Past and Future of Giving Access to Textual Data

DATA SERVICES AT THE KB

dr. Steven Claeyssens – Curator of Digital Collections

CLARIN

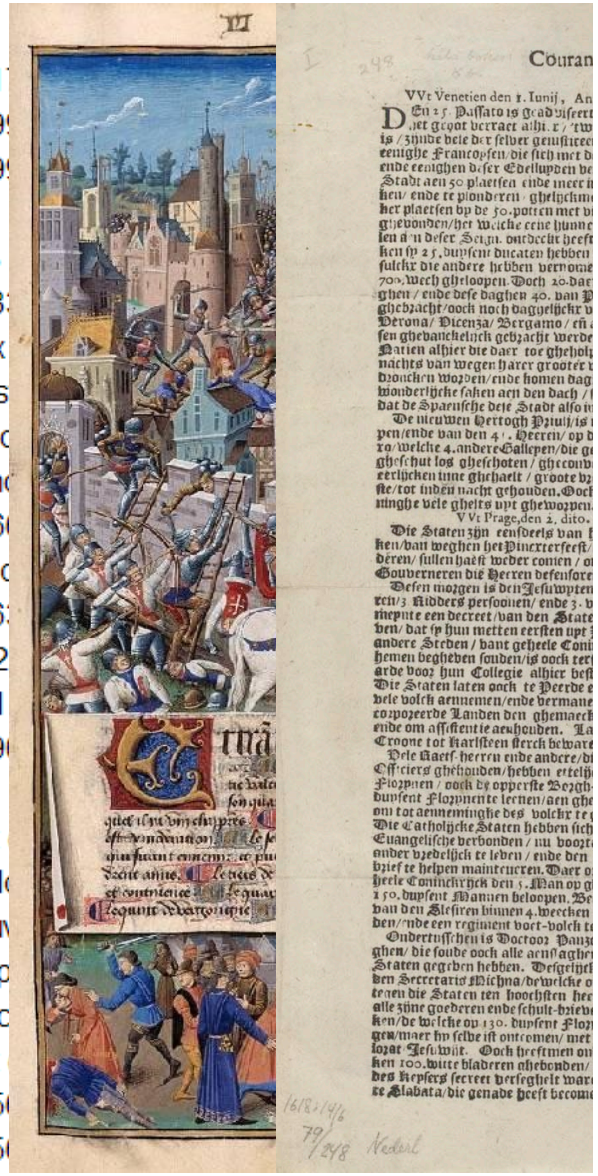
KB national library of the netherlands

#1
PAST

Data Services at the KB: a Short History

In 2012 the KB launched a service to give access to its collections of digitised publications for distant reading.

```
001@    $e2
001A    $0101
001B    $0199
001D    $0999
001F    $00
001U    $0utf8
001Z    $0108
002@    $0Aax
002C    $ateks
002D    $azonc
002E    $aband
003@    $0066
003O    $aOCc
004A    $0906
006C    $0B92
010@    $aned
011@    $a199
019@    $abe
021A    $aHet
028A    $dLudc
033A    $pLeuv
034D    $a63 p
034I    $a21 c
036E    $aHet
044Z/01 $9075
044Z/02 $9075
```



