LIBER Webinar: Generating Metadata with Artificial Intelligence

HOST

Jeannette Frey

LIBER President

Director, Bibliothèque Cantonale et Universitaire (BCU) Lausanne

Jeannette.frey@bcu.unil.ch

SPEAKER

Martijn Kleppe

Head of Research, National Library of the Netherlands (KB)

Martijn.Kleppe@kb.nl

NOTES

- **The webinar is being recorded.**
- **Slides and a recording will be shared** by email after the webinar.
- **Questions?** Put them in the chat box.
- **10-15 minutes of discussion** will take place following the presentations.

Generating metadata with AI

EXPERIENCE OF THE KB, NATIONAL LIBRARY OF THE NETHERLANDS

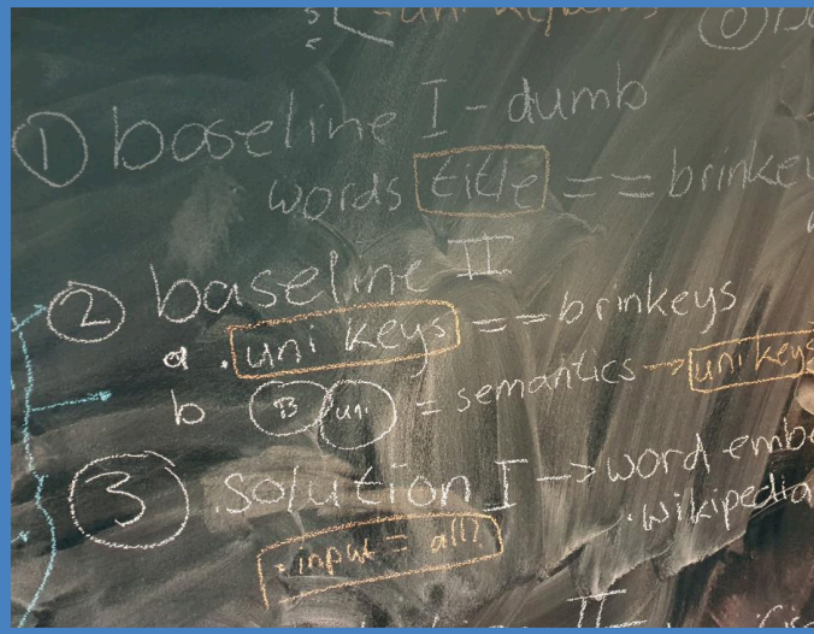
Martijn Kleppe – Head of Research
Martijn.kleppe@kb.nl | @martijnkleppe | www.kb.nl/martijnkleppe

KB nationale bibliotheek

https://www.kb.nl/en/news/2019/kb-explores-artificial-intelligence-to-generate-metadata

Exploration possibilities

AUTOMATED GENERATION OF METADATA



August 2019

KB national library of the netherlands

Sara Veldhoen

Meta van der Waal-Gentenaar

Dorien Haagsma

Brigitte den Oudsten

Kleppe, M., Veldhoen, S, Waal-Gentenaar, M., Oudsten, B. den, & Haagsma, D. (2019). *Exploration possibilities Automated Generation of Metadata*. http://doi.org/10.5281/zenodo.3375192

