



*LIBER Webinar:  
Generating Metadata  
with Artificial  
Intelligence*



# HOST



## **Jeannette Frey**

LIBER President

Director, Bibliothèque Cantonale et Universitaire (BCU) Lausanne

[Jeannette.frey@bcu.unil.ch](mailto:Jeannette.frey@bcu.unil.ch)

# SPEAKER



## Martijn Kleppe

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

[Martijn.Kleppe@kb.nl](mailto: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

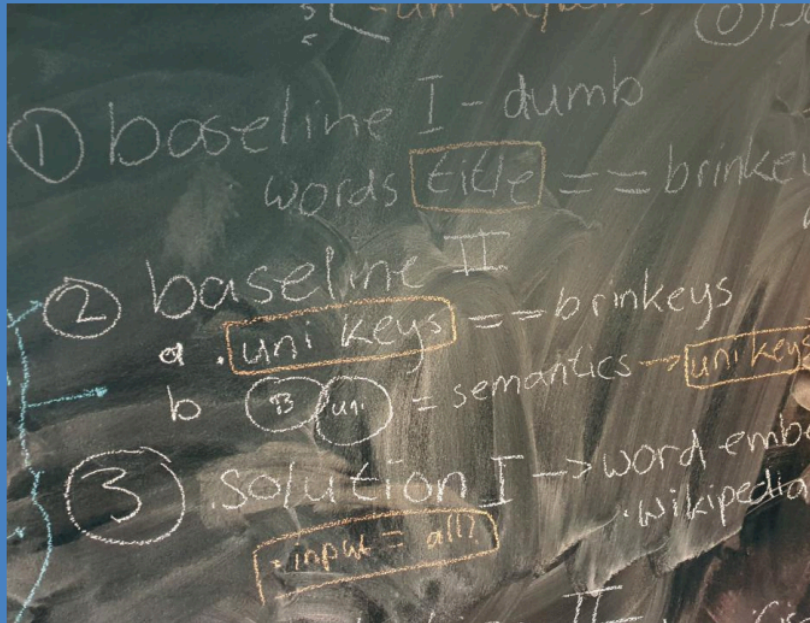
## EXPERIENCE OF THE KB, NATIONAL LIBRARY OF THE NETHERLANDS



**Martijn Kleppe – Head of Research**

[Martijn.kleppe@kb.nl](mailto:Martijn.kleppe@kb.nl) | [@martijnkleppe](https://twitter.com/martijnkleppe) | [www.kb.nl/martijnkleppe](http://www.kb.nl/martijnkleppe)

# Exploration possibilities AUTOMATED GENERATION OF METADATA



Sara Veldhoen



Meta van der  
Waal-Gentenaar



Dorien Haagsma



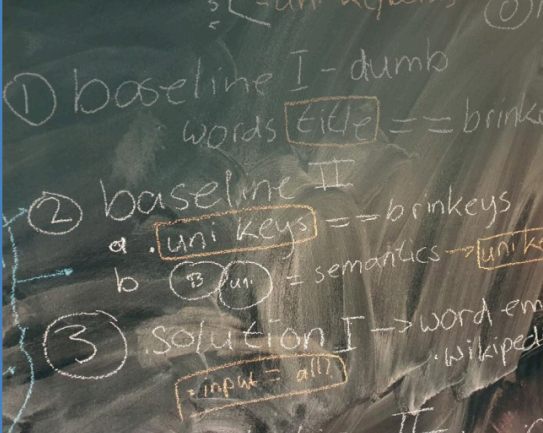
Brigitte den Oudsten

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>

# Outline

- I. Introduction
- II. Set-up experiment
- III. Lessons learned
- IV. Next steps

Exploration possibilities  
AUTOMATED GENERATION  
OF METADATA



August 2019

KB } national library  
of the netherlands

# 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**

zenodo

Koninklijke Bibliotheek | National Library of the Netherlands

Found 80 results.

< 1 2 3 4 >

Sort by: Most recent | asc.

