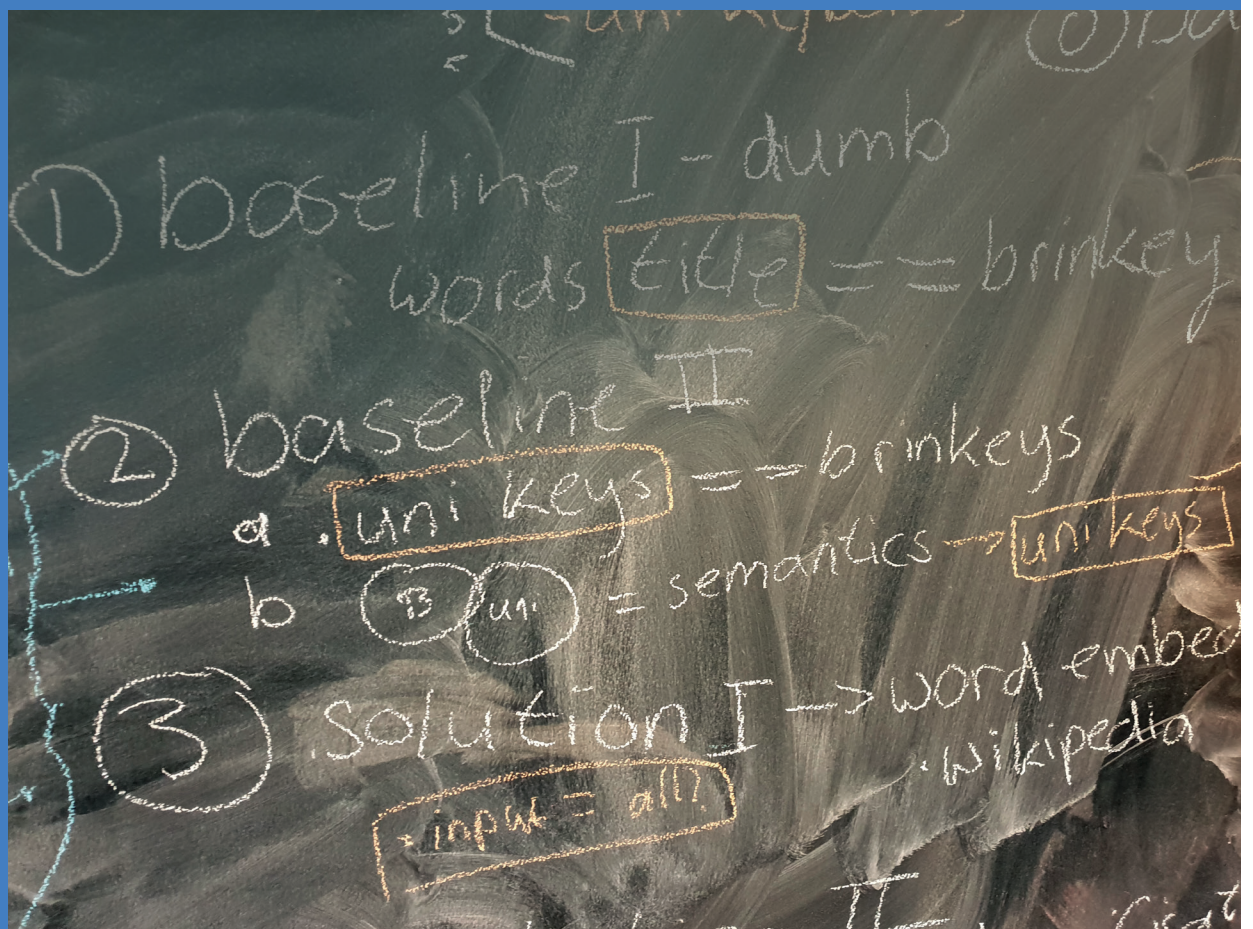


Exploration possibilities

AUTOMATED GENERATION OF METADATA



Colofon

Authors: Martijn Kleppe (Orcid 0000-0001-7697-5726; ISNI 0000 0003 8995 1058), Sara Veldhoen (Orcid 0000-0002-8376-0886; ISNI 0000 0004 7703 0857), Meta van der Waal-Gentenaar, (ISNI 0000 0003 9600 2854), Brigitte den Oudsten (ISNI 0000 0004 7703 0865), Dorien Haagsma (ISNI 0000 0004 7703 0873), all employees of KB

Editor-in-chief: Erik Jan Harmens (ISNI 0000 0000 3808 057X)

Design: Carlien Keilholtz

DOI: <http://doi.org/10.5281/zenodo.3375192>

Photo cover: Martijn Kleppe

Year of publication: 2019

Place of publication: Den Haag

License ©: This publication is under a CC-BY license. This means that anyone is free to copy, distribute and forward this publication by any medium and in any file format, except the photographs. This publication may also be altered or used to produce derivative works for any and all purposes, including commercial purposes. In the event of any use or reuse of this publication, the authors of the work must be stated, with a link to the license and an indication of whether the work has been altered. This is permitted to a reasonable extent but not to the point of creating the impression that the licensor has consented to the work or the use of the work.

See also: <https://creativecommons.org/licenses/by/4.0/deed.nl>



INHOUDSOPGAVE

Introduction	5
Automated attachment of metadata elsewhere	6
Media	7
Heritage	7
Libraries	8
Generating metadata and options for automation within the KB	10
Brief description of process	12
Quality of metadata	12
Subject interfacing	13
Options for automatically generating metadata	13
Results: automatic assignment of keywords	14
ICT With Industry Workshop	15
Challenges	16
Approach	17
Results and demo	18
Lessons and next steps	20
Data, data, data	21
Further action	22
Sources	23

INTRODUCTION

How can we use smart technologies to simplify the description of publications? This is one of the research questions on the KB Research Agenda that we intend to answer in the coming years.¹

At present, the Koninklijke Bibliotheek (KB), National Library of the Netherlands assigns the resource description (also referred to as “generation of metadata” or “creating bibliographic records”) by hand and partially by adopting the data we acquire through other channels. In part due to the growth in electronically generated material (“born digital”) and the growth in website storage, we expect a growing need for the retention of increasing numbers of publications in the coming years. For this reason we explore to optimise the options for manual description of publications. Two current developments offer opportunities: the growing volume of publications available in entirely electronic format, and the fact that smart technologies, for example artificial intelligence (AI) applications such as machine learning, are expanding the possibilities for having electronic texts be interpreted automatically (by computer).

In this white paper we describe the state of our initial explorations of the options for automated generation of metadata of publications. We first present an overview of the ways in which organizations and enterprises outside the KB are using smart technologies to analyse and describe sources such as news articles, books, television broadcasts and photographs. We then discuss how we within the National Library are currently describing titles in order to indicate where in the process we see opportunities for automated attachment of metadata. In the third chapter we discuss the result of our own experiments with the automated assignment of keywords to publications. We conclude with the lessons we have learned so far and discuss our next steps.

Our mission is to ensure that our network organisation uses the power of the written word to help make the Netherlands smarter, more competent and more creative. This is why we are sharing our findings not only within the KB, but with all interested parties who are working with or have an interest in these or related developments. Only by sharing knowledge and working together can we enable everyone to read, learn and research.