KB nationale bibliotheek

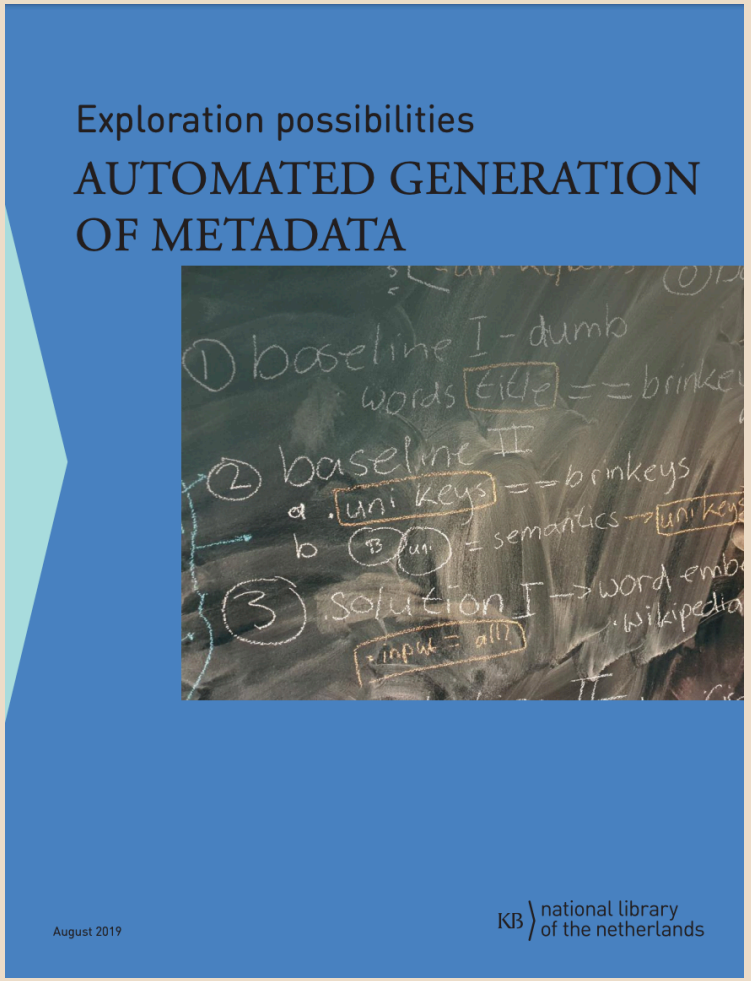
Outline

I. Introduction

II. Set-up experiment

III. Lessons learned

IV. Next steps

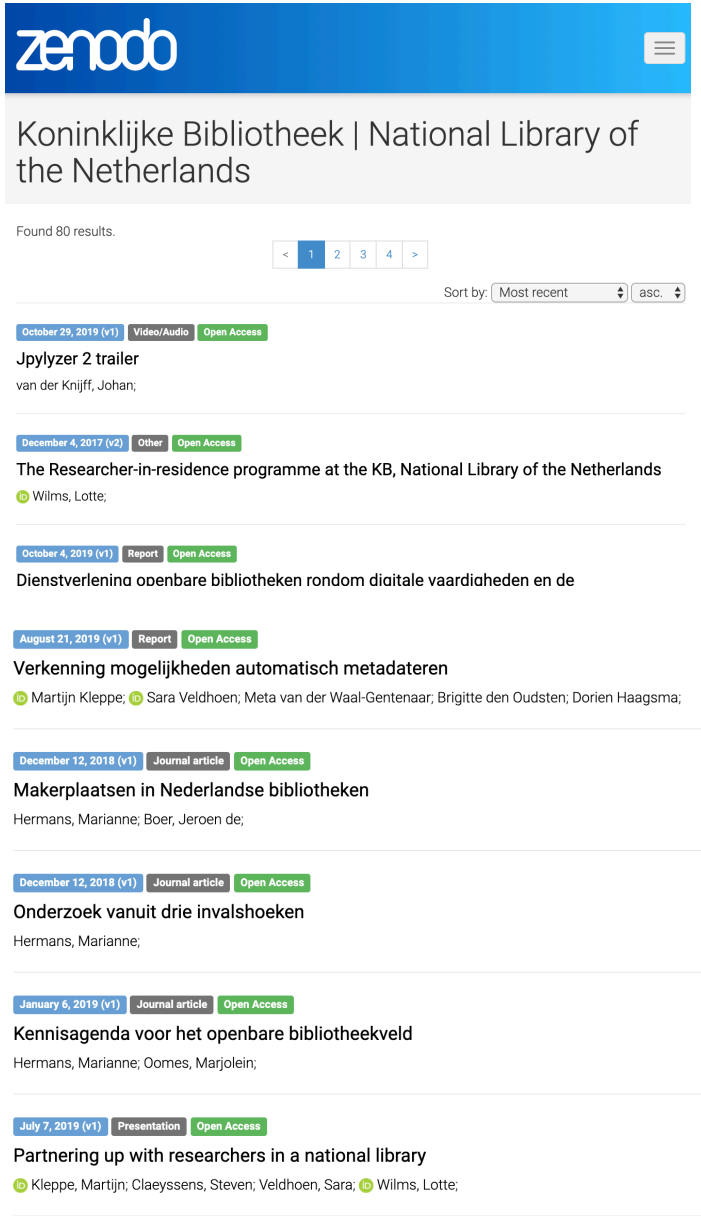


KB nationale bibliotheek

I. INTRODUCTION

Introduction

- About me
- **Research Department** at KB, National Library of the Netherlands (18 fte)
- **Topics**: Digital Preservation, Public Library Research, Copyright, Data Science
- **KB Researchagenda 2018-2022**



https://zenodo.org/communities/kbnl/search?page=1&size=20

https://www.kb.nl/en/organisation/research-expertise

KB nationale bibliotheek

Introduction - KB Research Agenda

- 5 Research themes:
  - Informationsociety
  - Publications
  - **Access & Sharing**
  - Customers
  - Impact

- **8 Researchgroups** with KB colleagues from the whole organisation

https://www.kb.nl/en/organisation/research-expertise/research-agenda-2018-2022
https://doi.org/10.5281/zenodo.1254226

KB nationale bibliotheek

Introduction - KB Research Agenda

- **Short term**: Proof of Concept, internships, researcher-in-residence, **workshops**
- **Long term:** Collaborate with partners: academic, libraries, industry
- 1 Researchgroup on **(Semi-) automated metadata**

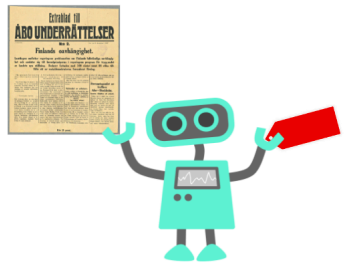
https://www.kb.nl/en/organisation/research-expertise/research-agenda-2018-2022
https://doi.org/10.5281/zenodo.1254226