October 29, 2019 (v1) Video/Audio Open Access  
Jpylyzer 2 trailer  
van der Knijff, Johan;

December 4, 2017 (v2) Other Open Access  
The Researcher-in-residence programme at the KB, National Library of the Netherlands  
Wilms, Lotte;

October 4, 2019 (v1) Report Open Access  
Dienstverlening openbare bibliotheken rondom digitale vaardigheden en de

August 21, 2019 (v1) Report Open Access  
Verkenning mogelijkheden automatisch metadateren  
Martijn Kleppe; Sara Veldhoen; Meta van der Waal-Gentenaar; Brigitte den Oudsten; Dorien Haagsma;

December 12, 2018 (v1) Journal article Open Access  
Makerplaatsen in Nederlandse bibliotheken  
Hermans, Marianne; Boer, Jeroen de;

December 12, 2018 (v1) Journal article Open Access  
Onderzoek vanuit drie invalshoeken  
Hermans, Marianne;

January 6, 2019 (v1) Journal article Open Access  
Kennissagenda voor het openbare bibliotheekveld  
Hermans, Marianne; Oomes, Marjolein;

July 7, 2019 (v1) Presentation Open Access  
Partnering up with researchers in a national library  
Kleppe, Martijn; Claeysens, Steven; Veldhoen, Sara; Wilms, Lotte;

<https://www.kb.nl/en/organisation/research-expertise>  
<https://zenodo.org/communities/kbnl/search?page=1&size=20>

# 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>

# 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>

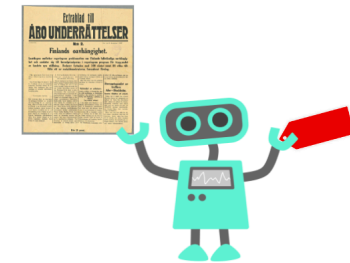
# Introduction – Research Group

- **Literature review:**

- Media sector
- Heritage institutes
- Libraries



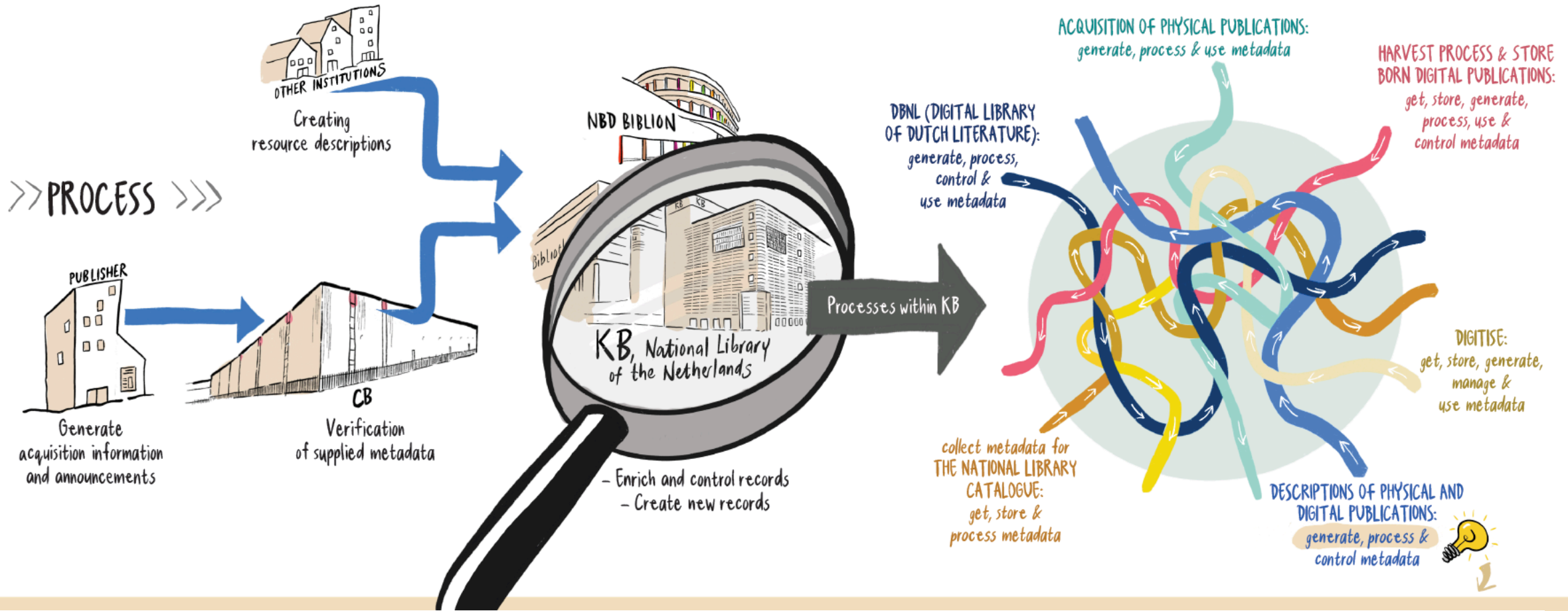
Bookarang )



