



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

Lotte Wilms | @lottewilms

[www.kb.nl](http://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 m<sup>2</sup>**

Library: **37,000 m<sup>2</sup>**, including **28,000 m<sup>2</sup>** storage

Other institutions: **15,000 m<sup>2</sup>**

## Capacity

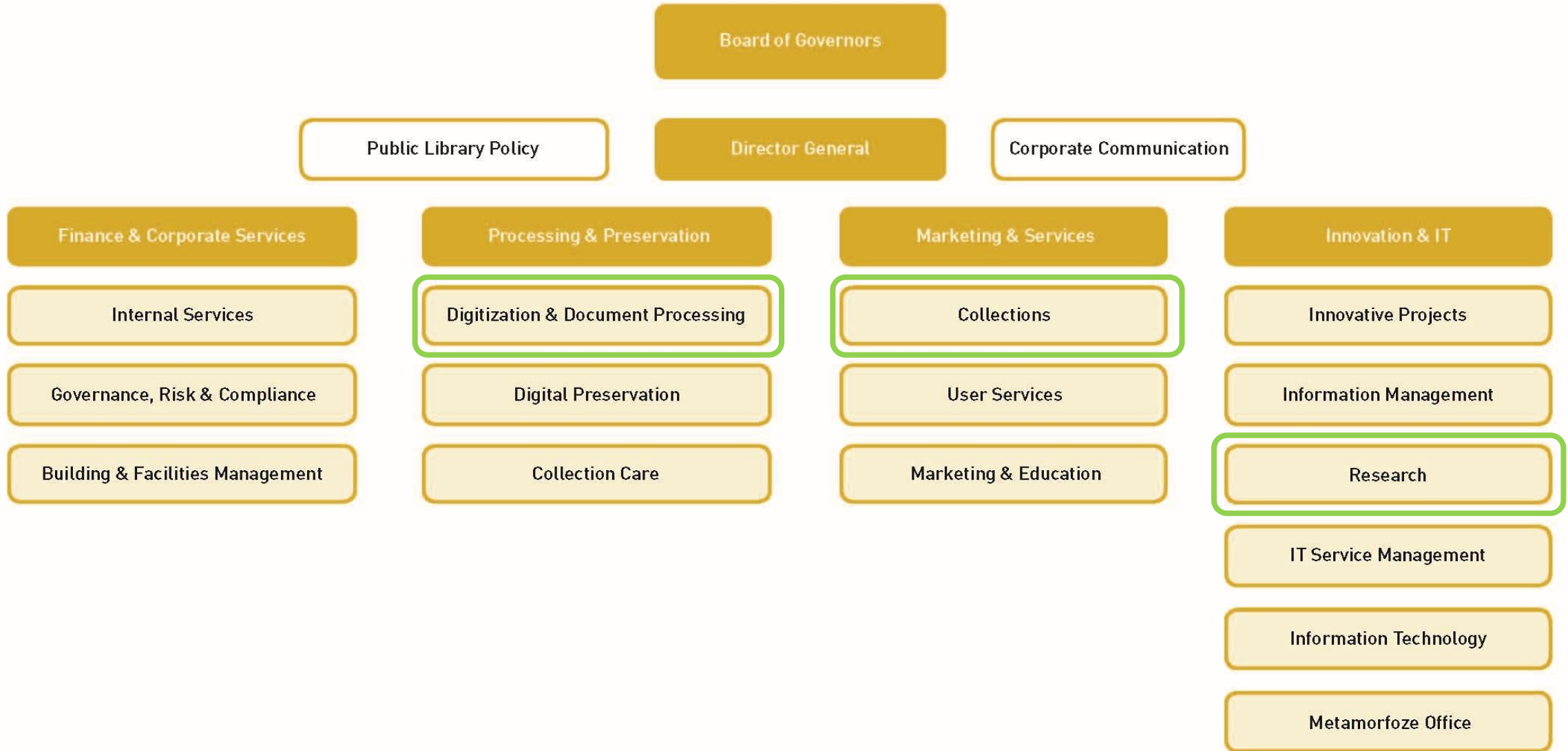
**500** study seats (including **125** with a work station)