KB nationale bibliotheek

Introduction – Research Group

- **Literature review:**
  - Media sector
  - Heritage institutes
  - Libraries

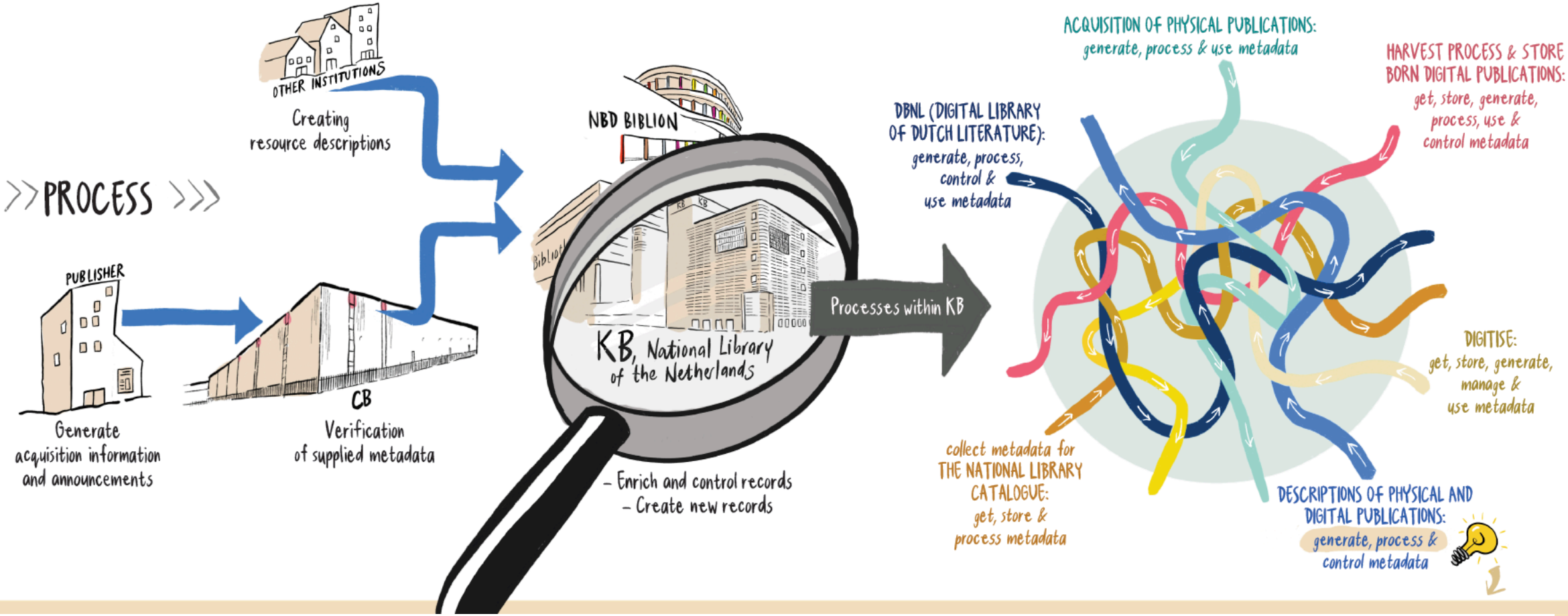
Bookarang)