Courant

VVt Venetien den 1. Iunij, An

VVt Prage, den 2. dito.


```
<?xml version="1.0" encodi
<text>
   <p>58</p>
   <p>veel zal doen afne
      een nog dieneren s
      Vastknellende aan
      (en in der daad toc
      gelijk nog nimmer
      internationale med
      wij dan met minde
      schrijvers overstro
      België aan de were
      maar ook hel goed
      koopt, en om de kv
      «"'ukken. Inderda
      over alle landen ge
   <p>Ook hier le lande h
      vooral uit den Boel
      achtingwaardige lo
      veroordeelden , in
      Franschen zeiven.
      noemenswaardiger
      internationale kont
      opofferingen doen,
      anderen zouden m
      verandering komer
      moeten ouderwerp
      aangewezen zijn, e
      schuilplaats te zijn
      waarmede inderda
      in een zeer naauw
   <p>Wij bevelen het bo
      publiek en aan onz
      eens, even goed al
      worden. Zij vervalt
      tegenovergestelde
      verdient beklaagd
   <p>INTERNATIONAAL
   <p>De Edinburgh Revi
      nummer van 1852
      opnemen , als uitd
      van dat, in een vro
      medegedeeld:</p>
   <p>ii Ofschoon letterd
      land, dat een lezen
      herdruk van Engel
      toeeigening van Fr
      België door de repr
      beide andere lande
      toch dat België tha
      nadrukken en nabo
```

```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.openarchives.org/OAI/2.0/">
    <responseDate>2014-04-30T12:45:59.368Z</responseDate>
    <request metadataPrefix="didl" verb="GetRecord">http://services.kb.nl/mdo/oai
      </request>
  - <GetRecord>
    - <record>
      - <header>
            <identifier>DDD:ddd:011010636:mpeg21</identifier>
            <datestamp>2012-06-26T20:53:37.545Z</datestamp>
            <setSpec>DDD</setSpec>
        </header>
      - <metadata>
        - <didl:DIDL xmlns="urn:mpeg:mpeg21:2002:02-DIDL-NS"
          xmlns:srw_dc="info:srw/schema/1/dc-v1.1"