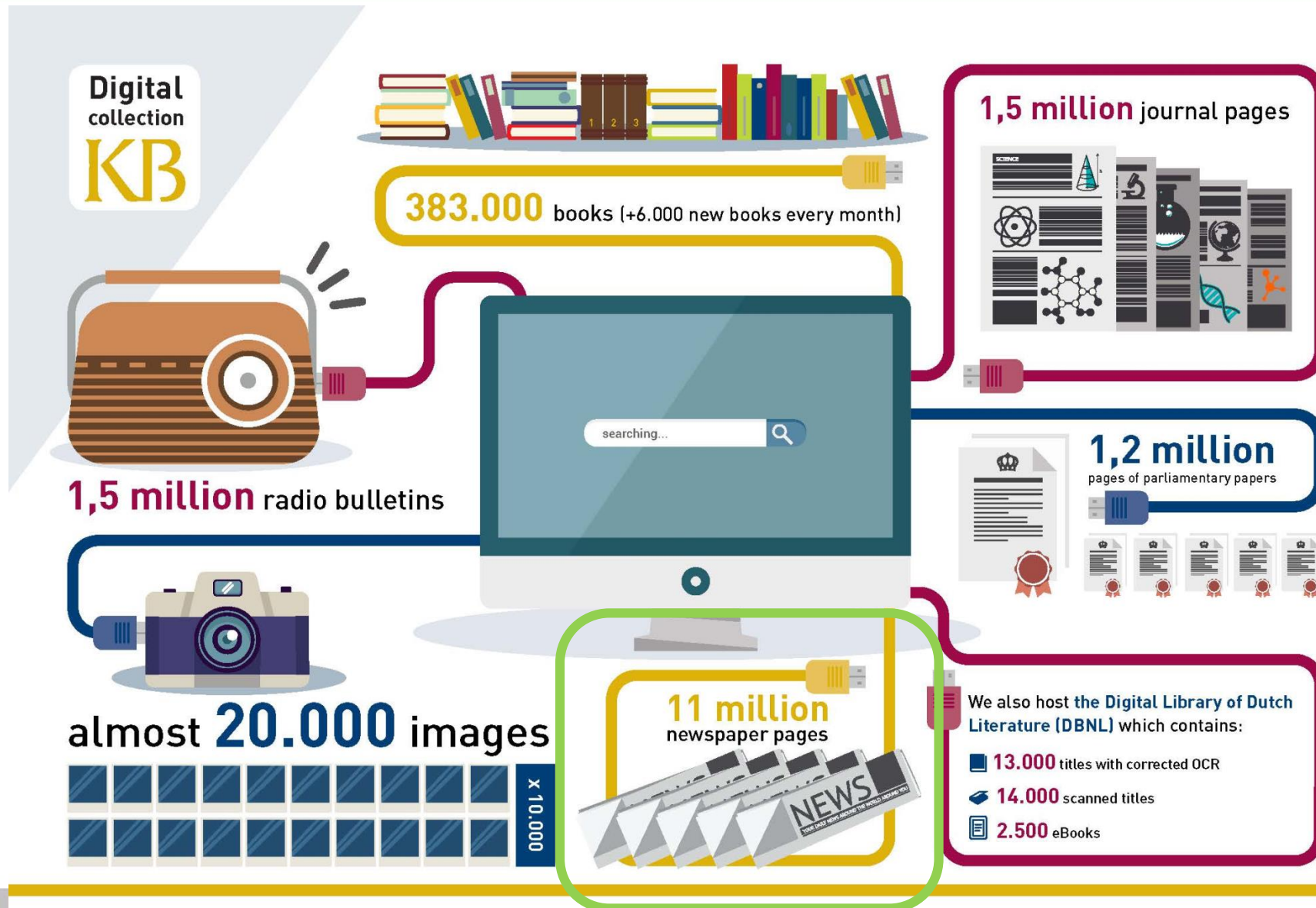
WiFi available

## 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.







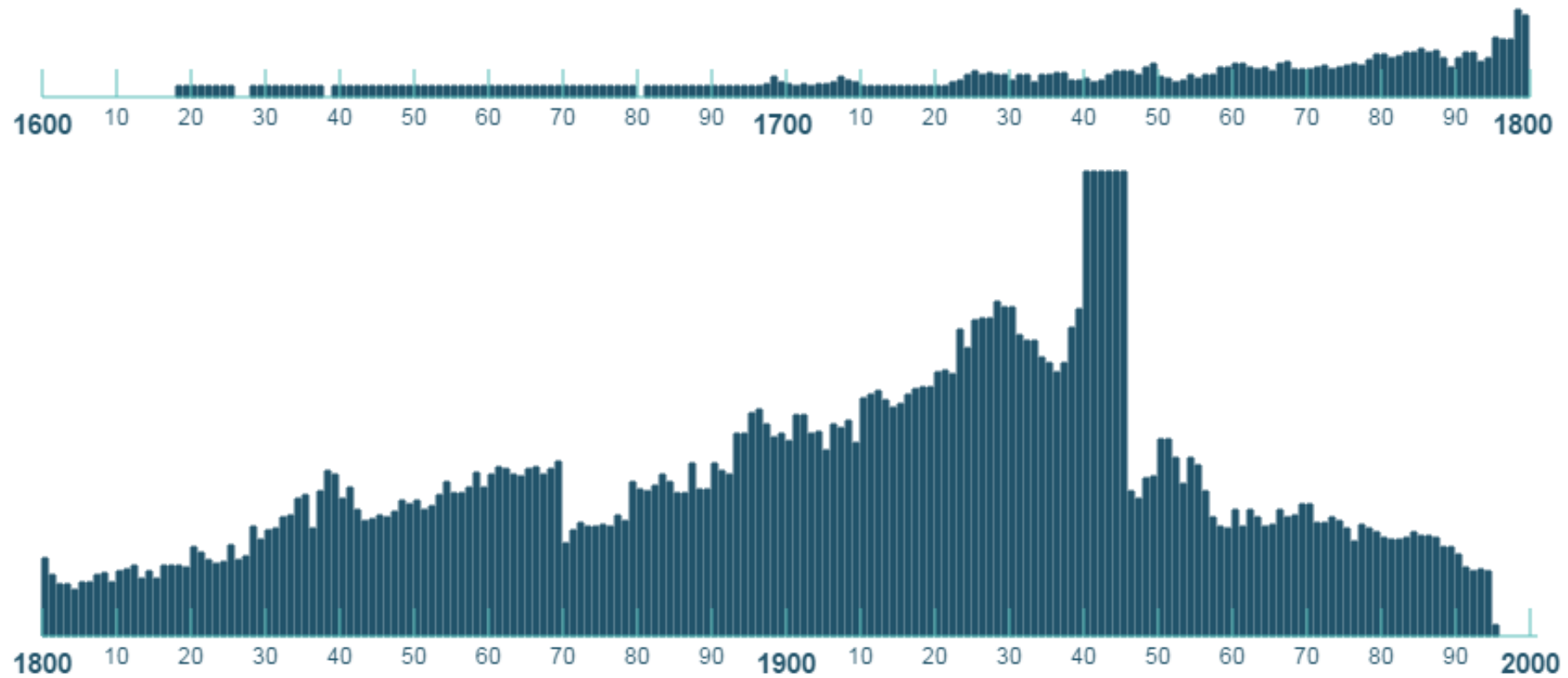
## **2. Digitised historical newspapers**