Metadateren

resource description. Also referred to as “generation of metadata” or “creating bibliographic records”

Automated GENERATION OF METADATA elsewhere

The automatic analysis of media content is something that has seen numerous applications over recent years. At present, perhaps the most familiar in the areas of artificial intelligence and machine learning are the voice-controlled digital assistants on mobile phones. Current smartphones are now capable of converting a spoken question into text and then analysing it for commonly used terms, persons, locations or concepts.

Until recently, this type of application was mainly used by major tech companies, but thanks to the availability of larger and larger data sets, open source software and advances in computing power, we are beginning to see these technologies applied in other domains. In this chapter we will review a number of these applications in neighbouring sectors of the KB, National Library of the Netherlands and describe what we have already done in this area.

Media

Journalists are already using various software applications in various ways to analyse their articles and photos to improve their findability. Leading American journal Forbes works with a content management system that assesses the content of an article as the journalist writes it in order to suggest possible keywords.² Getty uses a similar application to assist editors working on a story with finding appropriate photos for that story.³ The New York Times partners with Google Cloud on an application that both digitises and generates automatic descriptions for the photos in the NYT photo archive.⁴ In the Netherlands, media conglomerate RTL uses speech recognition and image recognition applications to index TV content.⁵

Alongside simplifying the production process and the search process for editorial content, a number of media parties are also working on consumer applications. One feature that has been worked on extensively in recent years is the “recommender”, which recommends similar or related articles, series or music based on the content of programs or news reports. Familiar examples are the recommendations that Spotify and Netflix make on the basis of music previously listened to or series and films previously viewed. There are also recommenders that make recommendations on the basis of text content. In the Netherlands, the country’s leading financial newspaper Het Financieele Dagblad is doing this in their online edition⁶, and Blendle is also doing this in its personalized newsletter.⁷ Additionally, the Dutch public broadcasting network NPO (Nederlandse Publieke Omroep) is working on a recommender to recommend television programs on the basis of what it is calling the “public value” of a program.⁸

We are also seeing similar applications used by book publishers; in the Netherlands, Bookarang is one example. Where recommenders in webshops are limited to statistics based on simple transactions (“People who bought this book also bought...”), Bookarang is capable of making recommendations on the basis of the content of book. The same technology is also used by AKO, a Dutch retail chain, and the Online Bibliotheek (Online Library).⁹ WPG Uitgevers, an indepent group of media companies, has worked with Driven by Data to create a similar application, the Thrill Seeker, specifically for readers seeking recommendation in the thriller genre.¹⁰ This recommender is unique in its transparency: the reader is offered insight into why the recommender is making its recommendation, and the components behind the recommendation. This may be subgenre, content of story, emotional arc, presence of certain keywords, or overall theme. Finally, while Google’s Talk to Books is not strictly a recommender, it does use similar technologies and enables the user to formulate a question on the basis of which the Google machine learning algorithms search for a quote from a book found in the Google Books index.¹¹

Heritage

Within the heritage sector, there are currently interesting movements underway for the automatic analysis of digitised and original digital heritage material. The National Archive has conducted a trial experiment with the classification of incoming e-mails, and is also working with Huygens ING and the *O*

(Network of War History Resources) within the TRIADO research project on a number of trials including automatic classification of documents from the CABR (Central Archive for Special Dispensation of Justice (*Centraal Archief Bijzondere Rechtspleging*)).¹² The subject of the documents are determined using deep learning technologies.¹³ Naturalis Biodiversity Center is using the same type of machine learning technologies for the automatic classification of images of insects.¹⁴ On this they are working with a number of other organisations including Waarneming.nl and Observation.org; in the project, volunteers upload their images of insects along with descriptions. This data is used as training material for developing an algorithm, known as the “classifier”, which will be able to independently identify insects based on the descriptions provided by the volunteers.¹⁵ KB is also working together with partners in both the academic and private sectors on similar research, not only within academic research projects but also in the context of our researcher-in-residence programme, the result of which we share in our KB Lab: <https://lab.kb.nl/>.¹⁶ As one example, together with Frank Harbers of the University of Groningen we have explored the potential for automatic assignment of journalistic genres to historical newspaper articles, and we are working further on this in a joint project called NEWSGAC¹⁷ with a number of other partners, including CLARIAH and the eScience Center. With Thomas Smits (Utrecht University) and Melvin Wevers (KNAW Humanities Cluster), we explored the potential for image recognition applications to, for example, classify images in historic newspapers as photos, drawings and cartoons, as well as the potential for processing these images for content.¹⁸ With Puck Wildschut (Radboud University) we experimented with recognising individual novels and charting their relationships.¹⁹

The institution in the Netherlands that is perhaps most advanced in this is the Netherlands Institute for Media Culture, Sound and Vision (*Nederlands Instituut voor Beeld en Geluid*). Drawing on their many years of experience in national and European research projects, since 2018 they have been able to go beyond the experiment and implement automatic description of TV and radio programmes in their production process.²⁰ They have implemented speech recognition software for converting TV and radio programmes into text, automatic speaker labelling for assigning names to the text generated, and facial recognition for identifying persons in television programmes.²¹ Finally, on the basis of the text generated, they assign subjects to the programmes on the basis of a keyword system (thesaurus) used by Sound and Vision.²² This last application would appear to be the most interesting for developments within the domain of library sciences.

Libraries

Ever since digital texts first began becoming available in the nineteen-fifties, the library sciences sector has seen the potential for automatic title description as a means of simplifying the description process for publications. For a good overview of the developments in these early years, read any article by Robert David Stevens (University of Manchester) or Karen Spärk Jones (Cambridge University Computer Laboratory). In this white paper, we will restrict ourselves to the most recent developments. For our purposes, good sources are the International Federation of Library Associations and Institutions (IFLA), the Ligue des Bibliothèques Européennes de Recherche (LIBER), the Dublin Core Metadata Initiative (DCMI) and Semantic Web in Libraries (SWIB).

With respect to simplifying the description process for publications, we see two developments: firstly, an assessment of the potential for automatically assigning keywords from an independently maintained thesaurus. At the 2018 IFLA conference, the German National Library presented a good overview of their progress in this area. At present they are using commercial software that relies on a support vector machine algorithm to identify patterns in texts and then assign keywords based on these patterns.²³ The German National Library is aware of potential errors and the users’ interest in obtaining complete and correct information, which is why they present the automatically generated keywords in their catalogue in a separate field with the label: *maschinell ermittelt* (“machine-generated”).

There are also plans in the Netherlands for using commercial software to assign keywords and other metadata. As one example, Bookarang (referred to above) is working with NBD Biblion to explore the possibilities.²⁴ Together they are working on a tool to generate a resource description, including keywords and a review, from the PDF of a book.

