, 0,5 fte | open call | early career*

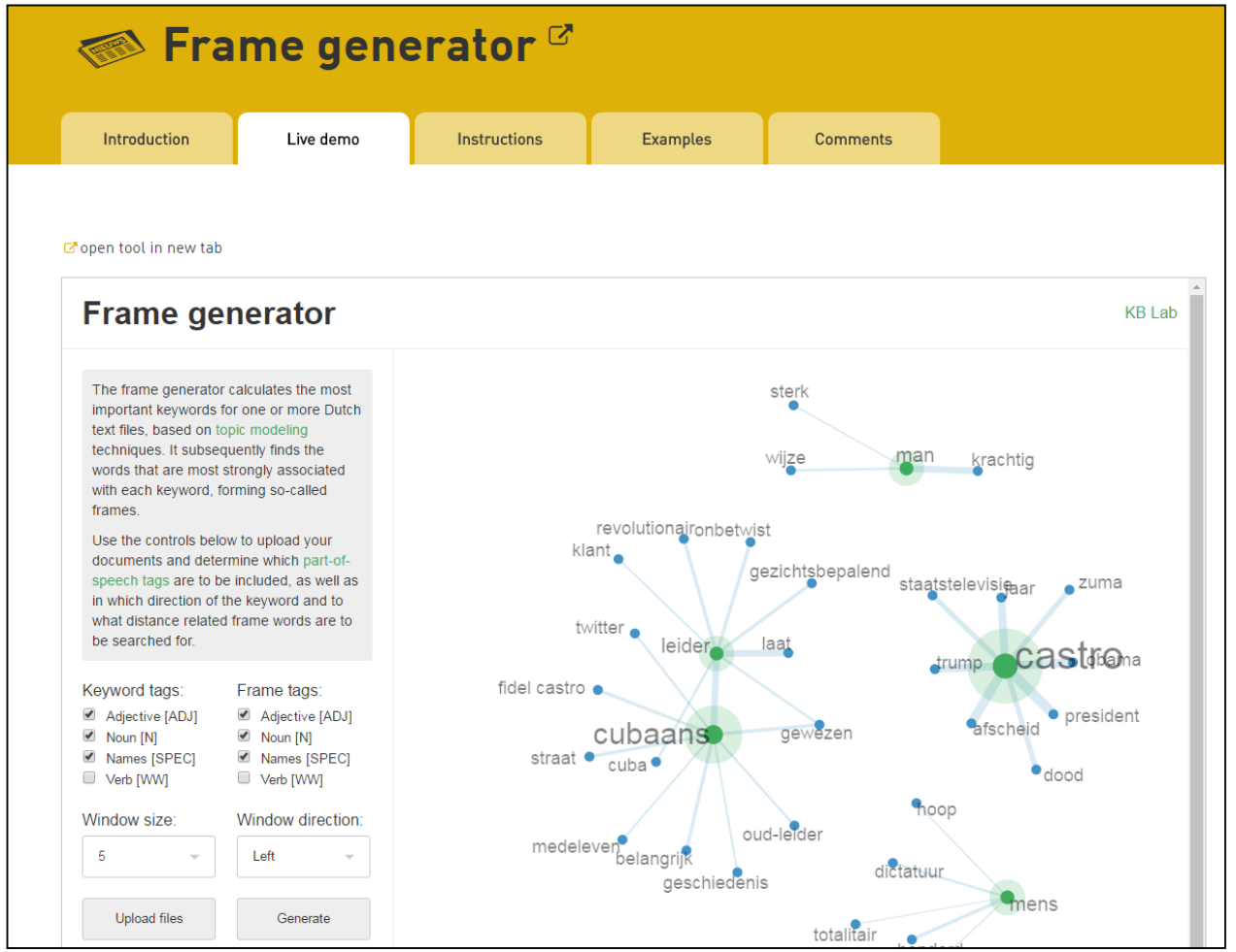
→ *Technical support from KB Lab*
→ *Tools and/or data always available for others*