          xmlns:didmodel="urn:mpeg:mpeg21:2002:02-DIDMODEL-NS"
          xmlns:dcx="http://krait.kb.nl/coop/tel/handbook/telterms.html"
          xmlns:dcmitype="http://purl.org/dc/dcmitype/"
          xmlns:dcterms="http://purl.org/dc/terms/"
          xmlns:dc="http://purl.org/dc/elements/1.1/"
          xmlns:ddd="http://www.kb.nl/namespaces/ddd"
          xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS">
          - <didl:Item dc:identifier="ddd:011010636:mpeg21">
            - <didl:Component dc:identifier="ddd:011010636:mpeg21:metadata">
              - <didl:Descriptor>
                  <didl:Statement dc:type="role"
                    mimeType="text/plain">metadata</didl:Statement>
                </didl:Descriptor>
              - <didl:Resource mimeType="text/xml">
                - <srw_dc:dcx>
                    <dc:title>Nieuwsblad van het Noorden</dc:title>
                    <dc:identifier
                      xsi:type="dcx:PPN">833013246</dc:identifier>
                    <dc:date>1983-02-15</dc:date>
                    <dcterms:temporal>Dag</dcterms:temporal>
                    <dcx:recordRights>Koninklijke Bibliotheek Den
                      Haag</dcx:recordRights>
                    <dc:publisher>Nieuwenhuis</dc:publisher>
                    <dcterms:spatial>Regionaal/lokaal</dcterms:spatial>
                    <dc:source>Groninger archieven </dc:source>
                    <dcx:volume>96</dcx:volume>
                    <dcx:issuenumber>38</dcx:issuenumber>
                    <dcx:recordIdentifier>ddd:011010636:mpeg21</dcx:recor
```

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

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...

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)

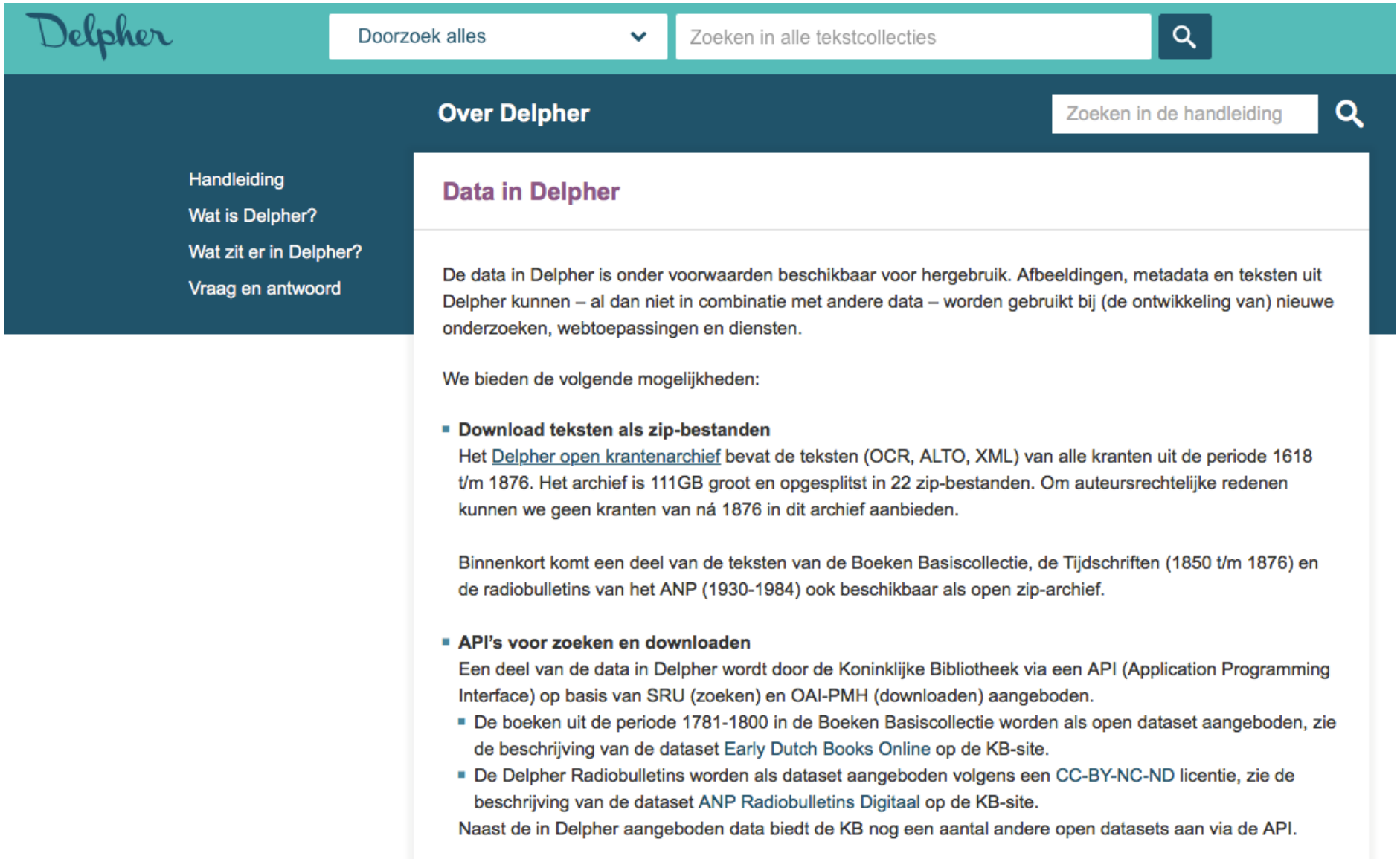
ANP Radiobulletins
1.5 million typoscripts of radio bulletins, 1937-1984 (anp.kb.nl)