## 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](http://www.delpher.nl)





## 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

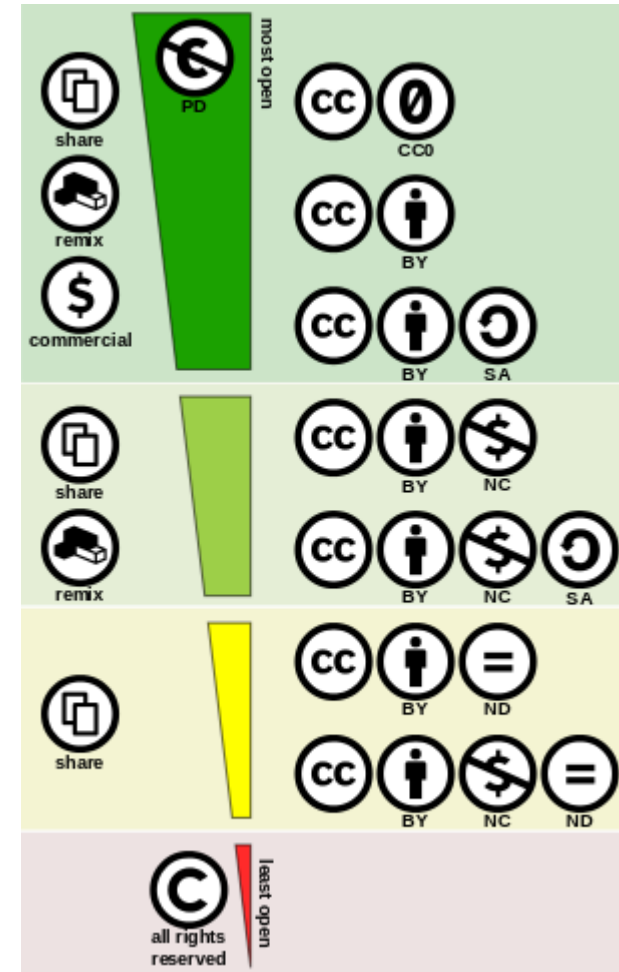
## Copyright

→ 1618 – 1876 = Public Domain

→ 1877 – 1995 = (Potentially) in copyright

*However:*

- All content online at [delpher.nl](http://delpher.nl)
- Available for research



## 3. Access



Delpher

Menu

Doorzoek alles

Zoeken in alle tekstcollecties

Handleiding

Ruim 60 miljoen pagina's uit  
Nederlandse kranten, boeken en  
tijdschriften

Vandaag 103 jaar geleden in de krant

## Delpher

- > 60 million digitised text pages
- Newspapers, books, magazines, radio bulletins
- 1618 – 1995
- Full text search
- Updated regularly
- [www.delpher.nl](http://www.delpher.nl)

Inhoud en structuur zip-bestand leest u hoe elk van deze 22 zip-bestanden opgebouwd.' At the bottom, it asks: 'Precies weten welke krantentitels, -edities, publicatiedatums en identifiers het Delpher open krantenarchief bevat? [Download dan de hele titellijst](#) (zip, 3MB. Uitgepakt 36MB, tsv = [tab seperated value](#))'." data-bbox="43 182 409 818"/>

Delpher Zoeken in alle tekstcollecties Menu

## Delpher open krantenarchief

### Wat zit er in het Delpher open krantenarchief? ^

Dit archief bevat in totaal 22 zip-bestanden:

- [1 zip-bestand](#) met 17e eeuwse kranten, 1618 t/m 1699.
- 10 zip-bestanden met 18e eeuwse kranten, 1700 t/m 1799, opgedeeld per 10 jaar.
- 11 zip-bestanden met 19e eeuwse kranten, 1800 t/m 1876, opgedeeld per 10 jaar (1800-1849) of per 5 jaar (1850-1876).

Onder [Inhoud en structuur zip-bestand](#) leest u hoe elk van deze 22 zip-bestanden opgebouwd.

Precies weten welke krantentitels, -edities, publicatiedatums en identifiers het Delpher open krantenarchief bevat? [Download dan de hele titellijst](#) (zip, 3MB. Uitgepakt 36MB, tsv = [tab seperated value](#))

## Delpher data

- 22 downloadable zip files
- All newspapers of 1618 - 1876
- Metadata + OCR + ALTO
- CC-BY
- Unzipped: 622 GB
- [www.delpher.nl/data](http://www.delpher.nl/data)