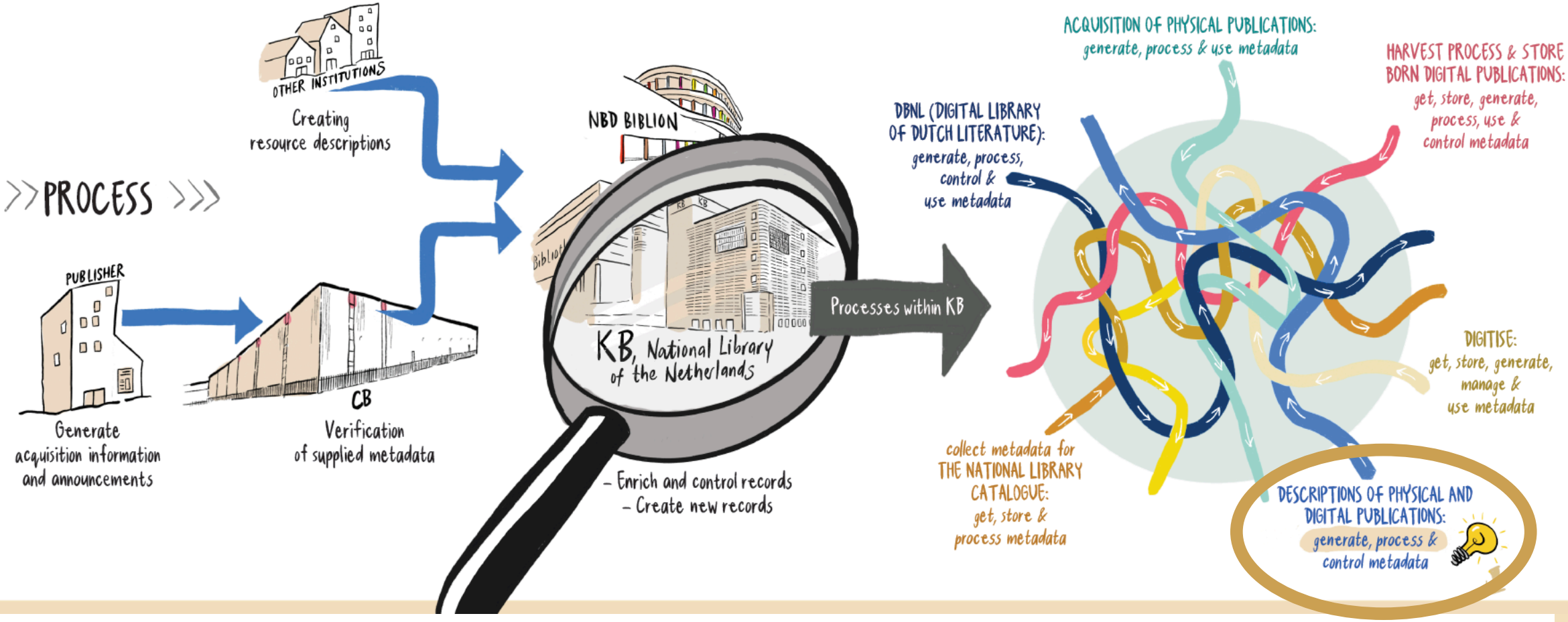
- **Site visits**

- **Which part of the process** do we focus on?

Introduction – Metadata process at KB



Introduction – Metadata process at KB



Introduction – Research Group

- Literature review:
  - Media sector
  - Heritage institutes
  - Libraries

- Sight visits

- Which part of the process do we focus on?

- **ICT with Industry Workshop**

KB nationale bibliotheek

Introduction – ICT with Industry Workshop

- Dutch Research Council (NWO)
- Formulate use-case & get selected
- Small funding required (1,5K EUR)
- Full week
- 13 participants
- Workingspace & hotel Lorentz Center Leiden

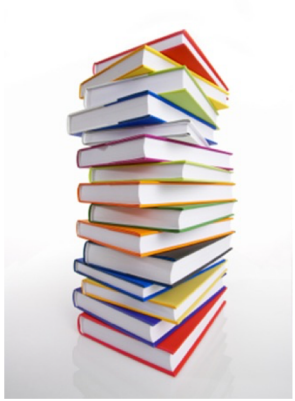
https://www.lorentzcenter.nl/lc/web/2019/1061/info.php3?wsid=1061&venue=Oort

