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

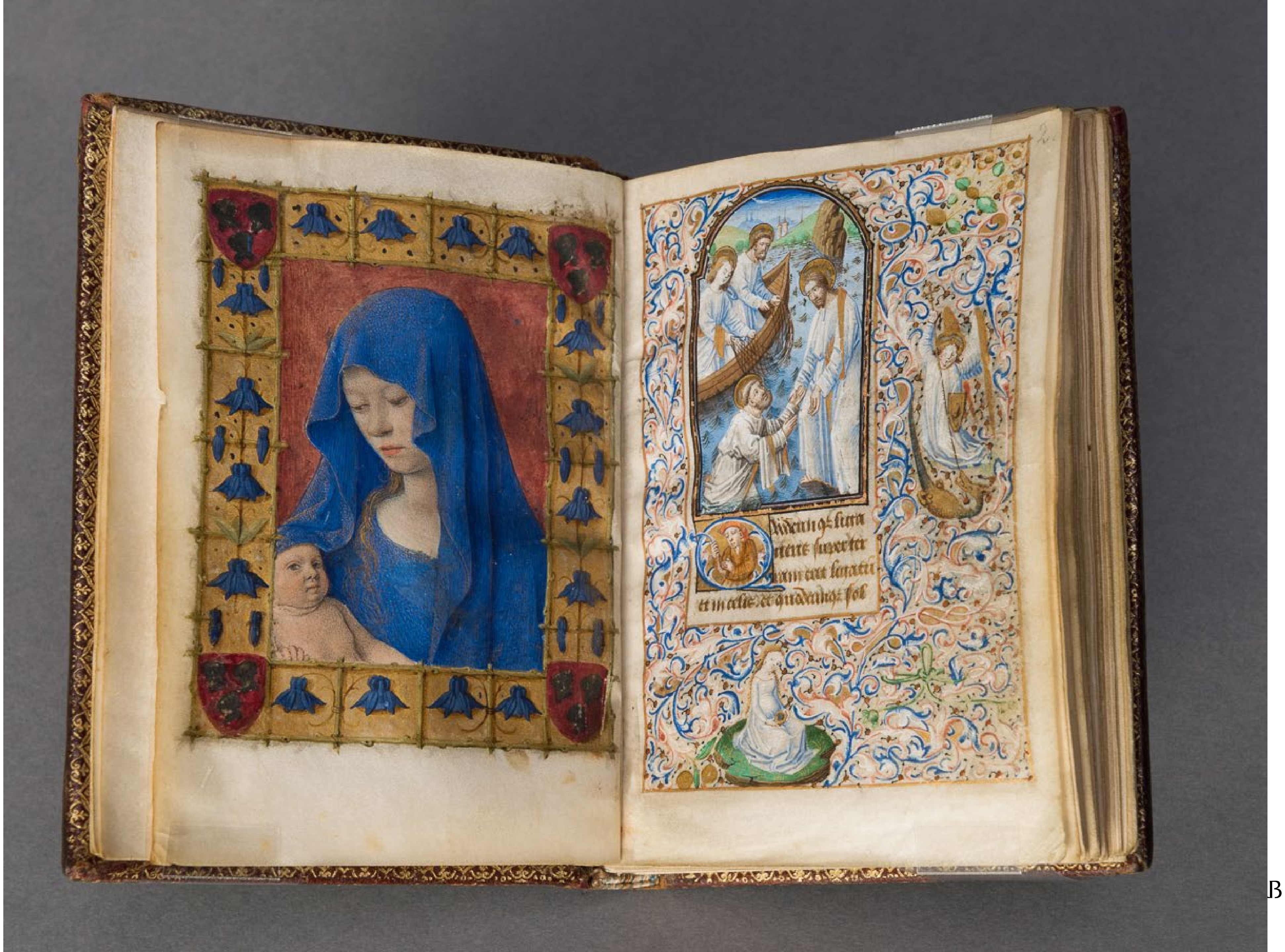
Dr. Martijn Kleppe
Board Member Research & Discovery
KB, national library of the Netherlands

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

KB nationale bibliotheek

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


2.
Videntes gl[illegible] signa
[illegible]icate super ter
[illegible]am cum signaturi
et in celis et quadraginta sol

158

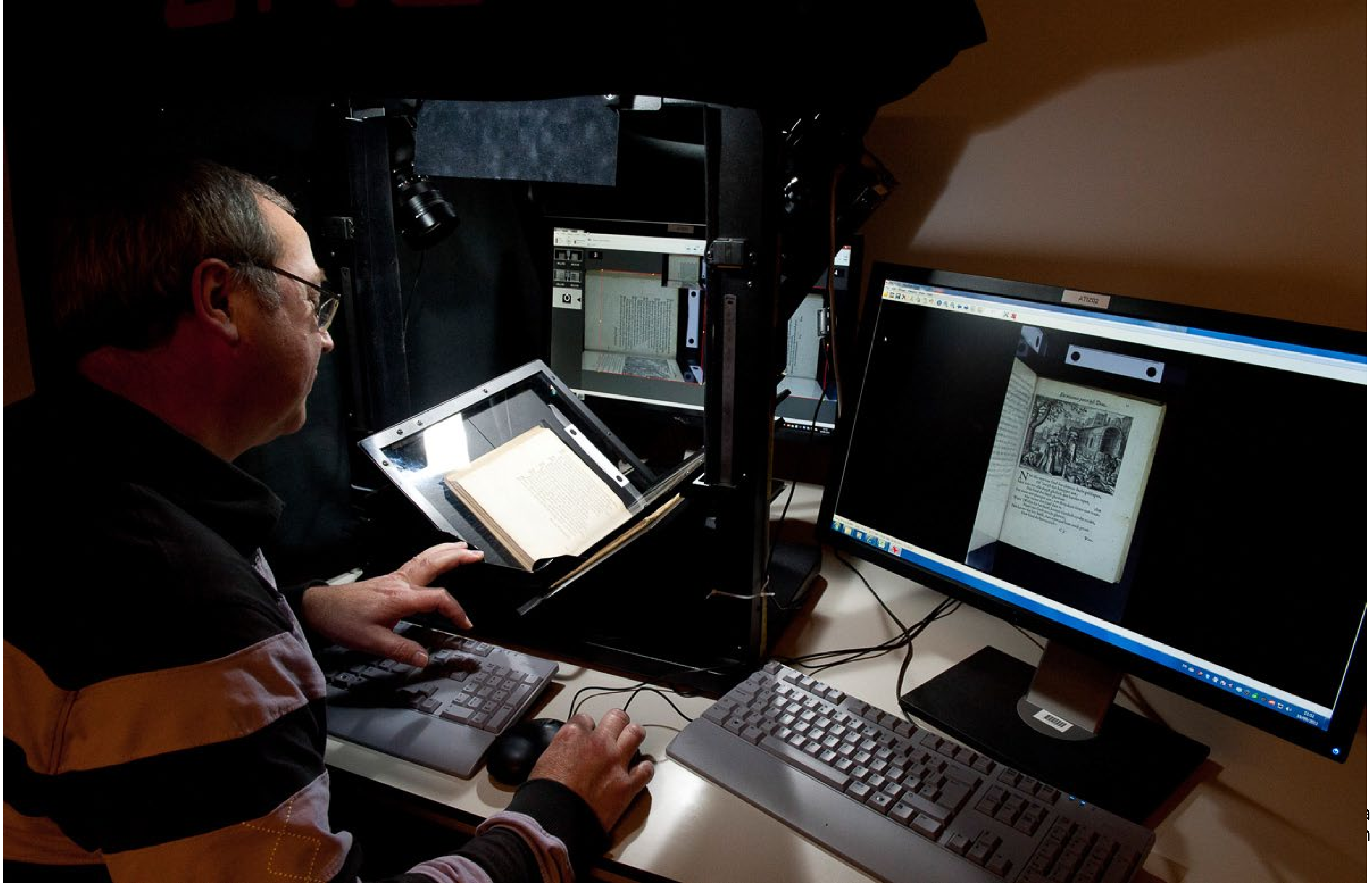
7 million items

120 kilometres publications

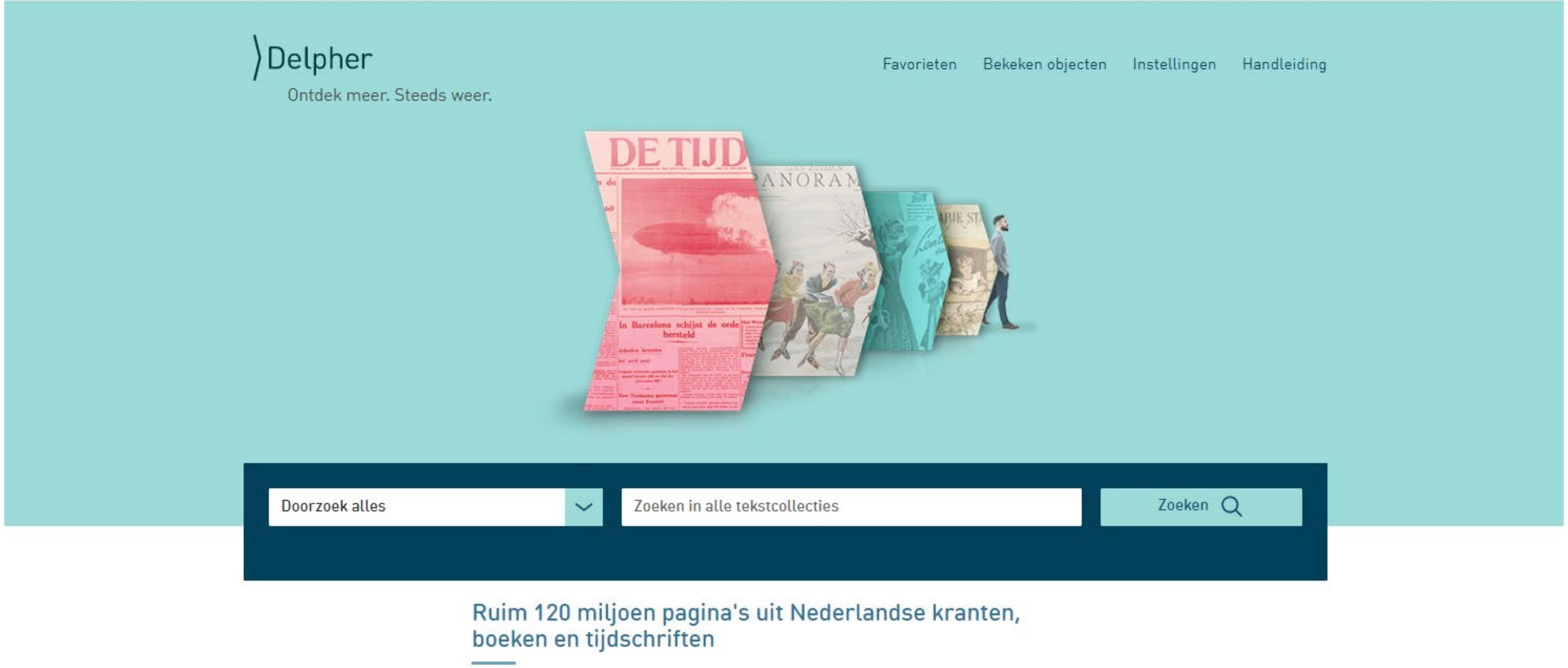
nationale
bibliotheek

aansCreoolsAn
liaansKaapver
aansEritreesIr
nsTurksMarok

le
eek



ale
heek



Ruim 120 miljoen pagina's uit Nederlandse kranten, boeken en tijdschriften

Full text (OCR) access to:

✓1.000.000 books (1486 – 2013)