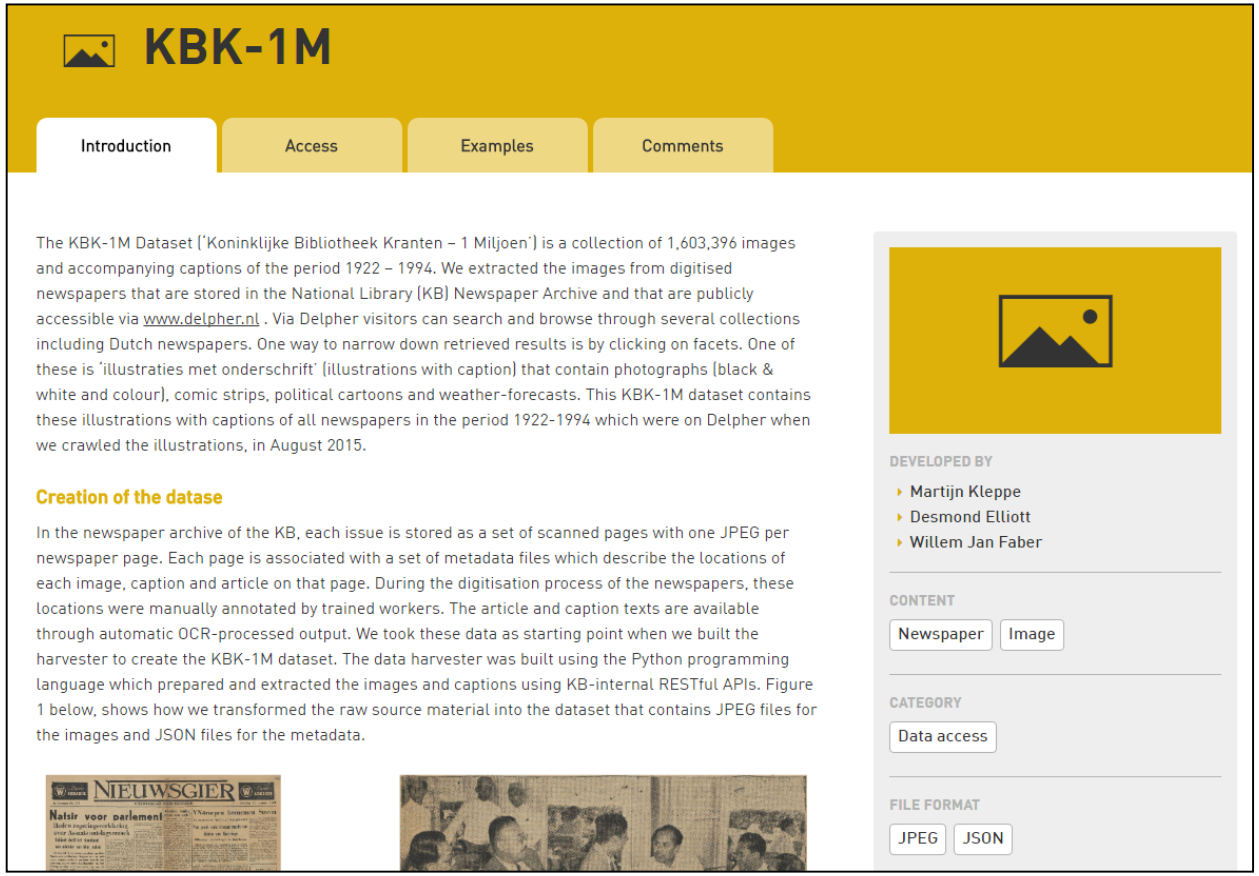


Tool: Frame generator

- 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?”