And more...

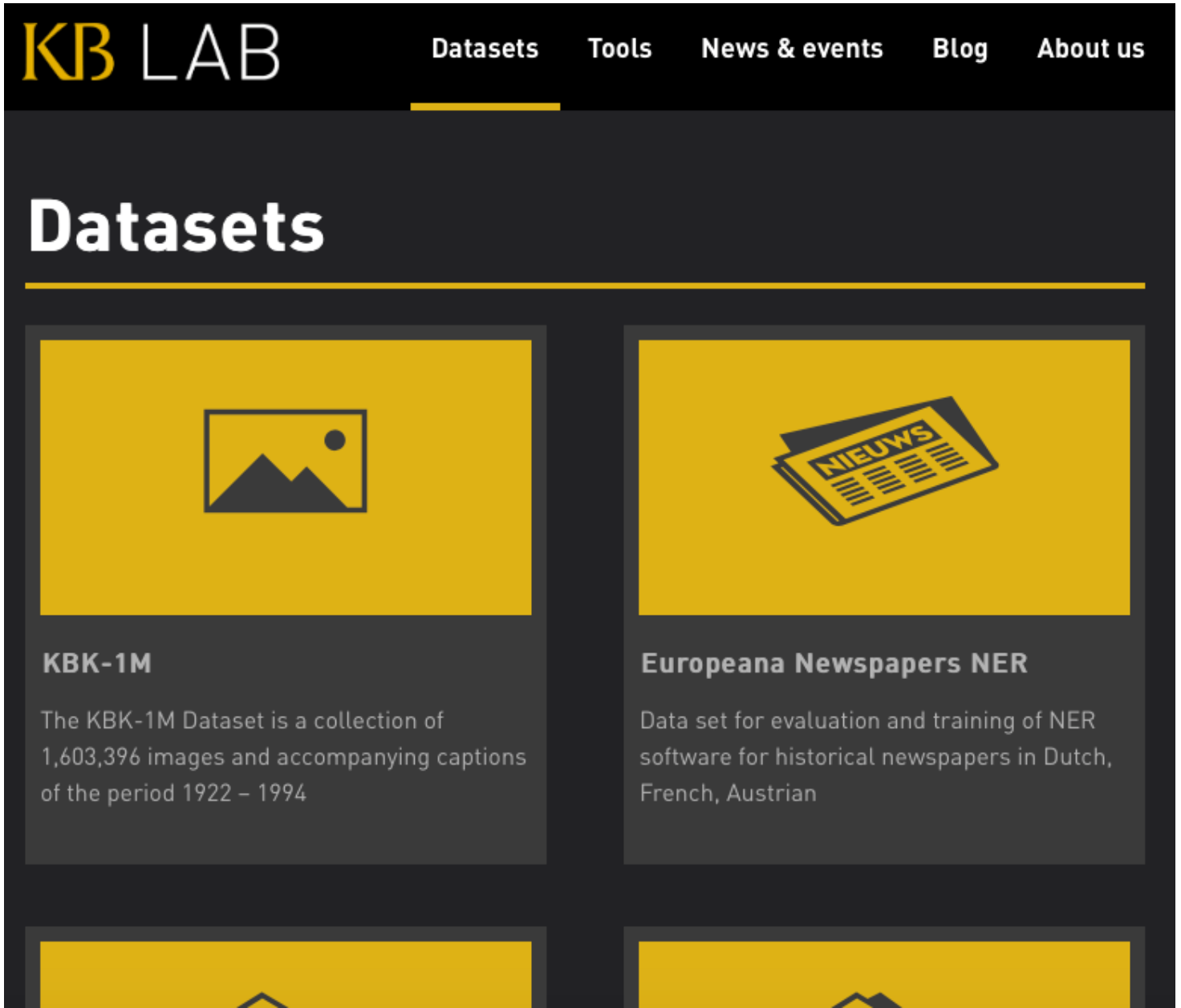
KB national library of the netherlands

kb.nl/dataservices

The following years we added a few options. In no particular order:



delpher.nl/data


KB LAB
Datasets
Tools
News & events
Blog
About us
Datasets
KBK-1M
The KBK-1M Dataset is a collection of 1,603,396 images and accompanying captions of the period 1922 – 1994
NIEUWS
Europeana Newspapers NER
Data set for evaluation and training of NER software for historical newspapers in Dutch, French, Austrian

lab.kb.nl

WIKIMEDIA COMMONS

Main page
Welcome
Community portal
Village pump
Help center

Taal
English
Selecteer

Participate
Upload file
Recent changes
Latest files
Random file
Contact us

Print/export
Create a book
Download as PDF
Printable version

In other projects
Wikipedia

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Wikidata item
Subpages
Nominate for deletion

In Wikipedia
العربية
Български
Català
Deutsch

Commons:Koninklijke Bibliotheek

Wikimedia Commons in het Nederlands [x]

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]

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.

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
"Alba amicorum van de Koninklijke Bibliotheek"@nl	http://data.bibliotheken.nl/id/dataset/rise-alba
"Brinkman trefwoordenthesaurus"@nl	http://data.bibliotheken.nl/id/dataset/brinkman
"Centsprenten"@nl	http://data.bibliotheken.nl/id/dataset/rise-centsprenten
"Gemeenschappelijke Trefwoordenthesaurus (GTT)"@nl	http://data.bibliotheken.nl/id/dataset/gtt
"Nederlandse Bibliografie Totaal (NBT)"@nl	http://data.bibliotheken.nl/id/dataset/nbt
"Organisaties uit de corporatiethesaurus van de Koninklijke Bibliotheek"@nl	http://data.bibliotheken.nl/id/dataset/corps
"Personen uit de Nederlandse Thesaurus van Auteursnamen (NTA)"@nl	http://data.bibliotheken.nl/id/dataset/persons
"Short-Title Catalogue Netherlands (STCN)"@nl	http://data.bibliotheken.nl/id/dataset/stcn
"Thesaurus Auteurs DBNL"@nl	http://data.bibliotheken.nl/id/dataset/dbnla
"Thesaurus KBcode"@nl	http://data.bibliotheken.nl/id/dataset/kbcode
"Titels DBNL"@nl	http://data.bibliotheken.nl/id/dataset/dbnlt

data.bibliotheken.nl

All of this resulted in (data driven) humanities research projects,

but also in the development of new research tools and environments.