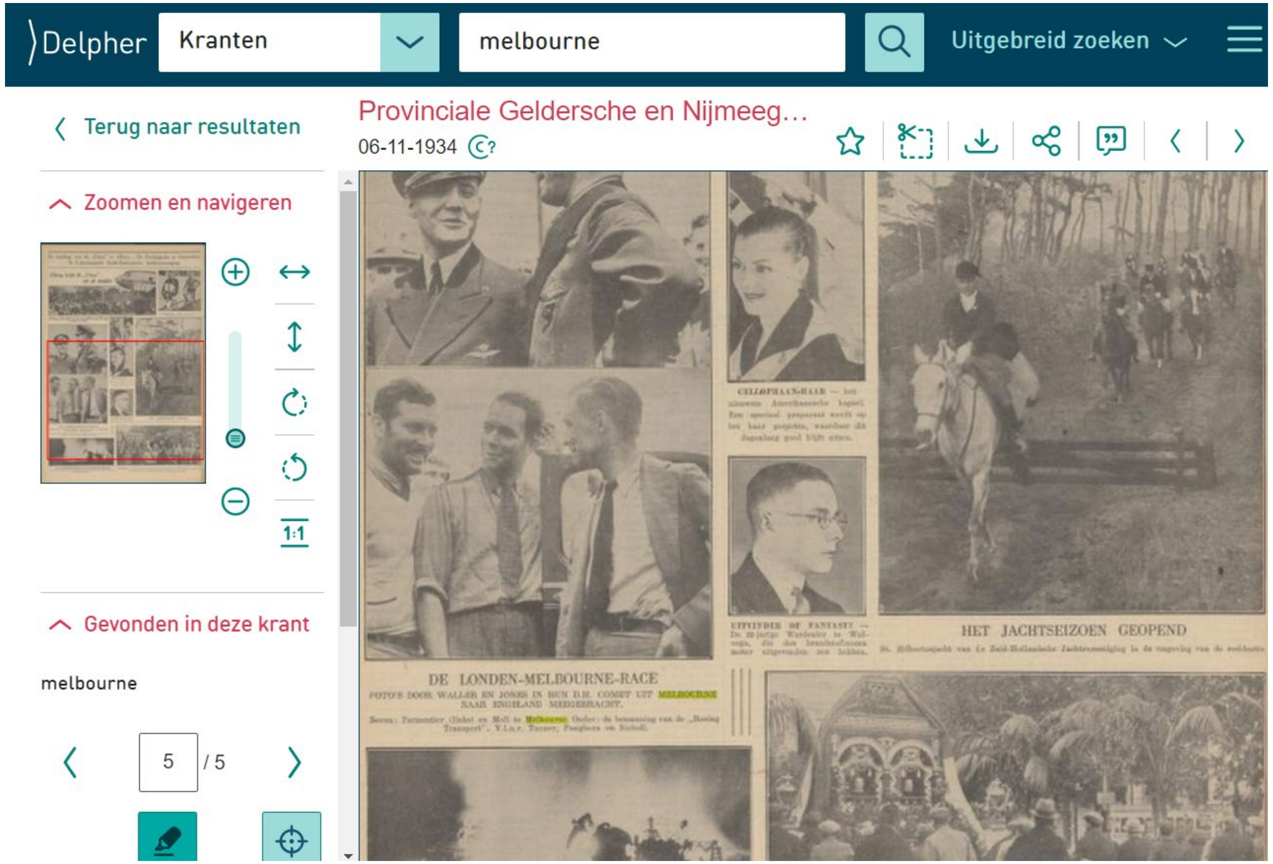
✓1.700.000 newspapers (1618 – 1995)

✓10.000.000 periodical pages (1840 – 1940)

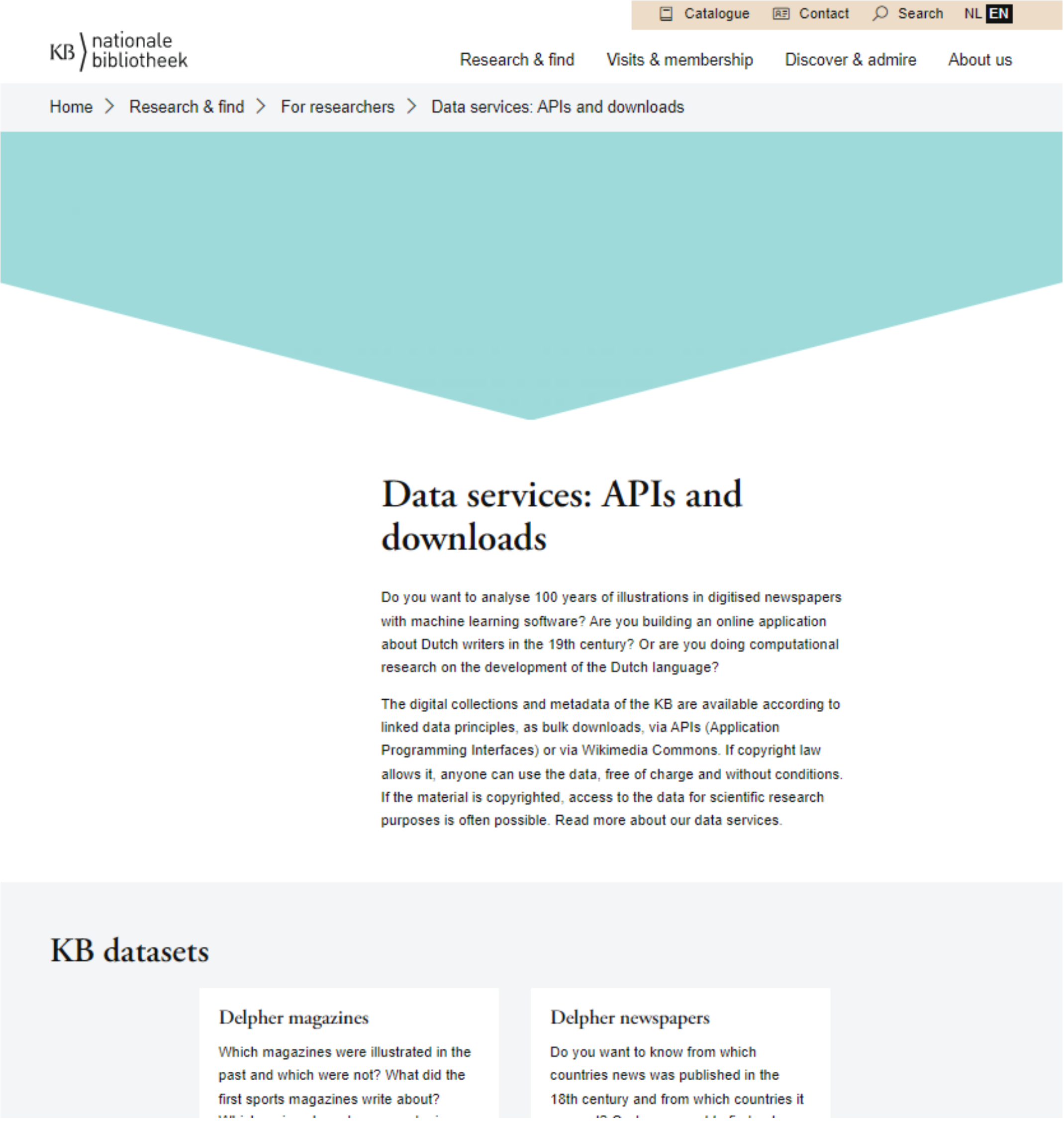
✓1.500.000 million ANP radio items (1937 – 1984)

https://www.delpher.nl/

KB nationale bibliotheek



www.delpher.nl



Data services: APIs and downloads

Do you want to analyse 100 years of illustrations in digitised newspapers with machine learning software? Are you building an online application about Dutch writers in the 19th century? Or are you doing computational research on the development of the Dutch language?

The digital collections and metadata of the KB are available according to linked data principles, as bulk downloads, via APIs (Application Programming Interfaces) or via Wikimedia Commons. If copyright law allows it, anyone can use the data, free of charge and without conditions. If the material is copyrighted, access to the data for scientific research purposes is often possible. Read more about our data services.

KB datasets

Delpher magazines

Which magazines were illustrated in the past and which were not? What did the first sports magazines write about?

Delpher newspapers

Do you want to know from which countries news was published in the 18th century and from which countries it

www.kb.nl/dataservices

Dataservices

= result of

more than 200 years of collecting

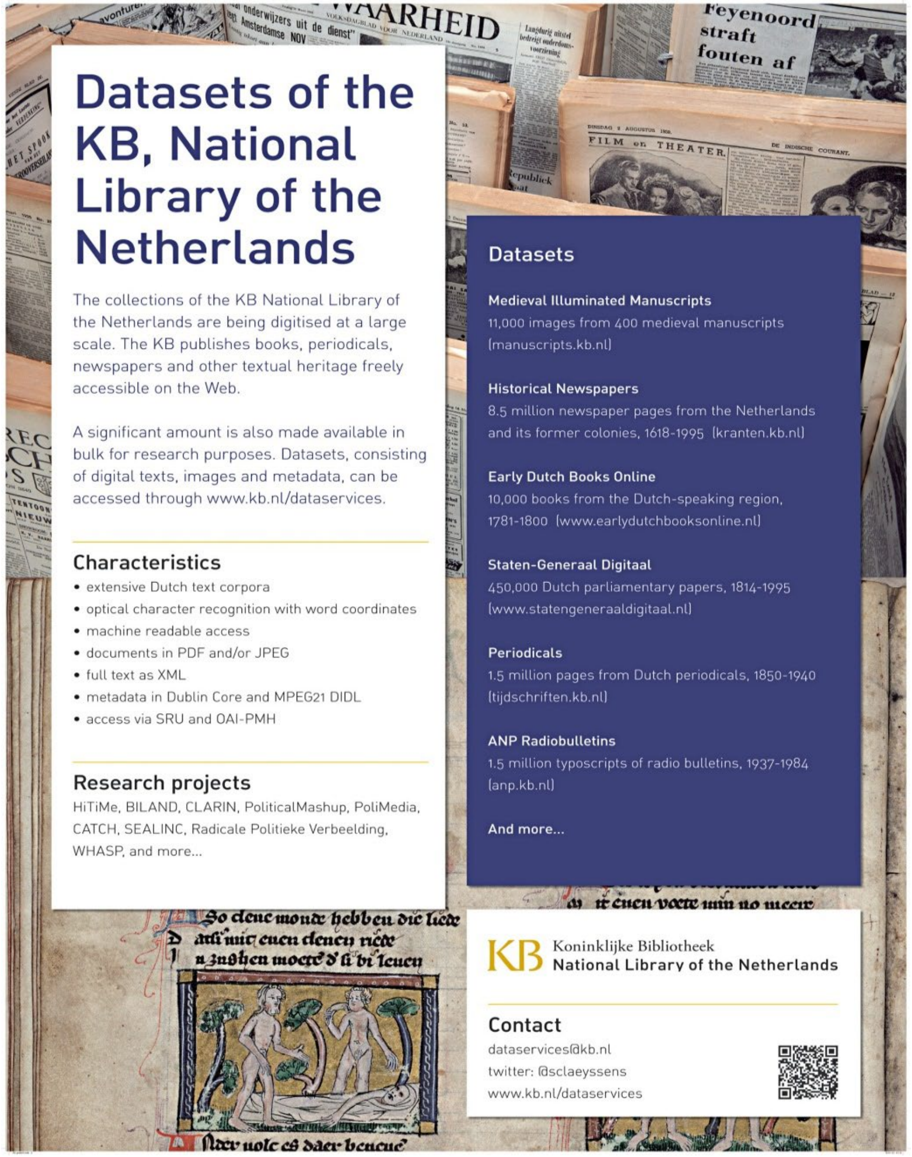
over 30 years of digitisation

10+ years of collecting born-digital publications

= machine readable, mostly textual

= structured or semi-structured

= legally as open as possible



Datasets of the KB, National Library of the Netherlands

The collections of the KB National Library of the Netherlands are being digitised at a large scale. The KB publishes books, periodicals, newspapers and other textual heritage freely accessible on the Web.

A significant amount is also made available in bulk for research purposes. Datasets, consisting of digital texts, images and metadata, can be accessed through www.kb.nl/dataservices.

