MEASURING TOOL BIAS & IMPROVING DATA QUALITY FOR DIGITAL HUMANITIES RESEARCH

MYRIAM C. TRAUB

Measuring Tool Bias & Improving Data Quality for Digital Humanities Research

Myriam Christine Traub

The research reported in this thesis has been carried out at CWI, the Dutch National Research Laboratory for Mathematics and Computer Science, within the Information Access Group.

The research reported in this thesis was supported by the Dutch national program COMMIT/.

SIKS Dissertation Series No. 2020-09

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The research reported in this thesis has been carried out in the context of the SEALINCmedia Project.

ISBN 978-3-00-065364-3

Cover by: **Aoife Dooley** is an award winning illustrator, author and comedian from Dublin. Aoife is best known for her *Your One Nikita* illustrations. She released her first children's book earlier this year. *123 Ireland* won *Specsavers Childrens book of the year* at the An Post book awards 2019. Aoife gigs regularly in clubs and at festivals. She won U Magazines 30 under 30 award for best comedian in 2017. Aoife openly shares her experiences of being diagnosed as autistic at the age of 27, neurodiversity and how a diagnosis helped her to truly understand herself. Aoife has helped dozens of men and women to seek and receive a diagnosis over the last year.
http://aoifedooleydesign.com

MEASURING TOOL BIAS & IMPROVING DATA QUALITY FOR DIGITAL HUMANITIES RESEARCH

METEN VAN TOOL BIAS & VERBETEREN VAN DATAKWALITEIT VOOR DIGITAAL GEESTESWETENSCHAPPELIJK ONDERZOEK

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 11 mei 2020 des ochtends te 10.30 uur

door

myriam christine traub

geboren op 23 augustus 1982 te Villingen-Schwenningen, Duitsland

promotor:
Prof. dr. L. Hardman

copromotor:
Dr. J. van Ossenbruggen

CONTENTS

ACRONYMS

AAT	Art & Architecture Thesaurus
AI	Artificial Intelligence
ANP	Algemeen Nederlands Persbureau
CC	Character Confidence
CS	Computer Science
CV	Confidence Value
CWI	Centrum Wiskunde & Informatica
DH	Digital Humanities
DQR	Document Query Ratio
IR	Information Retrieval
KB	National Library of the Netherlands
MRR	Mean Reciprocal Rank
NER	Named Entity Recognition
OCR	Optical Character Recognition
PC	Page Confidence
PS	Parameter Set
RMA	Rijksmuseum Amsterdam
RQ	Research Question
SERP	Search Engine Result Page
TFIDF	Term Frequency - Inverse Document Frequency
TREC	Text Retrieval Conference

1 INTRODUCTION

Many cultural heritage institutions worldwide maintain archives containing invaluable assets, such as historic documents, artworks or culture-historical items. The missions of these institutions are not only to preserve the assets themselves and the contextual knowledge that was collected about them, but also to grant access to these collections to users for (scientific) research.

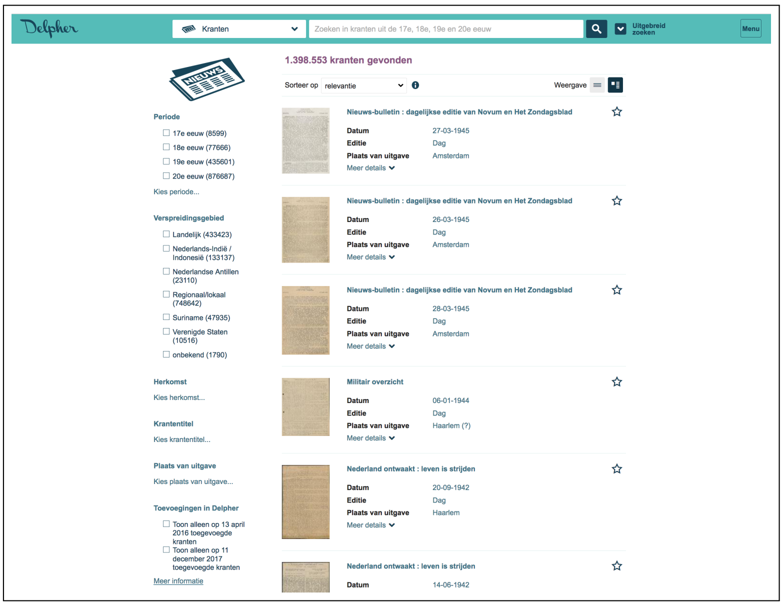
Since the advent of the WWW, more and more institutions have started to provide online access to (parts of) their collections. Individual institutions, such as the *Rijksmuseum Amsterdam*[1] (RMA) or the *National Library of the Netherlands*[2] (KB) have digitized large parts of their collections and set up online portals that allow users to search and browse the collections. On an international scale, initiatives such as *Europeana*[3] have successfully established a network of cultural heritage institutions that seeks to facilitate the general public's access to cultural heritage by interweaving previously isolated collections and enriching them with items and metadata contributed by the public[4].

Tools to access digital archives provide a rich resource for amateurs and professionals alike. Different user groups, however, have their own needs for interpreting results provided by the tools they use to access the collections. It is understanding users' tasks along with corresponding measures of tool reliability that form the inspiration for this thesis.

1.1 PROJECT CONTEXT

The research for this thesis was conducted at *Centrum Wiskunde & Informatica*[5], under the umbrella of the *SEALINCMedia*[6] project and the research framework *COMMIT/*[7]. One of the project goals was to find ways to efficiently and effectively collect trustworthy annotations for cultural heritage institutions using crowdsourcing. For this thesis, we closely collaborated with KB and RMA, organisations that both maintain large digitized archives and contributed invaluable expert knowledge and data for several of our studies.

[1] https://www.rijksmuseum.nl/nl/zoeken
[2] https://www.kb.nl
[3] https://www.europeana.eu/portal/en
[4] https://sourceforge.net/projects/steve-museum/
[5] https://www.cwi.nl/
[6] https://sealincmedia.wordpress.com/
[7] http://www.commit-nl.nl/



1.398.553 kranten gevonden

Periode

- □ 17e eeuw (8599)
- □ 18e eeuw (77666)
- □ 19e eeuw (435601)
- □ 20e eeuw (876687)

Kies periode...

Verspreidingsgebied

- □ Landelijk (433423)
- □ Nederlands-Indië / Indonesië (133137)
- □ Nederlandse Antillen (23110)
- □ Regionaal/lokaal (748642)
- □ Suriname (47935)
- □ Verenigde Staten (10516)
- □ onbekend (1790)

Herkomst

Kies herkomst...

Krantentitel

Kies krantentitel...

Plaats van uitgave

Kies plaats van uitgave...

Toevoegingen in Delpher

- □ Toon alleen op 13 april 2016 toegevoegde kranten
- □ Toon alleen op 11 december 2017 toegevoegde kranten

Nieuws-bulletin : dagelijkse editie van Novum en Het Zondagsblad

Datum	27-03-1945
Editie	Dag
Plaats van uitgave	Amsterdam

Nieuws-bulletin : dagelijkse editie van Novum en Het Zondagsblad

Datum	26-03-1945
Editie	Dag
Plaats van uitgave	Amsterdam

Nieuws-bulletin : dagelijkse editie van Novum en Het Zondagsblad

Datum	28-03-1945
Editie	Dag
Plaats van uitgave	Amsterdam

Militair overzicht

Datum	06-01-1944
Editie	Dag
Plaats van uitgave	Haarlem (?)

Nederland ontwaakt : leven is strijden

Datum	20-09-1942
Editie	Dag
Plaats van uitgave	Haarlem

Nederland ontwaakt : leven is strijden

Datum	14-06-1942

Figure 1: The KB maintains a digital (newspaper) archive that is accessible through full-text and faceted search.

The KB maintains several digitized collections of books, newspapers and magazines on their online portal *Delpher*[8]. Their newspaper collection spans more than 400 years, with the earliest issue dating back to 1618. With the passage of time, newspapers have changed considerably. The earliest issues[9] focus on providing concise reports on international political and economic developments. Only much later, other types of reports such as family notifications, images and advertisements were introduced. On top of the development of newspapers that are due to advanced manufacturing methods, they were also subject to changes in political and societal conditions. During World War II, Dutch resistants to the German occupation printed illegal newspapers which differ strongly from the official newspapers in terms of quality of print, layout and content.

The historic newspapers of the KB thus form a very diverse document collection that make it an interesting object for research. As a consequence, unfortunately, the KB's digitized versions of old newspaper pages suffer from (partially very) poor data quality due to limitations of Optical Character Recognition (OCR) and other technology. For cultural heritage institutions such as the KB, it is important to evaluate and improve data quality of their digital records.

The document collection of the KB is not only popular among the general public, it is also well-suited for research related to DH practices as it entails key problems that scholars face when using digitized corpora [35]: Documents are written in multiple languages and temporally very heterogeneous, both of which strongly affects the quality

[8] https://www.delpher.nl
[9] https://resolver.kb.nl/resolve?urn=ddd:010500649

of the digitization output. Since the content of the digitized documents is also used by the search engine of the archive, the result of any search task is influenced by errors in the text. In order to improve data quality, however, it is important to take users' requirements into account [25, 40]. The KB's newspaper collection is frequently accessed by the general public to look for genealogical information on members of their own family, and Humanities scholars who seek to find answers to their research questions.

While good search results matter for both groups, humanities scholars need a sufficient level of certainty about the correctness of their results in order to use them for their publications and missing out on relevant documents can therefore have serious ramifications for them. Therefore, it is important to know how, and for what types of tasks the scholars use digital sources and what level of data quality is required to support these tasks. From the way their data is used, digital archives can develop strategies for data quality management.

This thesis investigates how better support can be provided for humanities research for accessing digital archives by measuring tool bias and improving data quality. For this, we identified which research tasks humanities scholars typically perform using digital archives and evaluated how well they are supported by the archives' data and infrastructure. We measured the data quality for a subset of the KB's newspaper archive and evaluated its impact on the retrieval of relevant documents. In particular, we investigated potential bias in search results introduced by search tools and data quality. Finally, we studied, how metadata of cultural heritage data can be extended with accurate annotations by non-experts using a crowdsourcing approach based on gamification.

1.2 RESEARCH QUESTIONS

Searching a large digital archive is made easier for a user if the search interface allows to filter the results along different features. In order to facilitate these technologies, in some cases additional metadata may be needed. Unfortunately, experts to make these additional annotations as scarce and expensive. A study conducted by [57] showed that classification of paintings into subject types cannot be successfully done by automatic classifiers. They can, however, provide a set of candidates that is likely to contain the correct class.