CL@RIAH

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

Time Capsule

Making Pharmaceutical and Botanical Digital Heritage Accessible and Usable

HiTiME

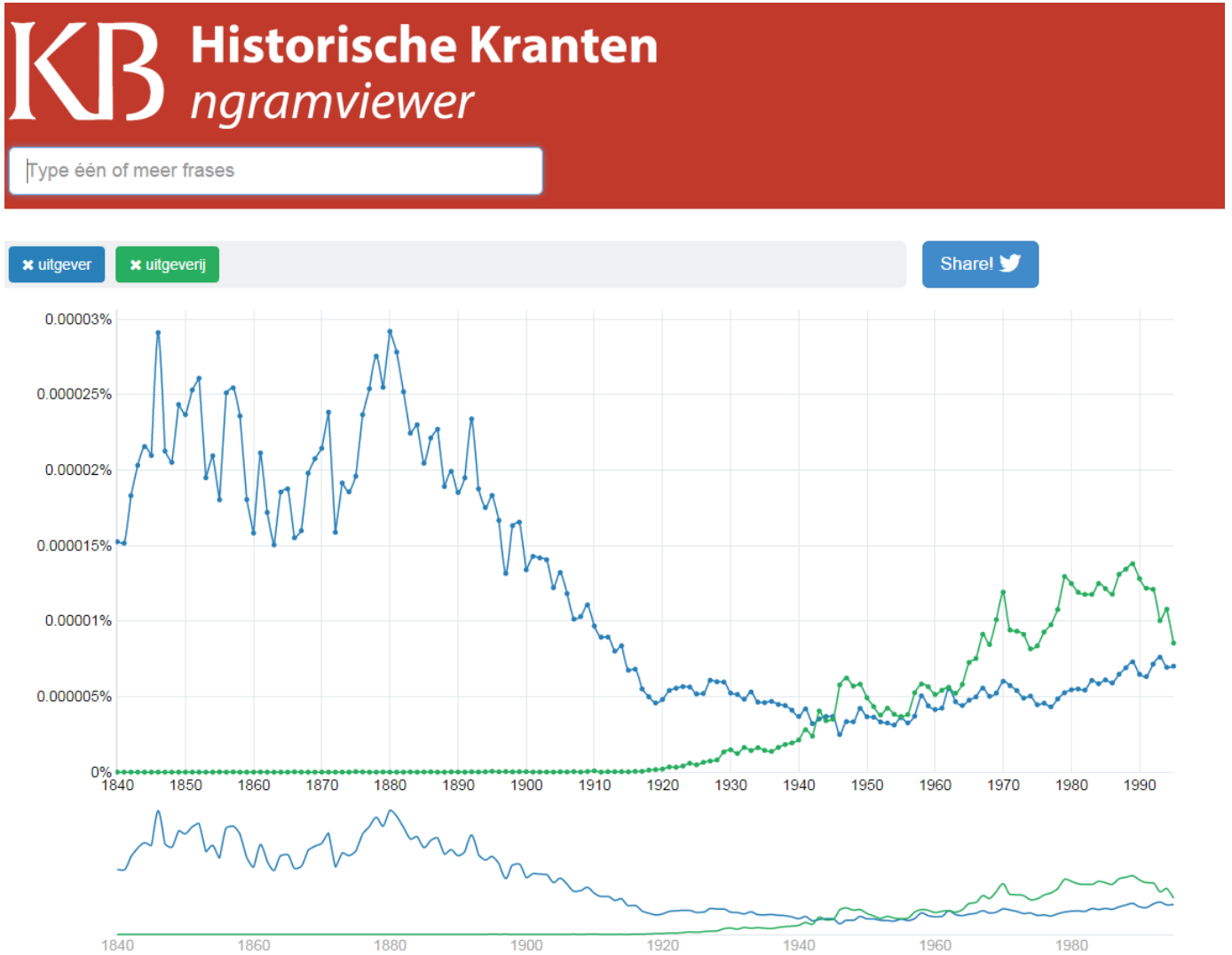
Historical Timeline Mining and Extraction