Characteristics

- extensive Dutch text corpora
- optical character recognition with word coordinates
- machine readable access
- documents in PDF and/or JPEG
- full text as XML
- metadata in Dublin Core and MPEG21 DIDL
- access via SRU and OAI-PMH

Research projects

HiTiMe, BILAND, CLARIN, PoliticalMashup, PoliMedia, CATCH, SEALINC, Radicale Politieke Verbeelding, WHASP, and more...

Datasets

Medieval Illuminated Manuscripts
11,000 images from 400 medieval manuscripts (manuscripts.kb.nl)

Historical Newspapers
8.5 million newspaper pages from the Netherlands and its former colonies, 1618-1995 (kranten.kb.nl)

Early Dutch Books Online
10,000 books from the Dutch-speaking region, 1781-1800 (www.earlydutchbooksonline.nl)

Staten-Generaal Digitaal
450,000 Dutch parliamentary papers, 1814-1995 (www.statengeneraaldigitaal.nl)

Periodicals
1.5 million pages from Dutch periodicals, 1850-1940 (tijdschriften.kb.nl)

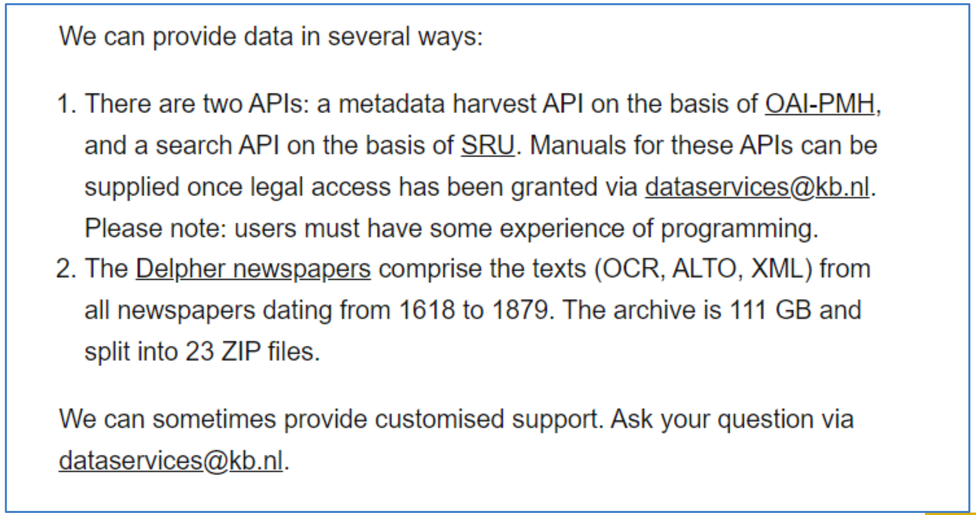
ANP Radiobulletins
1.5 million typoscripts of radio bulletins, 1937-1984 (anp.kb.nl)

And more...

KB Koninklijke Bibliotheek
National Library of the Netherlands

Contact

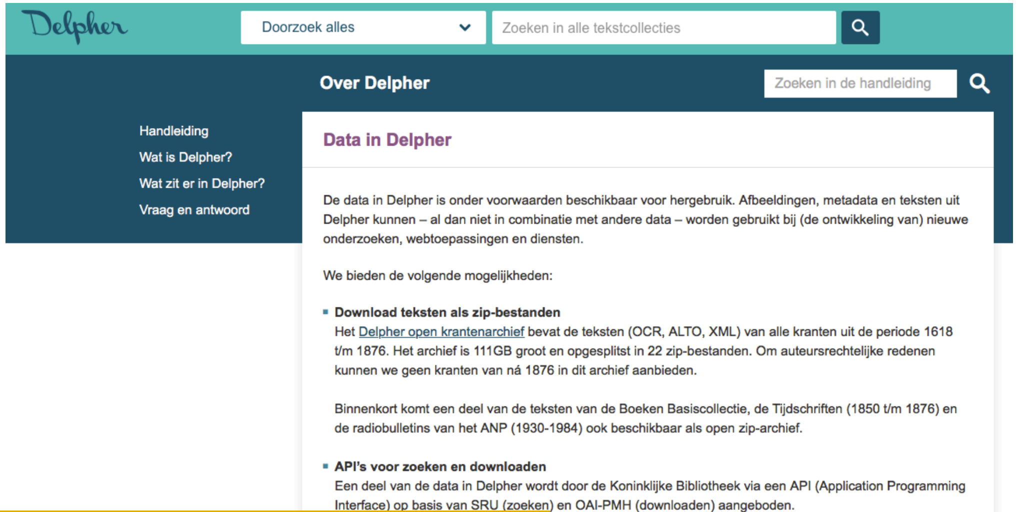
dataservices@kb.nl
twitter: @sclaeyssens
www.kb.nl/dataservices



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.



Data in Delpher

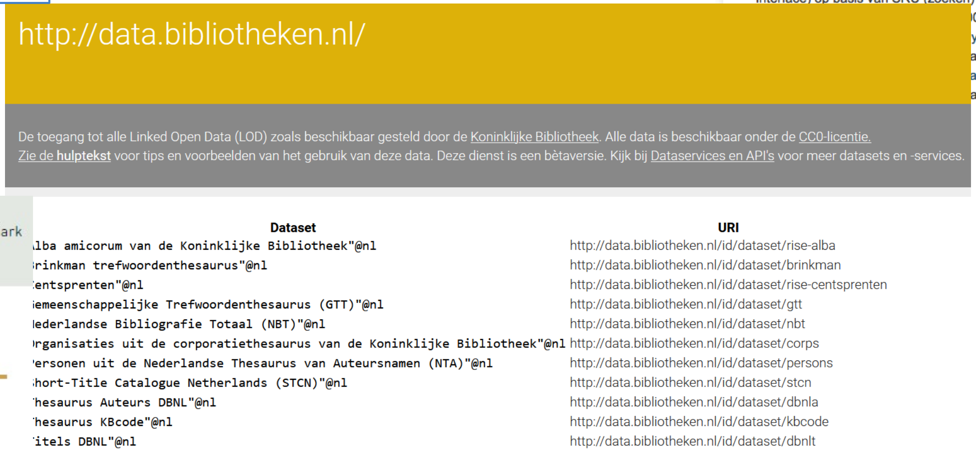
De data in Delpher is onder voorwaarden beschikbaar voor hergebruik. Afbeeldingen, metadata en teksten uit Delpher kunnen – al dan niet in combinatie met andere data – worden gebruikt bij (de ontwikkeling van) nieuwe onderzoeken, webtoepassingen en diensten.

We bieden de volgende mogelijkheden:

- **Download teksten als zip-bestanden**
  Het Delpher open krantenarchief bevat de teksten (OCR, ALTO, XML) van alle kranten uit de periode 1618 t/m 1876. Het archief is 111GB groot en opgesplitst in 22 zip-bestanden. Om auteursrechtelijke redenen kunnen we geen kranten van ná 1876 in dit archief aanbieden.

  Binnenkort komt een deel van de teksten van de Boeken Basiscollectie, de Tijdschriften (1850 t/m 1876) en de radiobulletins van het ANP (1930-1984) ook beschikbaar als open zip-archief.

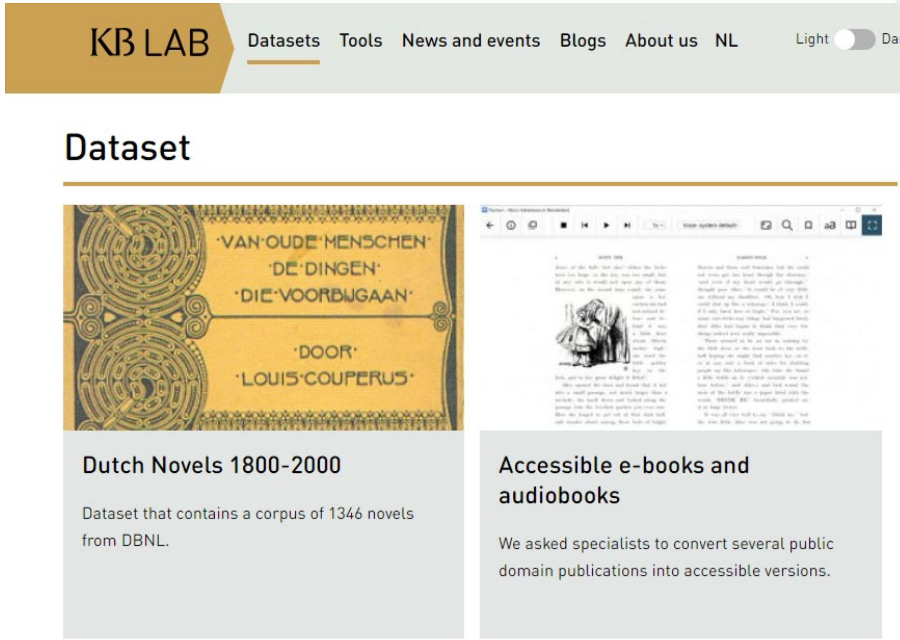
- **API's voor zoeken en downloaden**
  Een deel van de data in Delpher wordt door de Koninklijke Bibliotheek via een API (Application Programming Interface) op basis van SRU (zoeken) en OAI-PMH (downloaden) aangeboden.
  0 in de Boeken Basiscollectie worden als open dataset aangeboden, zie
  y Dutch Books Online op de KB-site.
  als dataset aangeboden volgens een CC-BY-NC-ND licentie, zie de
  adiobulletins Digitaal op de KB-site.
  a biedt de KB nog een aantal andere open datasets aan via de API.



http://data.bibliotheken.nl/

De toegang tot alle Linked Open Data (LOD) zoals beschikbaar gesteld door de Koninklijke Bibliotheek. Alle data is beschikbaar onder de CC0-licentie.
Zie de hulptekst voor tips en voorbeelden van het gebruik van deze data. Deze dienst is een bètaversie. Kijk bij Dataservices en API's voor meer datasets en -services.

Dataset	URI
lba amicorum van de Koninklijke Bibliotheek"@nl	http://data.bibliotheken.nl/id/dataset/rise-alba
rinkman trefwoordenthesaurus"@nl	http://data.bibliotheken.nl/id/dataset/brinkman
entsprenten"@nl	http://data.bibliotheken.nl/id/dataset/rise-centsprenten
emeenschappelijke Trefwoordenthesaurus (GTT)"@nl	http://data.bibliotheken.nl/id/dataset/gtt
ederlandse Bibliografie Totaal (NBT)"@nl	http://data.bibliotheken.nl/id/dataset/nbt
rganisaties uit de corporatiethesaurus van de Koninklijke Bibliotheek"@nl	http://data.bibliotheken.nl/id/dataset/corps
ersonen uit de Nederlandse Thesaurus van Auteursnamen (NTA)"@nl	http://data.bibliotheken.nl/id/dataset/persons
hort-Title Catalogue Netherlands (STCN)"@nl	http://data.bibliotheken.nl/id/dataset/stcn
hesaurus Auteurs DBNL"@nl	http://data.bibliotheken.nl/id/dataset/dbnla
hesaurus KBcode"@nl	http://data.bibliotheken.nl/id/dataset/kbcode
itels DBNL"@nl	http://data.bibliotheken.nl/id/dataset/dbnlt


KB LAB
Datasets
Tools
News and events
Blogs
About us
NL
Light
Dark
VAN OUDE MENSCHEN DE DINGEN DIE VOORBIJGAAN DOOR LOUIS COUPERUS

Dataset

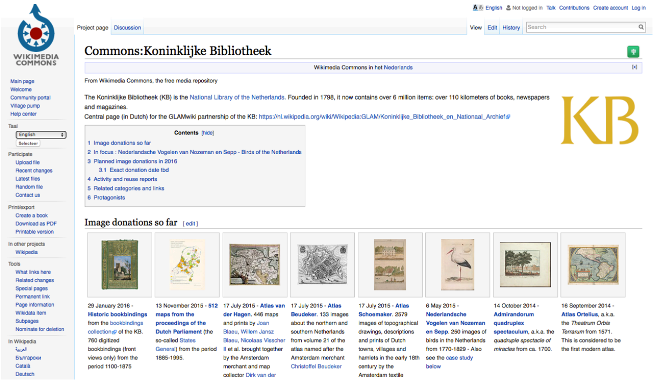
Dutch Novels 1800-2000

Dataset that contains a corpus of 1346 novels from DBNL.