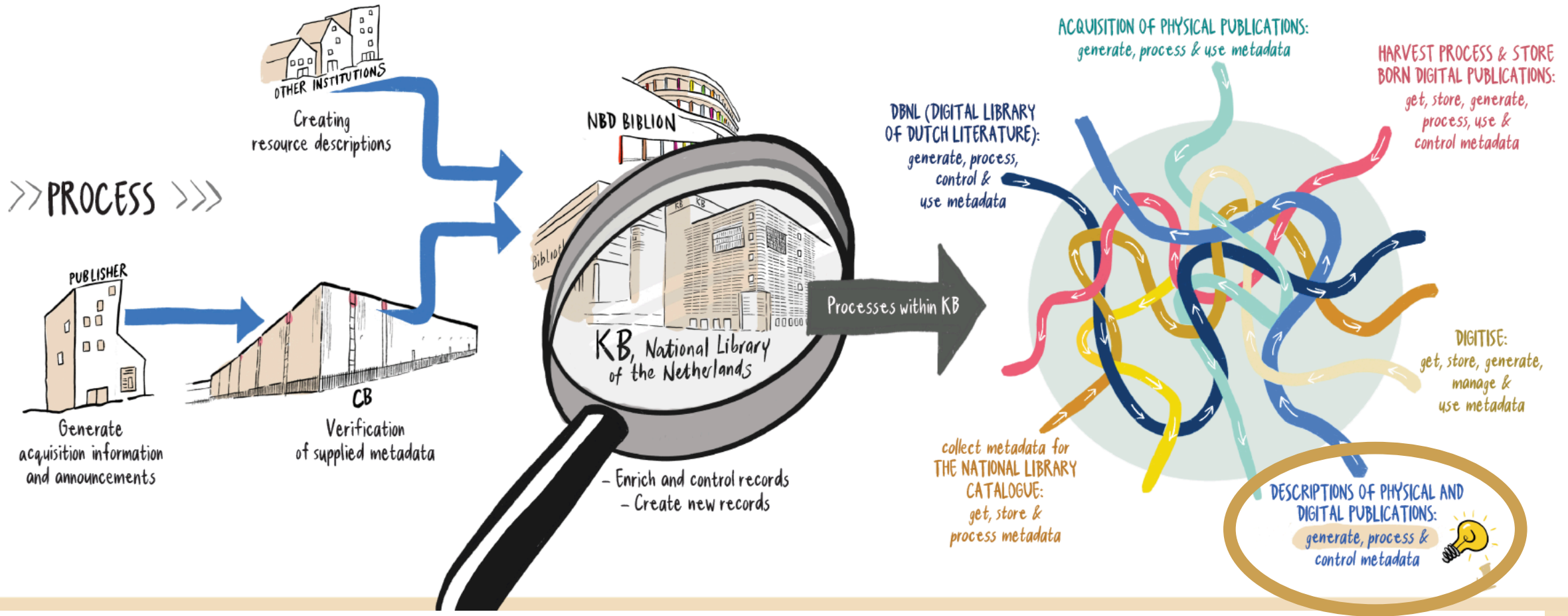
- **Site visits**

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

# Introduction – Metadata process at KB



# Introduction – Metadata process at KB



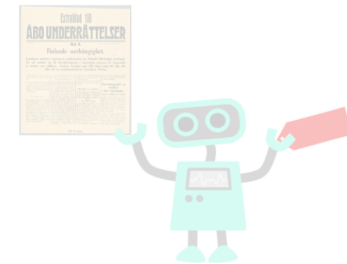
# Introduction – Research Group

- Literature review:

- Media sector
- Heritage institutes
- Libraries



Bookarang )



- Sight visits

- Which part of the process do we focus on?

- **ICT with Industry Workshop**

# 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



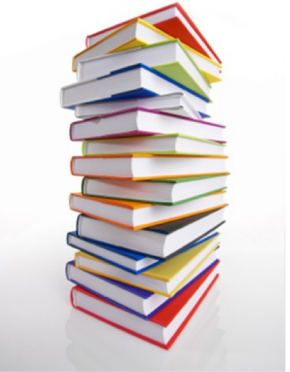


# II. SET-UP EXPERIMENT

# 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

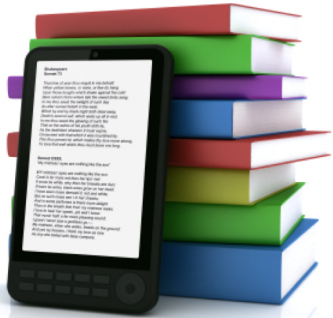
Search  
Interface

Physical &  
Digital  
Repository

Manual  
Annotation  
of keywords

Physical &  
full text  
digital  
publications

# Since 2019



Researcher

Search  
Interface

Physical &  
Digital  
Repository

**Lorentz center**  
Workshop @Oort  
**ICT with Industry 2019**  
50 Researchers Working on 5 Case Studies  
21 - 25 January 2019, Leiden, the Netherlands