Translantis

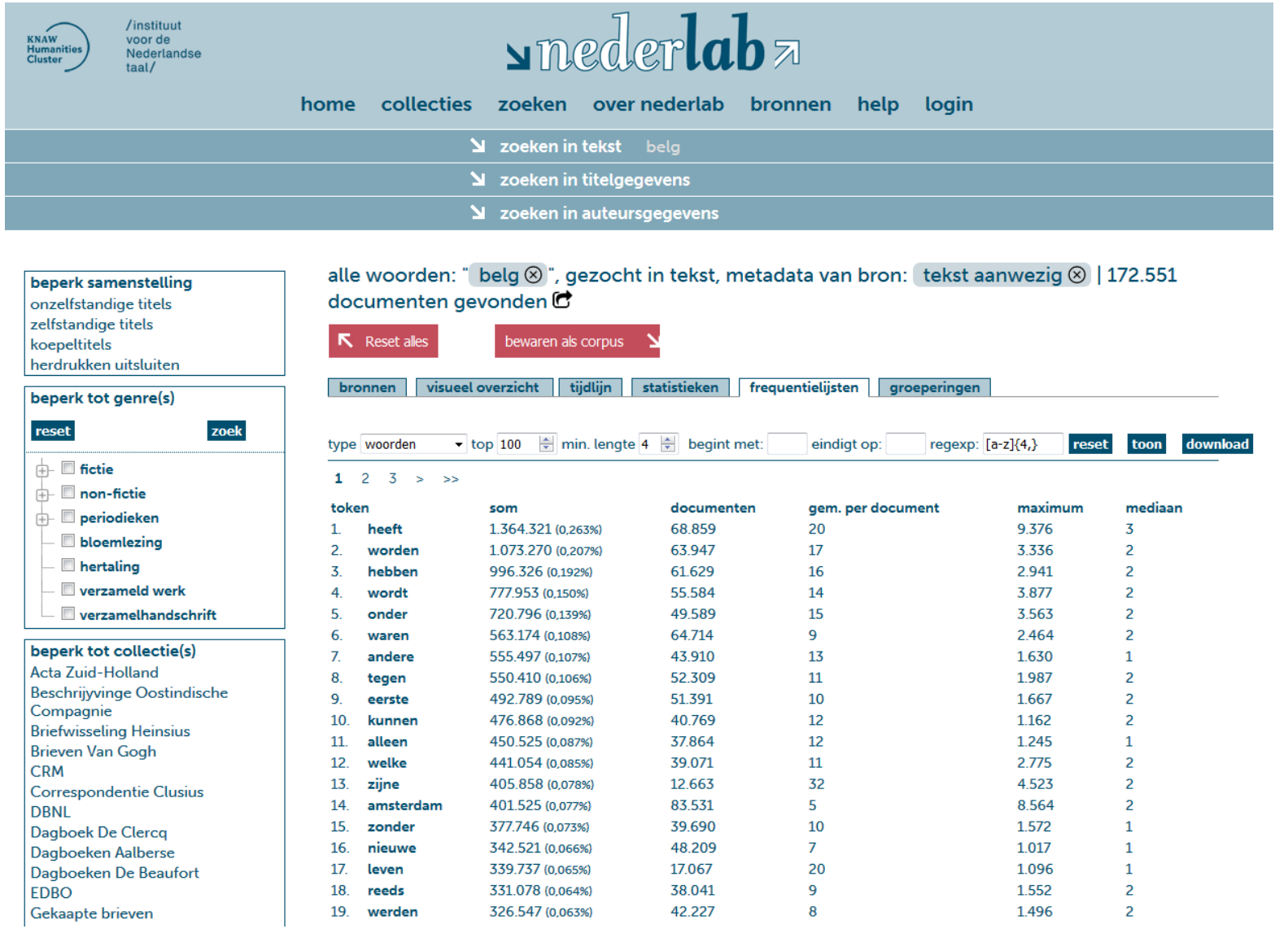
Digital Humanities Approaches to Reference Cultures: The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990