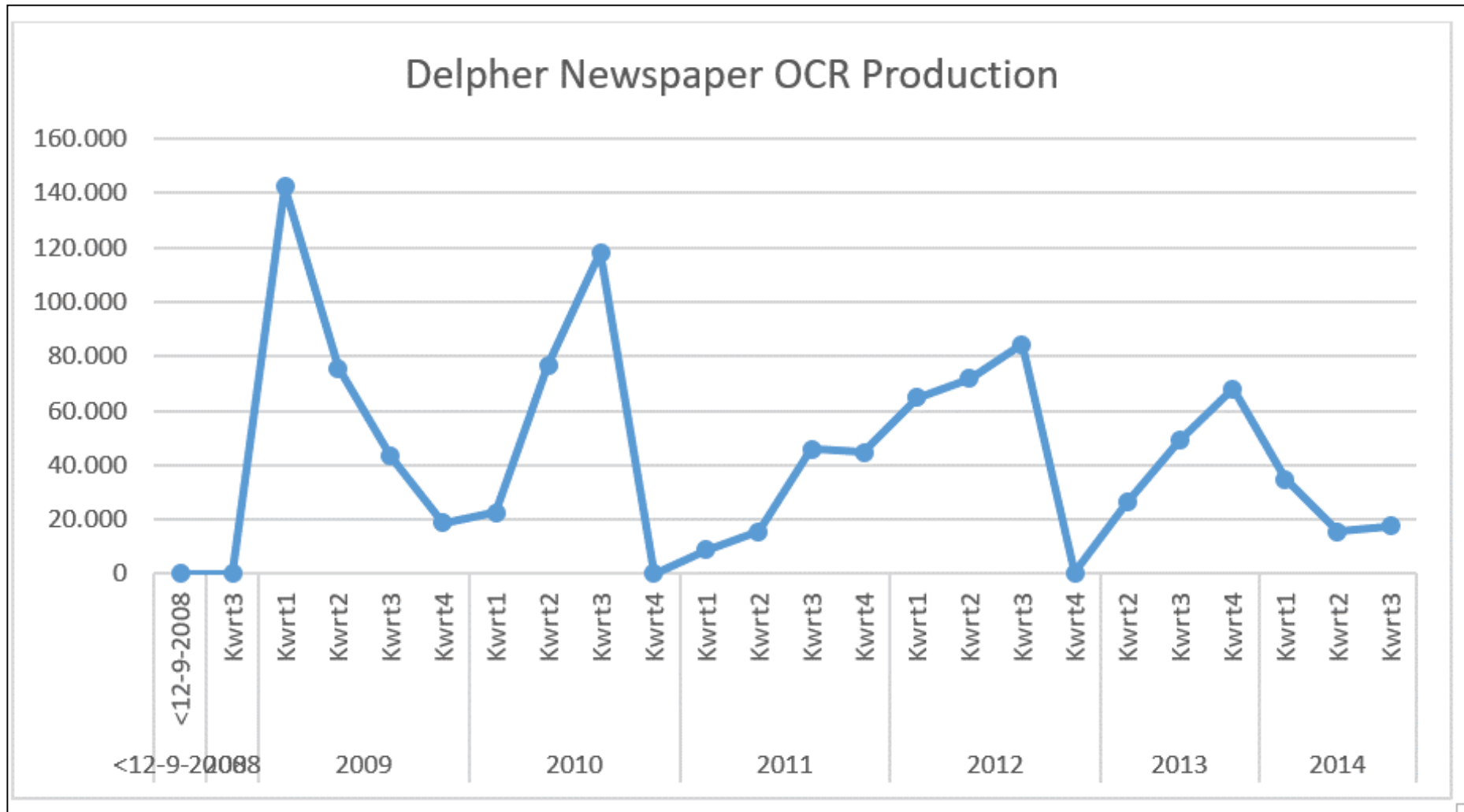
## Data services

- Access to data via APIs
- Open data
  - Set descriptions
  - Manual to access via API
- Data in copyright
  - Contact [dataservices@kb.nl](mailto:dataservices@kb.nl)
  - Access for research
- [www.kb.nl/dataservices](http://www.kb.nl/dataservices)

The screenshot shows the top navigation bar of the KB website with the text 'Koninklijke Bibliotheek National Library of the Netherlands'. Below the navigation bar, there are two dropdown menus: '& research guides' and 'Data services & APIs'. The main content area has a yellow header with the title 'Data services & APIs'. The text below the title states: 'KB data are available for research and other purposes. Digital illustrations, metadata and texts can often be made available through an API (Application Programming Interface) using [SRU](#) or [OAI-PMH](#). You can apply our data for new research, web applications and other services.' Under the heading 'Open data sets', it lists three data sets: 'Early Dutch Books Online (EDBO or DPO) (Dutch only) in cooperation with the university libraries of Amsterdam (UvA) and Leiden', 'Medieval Illuminated Manuscripts in cooperation with Museum Meermanno | Huis van het Boek', and 'The Dutch Digital Parliamentary Papers (Dutch only) in cooperation with the House of Representatives'.



## 4. OCR



## OCR correction

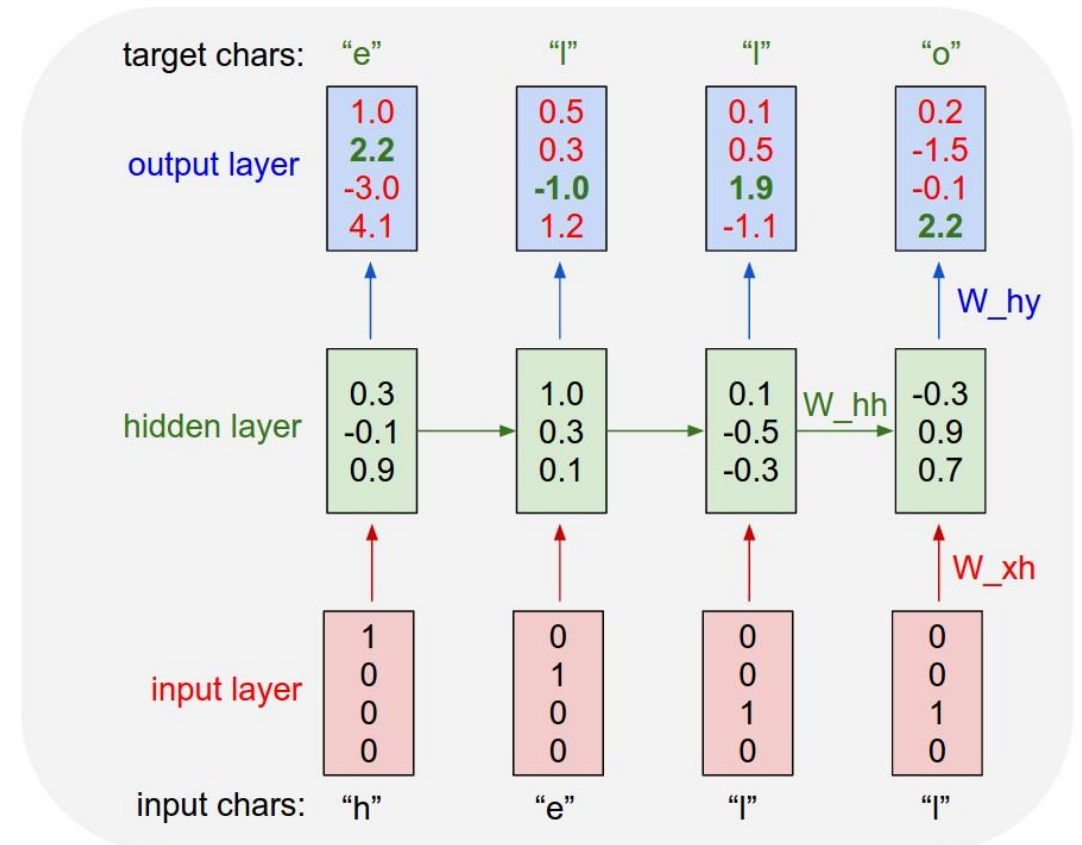
- Crowdsourced correction
- 17th century newspapers
- Partner: Meertens Instituut
- WWII newspapers
- Partner: NIOD