KB nationale bibliotheek

II. SET-UP EXPERIMENT

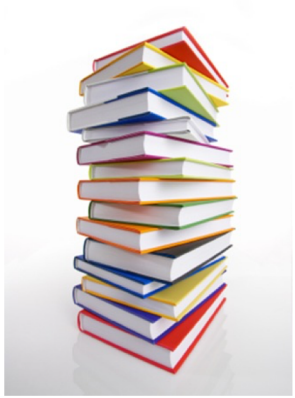
Since 1789

Researcher



Physical publications

Since 1789



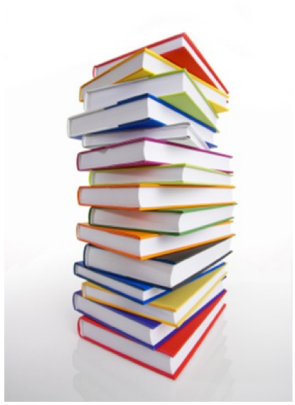
Researcher

Search Interface

Physical Repository

Physical publications

Since 1789



Researcher

Search Interface

Physical Repository

Manual Annotation of keywords

Physical publications

Since 2003

Researcher

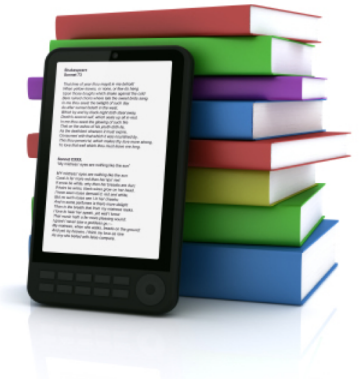
Search Interface

Digital Repository

Manual Annotation of keywords

Full text digital publications

Since 2019



Researcher

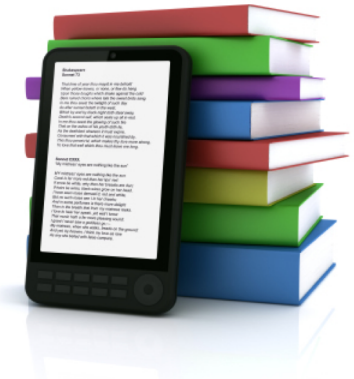
Search Interface

Physical & Digital Repository

Manual Annotation of keywords

Physical & full text digital publications

Since 2019



Researcher

Search Interface

Physical & Digital Repository

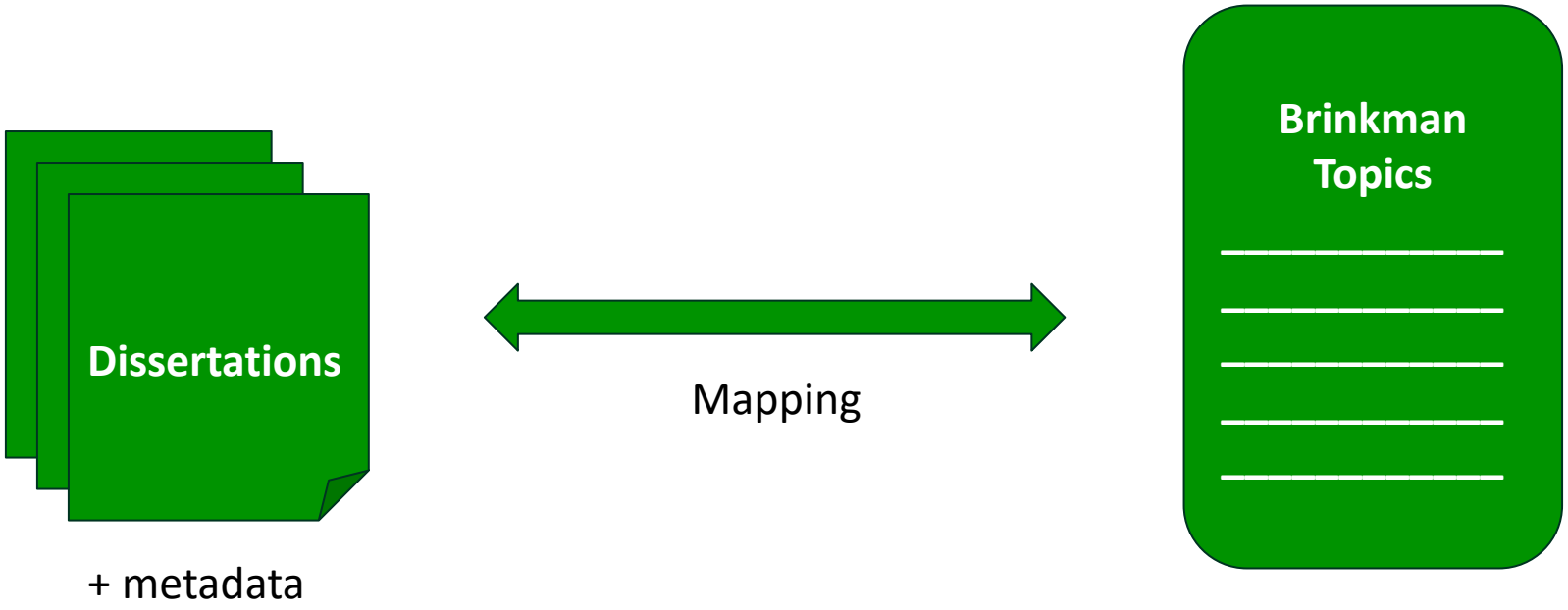
Annotation of keywords

Physical & full text digital publications

Set up - Research question

Research question:

How can we automatically label dissertations with relevant keywords from the Brinkman thesaurus?

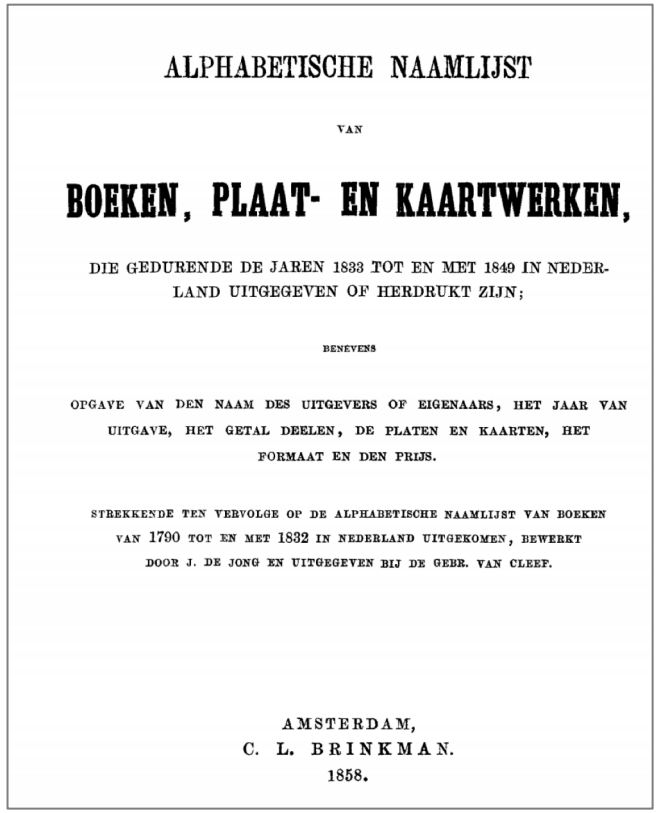


Set up – Data & Thesaurus

- Data – **Dissertations**: Full text and metadata via 6 university libraries

- Thesaurus Brinkman - '**Brinkeys**': 15K keywords, since 1885

➢ Challenge: Map dissertations of university libraries with titles in KB Catalog



ALPHABETISCHE NAAMLIJST

VAN

BOEKEN, PLAAT- EN KAARTWERKEN,

DIE GEDURENDE DE JAREN 1833 TOT EN MET 1849 IN NEDERLAND UITGEGEVEN OF HERDRUKT ZIJN;

BENEVENS

OPGAVE VAN DEN NAAM DES UITGEVERS OF EIGENAARS, HET JAAR VAN UITGAVE, HET GETAL DEELEN, DE PLATEN EN KAARTEN, HET FORMAAT EN DEN PRIJS.