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/data-services-apis-and-downloads



Commons:Koninklijke Bibliotheek

Wikimedia Commons in het Nederlands

From Wikimedia Commons, the free media repository

The Koninklijke Bibliotheek (KB) is the National Library of the Netherlands. Founded in 1798, it now contains over 6 million items: over 110 kilometers of books, newspapers and magazines.
Central page (in Dutch) for the GLAMwiki partnership of the KB: https://nl.wikipedia.org/wiki/Wikipedia:GLAM/Koninklijke_Bibliotheek_en_Nationaal_Archief

Contents [hide]
1 Image donations so far
2 In focus : Nederlandsche Vogelen van Nozeman en Sepp - Birds of the Netherlands
3 Planned image donations in 2016
3.1 Exact donation date tbd
4 Activity and reuse reports
5 Related categories and links
6 Protagonists

Image donations so far [edit]

29 January 2016 - Historic bookbindings from the bookbindings collection of the KB. 760 digitized bookbindings (front views only) from the period 1100-1875

13 November 2015 - 512 maps from the proceedings of the Dutch Parliament (the so-called States General) from the period 1885-1995.

17 July 2015 - Atlas van der Hagen. 446 maps and prints by Joan Blaeu, Willem Jansz Blaeu, Nicolaas Visscher II et al. brought together by the Amsterdam merchant and map collector Dirk van der

17 July 2015 - Atlas Beudeker. 133 images about the northern and southern Netherlands from volume 21 of the atlas named after the Amsterdam merchant Christoffel Beudeker

17 July 2015 - Atlas Schoemaker. 2579 images of topographical drawings, descriptions and prints of Dutch towns, villages and hamlets in the early 18th century by the Amsterdam textile

6 May 2015 - Nederlandsche Vogelen van Nozeman en Sepp. 250 images of birds in the Netherlands from 1770-1829 - Also see the case study below

14 October 2014 - Admirandorum quadruplex spectaculum, a.k.a. the *quadruple spectacle of miracles* from ca. 1700.

16 September 2014 - Atlas Ortelius, a.k.a. the *Theatrum Orbis Terrarum* from 1571. This is considered to be the first modern atlas.

PoliMedia

humanities and social sciences

Mining Shifting Concepts Through Time (ShiCo)

BEELD EN GELUID AVResearcher XL

GOLDEN AGENTS

CREATIVE INDUSTRIES AND THE MAKING OF THE DUTCH GOLDEN AGE

DIVE+

Dynamically Linking Collections on the Basis of Events

Time Capsule

Making Pharmaceutical and Botanical Digital Heritage Accessible and Usable

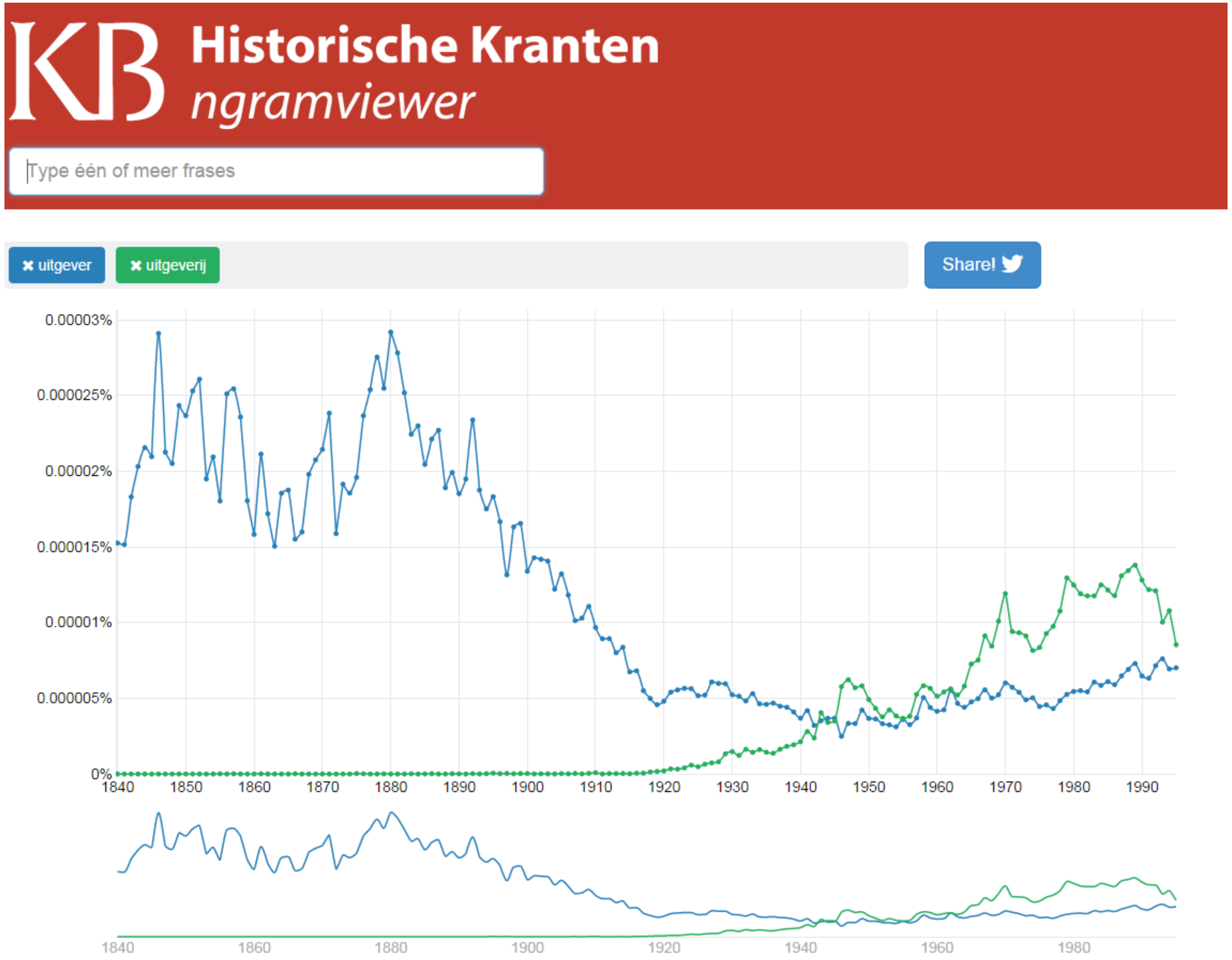
HiTiME

Historical Timeline Mining and Extraction

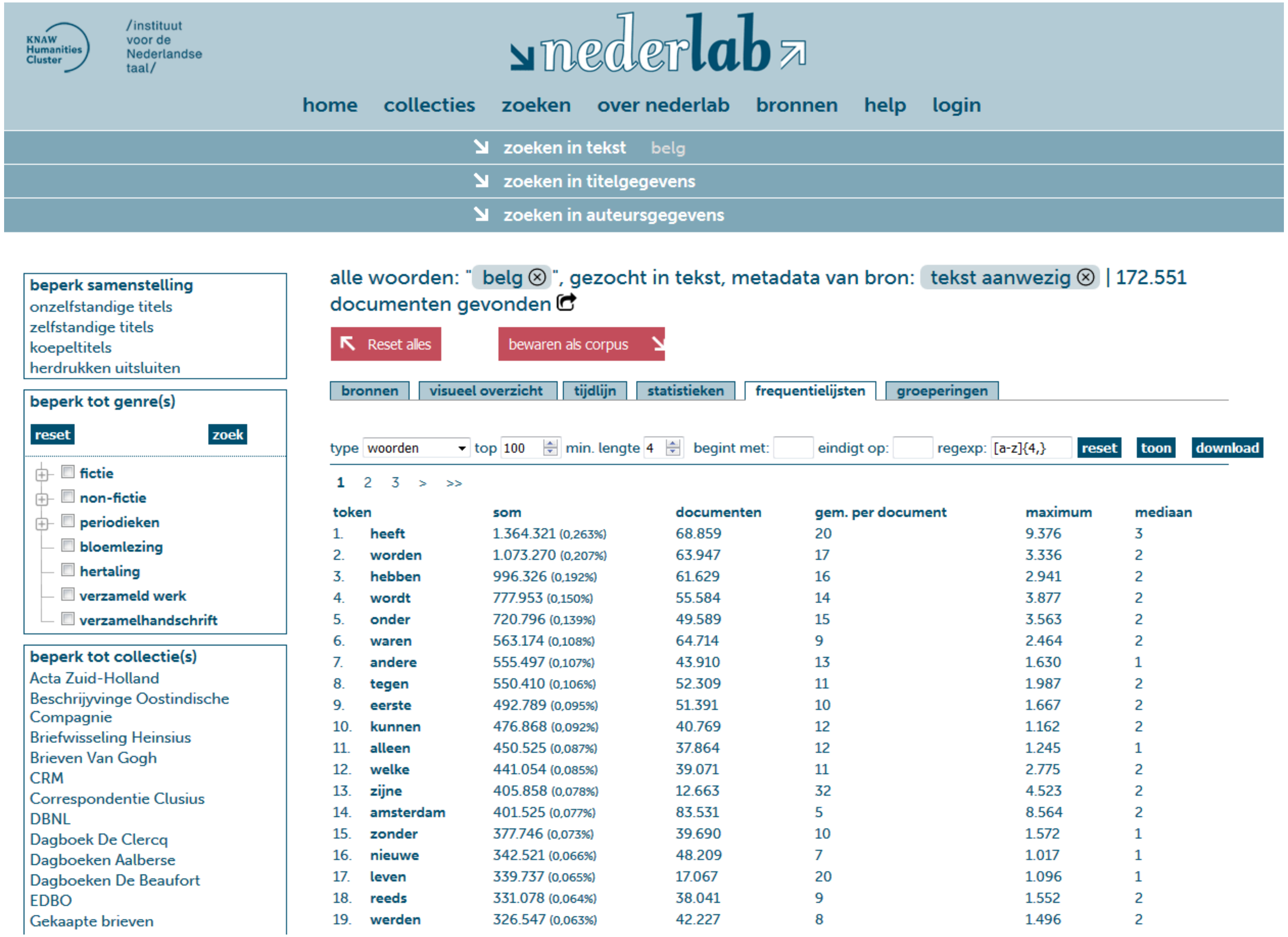
Translantis

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

PoliticalMashup



http://ngramviewer.kbresearch.nl/



beperk samenstelling
onzelfstandige titels
zelfstandige titels
koepeltitels
herdrukken uitsluiten

beperk tot genre(s)
reset zoek
□ fictie
□ non-fictie
□ periodieken
□ bloemlezing
□ hertaling
□ verzameld werk
□ verzamelhandschrift

beperk tot collectie(s)
Acta Zuid-Holland
Beschrijvinge Oostindische Compagnie
Briefwisseling Heinsius
Brieven Van Gogh
CRM
Correspondentie Clusius
DBNL
Dagboek De Clercq
Dagboeken Aalberse
Dagboeken De Beaufort
EDBO
Gekaapte brieven