The screenshot displays the 'KB Kranten' interface. At the top, it shows '2 afbeeldingen' with thumbnails 'afb.1' and 'afb.2', and the title 'Tijdinghe uyt verscheyde quartieren' dated 'Wt Ceulen den 5. dito.'. Below this is a preview of a newspaper page with Dutch text in a historical font. To the right of the preview is a correction tool with a toolbar containing 'B', 'I', 'inspringen', and a dropdown arrow. A list of correction suggestions is shown on the right, each with a checkbox:

- VVt Darnstadt den 2. Augusti.
- Alhier te lande openbaert hem wederom een
- schricklijck wonderleycken/ want to Ebersta
- dt/
- een halve mijle van hier/ de posten vande K
- ercke-
- deur/ alsmede die geheele muyre tot aent da
- ck toe/
- met bloet besprengt is/ gelijk sich mede a
- en allen
- 
-

## 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**

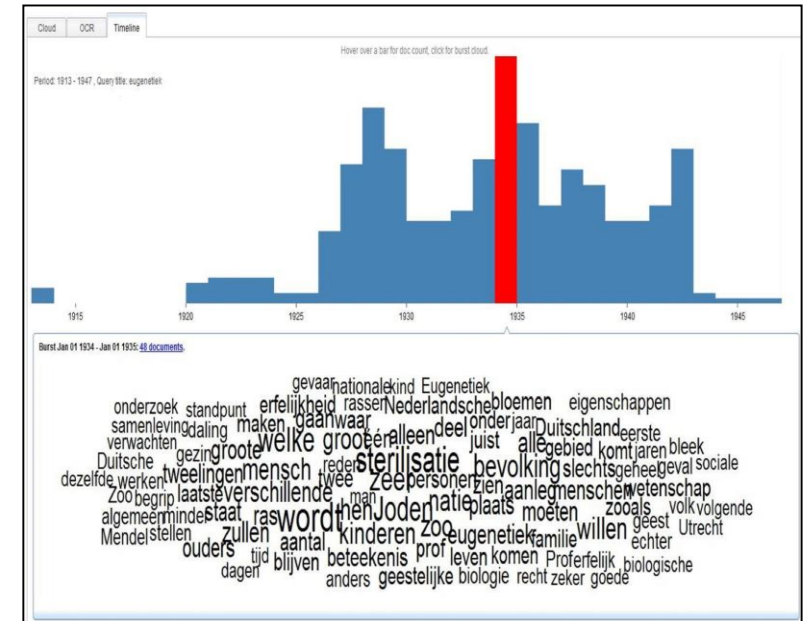
## Polimedia

- 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/>

The screenshot shows the Polimedia website interface. At the top, there is a search bar containing the text 'Oliecrisis' and a 'Zoek' button. Below the search bar, there are navigation links: 'Terug naar zoekresultaten', '1977-02-03', 'Handelingen Tweede Kamer 1976-1977 03 februari 1977, Pagina's 2953-3020.', and '▼ Sprekers'. The 'Sprekers' section lists several names with their respective counts: Abraham Stemerdink (14), Adrianus Ploeg (1), Albertus Johannes Verbrugh (1), Anne Vondeling (28), Annelien Kappeyne van de Coppello (5), Arie Kosto (1), Bastiaan de Gaay Fortman (4), and Abraham Stemerdink (14). The main content area displays a search result for 'Abraham Stemerdink' with a snippet of text: 'Mijnheer de Voorzitter ! Ik ben het eens met de laatste opmerking van de heer Van Win kel . Zijn vragen 1 , 2 en 3 kan ik met ' ja ' beantwoorden , maar vraag 4 moet iets genuanceerder worden benaderd . Het is de militair verboden , bepaalde algemeen aangeduide handelingen te verrichten . De kwestie is , wie in eerste instantie beoordeelt , of bepaalde handelingen al dan niet onder dit verbod vallen . In beginsel is'. To the right of the main content area, there is a sidebar with a search bar and a list of results, including 'ANP stemerdink heeft in de tweede kamer bevestigd, dat schout bij nacht \ ^ na' and 'De Arbeiderspers medewerking release mag marine-commandant tot de orde geroepen'.