The Swedish National Library is using a similar approach with their own, self-developed software.²⁵ An initial insight was that the data set for training material was too limited in scope, point highlighted by Joseph Busch in his paper "Categorization Ethics: Questions about Truth, Privacy and Big Data".²⁶ The Norwegian National Library has tried to solve this problem by automatically generating a number of documents.²⁷ Norway is also leading the field with efforts towards the smart analysis of multimedia collections, both for library clients and internal processes.²⁸

The Finnish National Library has put a great deal of energy into the creation of Annif: an open-source system for generating keywords on the basis of a proprietary thesaurus and training data.²⁹ Annif uses a number of different open-source tools for natural language processing and machine learning packages like Maui, FastText and Gensim. It is being actively developed, and has a growing user group.³⁰

Along with automatic assignment of keywords from a proprietary thesaurus, we see a second development: various libraries are exploring the options for adding extra metadata to make it easier for users to find publications. For example, since 2016 the National Library Board of Singapore has been working on recognizing and labelling names of persons from texts using Named Entity Recognition (NER). In 2017 subjects were also generated from texts using controlled databases such as Wikidata and Geonames, which means that now a number of collections can be searched using an integrated interface.³¹ The George A. Smathers Library of the University of Florida has chosen a similar approach but with a focus on special collections. Here, metadata fields have been expanded to include the thesaurus of the digital JSTOR collection, as well as geographic location data; this increased searchability and findability.³² The KB has also experimented significantly in these areas. For example, Theo van Veen, Juliette Lonij and Willem Jan Faber have been working hard in recent years on identifying entities such as names and places in historical newspapers using Named Entity Software. These entities are then linked to databases such as DBpedia and Wikidata. The extra information that this has produced has been placed in an internal, experimental environment on Delpher.nl. Users can now search for characteristics of a person even where those characteristics are not specifically named in the newspaper.³³ We have also seen a similar approach in the JSTOR Text Analyzer tool.³⁴ In this tool users can load a text and have the system analyse it, extract relevant entities and on the basis of the results make suggestions for relevant articles in the JSTOR database.

Generating metadata and OPTIONS FOR AUTOMATION within the KB

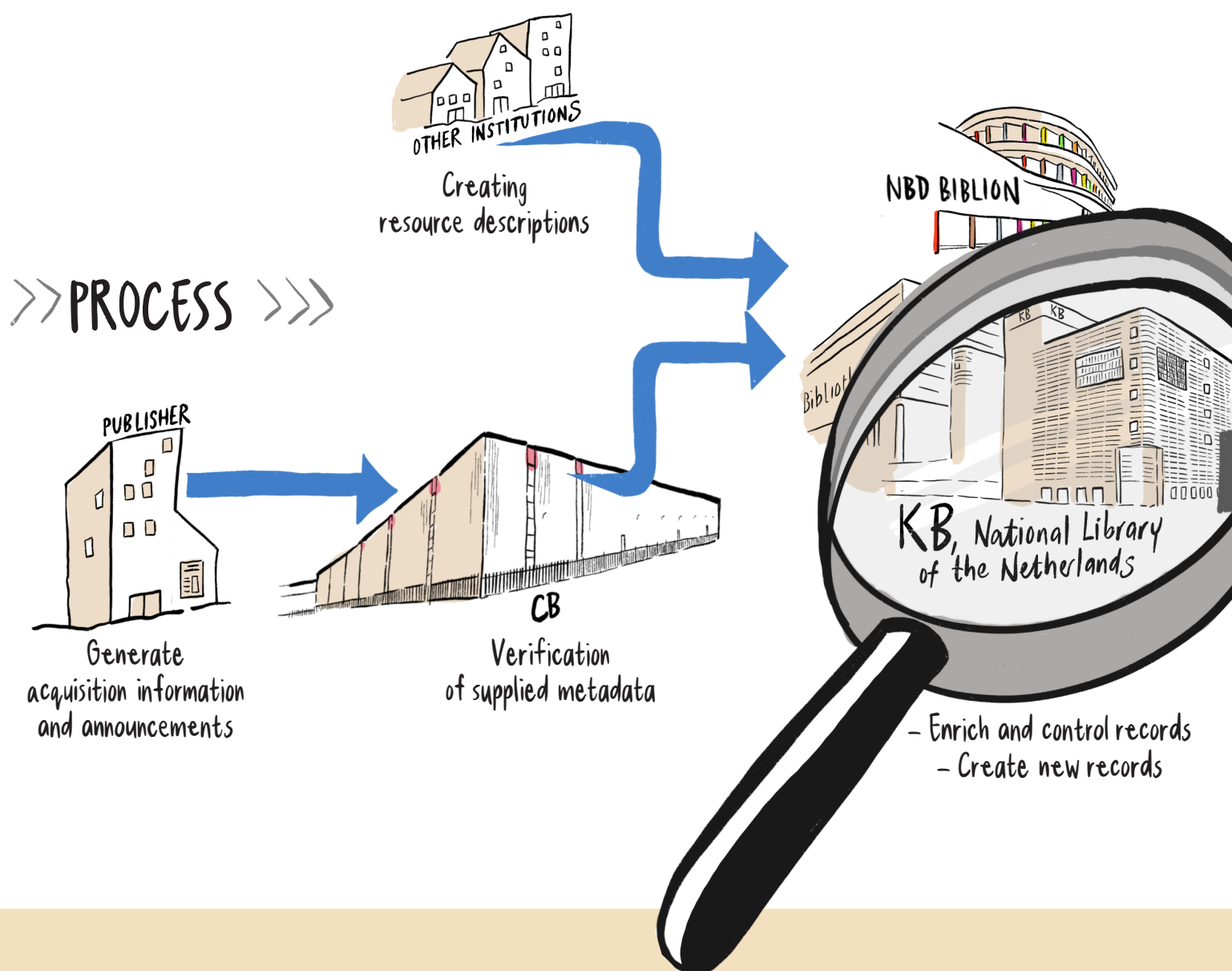
Following our inventory of the options for automatic metadata assignment, in this chapter we describe how we in the KB, National Library of the Netherlands are currently describing the publications and assigning metadata. Here we will indicate where in the process we are exploring the options for smarter generation of metadata.

Brief description of process

Within the KB we generate metadata in a number of ways. Some of the processes are conducted externally (see the very highly generalized diagram at the next pages). Of all the publications received by the KB, approximately 70% already have a record in the Shared Automated Catalogue System (*Gemeenschappelijk Geautomatiseerd Catalogiseersysteem* (GGC)). The bulk of these records consist of metadata originating from publishers and supplied via the *Centraal Boekhuis*³⁵; this is the “ONIX-metadata”.³⁶ A very small number of records originate from other organisations; for example, up to a few years ago the university libraries also catalogued in the GGC.