STREKKENDE TEN VERVOLGE OP DE ALPHABETISCHE NAAMLIJST VAN BOEKEN VAN 1790 TOT EN MET 1832 IN NEDERLAND UITGEKOMEN, BEWERKT DOOR J. DE JONG EN UITGEGEVEN BIJ DE GEBR. VAN CLEEF.

AMSTERDAM,
C. L. BRINKMAN.
1858.

KB nationale bibliotheek

Set up – Data & Thesaurus

In the Ideal World:

Every Thesis has an ISBN
Every Author has an ORCID
Every thesis is in Dutch (or English)
A title is always written consistently
Author names are written consistently
All text is in UTF-8
Every university uses the same keywords consistently

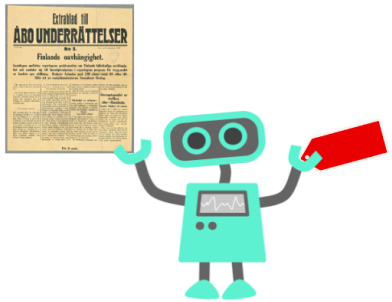
Set up – Approaches

- **Naive Baselines**:
  - Lexical overlap between titles and Brinkeys
  - Lexical overlap keywords universities and Brinkeys

- **Methods**:
  - **Naive Bayes:** simple machine learning algorithm that predicts a Brinkey on the basis of the words that appear in the title and/or a summary
  - **Word Embeddings:** neural networks that places the meaning of words in a continuous virtual "vector space"
  - **Fasttext**

Set up – Approaches

- **Annif**
  - Finnish National Library
  - Use own thesaurus
  - Open Source
  - Combination of techniques



16 40
THE NATIONAL LIBRARY OF FINLAND

http://annif.org/

- **Ariadne**
  - OCLC Research
  - Trained on a lot of data
  - Scores very well
  - Not open source

OCLC Research

https://www.oclc.org/research/themes/data-science/ariadne.html

KB nationale bibliotheek

Set up – Results

Method	Recall			Precision	
	At1	At10	At20	At1	At3
Baseline 1 (overlap titel - Brinkey)	16.9			30.5	
Baseline 2 (overlap Unikey - Brinkey)	11.6			14	
Methode 1 (Naive Bayes classifier)	3.5			6.5	
Methode 2 (Multi-lingual word embeddings)			24.8		6.6
Methode 3 (FastText classifier)			40.3		16.2
Tool 1 (Annif)		16.7			16.7
Tool 2 (Ariadne)			56,9		29.2

Focus on Recall: if the system outputs a list of twenty possible Brinkeys, are the correct Brinkeys according to our thesaurus among them?

KB nationale bibliotheek

III. LESSONS LEARNED

KB nationale bibliotheek

Lessons

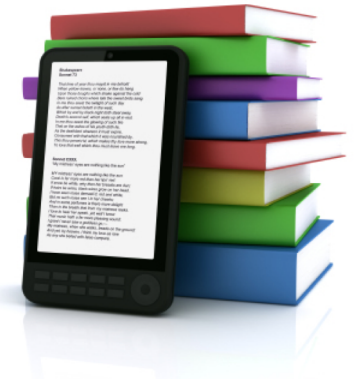
- **Data**, data, data: Quality, Amount
- Do not underestimate **preprocessing**
- **How to keep up** with researchers that go beyond state of the art?
- "The **human perspective, expertise and skill** will remain necessary for guaranteeing the quality that we as the KB, National Library of the Netherlands represent"
- Results still **vague** for cataloguers at KB