Research shows that crowds are able to perform simple tasks (e.g. estimating the weight of an ox) with a precision that is close or even better than judgements given by experts of the field [20]. We therefore explored how output from a machine learning algorithm can be used as input for a crowdsourcing classification task.

RQ: Can crowd workers contribute data that is in line with expert contributions?

a.) How do classifications obtained from crowd workers performing a simplified expert classification task compare to classifications done by experts?

b.) Do crowd workers become better at performing the task and, if so, is that only on repeated items or also on new items?

c.) How does the partial absence of the correct answer affect the performance of the crowd workers?

These research questions are answered in Chapter 2. The results from this study raised the question, what tasks users are conducting in digital archives that the data does not (yet) support sufficiently.

The KB closely collaborates with humanities researchers to support them in their research and, in return, learn about their interests and requirements with respect to their research. To better understand what types of research tasks scholars perform on Delpher, and what the key requirements for these tasks are, we interviewed humanities scholars who regularly use large digital collections. As we know that the documents in Delpher vary strongly in terms of data quality, we investigated whether working with digitized collections that contain errors influences their work.

RQ: How do professional users perceive the effect of data quality on (research) task execution?

a.) Which tasks do digital humanities scholars carry out in digital archives?

b.) What types of tasks can we identify and what is the potential impact of OCR errors on these tasks?

c.) What data do professional users require to be able to estimate the quality indicators for different task categories?

These research questions are answered in Chapter 3.

It is important to not only engage computer scientists in the discussion around tool bias, data quality and the impact they may have on end results, but also the users of the tools. We organized a workshop to raise awareness among humanities scholars about the pitfalls of digital tools and data, but more importantly, to find out which aspects of digital tool use require more research.

RQ: How can we better understand the impact of technology-induced bias on specific research contexts in the Humanties?

a.) What are good examples for typical research tasks affected by technology-induced bias or other tool limitations?

b.) What is the specific information, knowledge and skills required for scholars to be able to perform tool criticism as part of their daily research?

c.) What are useful guidelines or best practices to identify technology-induced bias systematically?

The workshop brought together researchers from different research domains in computer science and the humanities and inspired discussions between tool builders and tool users. These discussions were later continued in workshops at the Digital Humanities Benelux Conference 2017[10] and in the context of a symposium organized by the CLARIAH project[11]. The insights gained from this workshop inspired the development of the research questions for this thesis and thereby influenced its general direction.

While no direct scientific results were derived from the workshop, it provided context for the results presented in following chapters. A summary of the discussions that took place during the workshop and the findings are presented in Chapter 4.

The scholars we interviewed for the study presented in Chapter 3 agreed that the high error-rate in digitized archives make it very hard to obtain reliable results. Since the retrieval system of an archive has a major impact on the search results, we investigated retrieval bias in the KB's historic newspaper archive using queries collected from the archive's users.

RQ: What types of bias can typically be found in a digital newspaper archive?

a.) Is the access to the digitized newspaper collection influenced by a retrievability bias?

b.) Can we find a relation between features of a document (such as document length, time of publishing, and type of document) and its retrievability score?

c.) To what extent are retrievability experiments using simulated queries representative of the search behavior of real users of a digital newspaper archive?

These research questions are answered in Chapter 5.

The main criticism of the scholars in our interviews was the data quality in the archives and the fact that they do not know how it influences the access to documents. Digital libraries therefore set up projects to improve data quality by having (parts of) their collections transcribed by volunteers or crowd workers. We studied the effects of correcting OCR errors on the retrievability of documents in a historic newspaper corpus of a digital library.

RQ: How do crowd-sourced improvements of OCRed documents impact retrievability?

[10] https://dhbenelux2017.eu/programme/pre-conference-events/workshop-8-tool-criticism-workshop-dh-benelux-2017/

[11] https://www.clariah.nl/en/new/news/symposium-on-tool-criticism

A.) What is the relation between a document's OCR character error rate and its retrievability score?

B.) How does the correction of OCR errors impact the retrievability bias of the corrected documents (direct impact)?

C.) How does the correction of a fraction of error-prone documents influence the retrievability of non-corrected ones (indirect impact)?

These research questions are answered in Chapter 6.

In Chapter 7 we present a summary of the thesis, we draw the conclusions from the insights we gained in the studies and point out which aspects should be further investigated.

1.3 PUBLICATIONS

The chapters in this thesis are based on the following publications.

CHAPTER 1 is based on the doctoral consortium paper *Measuring and Improving Data Quality of Media Collections for Professional Tasks* presented at Information Interaction in Context 2014 (IIiX 2014) by Myriam C. Traub.

CHAPTER 2 is based on *Measuring the Effectiveness of Gamesourcing Expert Oil Painting Annotations* published at the European Conference on Information Retrieval 2014 by Myriam C. Traub, Jacco Ossenbruggen, Jiyin He, and Lynda Hardman. This work is based on the Fish4Knowledge game designed and described by Jiyin He in [23]. Myriam Traub adapted the game to the art domain, designed the experiment and analyzed the results. All authors contributed to the text.

CHAPTER 3 is based on *Impact Analysis of OCR Quality on Research Tasks in Digital Archives* published at TPDL 2015 by Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman.

CHAPTER 4 is based on the workshop report on the topic of *Tool Criticism for Digital Humanities* written by Myriam Traub and Jacco van Ossenbruggen. The workshop took place on May 22nd, 2015 in Amsterdam, NL, and was chaired by Sally Wyatt. The organizing committee further consisted of Victor de Boer, Serge ter Braake, Jackie Hicks, Laura Hollink, Wolfgang Kaltenbrunner, Marijn Koolen and Daan Odijk.

CHAPTER 5 is based on *Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus* published at ACM/IEEE Joint Conference on Digital Libraries 2016 by Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. Myriam Traub conducted the experiments and performed the data analysis. Thaer Samar performed the document pre-processing, the setup of the Indri experimental environment and contributed to the discussion of the results. All authors contributed to the text.

CHAPTER 6 is based on *Impact of Crowdsourcing OCR Improvements on Retrievability Bias* published at ACM/IEEE Joint Conference on Digital Libraries 2018 by Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, and Lynda Hardman. Myriam Traub conducted the experiments and performed the data analysis. Thaer Samar performed the document pre-processing. All authors contributed to the text.

A full list of publications by the author can be found at the end of this thesis on page 107.

2 MEASURING THE EFFECTIVENESS OF GAMESOURCING EXPERT OIL PAINTING ANNOTATIONS

Tasks that require users to have expert knowledge are difficult to crowdsource. They are mostly too complex to be carried out by non-experts and the available experts in the crowd are difficult to target. Adapting an expert task into a non-expert user task, thereby enabling the ordinary "crowd" to accomplish it, can be a useful approach. We studied whether such a simplified version of an expert annotation task can be carried out by non-expert users. Users conducted a gamified annotation task of oil paintings using categories from an expert vocabulary. The obtained annotations were compared with those from experts. Our results show a significant agreement between the annotations done by experts and non-experts, that users improve over time and that the aggregation of users' annotations per painting increases their precision.

2.1 INTRODUCTION

Cultural heritage institutions place great value in the correct and detailed description of the works in their collections. They typically employ experts (e.g. art-historians) to annotate artworks, often using predefined terms from expert vocabularies, to facilitate search in their collections. Experts are scarce and expensive, so that involving non-experts has become more common. For large image archives that have been digitized but not annotated, there are often insufficient experts available, so that employing non-expert annotations would allow the archive to become searchable (see for example ARTigo[1], a tagging game based on the ESP game[2]).

In the context of a project with the Rijksmuseum Amsterdam, we take an example annotation task that is traditionally seen as too difficult for the general public, and investigate whether we can transform it into a game-style task that can be played directly, or quickly learned while playing, by non-experts. Since we need to compare the judgments of non-experts with those of experts, we picked a dataset and annotation task for which expert judgments were available.

We conducted two experiments to investigate the following research questions.

[1] http://www.artigo.org/

[2] http://www.gwap.com/gwap/gamesPreview/espgame/

RQ: Can crowd workers contribute data that is in line with expert contributions?

A.) How do crowd workers performing a simplified expert classification task compare to experts?

B.) Do crowd workers become better at performing the task and, if so, is that only on repeated items or also on new items?

C.) How does the partial absence of the correct answer affect the performance of the crowd workers?

The results to these research questions allow us to estimate the suitability of the non-expert annotations as part of a professional workflow and to determine whether purely non-expert input is reliable.

2.2 RELATED WORK IN CROWDSOURCING

Increasing numbers of cultural heritage institutions initiate projects based on crowdsourcing to either enrich existing resources or create new ones [14]. Two well-known projects in this field are the *Steve Tagger*[3] and the *Your Paintings Tagger*[4]. Both constitute cooperations between museum professionals and website visitors to engage visitors with museum collections and to obtain tags that describe the content of paintings to facilitate search.

A previous study by Hildebrand et al. suggests that expert vocabularies that are used by professional cataloguers are often too limited to describe a painting exhaustively [27]. This gap can be closed by making use of external thesauri from domains other than art history (e.g. WordNet, a lexical, linguistic database[5]). The interface for this task, however, targets professional users.

Steve Tagger and the *Your Paintings Tagger* focus on enriching their artwork descriptions with information that is common knowledge (e.g. Is a flower depicted?). The SEALINCMedia project[6] focuses on finding precise information (e.g. the Latin name of a plant) about depicted objects. To achieve this, the crowd is searched for experts who are able to provide this very specific information [18] and a recommender system selects artworks that match the users' expertise.

Another example for crowdsourcing expert knowledge is *Umati*. Heimerl et al. transformed a vending machine into a kiosk that returns snacks for performing survey and grading tasks [24]. The restricted access to *Umati* in the university hallway ensured that the participants possessed the necessary background knowledge to solve the presented task. While their project also aims at getting expert

[3] http://tagger.steve.museum/

[4] http://tagger.thepcf.org.uk/

[5] http://wordnet.princeton.edu/

[6] http://sealincmedia.wordpress.com/

work done with crowdsourcing mechanisms, their approach is different from ours. Whereas they aim at attracting skilled users to accomplish the task, we give non-experts the support they need to carry out an expert task.

Since most of these approaches target website visitors or passers-by, rather than paid crowd workers on commercial platforms, they need to offer an alternative source of motivation for users. Luis von Ahn's *ESP Game* [50] inspired several art tagging games developed by the *ARTigo project*[7]. These games seek to obtain artwork annotations by engaging users in gameplay.