## 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

The screenshot shows the KB LAB website interface. At the top, there is a dark navigation bar with the KB LAB logo and links for Datasets, Tools, News & events, Blog, and About us. On the right side of the navigation bar, there are options for LIGHT/DARK theme and a search icon. Below the navigation bar is a large yellow banner with the text: "Join us and explore the KB's digital treasure trove" and "The KB Lab hosts all experimental tools and data sets based on the KB's digitised collection." Below the banner, there are three statistics: "4.878 lines of code", "40.330 MB files", and "5 events". The main content area is divided into two sections: "Datasets" and "Tools". Under "Datasets", there is a card for "KBK-1M" which features a yellow background with a mountain icon and the text: "The KBK-1M Dataset is a collection of 1,603,396 images and accompanying captions of the period 1922 – 1994". Under "Tools", there are two cards: "Frame generator" which is described as a "Tool for extracting topics, keywords and their co-occurrence patterns from a Dutch corpus" and "DBNL ngram viewer" which is described as "An ngram viewer counting terms and phrases in the Digital Library of Dutch Literature [DBNL]".

[www.lab.kb.nl](http://www.lab.kb.nl)

# Cooperation via KB initiated programs

## 1. Fellowship Program

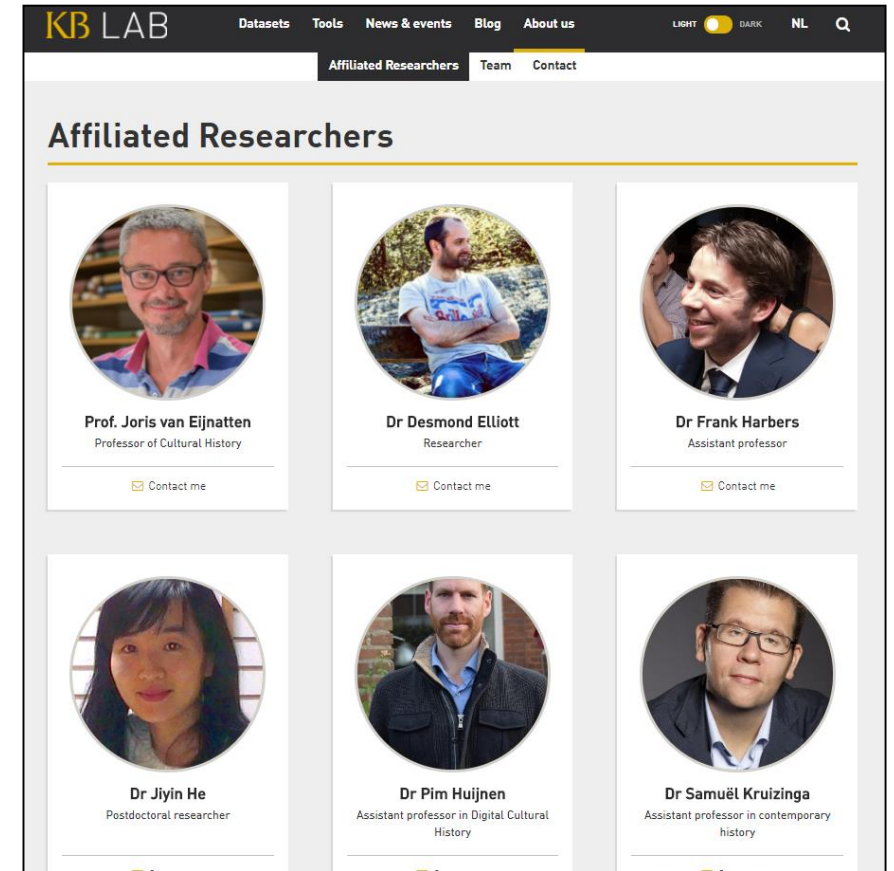
= 4 months, 1 fte | invited | successful academic

## 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?”

The screenshot shows the 'Frame generator' web application. At the top, there is a yellow header with the title 'Frame generator' and a share icon. Below the header are navigation tabs: 'Introduction', 'Live demo', 'Instructions', 'Examples', and 'Comments'. The 'Live demo' tab is active. Below the tabs, there is a link to 'open tool in new tab'. The main content area is titled 'Frame generator' and includes a 'KB Lab' logo in the top right corner. On the left side, there is a text box explaining the tool's function: 'The frame generator calculates the most important keywords for one or more Dutch text files, based on topic modeling techniques. It subsequently finds the words that are most strongly associated with each keyword, forming so-called frames.' Below this text are instructions for using the controls to upload documents and determine which part-of-speech tags are to be included, as well as in which direction of the keyword and to what distance related frame words are to be searched for. There are two columns of checkboxes for 'Keyword tags' and 'Frame tags'. The 'Keyword tags' column has checkboxes for Adjective [ADJ], Noun [N], Names [SPEC], and Verb [WW]. The 'Frame tags' column has checkboxes for Adjective [ADJ], Noun [N], Names [SPEC], and Verb [WW]. Below the checkboxes are dropdown menus for 'Window size' (set to 5) and 'Window direction' (set to Left). At the bottom of the controls are 'Upload files' and 'Generate' buttons. On the right side of the interface, there is a network diagram showing several clusters of words. The largest cluster is centered around 'cubaans' and 'leider', with other words like 'revolusionair', 'betwist', 'gezichtsbepalend', 'twitter', 'fidel castro', 'straat', 'cuba', 'medeleven', 'belangrijk', 'oud-leider', 'geschiedenis', 'gewezen', 'man', 'sterk', 'wijze', 'krachtig', 'staatstelevisie', 'jaar', 'zuma', 'trump', 'castro', 'afschied', 'president', 'dood', 'hoop', 'dictatuur', 'mens', 'totalitair', and 'man' connected to it. Other smaller clusters are visible around 'man' and 'mens'.