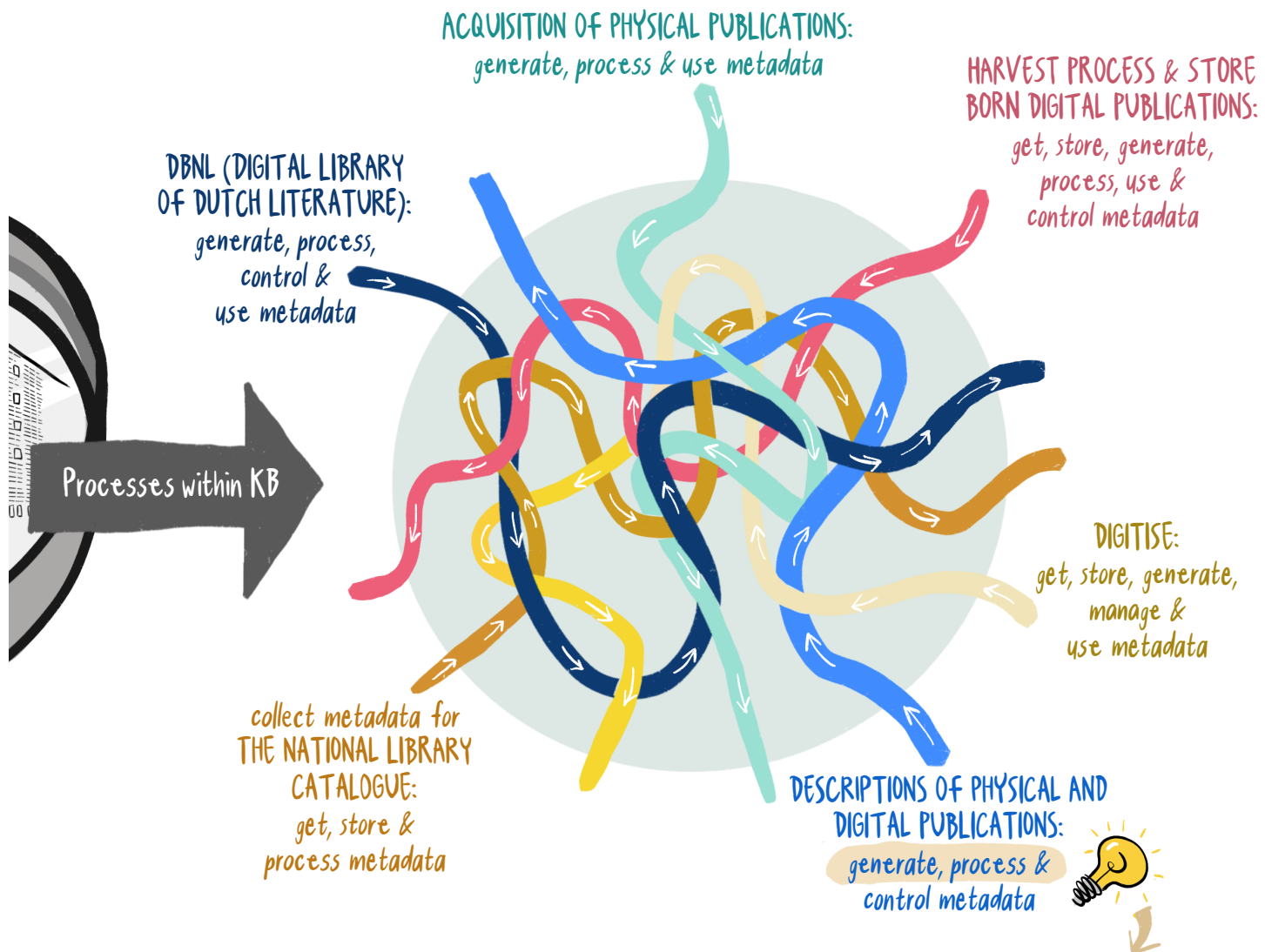
The metadata in the records supplied by the Centraal Boekhuis are not complete. For example, the data has not yet been linked to a thesaurus, and in some cases additional data concerning the content to be added. NBD Biblion³⁷ and the KcatlogiB are enriching and verifying the metadata. If there is no metadata available yet, the KB’s catalogers generate a new record. Along with generating metadata during the cataloguing process, new metadata is also generated in other processes.

EXPLORATION POSSIBILITIES



AUTOMATED GENERATION OF METADATA

within the KB, National Library of the Netherlands



SURVEY CONDUCTED:

Innovation in Brinkman keywords
(in generation and processing of metadata)

FUTURE SURVEYS:

to be determined...

Quality of metadata

For the KB, the quality of the metadata is important because we consider sustainable storage important and because we as a national library want to ensure that reliable information about author and title is available for at least every Dutch publication (the National Bibliography). At least in reference to Dutch titles, the National Bibliography must be qualified as the definitive reference file in the national and international library sector and within the academic world.

The goal of description or generation of metadata is to allow users to properly find, identify, select, obtain and explore information.³⁸ Metadata includes (but is not limited to) identifiers, title information, information on author, publication type and carrier, information about the metadata itself and administrative information such as location, annotations and subject cataloging. The KB defines metadata in accordance with the internationally recognized Resource Description and Access (RDA) standard.³⁹

Description remains necessary. It is possible to search the full text of digital publications, but by no means is every publication already available in digital format. Metadata makes it easy for a user to find multiple publications on the same subject, in multiple languages or in any other available spellings. A subject may not be set out in so many words in the publication itself, but it may still be very important within that subject area. Keywords make such publications easily findable for the user. Additionally, with metadata a user can easily find multiple editions of a publication or determine the most recent edition. Finally, metadata adds context to a publication.⁴⁰ A reliable National Bibliography can only be considered authoritative with good, verified metadata and a verified link to multiple thesauri.

Subject cataloging

Subject cataloging refers to making the material aspect of a publication accessible using keywords or classifications. The biggest added value of subject cataloging is when verified keywords, terms and codes can be used from a thesaurus or classification, because this makes it possible to create groups that give the user a view to all publications that may be relevant to him or her.

Subject cataloging improves the quality of the descriptive and bibliographic metadata. Manually adding this metadata is a time-consuming activity, which is why the KB assigns only non-fiction by subject with its own Dutch-language thesaurus: the “Brinkman subjects”.⁴¹ This is a thesaurus that is used for subject cataloging a significant portion of the National Library collection. Children’s books also have specific characteristics and genres assigned from a thesaurus for cataloging purposes for the Central Children’s Books Database (Centraal Bestand Kinderboeken (CBK)).⁴² Fiction, however, is only assigned a Brinkman subject keyword for genre or form. In the case of subject cataloging for fiction, these terms come primarily from NBD Biblion.

Options for automatically generating metadata

The research into options for automatically generating metadata looked at where the biggest needs are and what would deliver the most efficiency for the KB. Improvement of the supplied metadata by consulting with external organisations in the chain was outside the scope of this exploration, although the KB did discuss with various parties on various initiatives relating to the automatic generation of metadata, as described in the preceding chapter. This can produce better and more complete metadata, even though we will always wish to exert a certain degree of editorial monitoring in order to guarantee the quality of our metadata.

Existing processes that already automatically generate metadata also fell outside the scope of this exploration. This refers to, for example, the processing of born digital publications and digitised material that generates various types of metadata, such as technical (i.e. file format, check value) and administrative metadata.

We looked at which types of metadata could be generated automatically and which cannot, such as links with a thesaurus, annotation and administrative data. We describe the results of this research in the following chapter.



Results:

AUTOMATIC ASSIGNMENT of keywords

The first exploration that we conducted ourselves was focused on subject cataloging by means of the automatic assignment of keywords. For this exploration we used the Brinkman subjects. At present Brinkman subjects are assigned manually. A cataloger has the publication on his or her desk and makes an assessment of the keywords to be assigned. Because we also increasingly have these publications in fully digital formats, we investigated the extent to which we could train the computer to make automatic suggestions for these types of Brinkman subjects, referred to by the research team as “Brinkeys”.



ICT with industry workshop participants from left to right: Martijn Kleppe (KB), Rob Koopman (OCLC Research), Karin Goes (VU), Shenghui Wang (OCLC Research), Areumbyeol Kim (VU), Myrthe Reuver (Radboud Universiteit), Iris Hendrickx (Radboud Universiteit), Sara Veldhoen (KB), Alex Brandsen (Universiteit Leiden), Hugo de Vos (Universiteit Leiden), Sepideh Mesbah (TU Delft), Hugo Huurdeman (UvA), Richard Zijdemann (IISG).