Golbeck et al. showed that tagging behavior is significantly different for abstract compared with representational paintings [22]. Users were allowed to enter tags freely, without being limited to the use of expert vocabularies. Since our set of images showed a similar variety in styles and periods, we also investigated whether particular features of images had an influence on the user behavior.

He et al. investigated if and how the crowd is able to identify fish species on photos taken by underwater cameras [23]. This task is usually carried out by marine biologists. In the study, users were asked to identify fish species by judging the visual similarity between an image taken from video and images showing already identified fish species.

A common challenge of tagging projects lies in transforming the large quantity of tags obtained through the crowd to high quality annotations of use in a professional environment. As Galton proved in 1907, the aggregation of the *vox populi* can lead to surprisingly exact results that are "correct to within 1 per cent of the real value" [20]. Such aggregation methods can improve the precision of user judgments [30], a feature that can potentially be used to increase the agreement between users and experts of our tagging game.

2.3 experimental setup

We investigated the categorization of paintings into subject types (e.g. landscapes, portraits, still lifes, marines), which is typically considered to be an expert task. We simplified the task by changing it into a multiple choice game with a limited, preselected set of candidates to choose from. Each included the subject type's label, a short explanation of its intended usage and a representative example image. To investigate the influence of the pre-selection of the candidates on the performance of the users, we carried out two experiments: a baseline condition, which always had a correct answer among the presented candidate answers, and, to simulate a more realistic setting, a condition where in 25% of the cases the correct answers had been deliberately removed.

[7] http://www.artigo.org/

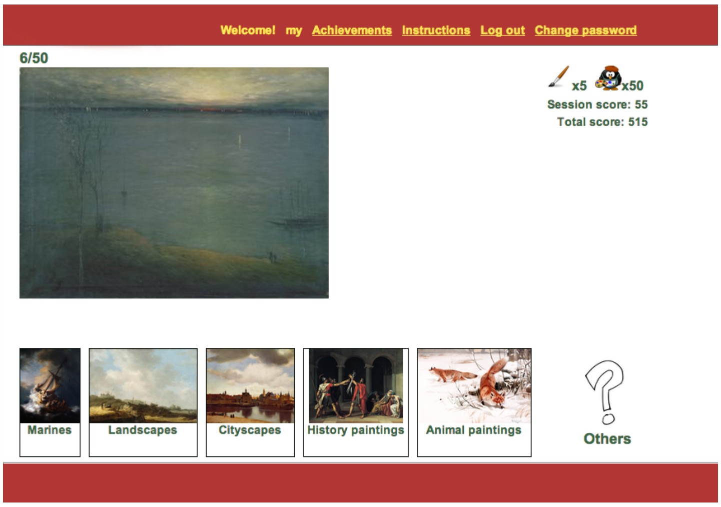


Figure 2: Interface of the art game with the large query image on the upper left. The five candidate subject types are shown below, together with the *others* candidate.

2.3.1 *Procedure*

Users were presented with a succession of images (referred to as *query images*) of paintings that they were asked to match with a suitable subject type (see Fig. 2). We supported users by showing them a preselection of six *candidates*. Five of these candidates represented subject types and one of them (labeled "others") could be used if the assumed correct subject type was not presented. To motivate users to annotate images correctly and to give them feedback about the "correctness"[8] of their judgments, they were awarded ten points for judgments that agree with the expert and one point for the attempt (even if incorrect).

The correct answer was always presented and users got direct feedback on every judgment they made. With this experiment we wanted to find out whether (and how well) users learn under ideal conditions. We use the data of the first experiment as a baseline for comparing the results of the second experiment.

In the second experiment, the correct answer is not always presented.

2.3.2 *Experiments conducted*

We adapted the online tagging game used for the Fish4Knowledge project [23]. On the login page of the game, we provide a detailed description of the game including screenshots, instructions and the rules of the game.

[8] By "correct" we mean that a given judgment agrees with the expert.

BASELINE CONDITION For each query image, we selected one candidate that, according to the expert ratings, represents a correct subject type and three candidates representing related, but incorrect, subject types. One candidate was chosen randomly from the remaining subject types. For cases, when there were only two related but incorrect subject types available, we showed two incorrect random ones, so the total number of candidates would remain six (including the *others* candidate). The categorization of similar subject types was done manually and is based on their similarity. An example of related subject types is *figure*, *full-length figure*, *half figure*, *portrait* and *allegory*.

IMPERFECT CONDITION In this setting, the correct candidate is not presented in 25% of the cases. This is used to find out how good the learning performance of users is when the candidate selection is done by an automated technique that may fail to find a correct candidate in its top five. The selection of the candidates was the same as in the baseline experiment, for the missing correct candidate we added another incorrect candidate.

2.3.3 *Materials*

The expert dataset [57] provides annotations of subject types for the paintings of the *Steve Tagger project* by experts from the Rijksmuseum Amsterdam. We selected 168 expert annotations for 125 paintings (Table 1). The number of annotations per painting ranged from four (for one painting) to only one (for 83 paintings). These multiple classifications are considered correct: a painting showing an everyday scene on a beach[9] can be classified as *seascapes*, *genre*, *full-length figure* and *landscapes*. This, however, makes our classification task more difficult.

QUERY IMAGES The images used as query images are a subset of the thumbnails of paintings from the Steve Tagger[10] data set. The paintings are diverse in origin, subject, degree of abstraction and style of painting. Apart from the image, we provided no further information about the painting. Within the first ten images that were presented to the user, there were no repetitions. Afterwards, images may have been presented again with a 50% chance. The repetitions gave us more insight on the performance of the users.

CANDIDATES A candidate consists of an image, a label (subject type) and a description. For each subject type we selected one representative image from the corresponding Wikipedia page[11]. The main criterion for the selection was that the painting should show typical

[9] http://tagger.steve.museum/steve/object/280

[10] http://tagger.steve.museum/

[11] E.g.: http://en.wikipedia.org/wiki/Maesta

Subject type	Annotations
full-length figures	40
landscapes	33
half figures	13
allegories, history paintings, portraits, animal paintings, genre, kacho, figures	8
townscapes	6
flower pieces	5
marines, cityscapes, maesta, seascapes, still lifes	3

Table 1: Used subject types and the number of expert annotations.

characteristics. The candidates were labeled with the names of the subject types from the Art & Architecture Thesaurus[12] (AAT) which comprises in total more than 100 subject types. The representative images were intended to give users a first visual indication of which subject type might qualify and it made it easier for users to remember it. If this was not sufficient for them to judge the image, they could verify their assumption by displaying short descriptions taken from the AAT, for example:

Marines:

"Creative works that depict scenes having to do with ships, shipbuilding, or harbors. For creative works depicting the ocean or other large body of water where the water itself dominates the scene, use 'seascapes'. "[13]

The descriptions of the subject types are important, as the differences between some subject types are subtle.

2.3.4 *Participants*

Participants were recruited over social networks and mailing lists. For the analysis we used 21 for the first experiment and 17 in the second one, in total 38, after removing three users who made fewer than five annotations. The majority of the participants have a technical professional background and no art-historical background. In the baseline condition, users who scored at least 400 received a small reward.

[12] http://www.getty.edu/research/tools/vocabularies/aat/index.html

[13] http://www.getty.edu/vow/AATFullDisplay?find=marines&logic=AND&note=&english=N&prev_page=1&subjectid=300235692

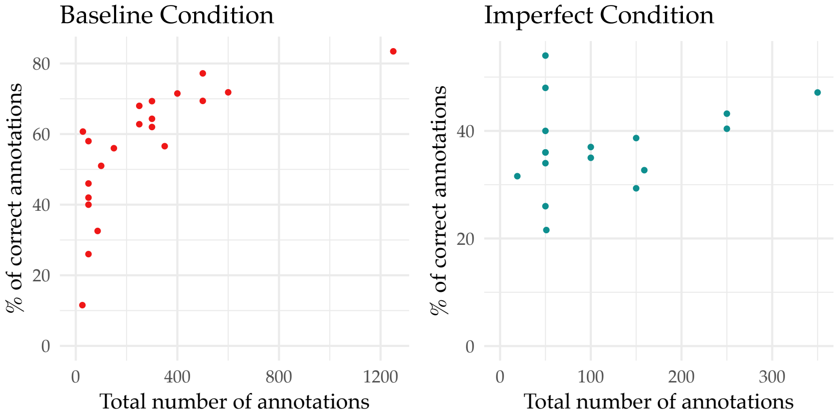


Figure 3: Percentage of correct annotations per user (y-axis) and the number of annotations (x-axis) for both experimental conditions. Each point represents the annotations from one user.

2.3.5 *Limitations*

Our image collection comprised 125 paintings, and compared with a museum's collection this is a small number. Because of the repetitions, the number of paintings that the user saw only increased gradually over time, which would have made it possible to successively introduce a larger number of images to the users. This, however, would have made it difficult to obtain the necessary ground truth.

In the available ground truth data, each painting was judged by only one expert, which prevents us from measuring agreement among experts. This measurement might have revealed inconsistencies in the data that influenced users' performance.

In realistic cases, ground truth will be available for only a small fraction of the data. To apply to such datasets, our setting needs other means of selecting the candidates. This can be realized, for example, by using the output of an imperfect machine learning algorithm, or by taking the results of another crowdsourcing platform. We think it is realistic to assume that in such settings the correct answer is not always among the results, and acknowledge that the frequency of this really happening may differ from the 25% we assumed in our second experiment.

The game did not go viral, which can mean that incentives for the users to play the game and/or the marketing could be improved.

2.4 RESULTS

An overview of the results of all users of both experiments shows a large variation in number of judgments and precision (Fig. 3). Users who judged more images also tend to have higher precision. This might suggest that users indeed learn to better carry out the task or that well-performing users played more.

In both conditions, all users who finished at least one round of 50 images performed much better than a random selection of the candidates (with a precision of 17%), suggesting that we do not have real spammers amongst our players. On average, the precision of the users in the baseline condition (56%) is higher than in the imperfect condition (37%). This indicates that the imperfect condition is more difficult. This is in line with our expectations: in order to agree with the expert, users in the imperfect condition sometimes need to select the *other* candidate instead of a candidate subject type that might look very similar to the subject type chosen by the expert.