KB nationale bibliotheek

III. NEXT STEPS

KB nationale bibliotheek

Next steps



Researcher

Search Interface

Physical & Digital Repository

Manual Annotation of keywords

Ingest physical & full text digital publications

Next steps

Follow up 1:
Interface that **suggests** keywords to annotators

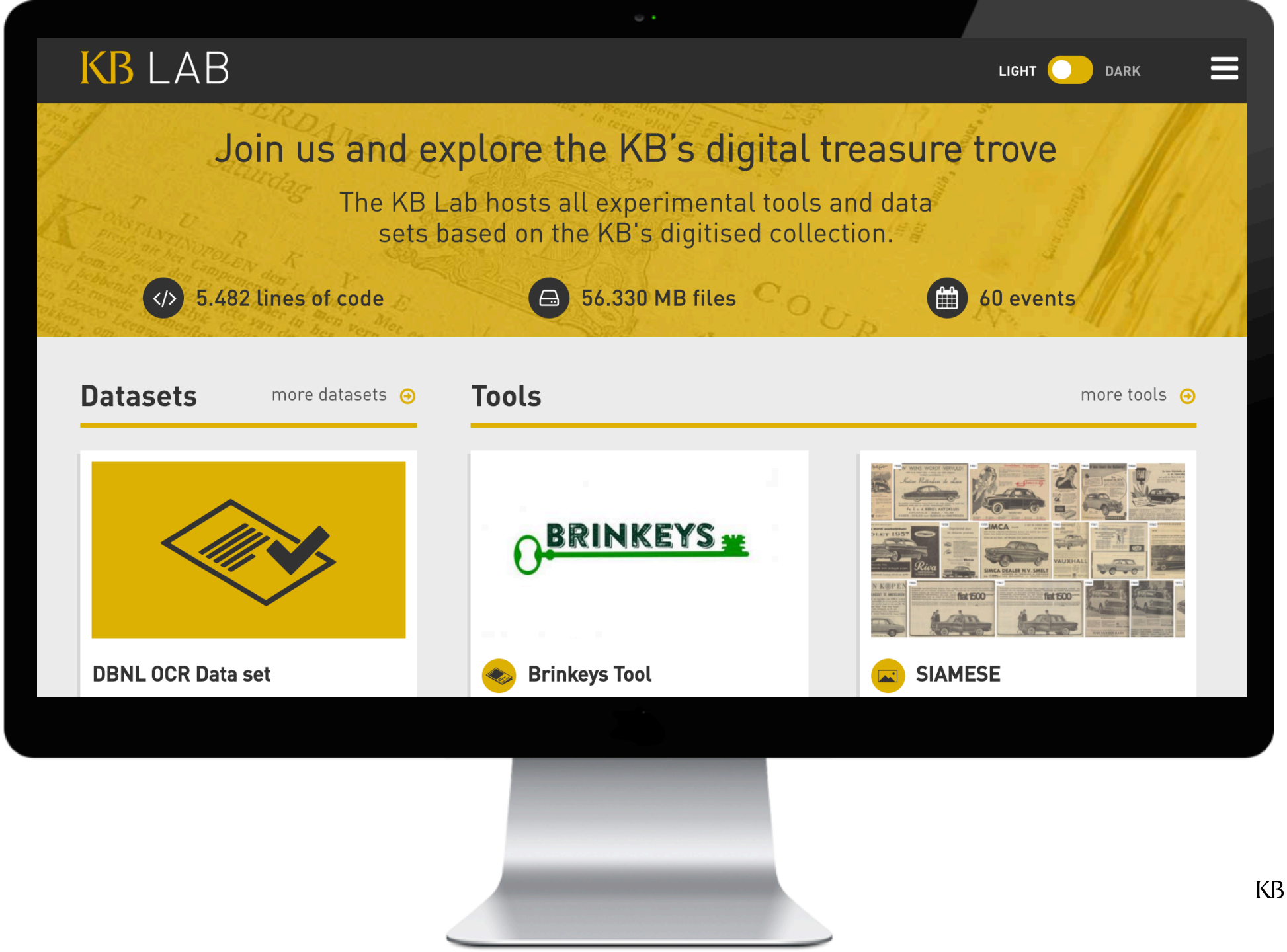
Researcher

Search Interface

Physical & Digital Repository

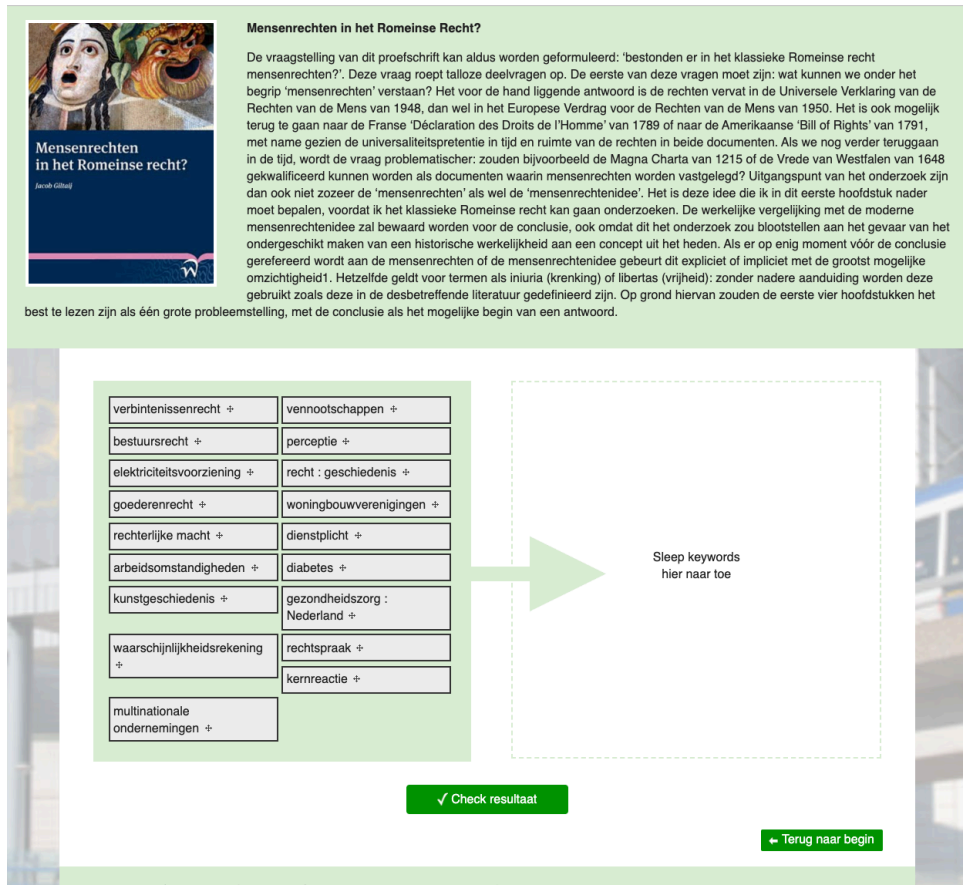
Manual Annotation of keywords

Ingest physical & full text digital publications



http://lab.kb.nl/

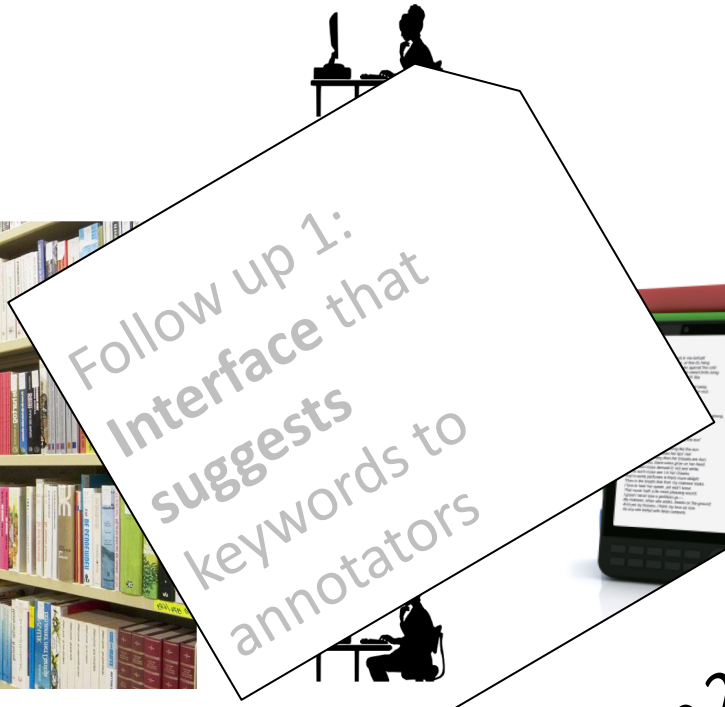
KB nationale bibliotheek

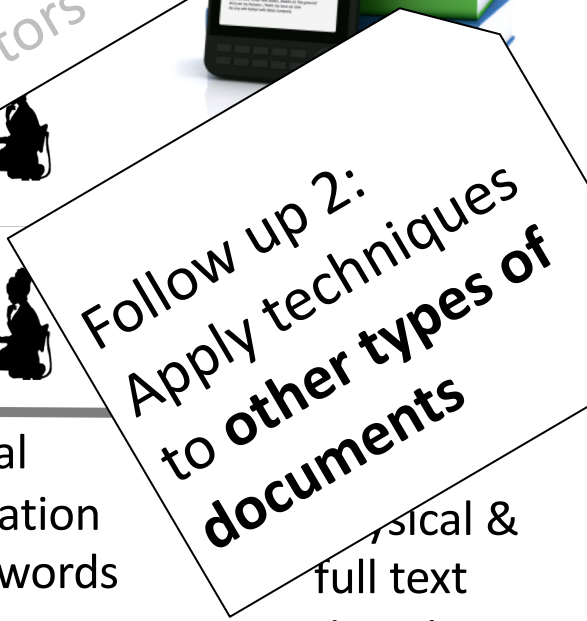


https://lab.kb.nl/tool/brinkeys-tool

KB nationale bibliotheek

Next steps





Researcher

Search Interface

Physical & Digital Repository

Manual Annotation of keywords

ysical & full text digital publications

Next steps

- Experiments with **full text data**:
  - Do we need full text or is title or summary sufficient?
  - Do we need different approaches per type of text?

- Set up a **dedicated & highly secure server** with full text files

- Main focus on **Annif:**
  - Open source
  - Use own thesaurus
  - Active user community (http://swib.org/swib19/programme.html)