ICT With Industry Workshop

We took the first steps towards this during the ICT with Industry Workshop in January 2019. This was a workshop organized by the Dutch Research Council (NWO, the Nederlandse Organisatie voor Wetenschappelijk Onderzoek) and held at the Lorentz Center in Leiden, where private sector and civil society organisations met to work out a case in collaboration with academic partners.⁴³ The team of researchers working on the KB case had a variety of backgrounds and areas of expertise, and varied in level from Master's students to senior researchers. Iris Hendrickx, a researcher affiliated with Radboud University in Nijmegen, was the academic leader of the project. The results that we describe in this chapter are based on the report that the team drafted after the conclusion of the workshop.⁴⁴

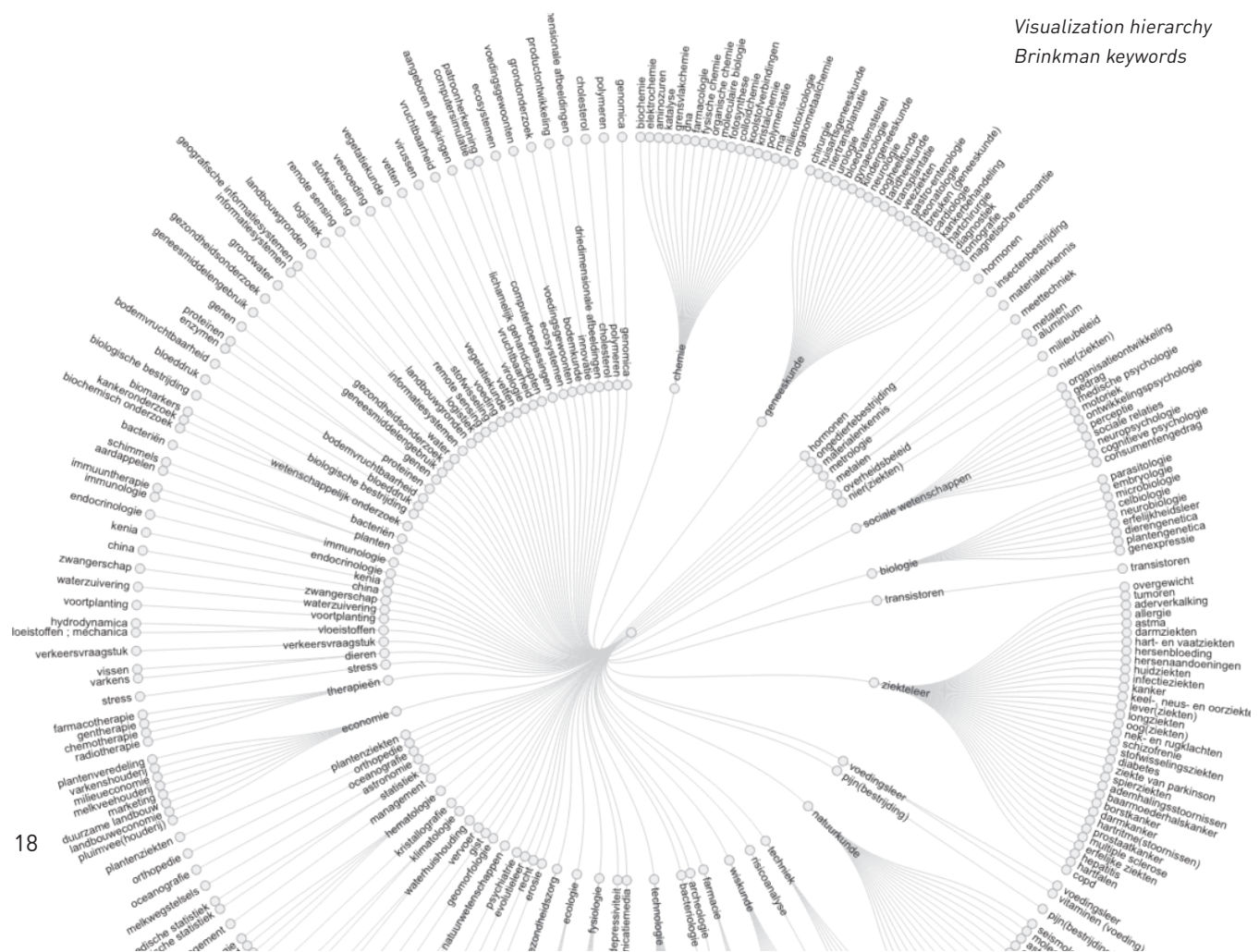
Our goal for the week was to explore various methods for the automatic allocation of keywords to digital publications. Because we worked with publications outside the walls of the KB, National Library of the Netherlands, we experimented with dissertations issued as open access publications and saved in the KB's Digital Depot. Because we had difficulty extracting the publications from the Digital Depot, with the permission of the university libraries we reacquired the dissertations from the digital archives of the relevant universities: Groningen, Delft, Rotterdam, Leiden, Utrecht and Wageningen. There was also metadata available from these sources, albeit of extremely varying nature and quality.

Challenges

Being able to work with this data presented a number of challenges:

- The linking of the resource descriptions from the GGC to the dissertations was no simple task, because based on the data sources a variety of identifiers had been assigned (i.e., titles and codes). Where known, we used the ISBN, and where this was unknown we used the author's name (surname) in combination with the title. For these latter two identifiers we used "string distance", because the lexical comparison is prone to causing errors as a result of special characters, accent marks, etc.
- While some of the dissertations were in Dutch, for the most part they were written in English, and this presented a problem in that the Brinkman subjects are in Dutch. A great deal of time and effort went into translating the terms and title words in an attempt to resolve this.
- The Brinkman subjects are, in principle, hierarchical, which is a usable characteristic for automating subject interfacing. However, the depth of the hierarchy was extremely unreliable, as the visualization below shows.
- A link was made between the Brinkman subjects and Wikidata, because the latter exhibited more connections between the subjects and additionally is available in multiple languages.

Visualization hierarchy
Brinkman keywords



Approach

Over the course of the week, a number of different approaches were tried. Firstly, to get an idea of how difficult the problem is, the “naïve” baselines were used; then, we investigated a number of more advanced methods, and finally we released two tools from external parties developed for this application onto the data.

In the first naïve baseline, we looked at the lexical overlap between title words and **Brinkeys**. If there are any exact matches to Brinkeys in the title, these are identified as a subject or potential subject. To check this, titles in English have to be translated, because the thesaurus is in Dutch. Translating the titles for this was no easy task, and we were surprised to find it produced lower scores than expected.

For the second naïve baseline, we looked at the lexical overlap between **Unikeys** (subjects as assigned by universities) and Brinkeys. Here again, we had to work with translations of English terms, and so we encountered the same problems and considerations that came along with that. This method scored lower than the previous method, which is in part explained by the fact that the Unikeys were not assigned according to a verified vocabulary. Additionally, it is likely that the cataloguers based themselves on the title, which would by definition lead to a high overlap.

The first method that we investigated was **Naïve Bayes**, a simple machine learning algorithm that predicts a Brinkey on the basis of the words that appear in the title and/or a summary. In contrast to the other methods, this method always assigns exactly one object. This method produced very poor results and was quickly set aside in favour of the second method.

That second method was **Word Embeddings**, a fairly recent technology based on neural networks that places the meaning of words in a continuous virtual “vector space”. This allows you to search by word or text string for other words (in this case: Brinkeys) that are most related in terms of meaning. With a few adjustments, this works for multiple languages simultaneously, which allowed the English dissertations to be linked to the Dutch-language Brinkeys. This method produce promising results and there are a great many ideas for adapting it and investigating it further.