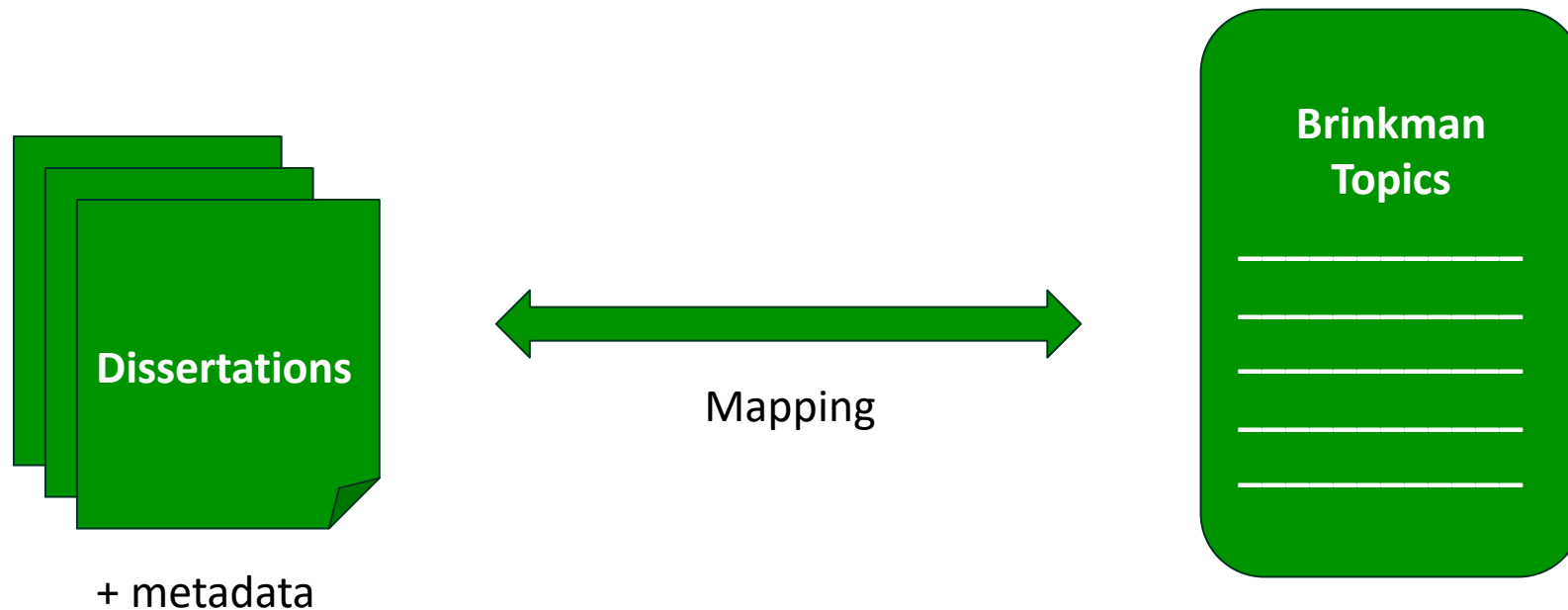
Annotation  
of keywords

Physical &  
full text  
digital  
publications

# Set up - Research question

## Research question:

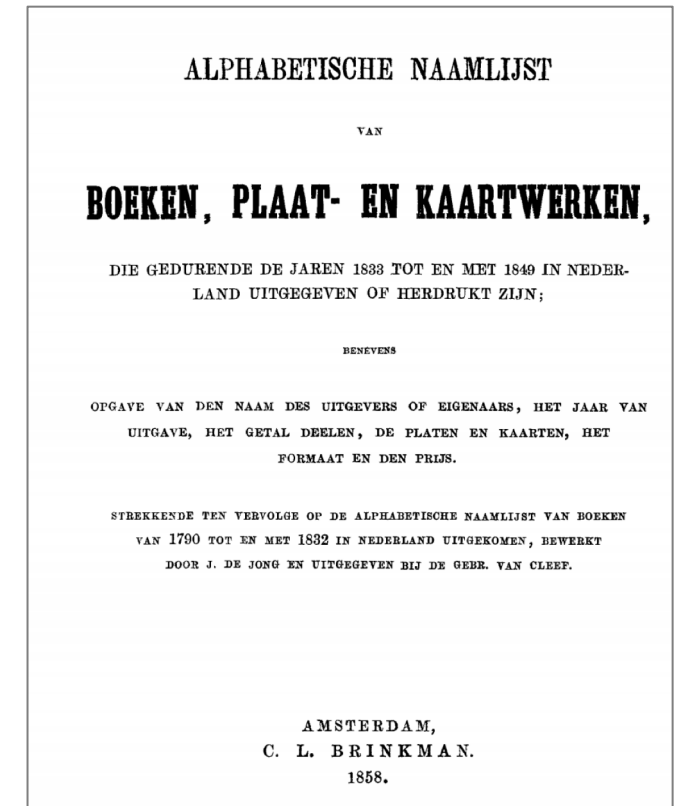
How can we automatically label dissertations with relevant keywords from the Brinkman thesaurus?





# 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



# 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

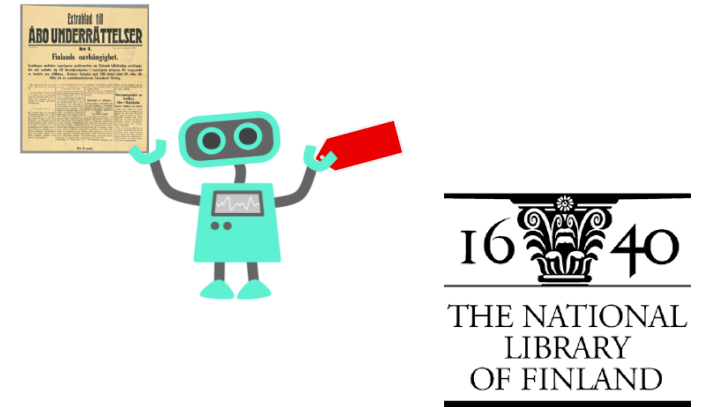
Every university uses the same keywords consistently

# 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



<http://annif.org/>

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



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

# 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?

# III. LESSONS LEARNED

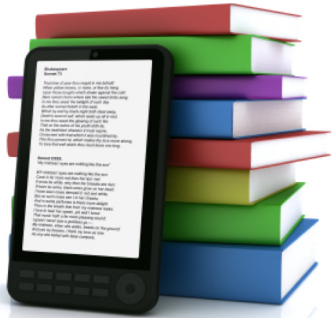
# Lessons

- **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

# III. NEXT STEPS



# Next steps



Researcher

Search  
Interface

Physical &  
Digital  
Repository

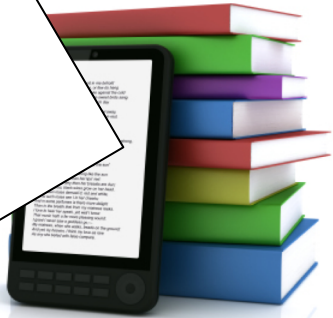
Manual  
Annotation  
of keywords

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

# Join us and explore the KB's digital treasure trove

The KB Lab hosts all experimental tools and data sets based on the KB's digitised collection.

5.482 lines of code

56.330 MB files

60 events

## Datasets

more datasets



DBNL OCR Data set

## Tools

more tools



Brinkeys Tool



SIAMESE

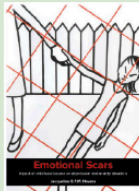
# BRINKEYS

Brinkeys is een porte-manteau van Brinkmanonderwerpen and Keywords. Brinkmanonderwerpen zijn het systeem dat de Koninklijke Bibliotheek (KB) gebruikt om al hun teksten te categoriseren.

In de ICT with Industry workshop hebben we een systeem gebouwd dat automatisch brinkmanonderwerpen kan suggereren voor wetenschappelijke dissertaties. We hebben verschillende methoden geëvalueerd en gekozen voor een systeem gebaseerd op [FastText](#).

Dit systeem zou in de toekomst gebruikt kunnen worden door werknemers van de KB om sneller en nauwkeuriger deze onderwerpen toe te kennen tijdens de metadata generatie.

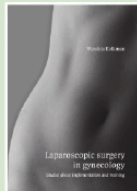
Probeer het hieronder zelf uit!