PoliticalMashup

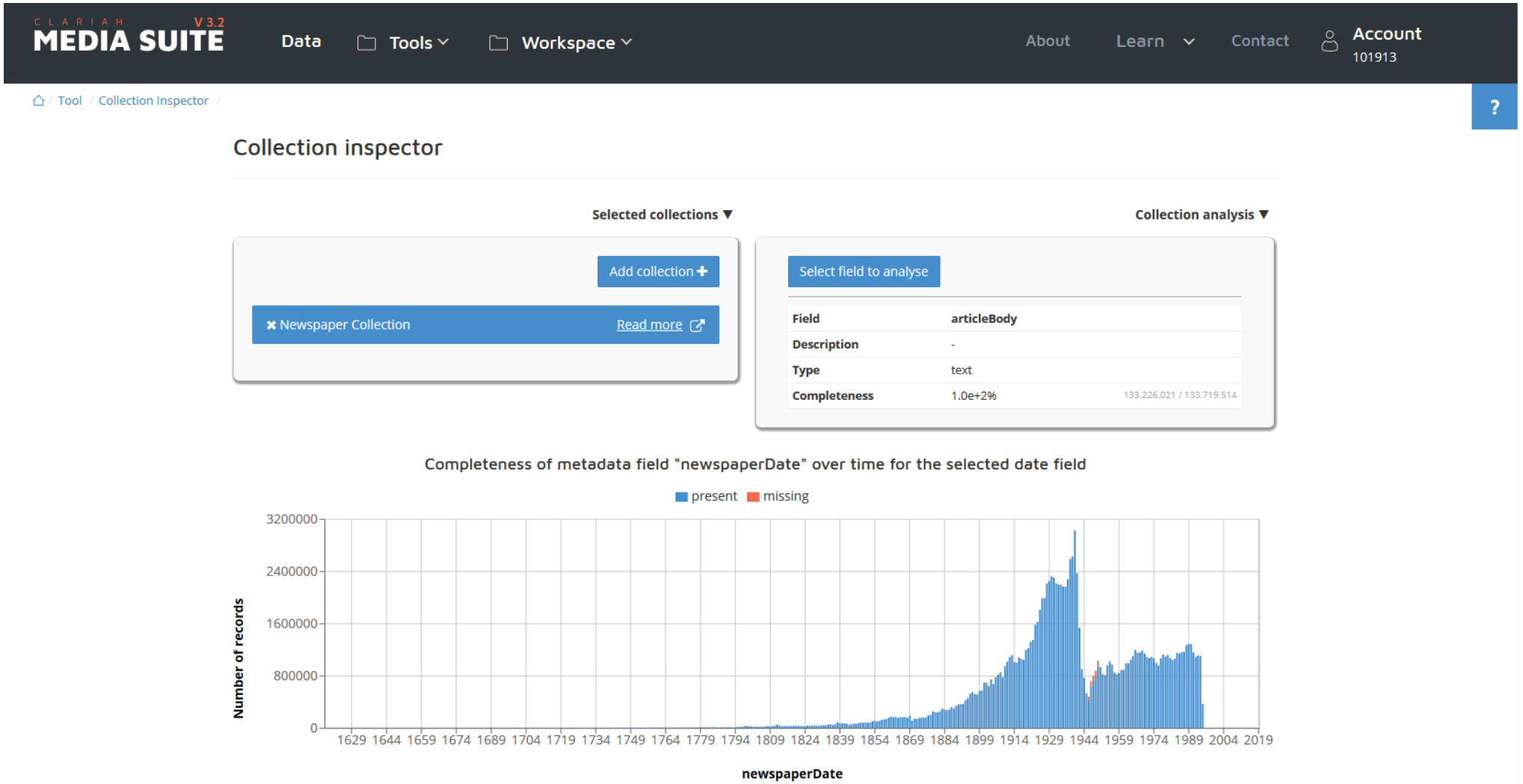
KB national library of the netherlands



ngramviewer.kbresearch.nl


/instituut voor de Nederlandse taal/
nederlab
home collecties zoeken over nederlab bronnen help login
zoeken in tekst belg
zoeken in titelgegevens
zoeken in auteursgegevens
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
bronnen visueel overzicht tijdlijn statistieken frequentielijsten groeperingen
type woorden top 100 min. lengte 4 begint met: eindigt op: regexp: [a-z]{4,} reset toon download
1 2 3 > >>
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

nederlab.nl



mediasuite.clariah.nl

CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

KB national library of the netherlands

ianalyzer.hum.uu.nl

DIGITAL HUMANITIES LAB

KB national library of the netherlands

#2
TODAY

Data Services at the KB: what’s missing

Benjamin Schmidt
@benmschmidt

Blog series. Too often cultural heritage organizations distribute texts idiosyncratically, making them all but unusable for researchers. There are better ways using modern data formats. Part I tries to repack some @Europeanaeu newspapers for normal humans.
benschmidt.org/post/2022-03-1...

< An API is not enough. Multiple ways to deliver the data are needed. >

Data

= result of

more than 200 years of collecting

over 30 years of digitisation

10+ years of collecting born-digital publications

< Can we find a way to provide access to our copyright protected born digital collections? >

Gale Digital Scholar Lab

OPEN NEW RESEARCH PATHWAYS

Gale Digital Scholar Lab creates new possibilities by offering solutions to the most common challenges facing researchers in the digital humanities today.

ProQuest

Part of Clarivate

Products & Se

What is Constellate?

Constellate is the text analytics service from the not-for-profit ITHAKA - the same people who brought you JSTOR and Portico. It is a platform for teaching, learning, and performing text analysis using the world's leading archival repositories of scholarly and primary source content.

TDM Studio

< Can we bridge the gap between simple search GUI's and advanced functionalities for (text) analysis? >

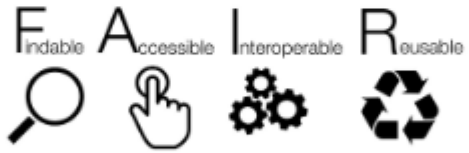
#3
FUTURE

Towards Data Services 2.0

FAIR @KB

Een manifest

november 2020



1. FAIR

[cf. CLARIAH FAIR Dataset Register]

2. Notebooks

4.7 Collectiezone

ONDERZOEK & WETENSCHAP

6. Text-dataminingruimte (TDM)

☼
Daglicht, maar ook zelf in te stellen verlichting.
♫
Geluidsdicht
Sfeer
Transparant, innovatief, futuristisch, digitaal i.c.m. nostalgie
Faciliteiten
Stopcontacten, wifi, LAN, presentatiescherm, touch table multifunctionele wanden.
Goed beveiligd (i.v.m. IT)

Text-datamining ruimte (8-12 personen)
De TDM ruimte is een belevingsruimte, waarin de onderzoeker en bezoeker kan werken met en aan de digitale ontwikkelingen van de KB.