**KBK-1M**

Introduction Access Examples Comments

The KBK-1M Dataset ('Koninklijke Bibliotheek Kranten - 1 Miljoen') is a collection of 1,603,396 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](http://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 'illustraties met onderschrift' (illustrations with caption) 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.

DEVELOPED BY

- ▶ Martijn Kleppe
- ▶ Desmond Elliott
- ▶ Willem Jan Faber

CONTENT

Newspaper Image

CATEGORY

Data access

FILE FORMAT

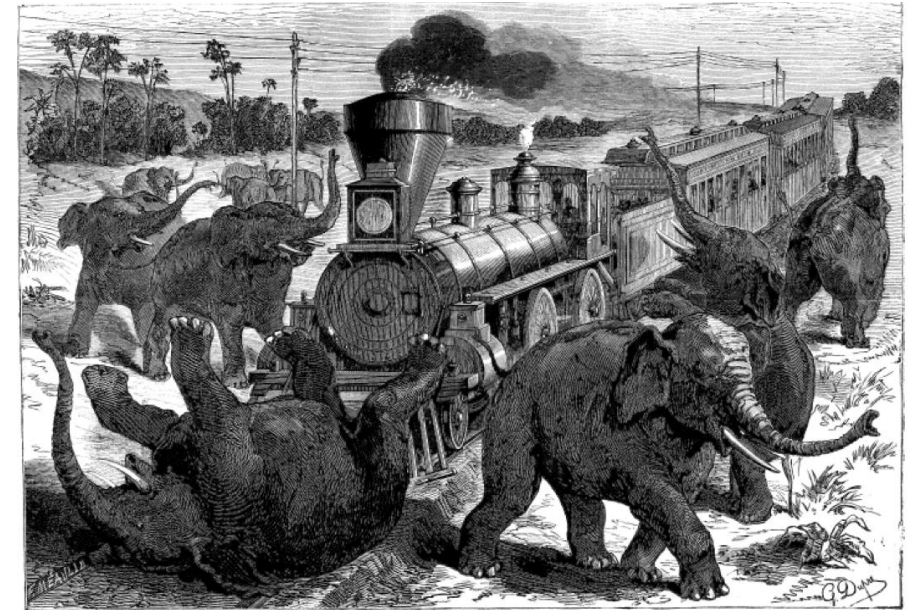
JPEG JSON

## 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 INDIE. — AANVAL VAN EEN KUDDE OLIFANTEN OP EEN SPOORTREIN. (Zie blz. 6.)

The screenshot shows the GitHub organization page for 'National Library of the Netherlands / Research'. At the top, there are navigation links for 'Features', 'Business', 'Explore', and 'Pricing', along with a search bar and 'Sign in or Sign up' options. The organization's profile includes a logo, name, location ('The Hague, The Netherlands'), website, and email. Below this, there are tabs for 'Repositories' and 'People 6'. A search bar for repositories is present, along with filters for 'Type: All' and 'Language: All'. The main content area displays two repositories: 'chatbot-builder-nl' (JavaScript, updated 10 minutes ago) and 'omSipCreator' (Python, 5 stars, updated an hour ago). A 'code4lib' badge is also visible. To the right, there are sections for 'Top languages' (Python, JavaScript, HTML, Java, XSLT) and 'People 6' with profile pictures.