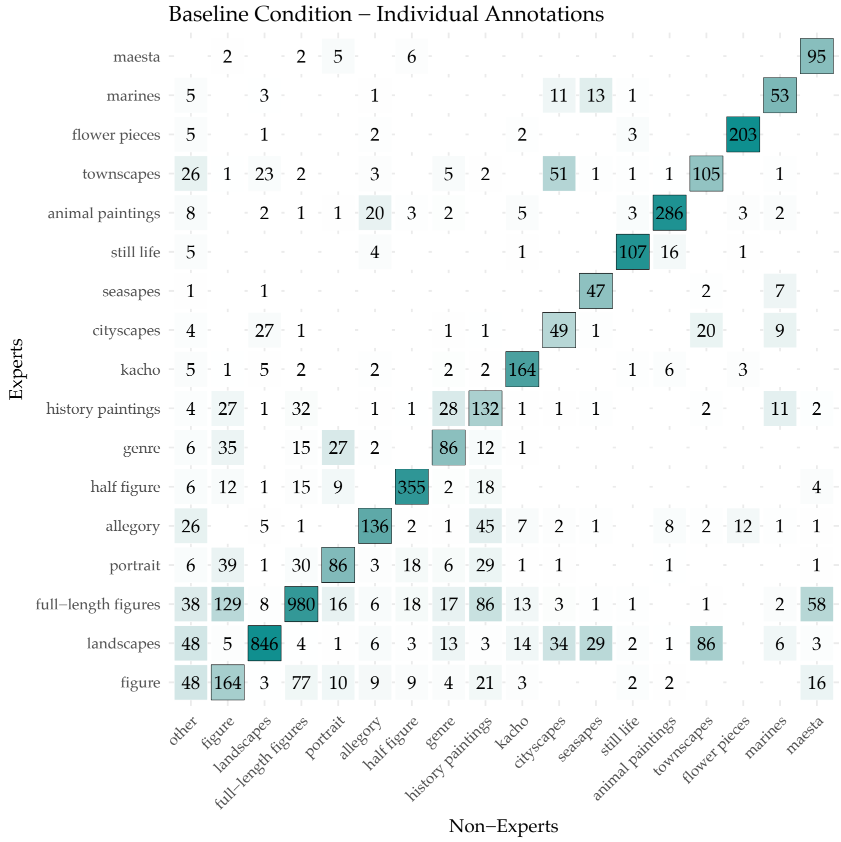
2.4.1 *Agreement per subject type*

To understand the agreement between experts and users, we measure precision and recall per subject type. *Precision* is the number of agreed-upon judgments for a subject type divided by the total judgments given by users for that subject type. *Recall* is the number of agreed-upon judgments for a subject type divided by the total judgments given by the expert for that subject type.

Both measures are visualized in confusion heatmaps (Fig.s 4 - 7). The rows represent the experts' judgements, while the columns show how the users classified the images. The shade of the cells visualizes the value of that cell as the fraction of the users' total votes for that specific subject type. Darker cells on the diagonal indicate higher agreement, while other dark cells indicate disagreement.

Some subject types score low on precision: *cityscapes* is frequently chosen by non-experts when the expert used *landscapes* or *townscapes*, while users select *history paintings* where the expert sees *figures* (Fig. 4). On the other hand, *flower pieces* and *animal paintings* score high on both precision and recall. Selecting the *others* candidate did not return points in the baseline condition, and some players reported to have noticed this and did not use this candidate afterwards. With 243 *others* judgements out of a total of 5640, it received relatively few clicks. The agreement between users and experts is substantial (Cohen's Kappa of 0.65), we see a clear diagonal of darker color.

Aggregating user judgements by using majority voting (Fig. 5), removes some deviations from the experts' judgments (Cohen's Kappa of 0.87) to almost perfect agreement. For example, all *cityscapes* judgments by users for cases where expert judged *landscapes* are overruled in the voting process and this major source of disagreement in Fig. 4 disappears. There is only one case where the expert judged *townscapes* and the majority vote of the users remained *cityscapes*. The painting description states that it shows "a dramatic bird's eye view of Broad-



Baseline Condition – Individual Annotations

Experts \ Non−Experts	other	figure	landscapes	full−length figures	portrait	allegory	half figure	genre	history paintings	kacho	cityscapes	seasapes	still life	animal paintings	townscapes	flower pieces	marines	maesta
maesta		2		2	5		6											95
marines	5		3			1					11	13	1				53	
flower pieces	5		1			2				2			3			203		
townscapes	26	1	23	2		3		5	2		51	1	1	1	105		1	
animal paintings	8		2	1	1	20	3	2		5			3	286		3	2	
still life	5					4				1			107	16		1		
seasapes	1		1									47			2		7	
cityscapes	4		27	1				1	1		49	1			20		9	
kacho	5	1	5	2		2		2	2	164			1	6		3		
history paintings	4	27	1	32		1	1	28	132	1	1	1			2		11	2
genre	6	35		15	27	2		86	12	1								
half figure	6	12	1	15	9		355	2	18									4
allegory	26		5	1		136	2	1	45	7	2	1		8	2	12	1	1
portrait	6	39	1	30	86	3	18	6	29	1	1			1				1
full−length figures	38	129	8	980	16	6	18	17	86	13	3	1	1		1		2	58
landscapes	48	5	846	4	1	6	3	13	3	14	34	29	2	1	86		6	3
figure	48	164	3	77	10	9	9	4	21	3			2	2				16

Figure 4: Despite many deviations, the graph shows a colored diagonal representing an agreement between non-experts and experts. The task therefore seems to be difficult but still manageable for users.

way and Wall Street"[14] in New York. Therefore, *townscapes* cannot be the correct subject type and users were right to disagree with the expert. Most *others* judgments are largely eliminated by the majority voting. However, three paintings remain classified as *others* by the majority which indicates a very strong disagreement with the experts' judgment. One of these paintings does not show a settlement, but in an abstract way depicts a bomb store in the "interior of the mine"[15]. The other two show a carpet merchant in Cairo[16] and the "Entry of Christ into Jerusalem"[17], both being representations of large cities and therefore incorrectly categorized as *townscapes* by the expert.

In the imperfect condition, the confusion heatmaps are similar, however, the disagreement between users and experts is higher. The *others* candidate was the correct option in 25% of the cases. The users made more use of it, as shown by the higher numbers in the first column of Fig. 7. The agreement in the *allegories* column is, with 13%, even below chance. Majority voting increases the precision, but only to 20%.

[14] http://www.clevelandart.org/art/1977.43
[15] http://www.tate.org.uk/art/artworks/bomberg-bomb-store-t06998
[16] https://collections.artsmia.org/index.php?page=detail&id=10361
[17] http://tagger.steve.museum/steve/object/172

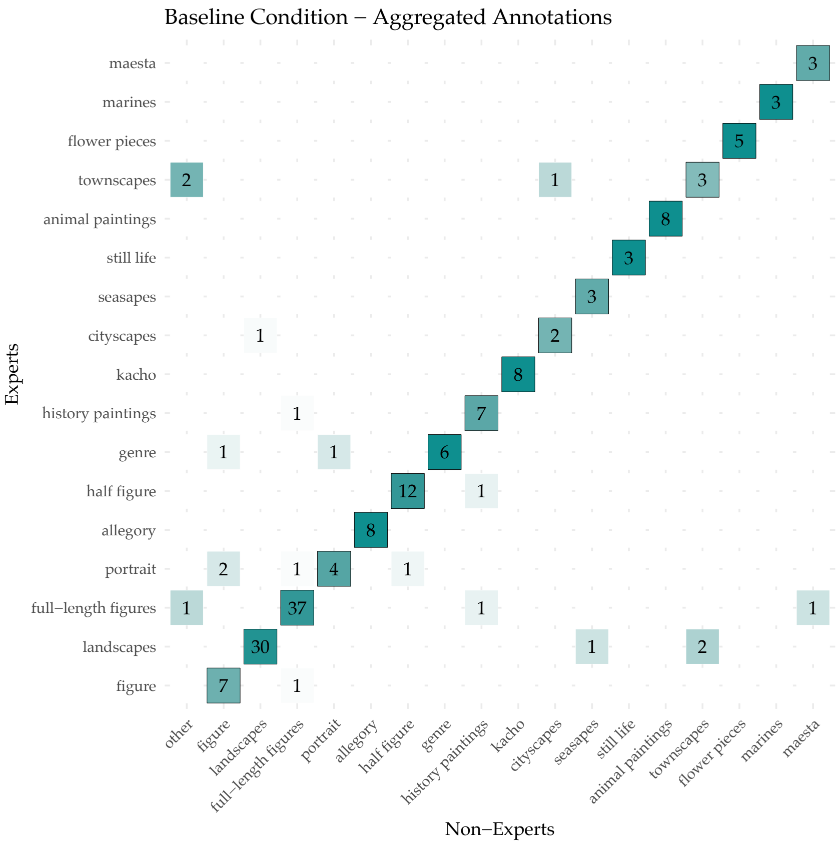
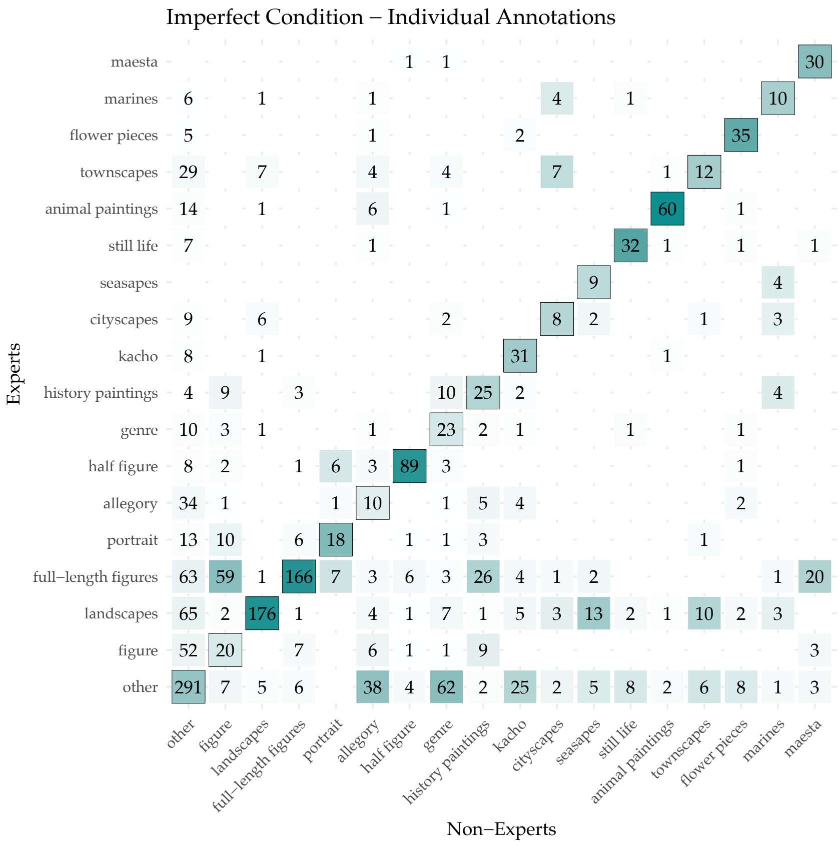


Figure 5: The "Wisdom of the Crowd" effect eliminates many deviations of the non-experts' judgements from the experts' judgements. However, there are still deviations for similar subject types such as *cityscapes* and *townscapes*.