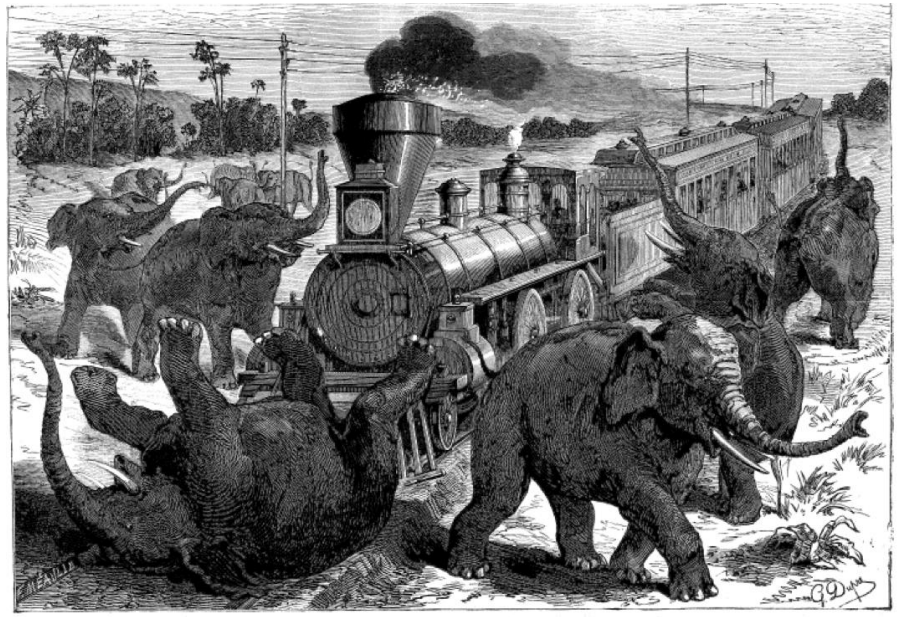


Data: KBK-1M

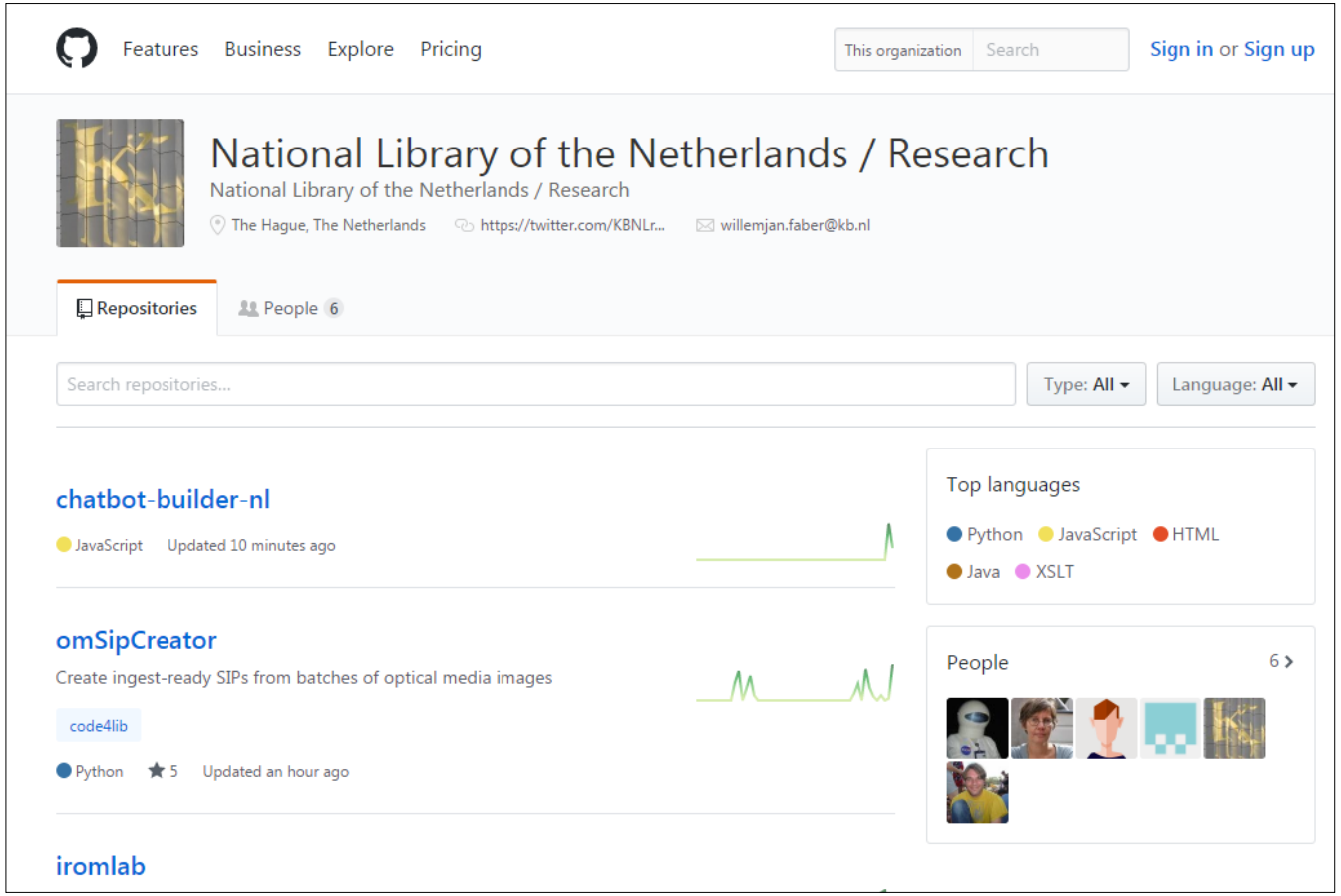
- 1,6 million images extracted from digitised newspapers
- 1922-1994
- Combination of image & byline
- Available for research purposes
- Computer vision

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



ENGELSCH INDIË. — AANVAL VAN EEN KUDDE OLIFANTEN OP EEN SPOORTREIN. (Zie blz. 6.)