## 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

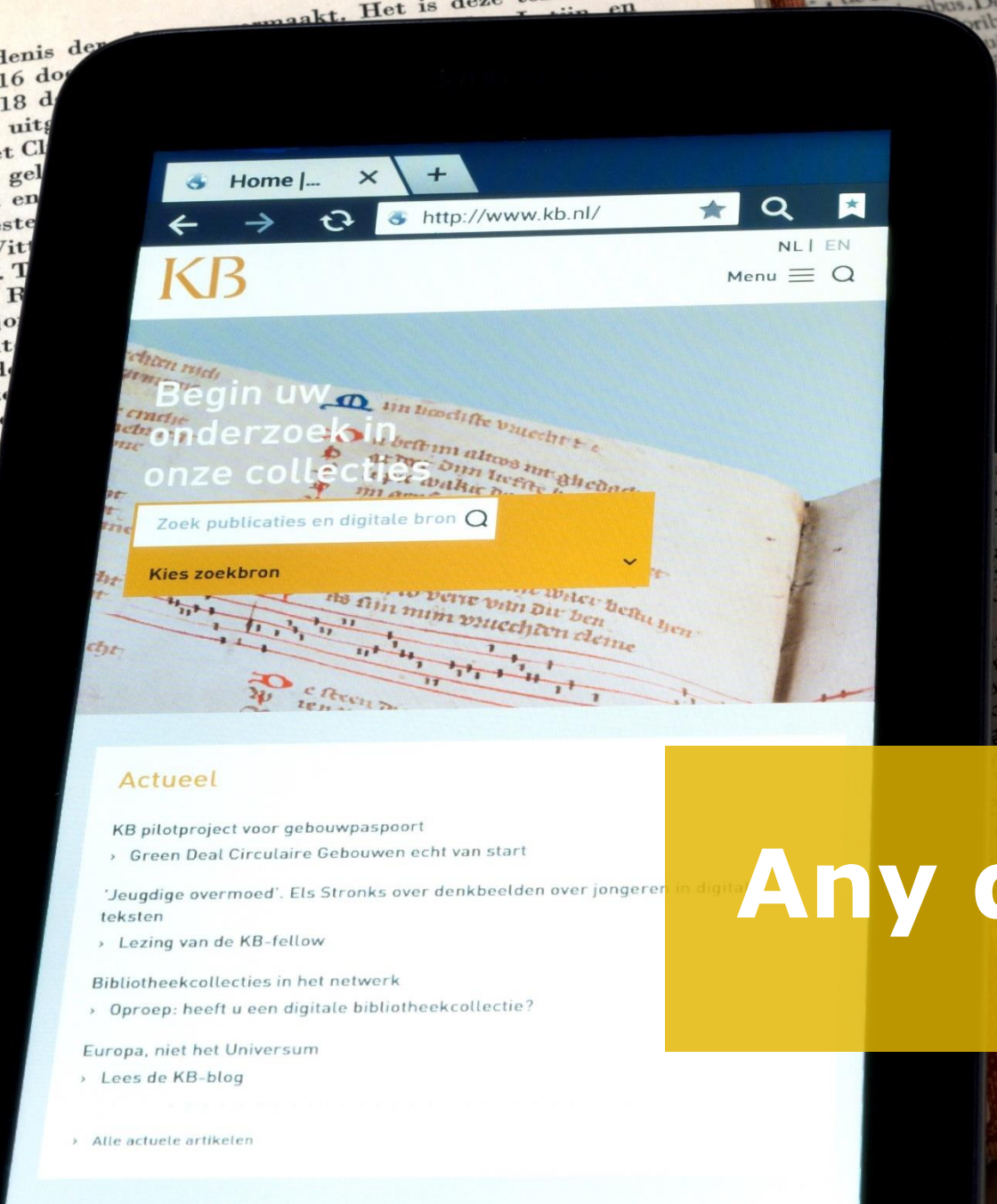
um odoratarumque nonnullarum  
Purgantium historiae libri IV,  
riae pemptades sive libri XXX.  
et groot Cruydtboeck, hetwelk  
zouden uitgeven en in gansch  
Plantin had zich langzamer-  
an al de houtblokken, die bij  
kers voor andere herbariums  
gd bij die van zijn eigen uit-  
lustratiemateriaal uit zonder  
het Museum Plantin Moretus  
van Pieter van der Borcht, in  
overwegend en niemand min-  
roote kunst, die de Mechelsche  
ag legde. „Het is mij”, schrijft  
en tijd in het bekijken van de  
nydtboeck van Dodoens ver-  
ik, daarna in den tuin rond-  
en bloem en heester geschaard  
eidscher dos meende te zien  
der takken, de levendigheid  
ormatie der wuivende kronen,  
voor mij geopenbaard, en ik  
ren in den eeuwig afwisselen-  
as. De prenten van Dodoens  
nepping duidelijker en dieper  
as de trouwe illustratie van  
nter der levende natuur ge-

tgever door de banden eener  
den. Toen Dodoens in 1584  
gedurende een paar jaren  
was, liet Plantin dit met  
emeenschappelijken vriend

, als blijk van vriendschap,  
verbeterden en vermeerder-  
dschen tekst van zijn ge-

schiedenis der  
in 1616 do  
in 1618 d  
werd uitg  
Met Cl  
Deze gel  
Gent en  
geweste  
te Wit  
ging. T  
van R  
de jo  
plant  
op de  
hij te  
wege  
bloed  
de l  
ver  
na  
vr  
tr  
k  
t  
g

maakt. Het is deze tekst, die  
in hoofdste vuercht  
best in altes m ghedoe  
Dijn wette  
m am  
wace beku hen  
sijn mijn vuerchten dene  
e steen



Any questions?

### Actueel

- KB pilotproject voor gebouwspaspoort
  - › 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

145  
34 plaats, in uitvoerige ac-  
d (nots. Knoll 1734 6 April.  
De nu volgende jaren staan in  
aise in het land en op de Haag-  
ise in den jaren 1737 tot 1747 staakte  
st „door ongelukkige tijden en  
ppers”. Ten einde aan het onheil  
den uitweg om — gebruik ma-  
n gecontroleerde „aucties onder de  
erkochte fondsen onderling in vervol-  
obligaties te betalen, en die bankiers.  
e transporteren aan de bankiers.  
nd, te transporteren van geld. Hierbij schijnt  
hebben gespeeld. Hij en zijn zoon Pieter,  
ing, die tegen 1740 uitliepen op de „Compag-  
ies, die Block, Swart, Beauregard, Moetjens.  
Gosse, hun speculatiën faalden. In 1744  
e de vier anderen benoemden om toezicht te hou-  
meischers personen moesten zich laten liquideeren en  
oedel en bewerken der fondsen.” Gosse had reeds in  
gegeven zijn zaken in Duitschland te liquideeren op zijn  
cht hij het tot stand dat heel zijn bezit, huis en  
ligaties natuurlijk, werd overgebracht op zijn  
Compagnie. — De oudste zoon, Henri Albert was  
de generatie van een belangrijken boekhandel, den Haag  
vestigd als chef de boutique du dit Libraire  
ter, Maria Catharina, ses fils, tous Nicolaas  
koopers voegen „la boutique de Genève et de  
Leipzig en de stad te Genève moest borg  
a Paris ook g. Riegers en de stad te Genève niet  
en er nog bijk. Van de stad te Genève niet  
den Haag). Van de stad te Genève niet  
zijn fonds dook J. C. Riegers en de stad te Genève niet  
Maar niettemin in Den Haag het blijkbaar voor be-  
schonoonzoon, die langzamerhand tot 51.000 zijn  
in de familie, die hield hij het blijkbaar voor be-  
are beslissingen in Den Haag het blijkbaar voor be-  
zoon Pieter zijn eigen huishouding had o-  
verliet hij de stad. Terwijl zijn vrou-  
in de oude omgeving bleven, brach  
zijn zonen in Genève door, en is a-  
khand