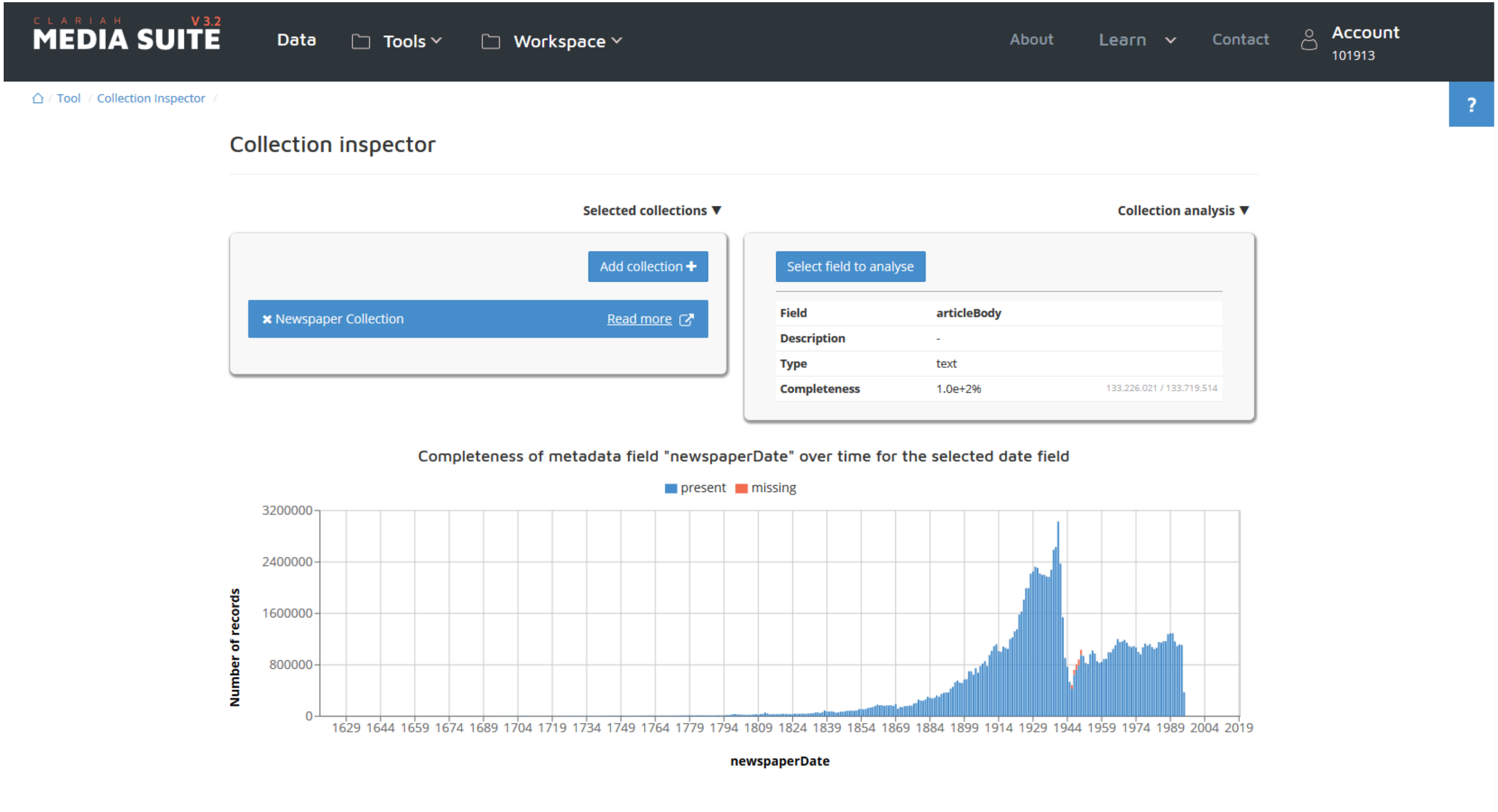
alle woorden: " belg ⊗", gezocht in tekst, metadata van bron: tekst aanwezig ⊗ | 172.551 documenten gevonden

Reset alles | bewaren als corpus

type: woorden | top: 100 | min. lengte: 4 | begint met: | eindigt op: | regexp: [a-z]{4,}

token	som	documenten	gem. per document	maximum	mediaan
1. heeft	1.364.321 (0,263%)	68.859	20	9.376	3
2. worden	1.073.270 (0,207%)	63.947	17	3.336	2
3. hebben	996.326 (0,192%)	61.629	16	2.941	2
4. wordt	777.953 (0,150%)	55.584	14	3.877	2
5. onder	720.796 (0,139%)	49.589	15	3.563	2
6. waren	563.174 (0,108%)	64.714	9	2.464	2
7. andere	555.497 (0,107%)	43.910	13	1.630	1
8. tegen	550.410 (0,106%)	52.309	11	1.987	2
9. eerste	492.789 (0,095%)	51.391	10	1.667	2
10. kunnen	476.868 (0,092%)	40.769	12	1.162	2
11. alleen	450.525 (0,087%)	37.864	12	1.245	1
12. welke	441.054 (0,085%)	39.071	11	2.775	2
13. zijne	405.858 (0,078%)	12.663	32	4.523	2
14. amsterdam	401.525 (0,077%)	83.531	5	8.564	2
15. zonder	377.746 (0,073%)	39.690	10	1.572	1
16. nieuwe	342.521 (0,066%)	48.209	7	1.017	1
17. leven	339.737 (0,065%)	17.067	20	1.096	1
18. reeds	331.078 (0,064%)	38.041	9	1.552	2
19. werden	326.547 (0,063%)	42.227	8	1.496	2

https://www.nederlab.nl/

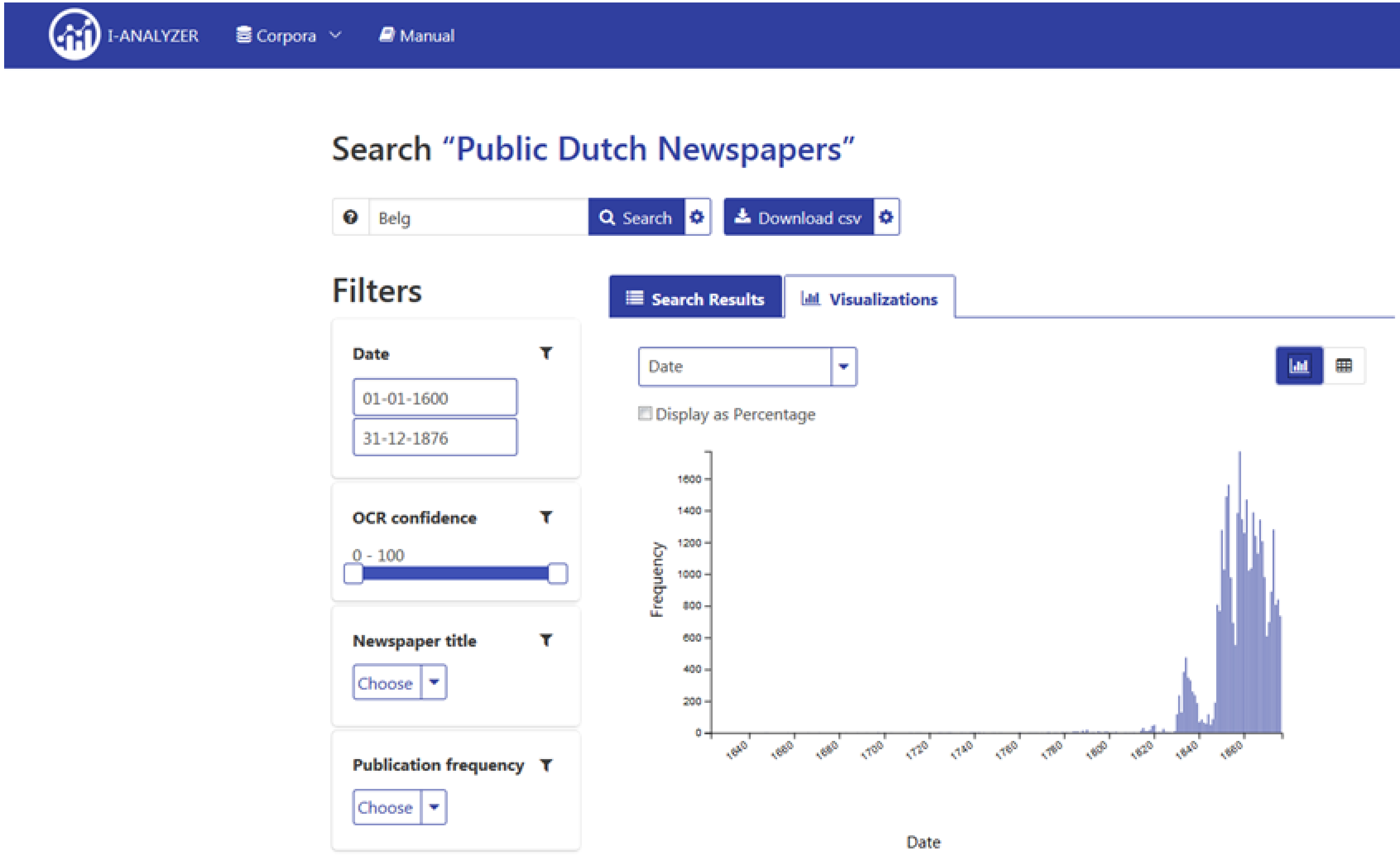


Collection inspector

Field	**articleBody**	
Description	-	
Type	text	
Completeness	1.0e+2%	133.226.021 / 133.719.514

Completeness of metadata field "newspaperDate" over time for the selected date field

https://mediasuite.clariah.nl/



I-ANALYZER   Corpora   Manual   claeyssens

Search "Public Dutch Newspapers"

Belg   Search   Download csv

Filters

Date

01-01-1600

31-12-1876

OCR confidence

0 - 100

Newspaper title

Choose

Publication frequency

Choose

Search Results   Visualizations

Date

□ Display as Percentage

https://ianalyzer.hum.uu.nl/

DIGITAL HUMANITIES LAB

KB national library of the netherlands

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?

dialogic
innovatie • interactie

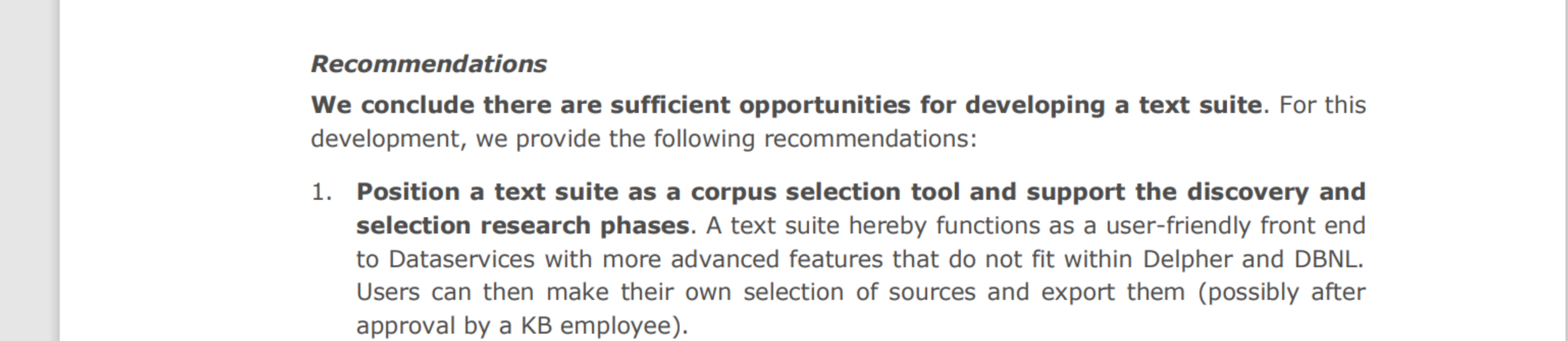
USER NEEDS FOR A TEXT SUITE FOR ADVANCED DIGITAL RESEARCH

dr. Max Kemman
ir. Nick Jelicic
Guido de Moor MSc
Marenne Massop MSc
ir. Tommy van der Vorst

COMMISSIONED BY
KB

PUBLICATION NUMBER
2022.024-2212

DATE
Utrecht, March 31st 2022



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 to Dataservices 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

And not only KB data but data from different heritage institutes

Case 1
FAIR Cultural Heritage Data

KB nationale bibliotheek

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.