The third method was **FastText**: not specifically developed for this task but rather a general tool for the classification of texts.⁴⁵ The input consisted of title, summary, Unikeys and the name of the institute where the research was carried out. This method scored fairly highly.

The first tool that we investigated was **Annif**: an existing solution developed by the Finish national library.⁴⁶ This tool offers the possibility to use your proprietary thesaurus of keywords.⁴⁷ Under the hood we find a combination of existing modules for natural language processing and machine learning. These modules can be combined in various ways, and the tool also allows the use of your own thesaurus as a list of possible subjects. We ran this tool on our data set with TF-IDF⁴⁸ as weighting factor and using Snowball.⁴⁹ As input, we entered the title and summary (if available) of each dissertation to produce ten subject predictions on the basis of the input. Doing this, the scores did not come out particularly high, but that could improve significantly with the input of more data and experimentation with other configurations.

We also investigated **Ariadne**, a tool recently developed by the OCLC (Online Computer Library Center) for searching and interpreting bibliographical information on the basis of text characteristics.⁵⁰ This approach achieved the highest scores (by far), although we still know fairly little about the exact methodology behind it and the material used to train the system.

Results and demo

In order to demonstrate the results, the team built a demo that is available on the KB Lab: <http://lab.kb.nl/tool/brinkeys-tool>. In the demo, you can drag an example dissertation to the “analysis box”, which will then display the Brinkman subjects that the system would assign to the dissertation and the Brinkman subjects actually assigned by the catalogers of the KB.



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](#)



Nederlandse Organisatie
voor Wetenschappelijk Onderzoek

© Alex Brandsen, 2019.

We also evaluated the systems. This was generally determined on the basis of *precision and recall*.⁵¹ In this case our focus was on recall: if the system outputs a list of twenty possible Brinkeys, are the correct Brinkeys according to our thesaurus among them? We think that a system with a high recall can help the catalogers to find the right subjects fast. In addition, we also measure precision: suppose that a system assigns three subjects completely autonomously; how often are these assignments correct?

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

Due to the explorative structure of the exploration, the results presented above cannot be interpreted in a definitive and unambiguous way; there were a great number of variations in the experiments. It is, however, clear that Ariadne performs particularly well. This is a tool specifically developed for this type of task, and additionally it was pre-trained on a large volume of academic literature. The disadvantage of this tool, however, is that unlike Annif it is not open source, which means we do not have enough of a grip on the details of its functioning and functionality. The FastText classifier also scored reasonably high, and can be used within Annif as an alternative backend. In any event, we can conclude from these high scores that it is certainly possible to automatically assign subjects in a meaningful way.

Lessons and NEXT STEPS

This initial exploration produced good results and valuable insight. However, we still have a long way to go before we arrive at a system that can be used in practice. In this chapter we identify the most important lessons from our exploration and the next steps.

Data, data, data

The most important aspect of experiments like these is the data being used. We experienced this at various levels; firstly, in the availability of the training data. As already described, we obtained this through various sources, which entailed various challenges in the area of data harmonisation as well as in the creation of the test set. Because in the Netherlands there are multiple parties working on similar analyses of publications, for our language area it may be worth exploring the possibilities of the construction that HathiTrust in the USA offers.⁵² This group not only saves digital publications permanently, but also (through its research centre) makes them available for various applications relating to text and data mining.⁵³

DaAn additional factor was that the amount of data was fairly limited, certainly when you consider that there were over 2200 different Brinkeys assigned, some of which only appear very sporadically. Also, the workshop only involved the use of titles and summaries, and not yet the full text. We can see from the performance of Ariadne that there is certainly room to further develop the system.

Finally, the results were evaluated on the basis of correspondence to the subjects in the catalogue. It may be that another system would suggest different Brinkeys that are just as appropriate, or even more appropriate, to the dissertation. To analyse this, we would ideally want to have a number of experts look at a number of these lists in order to further refine the training material.

Further action

On the basis of these lessons, we are currently working on a number of potential follow-up steps. Firstly, we are exploring Annif's potential with other types of data. Annif's advantages are that it is open source, combines multiple techniques, and has multiple applications. The results with dissertation texts were promising, but we do not yet know how this will work with other types of content. This is why we are now looking at the results that Annif produces when we work, for example, with summaries or full texts of books. We are interested in whether the quality of the suggested Brinkman subjects improves when a complete book is analysed, or whether the analysis of a portion of a book is sufficient.

Another direction we are currently looking at is the automatic addition of metadata to the texts on the website of the Digital Library of Dutch Literature DBNL.⁵⁴ On this website you will find texts from Dutch literature, linguistics and cultural history from the earliest time to the present. The DBNL digitises texts at a very high level of quality; these texts are then incorporated in their entirety into an XML-TEI format, and have a margin of error of less than 0.005%. A great deal of attention is also being devoted to enriching the texts, which ensures good access. There are references to relevant information, such as information about authors, titles, place names and data. This manually added metadata can be used as training and testing material.

In the next step of our exploration, we will assess the options for automatically assigning both content-related and structural metadata to these texts. For the content-related metadata we intend to use Named Entity Recognition to extract information from the texts. This information is then be linked to a thesaurus, taking into account multiple meanings (disambiguation) and spelling variants. The structural data relates to markers of structural elements such as headings, page numbers, poetry, tables, etc. With some titles (usually prose), the structure is clear, but this becomes somewhat more complicated with study books and periodicals. The question is whether we can automatically assign any or all of this type of information.

With this exploration and future explorations we hope to boost knowledge on the potential and limitations of automated generation of metadata both now and in the near future. The principle here is that this should facilitate the work of catalogers, rather than fully replacing them. The human perspective, expertise and skill will remain necessary for guaranteeing the quality that we as the KB, National Library of the Netherlands represent.