The AAT defines this subject type to "express complex abstract ideas, for example works that employ symbolic, fictional figures and actions to express truths or generalizations about human conduct or experience". Therefore, it is very difficult to recognize an *allegory* as such without context information about the painting. User judgments diverging from the expert's judgments are largely removed by majority vote. The "Wisdom of the Crowd" effect, however, is not as strong as in the baseline condition. It raised the Cohen's Kappa from 0.47 to a (still) moderate agreement of 0.55.

We further analyzed the agreement of the non-experts and the experts on image level in the baseline condition. The broad range from 2% to 98% indicates very strong (dis-)agreement for some cases. In the images with the highest agreement, the relation between the depicted scenes and the subject type is intuitively comprehensible: the images with 98% agreement show flowers (*flower pieces*), monkeys (*animal painting*) and a still life (*still lifes*). An entirely different picture emerges, when we look at the images with low agreement. We presented the most striking cases to an expert from the Rijksmuseum



	other	figure	landscapes	full–length figures	portrait	allegory	half figure	genre	history paintings	kacho	cityscapes	seasapes	still life	animal paintings	townscapes	flower pieces	marines	maesta
maesta							1	1										30
marines	6		1			1					4		1				10	
flower pieces	5					1				2						35		
townscapes	29		7			4		4			7			1	12			
animal paintings	14		1			6		1						60		1		
still life	7					1							32	1		1		1
seasapes												9					4	
cityscapes	9		6					2			8	2			1		3	
kacho	8		1							31				1				
history paintings	4	9		3				10	25	2							4	
genre	10	3	1			1		23	2	1			1			1		
half figure	8	2		1	6	3	89	3								1		
allegory	34	1			1	10		1	5	4						2		
portrait	13	10		6	18		1	1	3						1			
full–length figures	63	59	1	166	7	3	6	3	26	4	1	2					1	20
landscapes	65	2	176	1		4	1	7	1	5	3	13	2	1	10	2	3	
figure	52	20		7		6	1	1	9									3
other	291	7	5	6		38	4	62	2	25	2	5	8	2	6	8	1	3

Figure 6: The *others* candidate attracted many user votes. Compared to the baseline condition, the diagonal is less prominent, meaning that the agreement is lower in most cases.

Amsterdam to re-evaluate the experts' judgments and we identified two main reasons for disagreement: users would have needed additional information, such as the title, to classify the painting correctly; the expert annotations were incomplete or incorrect.

2.4.2 *Performance over time*

The improvement of the users' precision over time does not necessarily mean that they have learned how to solve the problem (*generalization*), but that they "only" have learned the correct solution for a concrete problem (*memorizing*).

MEMORIZING A learning effect is evident in the performance curve of the users for repeated images (Fig. 8). In the baseline condition, users had an initial success rate of 56% correct judgments. After seven repetitions, they judged 90% of the query images correctly. In the imperfect condition, the performance is consistently lower. The difference between the first appearance of an image (success rate of 36%) and the fifth appearance of an image (success rate of 46%) is lower

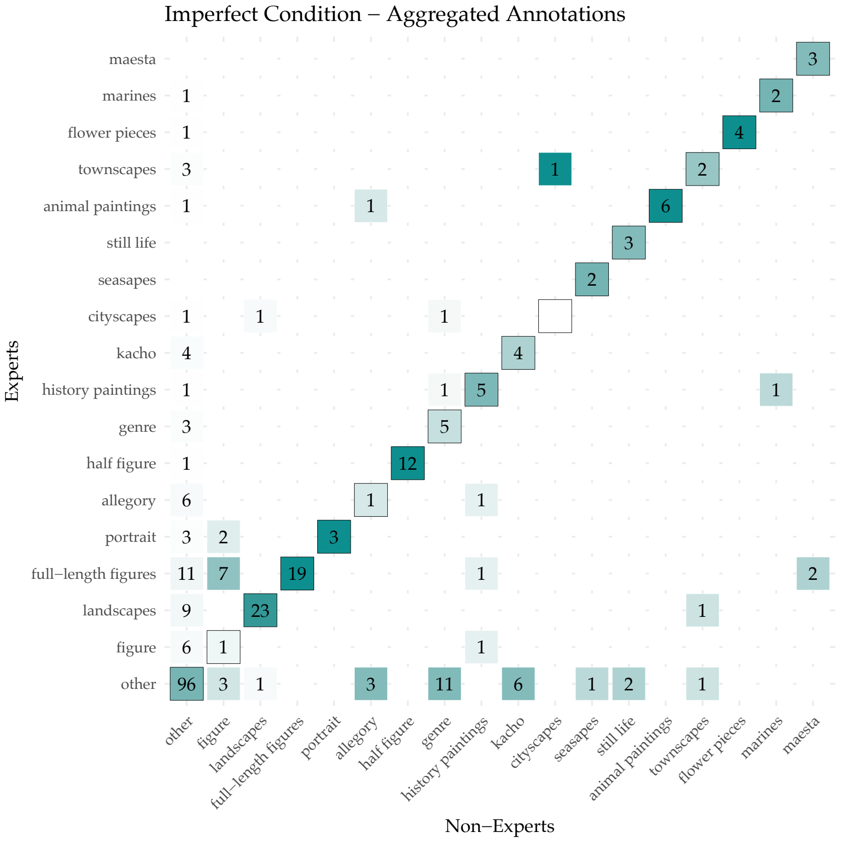


Figure 7: The aggregation of user votes could compensate some of the deviations from agreement, however the additional *others* candidate had a negative effect on the agreement for allegories, genre and kacho.

than in the baseline experiment where we see an increase of 25 percent units. The lines in Fig. 8 were cut off after eleven repetitions for the baseline condition and five repetitions for the imperfect condition because the number of judgments dropped below 15. We further analyzed the results of a fixed homogeneous population of seven (baseline) and eight (imperfect) users. The outcomes were nearly identical for both conditions. These results show that users in the baseline condition improve on memorizing the correct subject type for a specific image. The differences between the two conditions indicate that users found it more difficult to learn the subject types in the imperfect condition.

GENERALIZATION The judgement performance of users on the first appearances of images indicates whether they are able to generalize and apply the knowledge to unseen query images. If users learn to generalize, it is likely that they will improve over time at judging images that they have not seen before. Judgement precision increases throughout gameplay for both conditions (see Fig. 9). While users

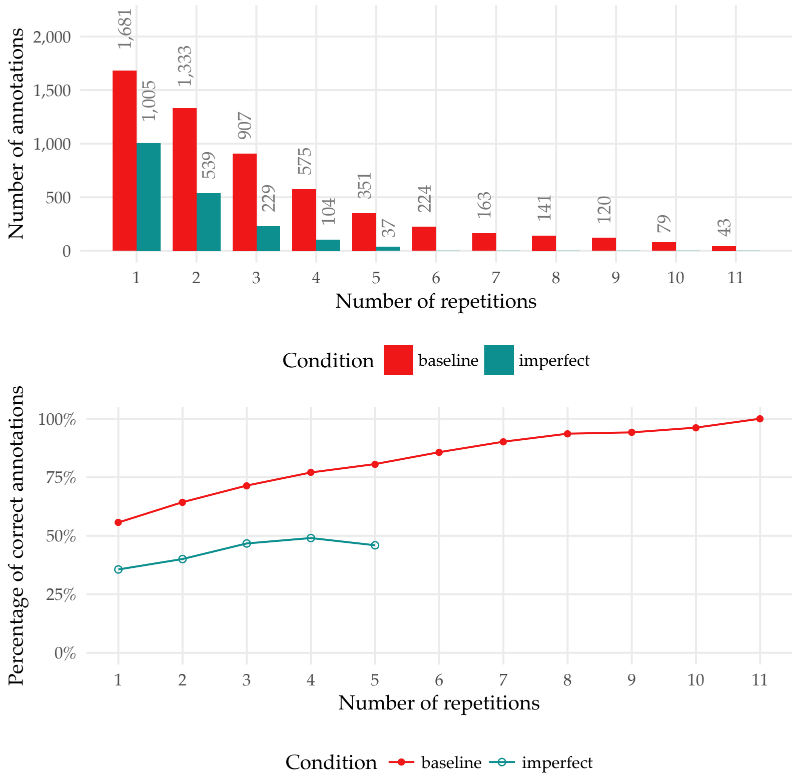


Figure 8: Learning curves (lines) for the memorization effect of repeated images and numbers on annotations (bars) per repetition.

in the baseline experiment started with a success rate of 44%, they reach 90% after about 250 images. Users in the imperfect condition started at a much lower rate of 33% and increase to 60%, after about 150 images. The declining number of images that are new to the user and the declining number of users that got so far in the game, lead to a drop in available judgments at later stages in the game. Therefore, we cut the graphs at sequence numbers 400 (baseline) and 160 (imperfect).

Our findings show that users can learn to accomplish the presented simplified expert task. This does not mean, however, that they would perform equally well if confronted with the "real" expert task. Users were given assistance by reducing the number of candidates from more than one hundred to six, they were provided a visual key (example image) to aid memorization and a short description of the subject type. A way to increase the success rate in a realistic setting would be to train users on a "perfect" data set and after passing a predefined success threshold, introduce "imperfect" data into the game.