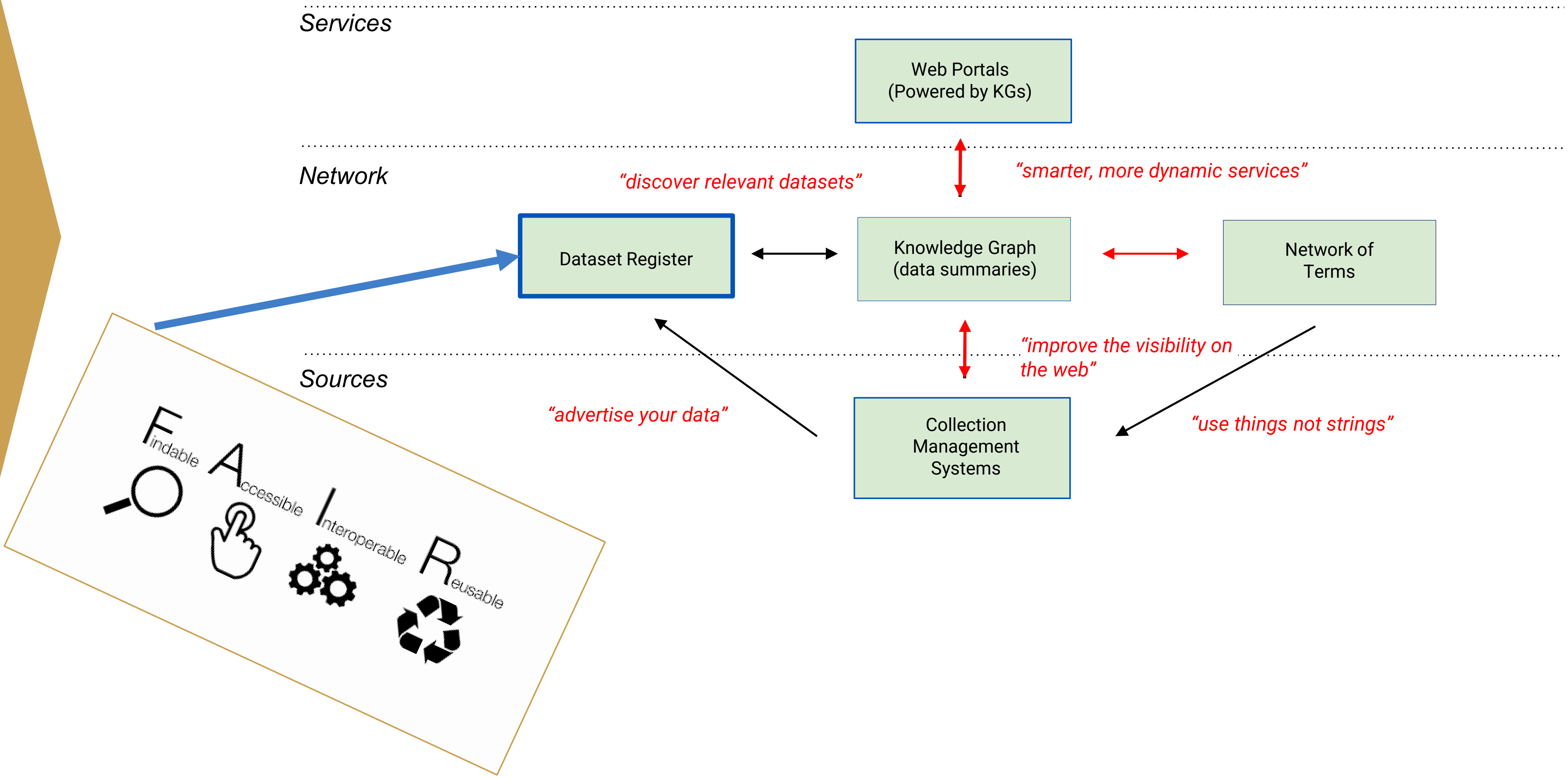
National Digital Heritage Strategy

March 2021

This is a publication of the Dutch Digital Heritage Network and the Ministry of Education, Culture and Science

https://netwerkdigitaalerfgoed.nl/en/

Roadmap for the NDE discovery infrastructure



Services

Web Portals
(Powered by KGs)

“smarter, more dynamic services”

Network

“discover relevant datasets”

Dataset Register

Knowledge Graph
(data summaries)

Network of
Terms

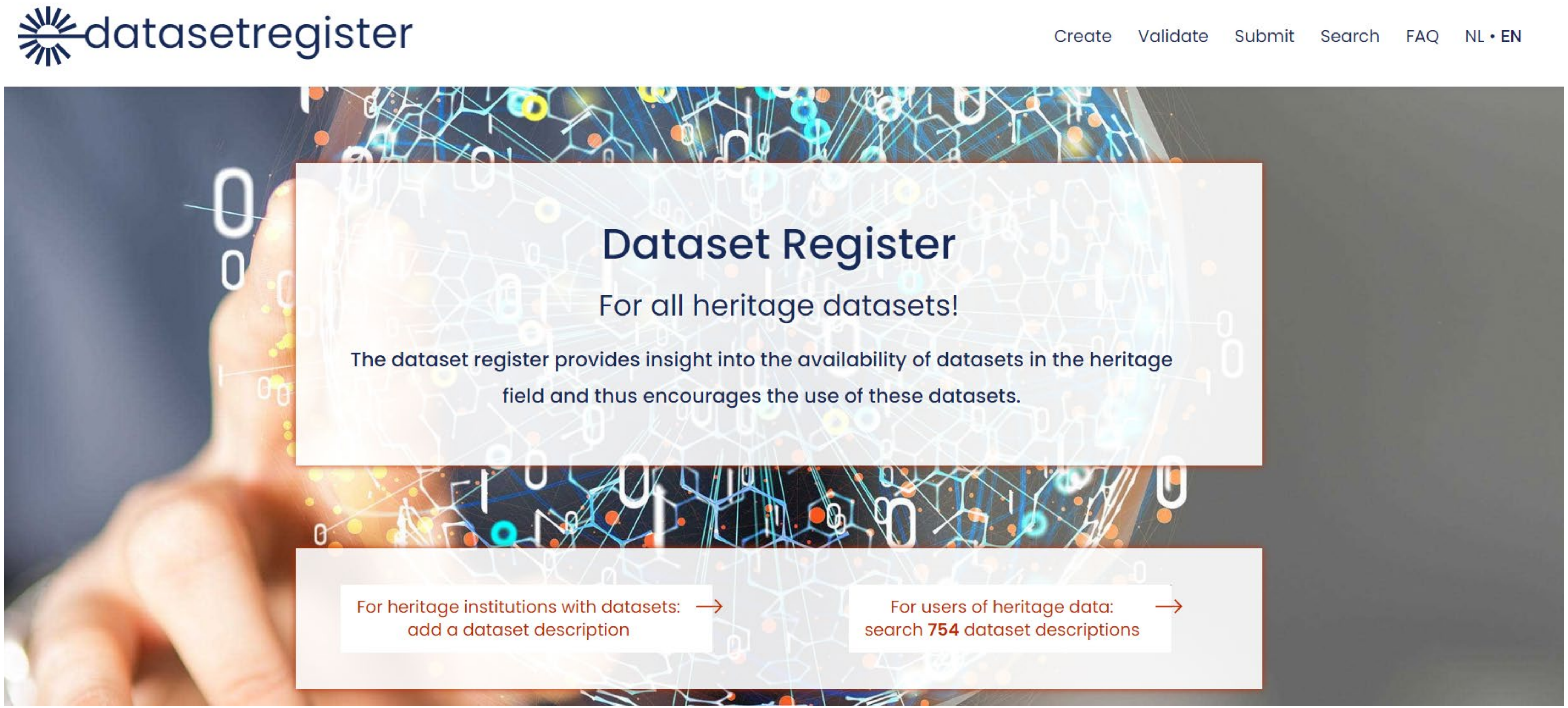
“improve the visibility on the web”

Sources

“advertise your data”

Collection
Management
Systems

“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.netwerkdigitaalerfgoed.nl/?lang=en

CLARIAH

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.

Clariah

https://www.clariah.nl/

BEELD& GELUID

Data Archiving and Networked Services DANS

huygens ing

IISG

/instituut voor de Nederlandse taal/

KB nationale bibliotheek

MAX PLANCK

Meertens Instituut

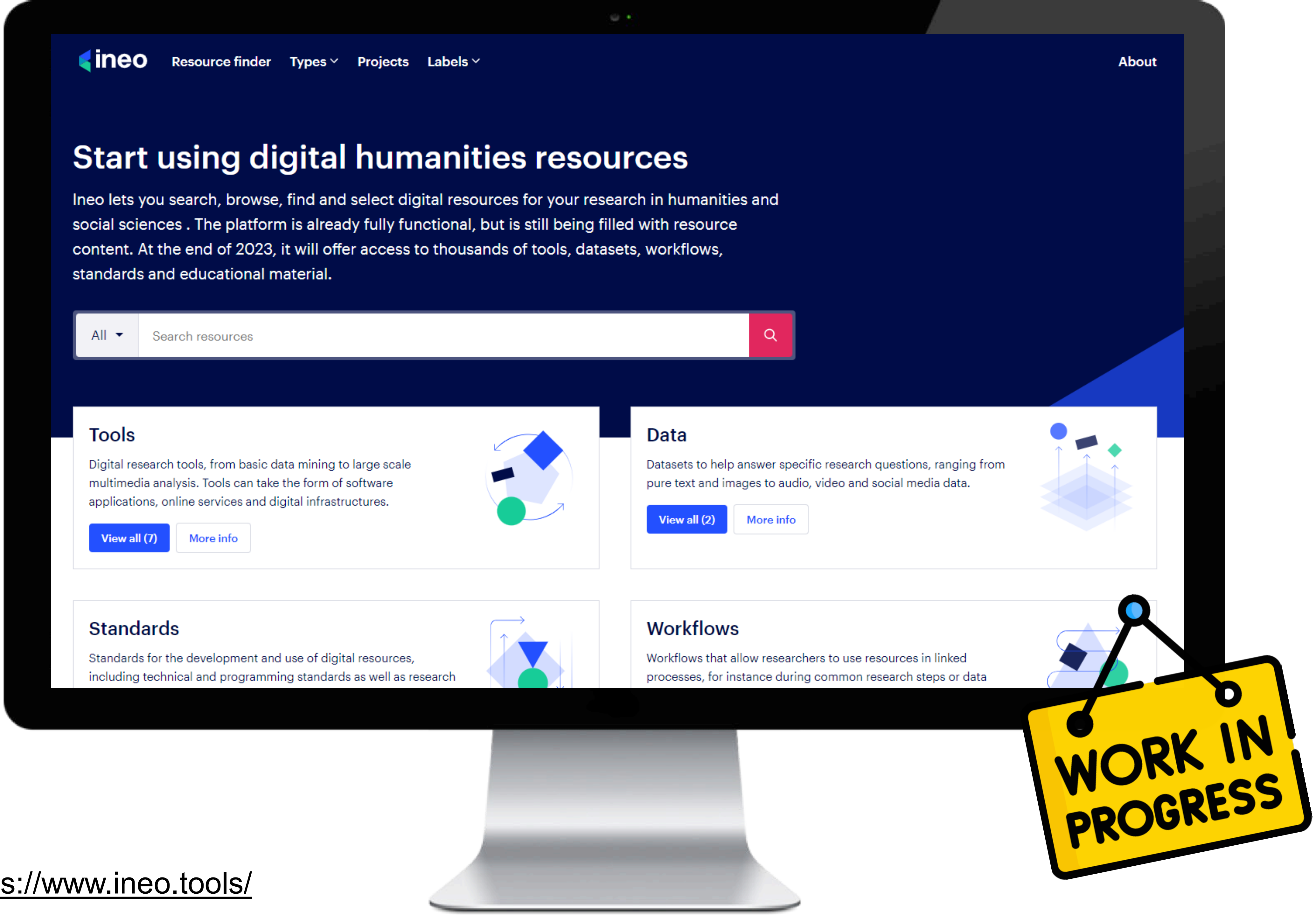
Radboud Universiteit

rijksuniversiteit groningen

Universiteit Utrecht

UNIVERSITEIT VAN AMSTERDAM

VU VRIJE UNIVERSITEIT AMSTERDAM



ineo Resource finder Types Projects Labels About

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.