Meer informatie



Meer informatie



Meer informatie



Meer informatie



Meer informatie



Sleep dissertatie hier naar toe voor analyse

Voor meer informatie, zie dit rapport, of email voor vragen naar Alex Brandsen

KB nationale bibliotheek

Lorentz center

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek

© Alex Brandsen, 2013



## Mensenrechten in het Romeinse Recht?

De vraagstelling van dit proefschrift kan aldus worden geformuleerd: 'bestonden er in het klassieke Romeinse recht mensenrechten?'. Deze vraag roept talloze deelvragen op. De eerste van deze vragen moet zijn: wat kunnen we onder het begrip 'mensenrechten' verstaan? Het voor de hand liggende antwoord is de rechten vervat in de Universele Verklaring van de Rechten van de Mens van 1948, dan wel in het Europese Verdrag voor de Rechten van de Mens van 1950. Het is ook mogelijk terug te gaan naar de Franse 'Déclaration des Droits de l'Homme' van 1789 of naar de Amerikaanse 'Bill of Rights' van 1791, met name gezien de universaliteitspretentie in tijd en ruimte van de rechten in beide documenten. Als we nog verder teruggaan in de tijd, wordt de vraag problematischer: zouden bijvoorbeeld de Magna Charta van 1215 of de Vrede van Westfalen van 1648 gekwalificeerd kunnen worden als documenten waarin mensenrechten worden vastgelegd? Uitgangspunt van het onderzoek zijn dan ook niet zozeer de 'mensenrechten' als wel de 'mensenrechtenidee'. Het is deze idee die ik in dit eerste hoofdstuk nader moet bepalen, voordat ik het klassieke Romeinse recht kan gaan onderzoeken. De werkelijke vergelijking met de moderne mensenrechtenidee zal bewaard worden voor de conclusie, ook omdat dit het onderzoek zou blootstellen aan het gevaar van het ondergeschikt maken van een historische werkelijkheid aan een concept uit het heden. Als er op enig moment vóór de conclusie gerefereerd wordt aan de mensenrechtenidee gebeurt dit expliciet of impliciet met de grootst mogelijke omzichtigheid<sup>1</sup>. Hetzelfde geldt voor termen als iniuria (krenking) of libertas (vrijheid): zonder nadere aanduiding worden deze gebruikt zoals deze in de desbetreffende literatuur gedefinieerd zijn. Op grond hiervan zouden de eerste vier hoofdstukken het

best te lezen zijn als één grote probleemstelling, met de conclusie als het mogelijke begin van een antwoord.

verbintissenrecht +	vennootschappen +
bestuursrecht +	perceptie +
elektriciteitsvoorziening +	recht : geschiedenis +
goederenrecht +	woningbouwverenigingen +
rechterlijke macht +	dienstplicht +
arbeidsomstandigheden +	diabetes +
kunstgeschiedenis +	gezondheidszorg : Nederland +
waarschijnlijkheidsrekening +	rechtspraak +
	kernreactie +
multinationale ondernemingen +	

Sleep keywords hier naar toe

✓ Check resultaat

← Terug naar begin

# Next steps



Follow up 1:  
Interface that  
suggests  
keywords to  
annotators



Follow up 2:  
Apply techniques  
to **other types of  
documents**

Researcher

Search  
Interface

Physical &  
Digital  
Repository

Manual  
Annotation  
of keywords

Physical &  
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

# Fantastic participants ICT with Industry Workshop



**Alex Brandsen**

**Hugo de Vos**

**Karen Goes**

**Lin Huang**

**Hugo Huurdeman**

**Aruembyeol Kim**

**Sepideh Mesbah**

**Myrthe Reuver**

**Shenghui Wang**

**Richard Zijdemans**

**Iris Hendrickx**

Leiden University

Leiden University

VU Amsterdam

Leiden University

University of Amsterdam

VU Amsterdam

TU Delft

Radboud University

University of Twente & OCLC

IISG & Stirling University

Radboud University



# 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

Interested in more?  
Working on similar challenges?

LET'S COLLABORATE!

# Generating metadata with AI

## EXPERIENCE OF THE KB, NATIONAL LIBRARY OF THE NETHERLANDS



**Martijn Kleppe – Head of Research**

[Martijn.kleppe@kb.nl](mailto:Martijn.kleppe@kb.nl) | [@martijnkleppe](https://twitter.com/martijnkleppe) | [www.kb.nl/martijnkleppe](http://www.kb.nl/martijnkleppe)



# THANKS!

## *Questions?*

Please put them in the chat box.

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