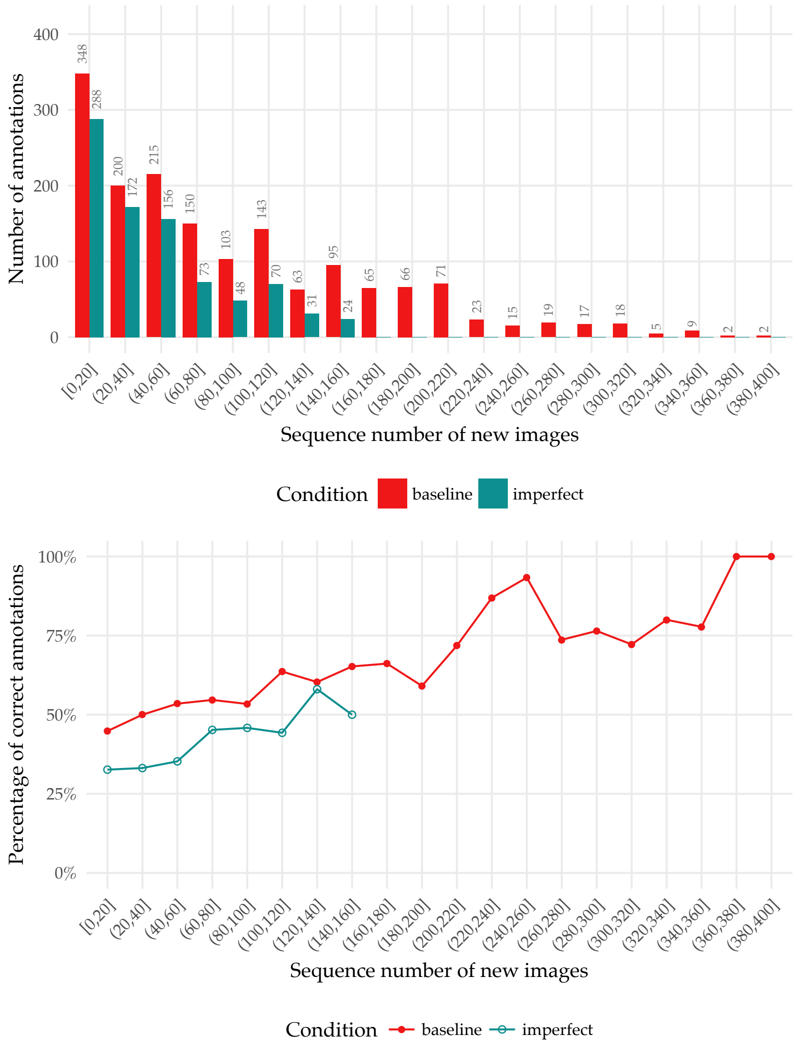


Figure 9: Users' performance for first appearances of images that occur in different stages of the game (lines) and number of annotations (bars).

2.5 CONCLUSIONS

Our study investigates the use of crowdsourcing for a task that normally requires specific expert knowledge. Such a task could be relevant to facilitate search by improving metadata on non-textual data sets, but also in crowdsourcing relevance judgments for more complex data in a more classic IR setting.

Our main finding is that non-experts are able to learn to categorize paintings into subject types of the AAT thesaurus in our simplified set-up. We studied two conditions, one with the expert choice always present, and one in which the expert choice had been removed in 25% of the cases. Although the agreement between experts of the Rijksmuseum Amsterdam and non-experts for the first condition is higher, the agreement in the imperfect condition is still acceptably high. We found that the aggregation of votes leads to a noticeable "Wisdom of the Crowds" effect and increases the precision of the users' votes. While this removed many deviations of the users' judgments from the experts' judgments, on some images, the disagreement remained. We consulted an expert and identified two main reasons: Either the annotations by the experts were incomplete or incorrect or the correct classification required knowing context information of the paintings that was not given to the users.

The analysis of user performance over time showed that users learned to carry out the task with higher precision the longer they play. This holds for repeated images (memorization) as well as new images (generalization).

The next step is to balance the interdependencies of the three players: experts, automatic methods and gamers. We hope that reducing their weaknesses (scarce, requiring much training data, insufficient expertise) by directing the interplay of their strengths (ability to provide: high quality data, high quantity data, high quality when trained and assisted) can lead to a quickly growing collection of high quality annotations.

3 IMPACT ANALYSIS OF OCR QUALITY ON RESEARCH TASKS IN DIGITAL ARCHIVES

Humanities scholars increasingly rely on digital archives for their research instead of time-consuming visits to physical archives. This shift in research method has the hidden cost of working with digitally processed historical documents: how much trust can a scholar place in noisy representations of source texts? In a series of interviews with historians about their use of digital archives, we found that scholars are aware that optical character recognition (OCR) errors may bias their results. They were, however, unable to quantify this bias or to indicate what information they would need to estimate it. This, however, would be important to assess whether the results are publishable. Based on the interviews and a literature study, we provide a classification of scholarly research tasks that gives account of their susceptibility to specific OCR-induced biases and the data required for uncertainty estimations. We conducted a use case study on a national newspaper archive with example research tasks. From this we learned what data is typically available in digital archives and how it could be used to reduce and/or assess the uncertainty in result sets. We conclude that the current knowledge situation on the users' side as well as on the tool makers' and data providers' side is insufficient and needs to be improved.

3.1 INTRODUCTION

Humanities scholars use the growing numbers of documents available in digital archives not only because they are more easily accessible but also because they support new research tasks, such as pattern mining and trend analysis. Especially for old documents, the results of OCR processing are far from perfect. While improvements in pre-/post-processing and in the OCR technology itself lead to lower error rates, the results are still not error-free. Scholars need to assess whether the trends they find in the data represent real phenomena or result from tool-induced bias. It is unclear to what extent current tools support this assessment task. To our knowledge, no research has investigated how scholars can be supported in assessing the data quality for their specific research tasks.

In order to find out what research tasks scholars typically carry out on a digital newspaper archive (*RQ1*) and to what extent scholars experienced OCR quality to be an obstacle in their research, we conducted interviews with humanities scholars (Section 3.2). From the information gained in the interviews, we were able to classify the research tasks and describe potential impact of OCR quality on these tasks (*RQ2*). With a literature study, we investigated, how digitization processes in archives influence the OCR quality, how Information Retrieval (IR) copes with error-prone data and what workarounds scholars use to correct for potential biases (Section 3.3). Finally, we report on insights we gained from our use case study on the digitization process within a large newspaper archive (Section 3.4) and we give examples of what data scholars need to be able to estimate the quality indicators for different task categories (*RQ3*).

3.2 INTERVIEWS: USAGE OF DIGITAL ARCHIVES BY HISTORIANS

We originally started our series of interviews to find out what research tasks humanities scholars typically perform on digital archives, and what innovative additions they would like to see implemented in order to provide (better) support for these research tasks. We were especially interested in new ways of supporting quantitative analysis, pattern identification and other forms of distant reading. We chose our interviewees based on their prior involvement in research projects that made use of digital newspaper archives and / or on their involvement in publications about digital humanities research. We stopped after interviewing only four scholars, for reasons we describe below. Our chosen methodology was a combination of a structured personal account and a time line interview as applied by Bron and Brown, [11, 12]. The former was used to stimulate scholars to report on their research and the latter to stimulate reflection on differences in tasks used for different phases of research. The interviews were recorded either during a personal meeting (*P1, P2, P4*) or during a Skype call (*P3*), transcribed and summarized. We sent the summaries to the interviewees to make sure that we covered the interviews correctly.

We interviewed four experts. (*P1*) is a Dutch cultural historian with an interest in representations of World War II in contemporary media. (*P2*) is a Dutch scholar specializing in modern European Jewish history with an interest in the implications of digital humanities on research practices in general. (*P3*) is a cultural historian from the UK, whose focus is the cultural history of the nineteenth century. (*P4*) is a Dutch contemporary historian who reported to have a strong interest in exploring new research opportunities enabled by the digital humanities.

All interviewees reported to use digital archives, but mainly in the early phases of their research. In the exploration phase the archives were used to get an overview of a topic, to find interesting research questions and relevant data for further exploration. In case they had never used an archive before, they would first explore the content the archive can provide for a particular topic (see Table 2, *E9*). At later

ID	Interview	Example	Category
E1	P1	Representation of Anne Frank in post-war media	T2
E3	P1	Contextualizing LDJ with sources used	T4
E4	P2	Comparisons of two digitized editions of a book to find differences in word use	T4
E5	P3	Tracing jokes through time and across newspapers	T3
E6	P3	Plot ngrams frequencies to investigate how ideas and words enter a culture	T1/T3
E7	P3	Sophisticated analysis of language in newspapers	T4
E8	P3	First mention of a newly introduced word	T1
E9	P3 /P4	Getting an overview of the archive's contents	T2
E11	P4	Finding newspaper articles on a particular event	T2

Table 2: Categorization of the examples for research tasks mentioned in the interviews. Task type *T1* aims to find the first mention of a concept. Tasks of type *T2* aim to find a subset with relevant documents. *T3* includes tasks investigating quantitative results over time and *T4* describes tasks using external tools on archive data.

stages, more specific searches are performed to find material about a certain time period or event. The retrieved items would later be used for close reading. For example, *P1* is interested in the representations of Anne Frank in post-war newspapers and tried to collect as many relevant newspaper articles as possible *E1*. *P3* reports on studies of introductions of new words into the vocabulary *E8*. Three of the interviewees (*P1*, *P3*, *P4*) mentioned that low OCR quality is a serious obstacle, an issue that is also reflected extensively in the literature [10, 16, 38]. For some research tasks, the interviewees reported to have come up with workarounds. *P1* sometimes manages to find the desired items by narrowing down search to newspaper articles from a specific time period, instead of using keyword search. However, this strategy is not applicable to all tasks.

Due to the higher error rate in old material and the absence of quality measures, they find it hard to judge whether a striking pattern in the data represents an interesting finding or whether it is a result of a systematic error in the technology. According to *P1*, the print quality of illegal newspapers from the WWII period is significantly worse than the quality of legal newspapers because of the conditions under which they were produced. As a consequence, it is very likely that they will suffer from a higher error rate in the digital archive, which in turn may cause a bias in search results. When asked how this uncertainty is dealt with, *P4* reported to try to explain it in the publications. The absence of error measures and information about possible preconceptions of the used search engine, however, made this very difficult. *P3* reported to have manually collected data for a publication to generate graphs tracing words and jokes over time (see *E5, E6* in Table 2) as the archive did not provide this functionality. Today, *P3* would not trust the numbers enough to use them for publications again.

P2 and *P3* stated that they would be interested in using the data for analysis independently from the archive's interfaces. Tools for text analysis, such as Voyant[1], were mentioned by both scholars (*E3, E4, E7*). The scholars could not indicate how such tools would be influenced by OCR errors. We asked the scholars whether they could point out what requirements should be met in order to better facilitate research tasks in digital archives. *P3* thought it would be impossible to find universal methodological requirements, as the requirements vary largely between scholars of different fields and their tasks.