All ▾ Search resources

Tools

Digital research tools, from basic data mining to large scale multimedia analysis. Tools can take the form of software applications, online services and digital infrastructures.

View all (7) More info

Data

Datasets to help answer specific research questions, ranging from pure text and images to audio, video and social media data.

View all (2) More info

Standards

Standards for the development and use of digital resources, including technical and programming standards as well as research

Workflows

Workflows that allow researchers to use resources in linked processes, for instance during common research steps or data

https://www.ineo.tools/

KB national library of the netherlands

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

Case 2
Secure Data Analysis Environment

Data

= result of

more than 200 years of collecting

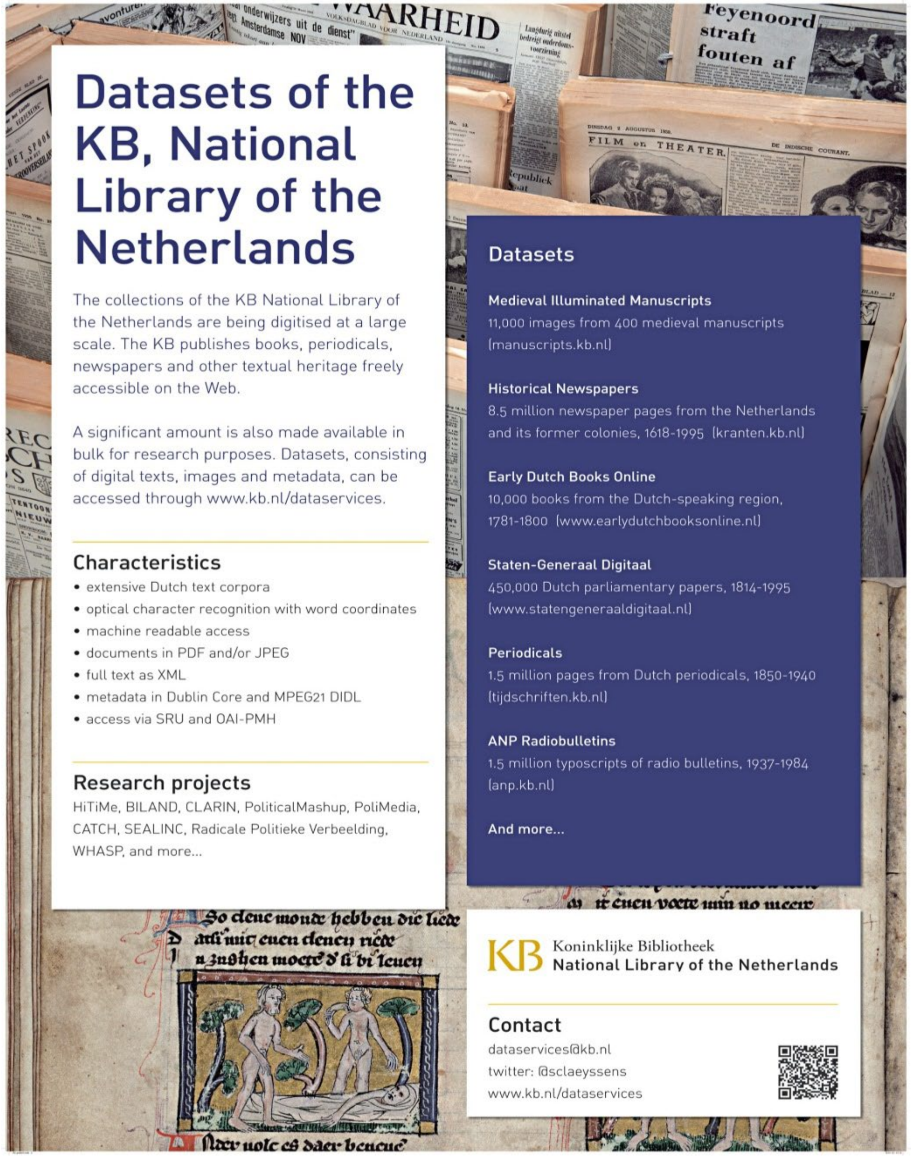
over 30 years of digitisation

10+ years of collecting born-digital publications

= machine readable, mostly textual

= structured or semi-structured

= legally as open as possible


WAARHEID
Feyenoord straft fouten af
FILM en THEATER

Datasets of the KB, National Library of the Netherlands

The collections of the KB National Library of the Netherlands are being digitised at a large scale. The KB publishes books, periodicals, newspapers and other textual heritage freely accessible on the Web.

A significant amount is also made available in bulk for research purposes. Datasets, consisting of digital texts, images and metadata, can be accessed through www.kb.nl/dataservices.

Characteristics

- extensive Dutch text corpora
- optical character recognition with word coordinates
- machine readable access
- documents in PDF and/or JPEG
- full text as XML
- metadata in Dublin Core and MPEG21 DIDL
- access via SRU and OAI-PMH

Research projects

HiTiMe, BILAND, CLARIN, PoliticalMashup, PoliMedia, CATCH, SEALINC, Radicale Politieke Verbeelding, WHASP, and more...

Datasets

Medieval Illuminated Manuscripts
11,000 images from 400 medieval manuscripts (manuscripts.kb.nl)

Historical Newspapers
8.5 million newspaper pages from the Netherlands and its former colonies, 1618-1995 (kranten.kb.nl)

Early Dutch Books Online
10,000 books from the Dutch-speaking region, 1781-1800 (www.earlydutchbooksonline.nl)

Staten-Generaal Digitaal
450,000 Dutch parliamentary papers, 1814-1995 (www.statengeneraaldigitaal.nl)

Periodicals
1.5 million pages from Dutch periodicals, 1850-1940 (tijdschriften.kb.nl)

ANP Radiobulletins
1.5 million typoscripts of radio bulletins, 1937-1984 (anp.kb.nl)

And more...

KB Koninklijke Bibliotheek
National Library of the Netherlands

Contact

dataservices@kb.nl
twitter: @sclaeyssens
www.kb.nl/dataservices

Data

= result of

more than 200 years of collecting

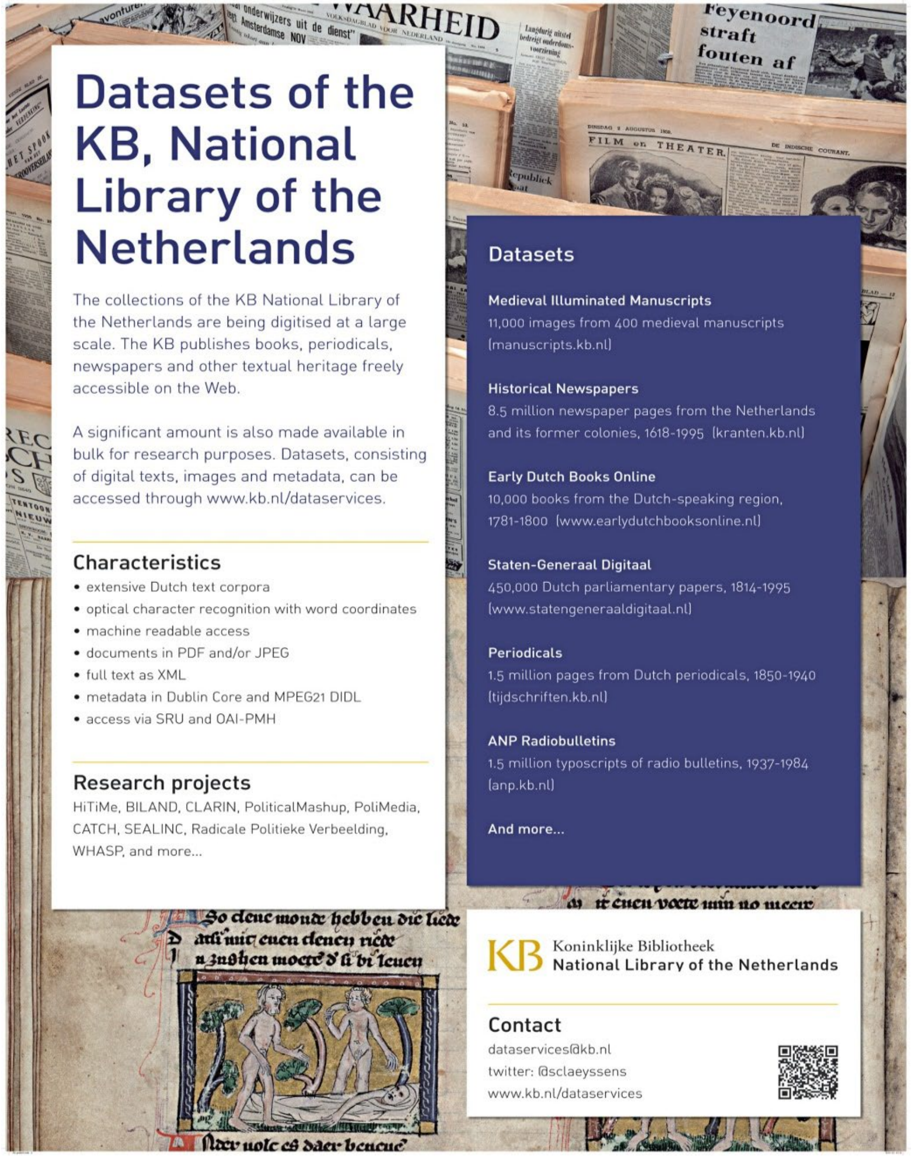
over 30 years of digitisation

10+ years of collecting born-digital publications

= machine readable, mostly textual

= structured or semi-structured

= legally as open as possible



Datasets of the KB, National Library of the Netherlands

The collections of the KB National Library of the Netherlands are being digitised at a large scale. The KB publishes books, periodicals, newspapers and other textual heritage freely accessible on the Web.

A significant amount is also made available in bulk for research purposes. Datasets, consisting of digital texts, images and metadata, can be accessed through www.kb.nl/dataservices.

Characteristics

- extensive Dutch text corpora
- optical character recognition with word coordinates
- machine readable access
- documents in PDF and/or JPEG
- full text as XML
- metadata in Dublin Core and MPEG21 DIDL
- access via SRU and OAI-PMH

Research projects

HiTiMe, BILAND, CLARIN, PoliticalMashup, PoliMedia, CATCH, SEALINC, Radicale Politieke Verbeelding, WHASP, and more...

Datasets

Medieval Illuminated Manuscripts
11,000 images from 400 medieval manuscripts (manuscripts.kb.nl)

Historical Newspapers
8.5 million newspaper pages from the Netherlands and its former colonies, 1618-1995 (kranten.kb.nl)

Early Dutch Books Online
10,000 books from the Dutch-speaking region, 1781-1800 (www.earlydutchbooksonline.nl)

Staten-Generaal Digitaal
450,000 Dutch parliamentary papers, 1814-1995 (www.statengeneraaldigitaal.nl)

Periodicals
1.5 million pages from Dutch periodicals, 1850-1940 (tijdschriften.kb.nl)

ANP Radiobulletins
1.5 million typoscripts of radio bulletins, 1937-1984 (anp.kb.nl)

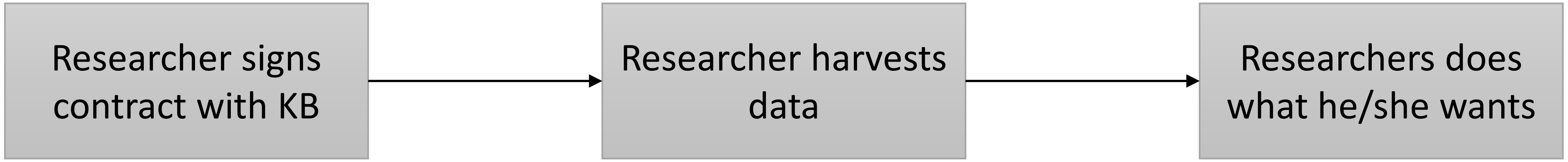
And more...