SOURCES

1. *Research Agenda for the National Library of the Netherlands 2018-2022*. (not dated). Retrieved from <https://www.kb.nl/organisatie/onderzoek-expertise/onderzoeksagenda-2018-2022>.
2. Willens, M. (2019, 3 January). *Forbes is building more AI tools for its reporters*. Retrieved from <https://digiday.com/media/forbes-built-a-robot-to-pre-write-articles-for-its-contributors/>.
3. Kunova, M. (2018, 6 August). *Getty Images launches a new AI tool that helps publishers find the right picture for the story*. Retrieved from <https://www.journalism.co.uk/news/getty-images-launches-a-new-ai-tool-that-helps-publishers-find-the-right-picture-for-the-story/s2/a725797/>.
4. He, A. (2018, 9 November). *The New York Times Digitizes Millions of Historical Photos Using Google Cloud Technology | The New York Times Company*. Retrieved from <https://www.nytc.com/press/new-york-times-google-cloud/> & Greenfield, S. (2018, 9 November). *Picture what the cloud can do: How the New York Times is using Google Cloud to find untold stories in millions of archived photos | Google Cloud Blog*. Retrieved from <https://cloud.google.com/blog/products/ai-machine-learning/how-the-new-york-times-is-using-google-cloud-to-find-untold-stories-in-millions-of-archived-photos>.
5. ICT with Industry 2019 – ICT Research Platform Netherlands. (undated). Retrieved from <https://ict-research.nl/ict-with-industry/ictwi2019/>.
6. ICT with Industry 2019 – ICT Research Platform Netherlands. (undated). <https://ict-research.nl/ict-with-industry/ictwi2019/> & Sappelli, M., Chu, D., Cambel, B., Graus, D. and Bressers, P. (2018). *Smart journalism: personalizing, summarizing, and recommending financial economic news*. The algorithmic personalization and news (apen18) workshop, The International AAAI Conference on Web and Social Media 2018. Retrieved from <https://graus.nu/publications/smart-journalism-position-paper/>.
7. Kuiken, J., Schuth, A., Spitters, M. & Marx, M. (2017, 2 February). *Effective Headlines of Newspaper Articles in a Digital Environment*. Digital Journalism. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1279978>.
8. Kraak, H. (2019, 31 January). *NPO Start wil je niet meer programma's bieden van jouw smaak, maar je smaak verbreden. Dit is hoe ze dat aanpakken. ("NPO Start doesn't want to offer you programs to suit your taste, but wants to expand your taste. Here's how they are doing it")*. De Volkskrant [Dutch newspaper with national distribution]. Retrieved from https://www.volkskrant.nl/cultuur-media/npo-start-wil-je-niet-meer-programma-s-bieden-van-jouw-smaak-maar-je-smaak-verbreden-dit-is-hoe-ze-dat-aanpakken-b59b54c9/?utm_campaign=shared_earned.
9. *KB zet Bookarang in om lezers door te leiden naar volgend boek ("Netherlands National Library uses Bookarang to lead readers to their next book")*. (undated). Retrieved from <https://www.kb.nl/ob/nieuws/2018/kb-zet-bookarang-in-om-lezers-door-te-leiden-naar-volgend-boek>.

10. Thrillseeker. (undated). Retrieved from <https://thrillseeker.io>. & Huijzer, D. (2019, 24 April). *WPG lanceert aanbevelingstool "Thrill Seeker"* ("WPG launches recommendation tool 'Thrill Seeker'"). Retrieved from <https://inct.nl/nieuws/6661/wpg-lanceert-aanbevelingstool-thrill-seeker> & Huijzer, D. (2019, 11 June). *Schwung bij WPG: "Wij leren elke dag nieuwe dingen."* ("WPG's Schwung: 'We're learning new things every day'") Retrieved from <https://inct.nl/news/6749/schwung-bij-wpg-lsquo-wij-leren-elke-dag-nieuwe-din-gen-rsquo->.
11. *Talk to Books*. (undated). Retrieved from <https://books.google.com/talktobooks/> & *Talk to Books by Google AI | Experiments with Google*. (undated). Retrieved from <https://experiments.withgoogle.com/talk-to-books>.
12. van Essen, M. (2019, 27 February). *Machine Learning en Automatische Classificatie - Evaluatierapport* ("Machine Learning and Automatic Classification - Evaluation Report"). Retrieved from <https://kia.pleio.nl/groups/view/53406652/kennisplatform-innovatie/blog/view/55809165/machine-learning-en-automatische-classificatie-evaluatierapport>.
13. *Tribunaalarchieven als digitale onderzoeksfaciliteit* ("Tribunal Archives as digital research facility"). (undated). Retrieved from <https://www.oorlogsbronnen.nl/tribunaalarchieven-als-digitale-onderzoeksfaciliteit> & Klijn, E. (2019 April). *Enorme stap voorwaarts om archieven toegankelijk te maken* ("Enormous step forward in making archives accessible"). *IP* (professional journal for information professionals). Retrieved from <https://www.oorlogsbronnen.nl/sites/default/files/IP%20mei%202019%20TRIADO%20Edwin%20Klijn.pdf>.
14. Hogeweg, L. (2018, 27 March). *Collaboration started to support biodiversity research through artificial intelligence*. Retrieved from <https://science.naturalis.nl/en/about-us/news/onderzoek/collaboration-started-support-biodiversity-research-through-artificial-intelligence/>.
15. Speksneijder, C. (2018, 7 September). *Deze slimme camera telt en herkent insecten* ("This smart camera counts and recognises insects"). *De Volkskrant* (Dutch newspaper with national distribution). Retrieved from <https://www.volkskrant.nl/wetenschap/deze-slimme-camera-telt-en-herkent-insecten-bd3b152d/>.
16. *Researcher-in-residence*. (undated). Retrieved from <https://www.kb.nl/organisatie/onderzoek-expertise/researcher-in-residence>.
17. Lonij, J., Harbers, F. (2016) *Genre classifier*. (2016). Retrieved from <http://lab.kb.nl/tool/genre-classifier>. & Broersma, M., Attema, J., Tjong Kim Sang, E. & Klaver, T. (undated). *Advancing Media History by Transparent Automatic Genre Classification*. Retrieved from <https://www.esciencecenter.nl/project/newsgac>.
18. Smits, T., Faber, W.J. (2018) *CHRONReader*. Retrieved from <http://lab.kb.nl/tool/chronreader>. & Lonij, J., Wevers, M. (2017) *SIAMESE*. Retrieved from <http://lab.kb.nl/tool/siamese>. & Wevers, M., & Smits, T. (2019). *The visual digital turn: Using neural networks to study historical images*. *Digital Scholarship in the Humanities*. Retrieved from <https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqy085/5296356>.
19. Wildschut, P., Faber, W.J. (2017) *Narralyzer*. Retrieved from <http://lab.kb.nl/tool/narralyzer>.
20. de Jong, A. (2018). *De evolutie van het media asset management bij Beeld en Geluid* ("The evolution of media asset management at Sound and Vision"). Retrieved from <https://publications.beeldengeluid.nl/pub/667>.