We classified the tasks that were mentioned by the scholars in the interviews according to their similarities and requirements towards OCR quality. The first mention of a concept, such as a new word or concept would fall into category *T1*. *T2* comprises tasks that aim to create a subcollection of the archive's data, e.g. to get to know the content of the archive or to select items for close reading. Tasks that relate word occurrences to a time period or make comparisons over different sources or queries are summarized in *T3*. Some archives allow the extraction of (subsets of) the collection data. This allows the use of specialized tools, which constitutes the last category *T4*.

We asked *P1*, *P2* and *P4* about the possibilities of more quantitative tools on top of the current digital archive, and in all cases the interviewees' response was that no matter what tools were added by the archive, they were unlikely to trust any quantitative results derived from processing erroneous OCRed text. *P2* explicitly stated that while he did publish results based on quantitative methods in the past, he would not use the same methods again due to the potential of technology-induced bias.

None of our interviews turned out to be useful with respect to our quest into innovative analysis tools. The reason for this was the perceived low OCR quality, and the not well-understood susceptibility of the interviewees' research tasks to OCR errors. Therefore, we decided

[1] http://voyant-tools.org/

to change the topic of our study to better understanding the impact of OCR errors on specific research tasks. We stopped our series of interviews and continued with a literature study on the impact of OCR quality on specific research tasks.

3.3 LITERATURE STUDY

To find out how the concerns of the scholars are addressed by data custodians and by research in the field of computer science, we reviewed available literature.

The importance of OCR in the digitization process of large digital libraries is a well-researched topic [28, 34, 47, 51]. However, these studies are from the point of view of the collection owner, and not from the perspective of the scholar using the library or archive. User-centric studies on digital libraries typically focus on user interface design and other usability issues [19, 58, 59]. To make the entry barrier to the digital archive as low as possible, interfaces often try to hide technical details of the underlying tool chain as much as possible. While this makes it easier for scholars to use the archive, it also denies them the possibility to investigate potential tool-induced bias.

There is ample research into how to reduce the error rates of OCRed text in a post-processing phase. For example, removing common errors, such as the "long s"-to-f confusion or the soft-hyphen splitting of word tokens, has shown to improve Named Entity Recognition. This, however, did not increase the overall quality to a sufficient extent as it addressed only 12% of the errors in the chosen sample [2]. Focusing on overall tool performance or performance on representative samples of the entire collection, such studies provide little information on the impact of OCR errors on specific queries carried out on specific subsets of a collection. It is this specific type of information we need, however, to be able to estimate the impact on our interviewees' research questions. We found only one study that aimed at generating high-quality OCR data and evaluating the impact of its quality on a specific set of research questions [42]. Strange et al. found that the impact of OCR errors is not substantial for a task that compares two subsets of the corpus [42]. For a different task, the retrieval of a list of the most significant words (in this case, describing moral judgement), however, recall and precision were considered too low.

Another line of research focuses on how to improve OCR tools or on using separate tools for improving OCR output in a post-processing step [32], for example by using input from the public [29]. Unfortunately, the actual extent, to which this crowdsourcing initiative has contributed to a higher accuracy has not been measured. While effective use of such studies may reduce the error rate, they do not help to better estimate the impact of the remaining errors on specific cases. Even worse, since such tools (and especially human input) add another layer of complexity and potential errors, they may also add more uncertainty to these estimates. Most studies on the impact of OCR errors are in the area of ad-hoc IR, where the consensus is that for long texts and noisy OCR errors, retrieval performance remains remarkably good for relatively high error rates [43]. On short texts, however, the retrieval effectiveness drops significantly [17, 36]. In contrast, information extraction tools suffer significantly when applied to OCR output with high error rates [45]. Studies carried out on unreliable OCR data sets often leave the OCR bias implicit. Some studies explicitly protect themselves from OCR issues and other technological bias by averaging over large sets of different queries and by comparing patterns found for a specific query set to those of other queries sets [1]. This method, however, is not applicable to the examples given by our interviewees, since many of their research questions are centered around a single or small number of terms.

Many approaches aiming at improving the data quality in digital archives have in common that they partially reduce the error rate, either by improving overall quality, or by eliminating certain error types. None of these approaches, however, can remove all errors. Therefore, even when applying all of these steps to their data, scholars still need to be able to quantify the remaining errors and assess their impact on their research tasks.

3.4 USE CASE: OCR IMPACT ON RESEARCH TASKS IN A NEWSPAPER ARCHIVE

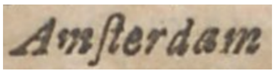
To study OCR impact on specific scholarly tasks in more detail, we investigated OCR-related issues of concrete queries on a specific digital archive: the historic newspaper archive[2] of the National Library of The Netherlands (KB). It contains over 10 million Dutch newspaper pages from the period 1618 to 1995, which are openly available via the Web. For each item, the library publishes the scanned images, the OCR-ed texts and the metadata records. Its easy access and rich content make the archive an extremely rich resource for research projects[3].

3.4.1 *Task: First mention of a concept*

One of the tasks often mentioned during our interviews was finding the first mention of a term (task *T1* in Section 3.2). For this task, scholars can typically deal with a substantial lack of precision caused by OCR errors, since they can detect false positives by manually checking the matches. The key requirement is recall. Scholars want to be sure that the document with the first mention was not missed due to

[2] www.delpher.nl/kranten

[3] See lab.kbresearch.nl for examples.



Amſterdam

Figure 10: Confusing the "long s" for an "f" is a common OCR error in historic texts.

OCR errors. This requires a 100% recall score, which is unrealistic for large digital archives. As a second best, they need to minimize the risk of missing the first mention to a level that is acceptable in their research field. The question remains how to establish this level, and to what extent archives support achieving this level. To understand how a scholar could assess the reliability of their results with currently available data, we aim to find the first mention of "Amsterdam" in the KB newspaper archive. A naive first approach is to simply order the results on the query "Amsterdam" by publication date. This returned a newspaper dated October 25, 1642 as the earliest mention. We then explore different methods to assess the reliability of this result. We first tried to better understand the corpus and the way it was produced, then we tried to estimate the impact of the OCR errors based on the confidence values reported by the OCR engine, and finally we tried to improve our results by incremental improvement our search strategy.

3.4.1.1 *Understanding the digitization pipeline*

We started by obtaining more information on the archive's digitization pipeline, in particular details about the OCR process, and potential post-processing steps.

Unfortunately, little information about the pipeline is given on the KB website. The website warns users that the OCR text contains errors[4], and as an example mentions the known problem of the "long s" in historic documents (see Fig. 10), which causes OCR software to mistake the 's' for an 'f'. The page does not provide quantitative information on OCR error rates.

After contacting library personnel, we learned that formal evaluation on OCR error rates or on precision/recall scores of the archive's search engine had not been performed so far. The digitization had been a project spanning multiple years, and many people directly involved no longer worked for the library. Parts of the process had been outsourced to a third party company, and not all details of this process are known to the library. We believe this practice is typical for many archives. We further learned that article headings had been manually corrected for the entire archive, and that no additional error correction or other post-processing had been performed. We concluded that for the first mention task, our inquires provided insufficient information to be directly helpful.

[4] http://www.delpher.nl/nl/platform/pages/?title=kwaliteit+(ocr)

3.4.1.2 *Uncertainty estimation: using confidence values*

Next, we tried to use the confidence values reported by the OCR engine to assess the reliability of our result. The ALTO XML[5] files used to publish the OCR texts do not only contain the text as it was output by the OCR engine, they also contain confidence values generated by the OCR software for each page, word and character. For example, this page[6], contains:

```
<Page ID="P2" ... PC="0.507">
```

Here, *PC* is a confidence value between 0 (low) and 1 (high confidence). Similar values are available for every word and character in the archive:

```
<String ID="P2_ST00800" ... CONTENT="AM" ...
        SUBS_CONTENT="AMSTERDAM." WC="0.45" CC="594"/>
<String ID="P2_ST00801" ... CONTENT="STERDAM." ...
        SUBS_CONTENT="AMSTERDAM." WC="0.30" CC="46778973"/>
```

Here, *WC* is the word-level confidence, again expressed as a value between 0 and 1. CC is the character-level confidence, expressed as a string of values between 0-9, with one digit for each character. In this case, 0 indicates high, and 9 indicates low confidence. This is an example for a word that was split by a hyphen. The representation of its two parts as "subcontent" of "AMSTERDAM" assures its retrieval by the search engine of delpher.

```
<String ID="P2_ST00766" ... CONTENT="Amfterdam,"
        WC="0.36" CC="0866869771"/>
```

For the last example, this means the software has lower confidence in the correct "m", than in the incorrect "f". Note that since the above XML data is available for each individual word, it is a huge dataset in absolute size, that could, potentially, provide uncertainty information on a very fine-grained level. For this, we need to find out what these values mean and/or how they have been computed. However, the archive's website provides no information about how the confidence values have been calculated.

Again, from the experts in the library, we learned that the default word level confidence scores were increased if the word was found in a given list with correct Dutch words. Later, this was improved by replacing the list with contemporary Dutch words by a list with historic spelling. Unfortunately, it is not possible to reproduce which word lists have been used on what part of the archive.

Another limitation is that even if we could calibrate the OCR confidence values to meaningful estimates, they could only be used to estimate how many of the matches found are likely false positives.

[5] http://www.loc.gov/standards/alto/

[6] http://resolver.kb.nl/resolve?urn=ddd:010633906:mpeg21:p002:alto

Category	Confusion matrix	CV output	CV alternatives
available for:	sample only	full corpus	not available
T1 1st mention of *x*	find all queries for *x*, impractical	estimated precision not helpful	improve recall
T2 Selecting subset relevant to *x*	as above	estimated precision, requires improved UI	improve recall
T3. Pattern over time *x*	pattern summarized over set of alt. queries	estimates of corrected precision	estimates of corrected recall
T3.a Compare x_1 and x_2	warn for diff. susceptibility to errors	as above, warn for diff. distribution of CVs	as above
T3.b Compare $corpus_1$ and $corpus_2$	as above	as above	as above

Table 3: The different types of tasks require different levels of quality. Quality indicators can be used to generate better estimates of the quality and also (to some extent) to compensate low quality. *x* stands for an abstract concept that is the focus of interest in the research task.

They provide little or no information on the false negatives, since all confidence values related to characters that were considered as potential alternatives to the character chosen by the OCR engine have not been preserved in the output and are lost forever. For this research task, this is the information we would need to estimate or improve recall. We thus conclude that we failed in using the confidence values to estimate the likelihood that our result indeed represented the first mention of "Amsterdam" in the archive. We summarized our output in Table 3, where for *T1* we indicate that using the confusion matrix is impractical, using the out confidence values (CV output) is not helpful, and using the confidence values of the alternatives (CV alternatives) could have improved recall, but we do not have the data.