KB Koninklijke Bibliotheek
National Library of the Netherlands

Contact

dataservices@kb.nl
twitter: @sclaeyssens
www.kb.nl/dataservices

Current workflow



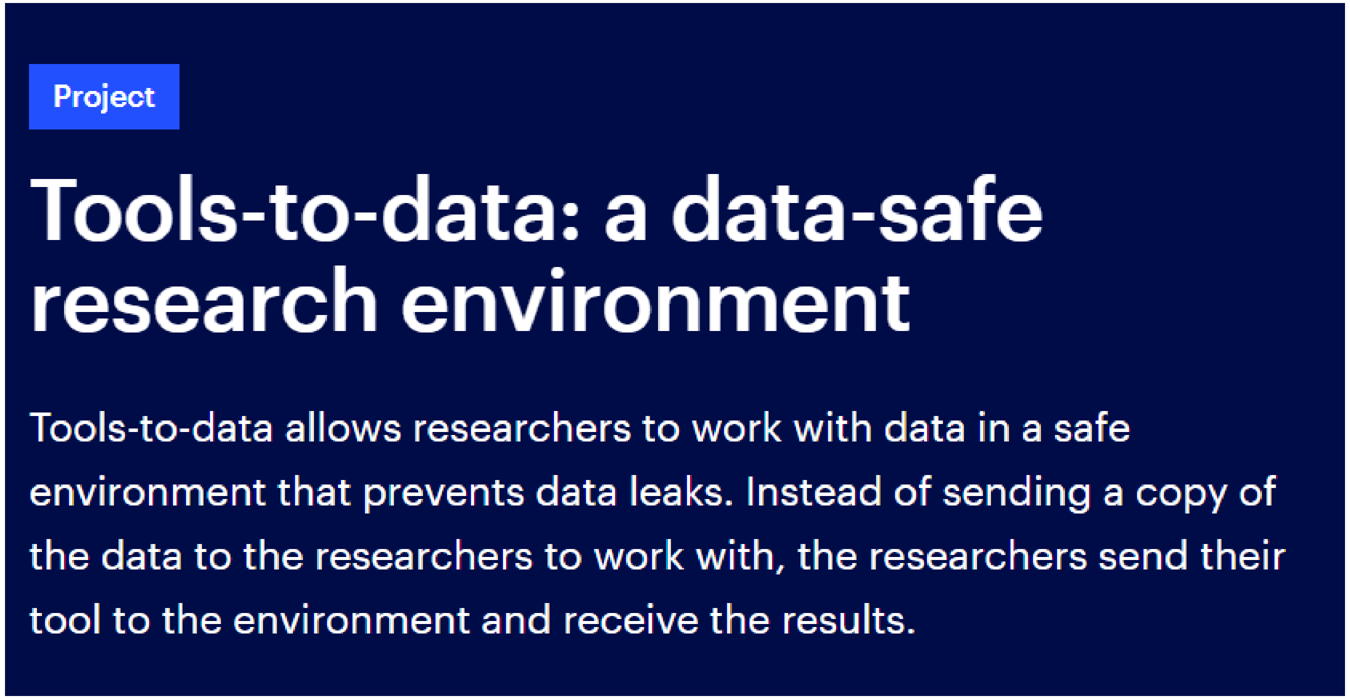
Researcher signs contract with KB

Researcher harvests data

Researchers does what he/she wants

Challenges for both researcher as well as KB:

- More data
- Better & richer data
- More recent data (legal issues)



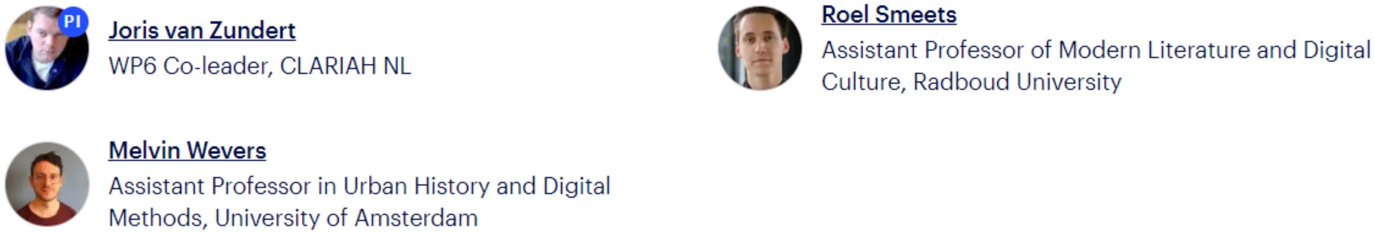
Project

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.

SURF

Clariah



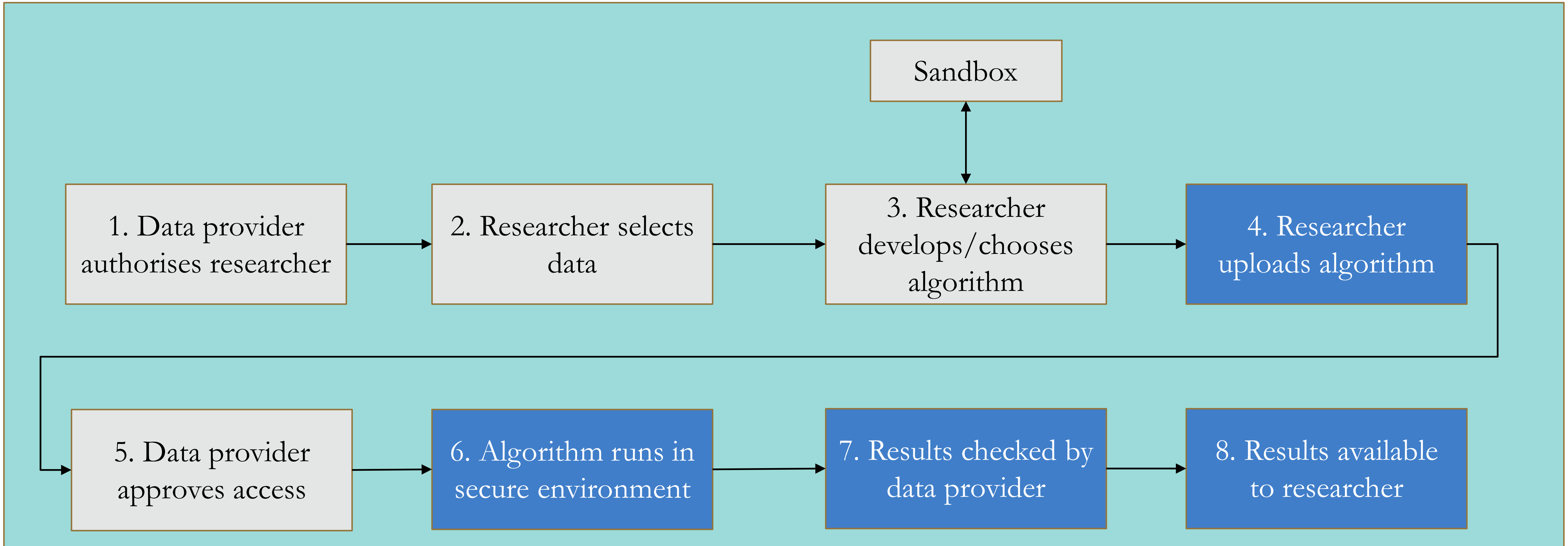
PI [Joris van Zundert](#)
WP6 Co-leader, CLARIAH NL

[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

Future workflow



Sandbox

1. Data provider authorises researcher

2. Researcher selects data

3. Researcher develops/chooses algorithm

4. Researcher uploads algorithm

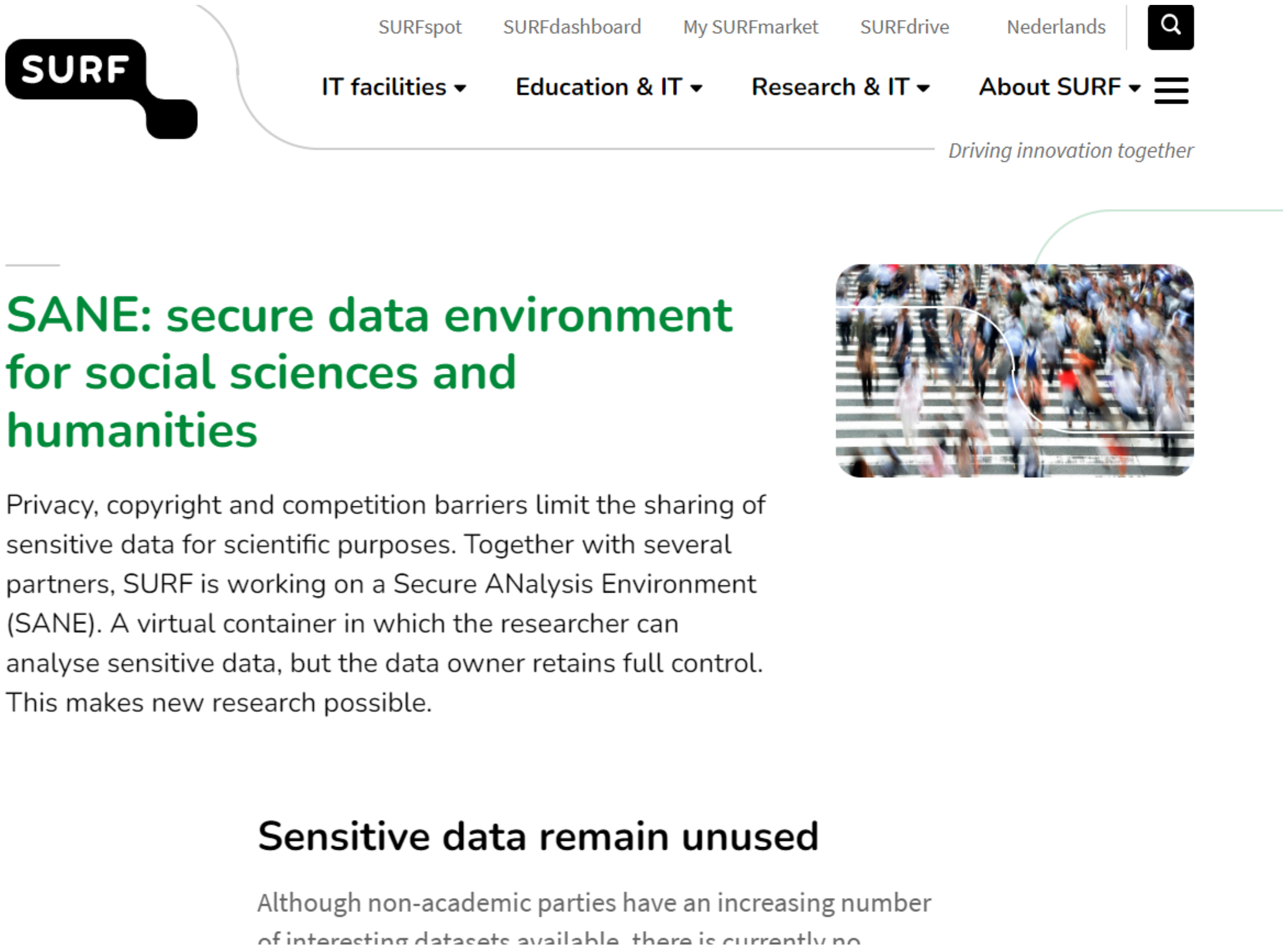
5. Data provider approves access

6. Algorithm runs in secure environment

7. Results checked by data provider

8. Results available to researcher

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

https://www.surf.nl/en/news/sane-secure-data-environment-for-social-sciences-and-humanities

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

Thanks fantastic colleagues 👏

Marian Hellema

Enno Meijers

Steven Claeyssens

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

KB ⟩ nationale bibliotheek

KB national library of the netherlands