Functie
- Hier wordt het digitale (o.a. web-archief) en gedigitaliseerde woord in historische context geplaatst.
- Door gebruik van oude computers en nieuw technologie, trekken wij de lijn van het verleden naar het heden.
- Het KB-lab kan hier digitale tools ontwikkelen
- Ruimte is geschikt voor langdurig gebruik (bijvoorbeeld Hackathons)
- Fysiek en digitaal zijn allebei aanwezig en versterken elkaar.
- In eerste instantie zijn de presentaties voor in de colloquiumruimte, maar. collectiespecialisten kunnen ook hier presentaties houden op het snijvlak van fysiek en digitaal.

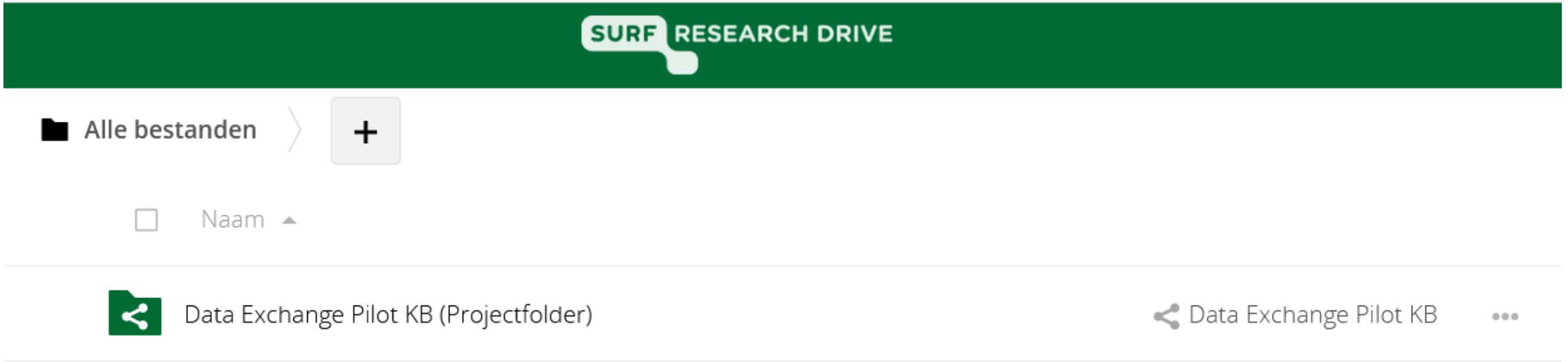
Locatie
- In de collectiezone.
- Bij de digitale collectieplekken in de buurt.

Faciliteiten
- Goede ventilatie moet mogelijk zijn (i.v.m. bijv. een hackathon).
- Laptops voor flexibiliteit.
- De luchtvochtigheid van de ruimtes moet controleerbaar en hanteerbaar zijn.
- Elementen van het fysieke boek terug laten komen (Zowel functioneel als ook ter decoratie).

Presentatiescherm in de sfeer van minority report (de film)

KB nationale bibliotheek

3. TDM Room



4. ‘tools to data’ – SANE Project

ODISSEI CLARIAH Common Lab Research Infrastructure for the Arts and Humanities KB national library of the netherlands

dialogic
innovatie • interactie

USER NEEDS FOR A TEXT SUITE FOR ADVANCED DIGITAL RESEARCH

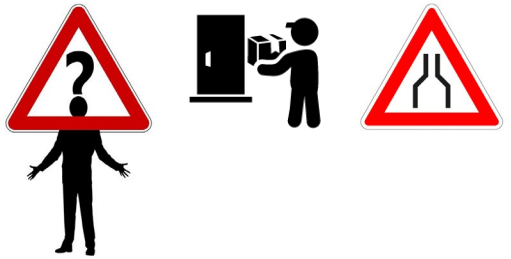
dr. Max Kemman
ir. Nick Jelicic
Guido de Moor MSc
Marenne Massop MSc
ir. Tommy van der Vorst

COMMISSIONED BY
KB

PUBLICATION NUMBER
2022.024-2212

DATE
Utrecht, March 31st 2022

KB national library of the netherlands



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).

5. Corpus (or collection?) selection tool

15:30 – 15:55 User demand for supporting advanced analysis of historical text collections [long]

Max Kemman and Steven Claeyssens

KB national library of the netherlands

KB national library of the netherlands

Steven Claeyssens - @sclaeyssens