21. Adrianus J. van Hessen. (undated). Retrieved from <https://research.utwente.nl/en/persons/adrianus-j-van-hessen> & Ordelman, R. J. F., & van Hessen, A. J. (2018). *Speech Recognition and Scholarly Research: Usability and Sustainability*. In I. Skadina, & M. Eskevich (Eds.), CLARIN 2018 Annual Conference (pp. 163-168.) Retrieved from <https://research.utwente.nl/en/publications/speech-recognition-and-scholarly-research-usability-and-sustainab>. & SpraakLab. (undated). Retrieved from <https://www.spraaklab.nl/>.
22. de Boer, V., Ordelman, R. & Schuurman, (2016) *Evaluating unsupervised thesaurus-based labeling of audiovisual content in an archive production environment*. International Journal on Digital Libraries. Retrieved from <https://doi.org/10.1007/s00799-016-0182-6> & de Boer, V., Priem, M., Hildebrand, M., Verplancke, N., de Vries, A., & Oomen, J. (2016). *Exploring Audiovisual Archives Through Aligned Thesauri*. In E Garoufallou, I Subirats Coll, A Stellato, & J Greenberg (Eds.), *Metadata and Semantics Research*. MTSR 2016. Springer, Cham. Retrieved from <http://publications.beeldengeluid.nl/pub/632>.
23. Junger, U. (2018). *Automation first – the subject cataloguing policy of the Deutsche Nationalbibliothek*. Retrieved from <http://library.ifla.org/2213/1/115-junger-en.pdf>.
24. NBD Biblion maakt efficiëntieslag door inzet van kunstmatige intelligentie in samenwerking met Bookarang ("NBD Biblion improves efficiency by using artificial intelligence in cooperation with Bookarang"). (2018, 19 June). [Press release]. Retrieved from https://www.nbdbiblion.nl/sites/nbdbiblion.nl/files/Persbericht+Samenwerking+Bookarang_def.pdf.
25. Golub, K., Hagelbäck, J., & Ardö, A. (undated). *Automatic Classification Using DDC on the Swedish Union Catalogue*. Retrieved from <http://ceur-ws.org/Vol-2200/paper1.pdf>.
26. Busch, J. (2018). *Automatic Classification Using DDC on the Swedish Union Catalogue*. International Conference on Dublin Core and Metadata Applications. Retrieved from <http://dcevents.dublincore.org/Int-Conf/dc-2018/paper/view/556/669>.
27. Brygfjeld, S., Wetjen, F., & Walsøe, A. (2018). *Machine learning for production of Dewey Decimal*. Retrieved from <http://library.ifla.org/2216/1/115-brygfjeld-en.pdf>.
28. Brygfjeld, S. (2019, 13 juni). *Codename Nancy - AI: lessons from the National Library of Norway* [Slides]. Retrieved from <https://www.slideshare.net/sconul/artificial-intelligence-the-national-library-of-norway-svein-arne-brygfjeld-national-library-of-norway>.
29. Annif - tool for automated subject indexing and classification. (undated). Retrieved from <http://annif.org> & Suominen, O., (2019). *Annif: DIY automated subject indexing using multiple algorithms*. LIBER Quarterly, 29(1), pp.1–25 Retrieved from <http://doi.org/10.18352/lq.10285>.
30. Google Groepen. (undated). [Forum-post]. Retrieved from <https://groups.google.com/forum/#!forum/annif-users>.
31. Goh, R. (2018). *Using Named Entity Recognition for Automatic Indexing*. Retrieved from <http://library.ifla.org/2214/1/115-goh-en.pdf>.
32. Hlava, M., Russell, J., & Hansen, D. (2018). *Inverting the Library Cataloguing Process to Streamline Technical Services and Significantly Increase Discoverability and Search for Special Collections*. Retrieved from <http://library.ifla.org/2219/1/115-hlava-en.pdf>.

33. van Veen, T., Lonij, J., & Faber, W. (2016). *Linking Named Entities in Dutch Historical Newspapers*. Zenodo. Retrieved from <http://doi.org/10.5281/zenodo.843504> & van Veen, T. (2019). Wikidata. *Information Technology and Libraries*, 38(2), 72-81. Retrieved from <https://doi.org/10.6017/ital.v38i2.10886> & van Veen, T. (2019, 5 april). *Using Wikidata for entity search in historical newspapers* [YouTube]. Retrieved from <https://www.youtube.com/watch?v=J5mCem-hEMg>.
34. *JSTOR: Text Analyzer*. (undated). Retrieved from <https://www.jstor.org/analyze/>.
35. "CB distributes physical books and e-books in the retail channel and online in the Netherlands and Belgium. CB also offers logistics solutions for the Healthcare market." (undated). Retrieved from <https://www.cb.nl/over-ons>.
36. "All ONIX standards are designed to support computer-to-computer communication between parties involved in creating, distributing, licensing or otherwise making available intellectual property in published form, whether physical or digital." (undated). Retrieved from <https://www.editeur.org/8/ONIX/>.
37. *NBD Biblion supplies "a complete package of products and services that contribute to the success of your school's or institution's library or media library*. NBD Biblion. (undated). Retrieved from <https://www.nbdbiblion.nl/product/nbd-biblion>.
38. Riva, P., Le Bœuf, P., & Žumer, M. (2017). *IFLA Library Reference Model. User Tasks Summary*, p.15. Retrieved from https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf.
39. "RDA is a package of data elements, guidelines, and instructions for creating library and cultural heritage resource metadata that are well-formed according to international models for user-focussed linked data applications." About RDA (undated). Retrieved from <http://www.rda-rsc.org/content/about-rda>.
40. Riley, J. (2017). Understanding Metadata. *What is metadata and what is it for?* Retrieved from https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf & Kuipers, A. (2016, 5 juli). *De comeback van de KB-code ("The comeback of the KB code")*. Retrieved from <https://www.kb.nl/blogs/over-de-kb/de-comeback-van-de-kb-code> & Miller, L. (2010, 9 september). *The trouble with Google Books*. Retrieved from https://www.salon.com/2010/09/09/google_books/.
41. *Geschiedenis van de Nederlandse Bibliografie ("History of Dutch Bibliography")*. (undated). Retrieved from <https://www.kb.nl/organisatie/voor-uitgevers/informatie-over-de-nederlandse-bibliografie/geschiedenis-van-de-nederlandse-bibliografie>.
42. *Over het Centraal Bestand Kinderboeken ("About the Central Children's Books Database")*. Retrieved from <https://www.kb.nl/bronnen-zoekwijzers/kb-collecties/moderne-gedrukte-werken-vanaf-1801/kinderboeken/over-het-centraal-bestand-kinderboeken>.
43. Lorentz Center - *ICT with Industry 2019 from 21 Jan 2019 through 25 Jan 2019*. (undated). Retrieved from <https://www.lorentzcenter.nl/lc/web/2019/1061/info.php3?wsid=1061>.
44. Kleppe, M., Hendrickx, I., Veldhoen, S., Brandsen, A., de Vos, H., Goes, K., ... Zijdemann, R. (undated). KB (National Library of the Netherlands): *(Semi-) Automatic Cataloguing of Textual Cultural Heritage Objects*. Retrieved from <https://kbresearch.nl/brinkeys/report.pdf>.

45. *fastText* (undated). Retrieved from <https://fasttext.cc>.
46. *Annif - tool for automated subject indexing and classification*. (undated). Retrieved from <http://annif.org>.
47. Suominen, O., (2019). *Annif: DIY automated subject indexing using multiple algorithms*. LIBER Quarterly, 29(1), pp.1–25 Retrieved from <http://doi.org/10.18352/lq.10285>
48. "TF-IDF is the degree of importance of a specific word in a text that is produced by comparing the frequency of that word in other texts." Groenewoud, R. (2016, 2 November). TF-IDF: ETF-IDF: Een nieuwe rage in SEO? ("TF-IDF: A new rage in SEO?") - Emerge. Retrieved from <https://www.emerge.nl/best-practice/tfidf-nieuwe-rage-seo>.
49. "Snowball is a small string processing language designed for creating stemming algorithms for use in Information Retrieval." Snowball. (undated). Retrieved from <https://snowballstem.org>.
50. *Ariadne's Thread: Interactive Context Explorer*. (undated). Retrieved from <https://www.oclc.org/research/themes/data-science/ariadne.html> & Wang, S. & Koopman, R. (2017), *Clustering articles based on semantic similarity*, *Scientometrics* (2017) 111: 1017. Retrieved from <https://doi.org/10.1007/s11192-017-2298-x>.
51. "In pattern recognition, information retrieval and binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance in pattern recognition and information retrieval. (2019). Retrieved from https://en.wikipedia.org/wiki/Precision_and_recall.
52. HathiTrust Digital Library | Millions of books online. (undated). Retrieved from <https://www.hathitrust.org>.
53. *HTRC Analytics*. (undated). Retrieved from <https://analytics.hathitrust.org>.
54. *DBNL · Digitale Bibliotheek voor de Nederlandse Letteren*. (undated). Retrieved from <https://www.dbnl.org>.

KB } national library
of the netherlands