  - Experiment with other types of materials: documents of 16&17th century

ACKNOWLEDGEMENT

KB nationale bibliotheek

Fantastic participants ICT with Industry Workshop

Alex Brandsen	Leiden University
Hugo de Vos	Leiden University
Karen Goes	VU Amsterdam
Lin Huang	Leiden University
Hugo Huurdeman	University of Amsterdam
Aruembyeol Kim	VU Amsterdam
Sepideh Mesbah	TU Delft
Myrthe Reuver	Radboud University
Shenghui Wang	University of Twente & OCLC
Richard Zijdeman	IISG & Stirling University
Iris Hendrickx	Radboud University

KB nationale bibliotheek

Great colleagues at KB

Arjan Dekker

Lida Zoutewelle

Angelique Tempels

Meta van der Waal Gentenaar

Irene Wolters

Brigitte den Oudsten

Erik Vos

Enno Meijers

Rene van der Ark

Willem Jan Faber

Sara Veldhoen

Dorien Haagsma

KB nationale bibliotheek

Interested in more?
Working on similar challenges?

LET'S COLLABORATE!

KB nationale bibliotheek

Generating metadata with AI

EXPERIENCE OF THE KB, NATIONAL LIBRARY OF THE NETHERLANDS

Martijn Kleppe – Head of Research
Martijn.kleppe@kb.nl | @martijnkleppe | www.kb.nl/martijnkleppe

KB nationale bibliotheek

THANKS!

Questions?

Please put them in the chat box.

Slides and a recording will be sent to all registered delegates.