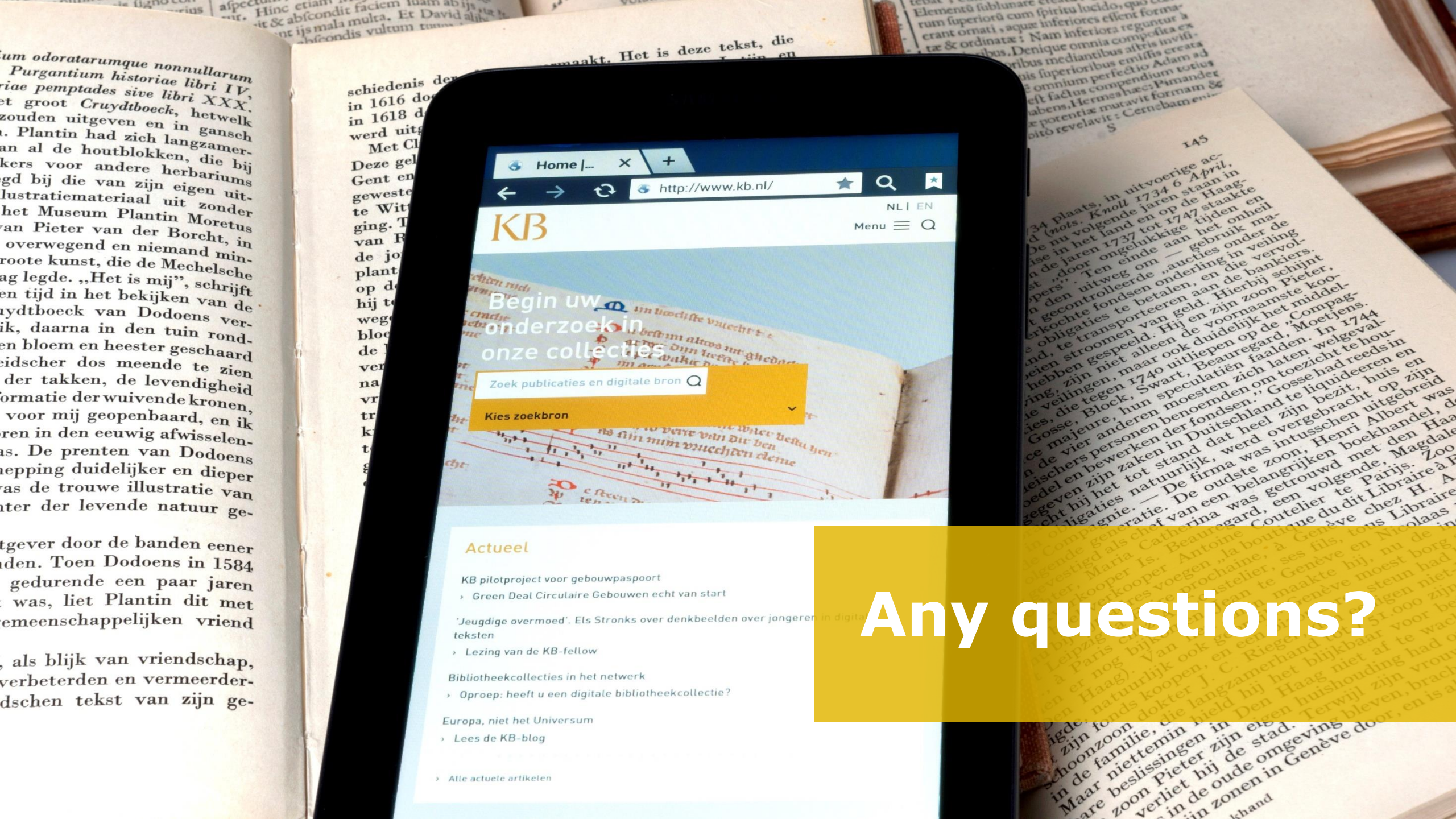


Open: Github

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

Future plans

- Be more transparant about collections
- Research own collections & publish about findings
- Make access to open sets easier
- Find new user groups such as social sciences


Home |...
http://www.kb.nl/
KB
NL | EN
Menu
Begin uw onderzoek in onze collecties
Zoek publicaties en digitale bron
Kies zoekbron
Actueel
KB pilotproject voor gebouwpaspoort
› Green Deal Circulaire Gebouwen echt van start
'Jeugdige overmoed'. Els Stronks over denkbeelden over jongeren in digitale teksten
› Lezing van de KB-fellow
Bibliotheekcollecties in het netwerk
› Oproep: heeft u een digitale bibliotheekcollectie?
Europa, niet het Universum
› Lees de KB-blog
› Alle actuele artikelen
Any questions?