3.4.1.3 *Incremental improvement of the search strategy*

We observed that the "long s" warning given on the archive's website is directly applicable to our query. Therefore, to improve on our original query, we also queried for "Amfterdam". This indeed results in an earlier mention: July 27, 1624. This result, however, is based on our anecdotal knowledge about the "long s problem". It illustrates the need for a more systematic approach to deal with spelling variants. While the archive provides a feature to do query expansion based on historic spelling variants, it provides no suggestions for "Amsterdam". Querying for known spelling variants mentioned on the Dutch history of Amsterdam Wikipedia page also did result in earlier mentions.

To see what other OCR-induced misspellings of Amsterdam we should query for, we compared a ground truth data set with the associated OCR texts. For this, we used the dataset[7] created in the context of the European IMPACT project. It includes a sample of 1024 newspaper pages, but these had not been completely finished by end of the project. This explains why this data has not been used in a evaluation of the archive's OCR quality. Because of changes in the identifier scheme used, we could only map 265 ground truth pages to the corresponding OCR text in the archive. For these, we manually corrected the ground truth for 134 pages, and used these to compute a confusion table[8]. This matrix could be used to generate a set of alternative queries based on all OCR errors that occur in the ground truth dataset. Our matrix contains a relatively small number of frequent errors, and it seems doable to use them to manually generate a query set that would cover the majority of errors. We decided to look at the top ten confusions and use the ones applicable to our query. All combinations of confusions resulted in 23 alternative spelling variations of "Amsterdam". When we queried for the misspellings, we found hits for all variations, except one, "Amfcordam". None, however, yielded an earlier result than our previous query.

This method could, however, be implemented as a feature in the user interface, the same way as historic spelling variants are supported[9]. Again, the issue is that for a specific case, it is hard to predict whether such a future would help, or merely provide more false positives.

Our matrix also contains a very long tail with infrequent errors, and for this specific task, it is essential to take all of them into account. This makes our query set very large and while this may not be a technical problem for many state of the art search engines, the current user interface of the archive does not support such queries. More importantly, the long tail also implies that we need to assume that our ground truth does not cover all OCR errors that are relevant for our task.

We conclude that while the use of a confusion matrix does not guarantee finding the first mention of a term, it would be useful to publish such a matrix on each digital archive's website. Just using the most frequent confusions can already help user to avoid the most

[7] lab.kbresearch.nl/static/html/impact.html

[8] available on http://dx.doi.org/10.6084/m9.figshare.1448810

[9] http://www.delpher.nl/nl/platform/pages/?title=zoekhulp

frequent errors, even in a manual setting. Systematic queries for all known variants would require more advanced backend support.

Fortunately, it lies in the nature of our task that with every earlier mention we can confirm, we can also narrow the search space by defining a new upper bound. In our example, the dataset with pages published before our 1624 upper bound is sufficiently small to allow manual inspection. The first page in the archive of the same title as the 1624 page, is published in 1619, and has a mention of "Amsterdam". It is on the very bottom of the page in a sentence that is completely missing in the OCR text. This explains why our earlier strategy has missed it. The very earliest page in the archive at the time of writing is from June 1618. Its OCR text contains "Amfterftam". Our earlier searches missed this one because it is a very rare variant which did not occur in the ground truth data. While we now have found our first mention in the archive with 100% certainty, we found it by manual, not automatic means. Our strategy would not have worked when the remaining dataset would have been too large to allow manual inspection.

3.4.2 *Analysis of other tasks*

We also analyzed the other tasks in the same way. For brevity, we only report our findings to the extent they are different from task *T1*. For *T2*, selecting a subset on a topic for close reading, the problem is that a single random OCR error might cause the scholar to miss a single important document as in *T1*. In addition, a systematic error might result in a biased selection of the sources chosen for close reading, which might be an even bigger problem. Unfortunately, using the confusion matrix is again not practical. The CV output could be useful to improve precision for research topics where the archive contains too many relevant hits, and selecting only hits above a certain confidence threshold might be useful. This requires, however, the user interface to support filtering on confidence values. For the CV alternatives, they again could be used to improve recall, but it is unclear against what precision.

For task *T3*, plotting frequencies of a term over time, the issue is no longer whether or not the system can find the right documents, as in *T1* and *T2*, but if the system can provide the right counts of term occurrences despite the OCR errors. Here, the long tail of the confusion matrix might be less of a problem, as we may choose to only query for the most common mistakes, assuming that the pattern in the total counts will not be affected much by the infrequent ones. CV output could be used to lower counts for low precision results, while CV alternatives could be used to increase counts for low recall matches. For *T3.a*, a variant of *T3* where the occurrence over time of one term is compared to another, the confusion matrix could also be used to warn scholars if one term is more susceptible to OCR errors than the other. Likewise, a different distribution of the CV output for the two terms might be flagged in the interface to warn scholars about potential bias. For *T3.b*, a variant where the occurrence of a term in different newspapers is analyzed, the CV values could likely be used to indicate different distributions in the sources, for example to warn for systematic errors caused by differences in print quality or fonts between the two newspapers.

For task *T4* (not in the table), the use of OCRed texts in other tools, our findings are also mainly negative. Very few text analysis tools can, for example, deal with different confidence values in their input, apart from the extensive standardization these would require for the input/output formats and interpretation of these values. Additionally, many tools suffer from the same limitation that only their overall performance on a representative sample of the data has been evaluated, and little is known about their performance on a specific use case outside that sample. By stacking this uncertainty on top of the uncertain OCR errors, predicting its behavior for a specific case will be even harder.

3.5 CONCLUSIONS

Through interviews we conducted with scholars, we learned that while the uncertain quality of OCRed text in archives is seen as a serious obstacle to wider adaption of digital methods in the humanities, few scholars can quantify the impact of OCR errors on their own research tasks. We collected concrete examples of research tasks, and classified them into categories. We analyzed the categories for their susceptibility to OCR errors, and illustrated the issues with an example attempt to assess and reduce the impact of OCR errors on a specific research task. From our literature study, we conclude that while OCR quality is a widely studied topic, this is typically done in terms of tool performance. We claim to be the first to have addressed the topic from the perspective of impact on specific research tasks of humanity scholars.

Our analysis shows that for many research tasks, the problem cannot be solved with better but still imperfect OCR software. Assessing the impact of the imperfections on a specific use case remains important.

To improve upon the current situation, we think the communities involved should begin to approach the problem from the user perspective. This starts with understanding better how digital archives are used for specific tasks, by better documenting the details of the digitization process and by preserving all data that is created during the process. Finally, humanity scholars need to transfer their valuable tradition of source criticism into the digital realm, and more openly criticize the potential limitations and biases of the digital tools we provide them with.

4 WORKSHOP ON TOOL CRITICISM IN THE DIGITAL HUMANITIES

In May 2015 we organized a workshop on *Tool Criticism for Digital Humanities* together with the eHumanities group of KNAW[1] and the Amsterdam Data Science Center[2]. The goal of this workshop was to bring together people with an interest in Digital Humanities research for focused discussions about the need for *tool criticism* in DH research.

We aimed to identify

- typical research tasks affected by by technology-induced bias or other tool limitations
- the specific information, knowledge and skills required for researchers to be able to perform tool criticism as part of their daily research
- guidelines or best practices for systematic tool and digital source criticism[3]

The following pages summarize the results of the workshop.

4.1 MOTIVATION AND BACKGROUND

In digital humanties (DH) research there is a trend to the use of larger datasets and mixing hermeneutic/interpretative with computational approaches. As the role of digital tools in these type of studies grows, it is important that scholars are aware of the limitations of these tools, especially when these limitations might bias the outcome of the answers to their specific research questions. While this potential bias is sometimes acknowledged as an issue, it is rarely discussed in detail, quantified or otherwise made explicit.

On the other hand, computer scientists (CS) and most tool developers tend to aim for generic methods that are highly generalisable, with a preference for tools that are applicable to a wide range of research questions. As such, they are typically not able to predict the performance of their tools and methods in a very specific context. This is often the point where the discussion stops.

The aim of the workshop was to break this impasse, by taking that point as the start, not the end, of a conversation between DH and CS researchers. The goal was to better understand the impact

[1] https://www.ehumanities.nl/archive/2013-2016/

[2] http://amsterdamdatascience.nl/

[3] https://event.cwi.nl/toolcriticism/

of technology-induced bias on specific research contexts in the humanties. More specifically, we aimed to identify:

- typical research tasks affected by by technology-induced bias or other tool limitations
- the specific information, knowledge and skills required for scholars to be able to perform tool criticism as part of their daily research
- guidelines or best practices for systematic tool and digital source criticism

4.1.1 *Tool Criticism*

With *tool criticism* we mean the evaluation of the suitability of a given digital tool for a specific task. Our goal is to better understand the impact of any bias of the tool on the specific task, not to improve the tools performance.

While source criticism is common practice in many academic fields, the awareness for biases of digital tools and their influence on research tasks needs to be increased. This requires scholars, data custodians and tool providers to understand issues from different perspectives. Scholars need to be trained to anticipate and recognize tool bias and its impact on their research results. Data custodians, tool providers and computer scientists, on the other hand, have to make information about the potential biases of the underlying processes more transparent. This includes processes such as collection policies, digitization procedures, optical character recognition (OCR), data enrichment and linking, quality assessment, error correction and search technologies.

4.1.2 *Organisation and format*

The scope and format of the workshop was developed during an earlier meeting of the workshop organisers at CWI in Amsterdam. Participants were asked to use the workshop website to submit use cases in advance, and we received seven use cases in total.

The program of the workshop was split in several parts. The morning was dedicated to introducing the concept of *tool criticism*, pointing out the goals and non-goals of the workshop and a short presentation of the use cases (see 4.2. During an informal lunch, participants could express interest in a specific use case. The participants choose 4 out of all 7 use cases for the afternoon sessions, and formed teams around these 4 cases. After lunch, each of the four breakout groups were asked to work out their use cases further. The organizers provided a list of questions to guide and inspire the breakout sessions (see Appendix 4.4). Afterwards, the results were presented and discussed in the plenary. All use case leaders were so kind as to send us their notes by email. These notes as well as notes taken during the presentations were used as input for section 4.2.

4.1.3 *Workshop opening*

Before the use cases were presented, we briefly explained the goals (see Section 4.1) and non-goals of the workshop. The non-goals included: discussions on how to *reduce* tool-induced bias (i.e. by improving the tool), to down-play the role of the tools ("the tool is only used in exploratory phase of research") or discussions about the pros and cons of digital versus non-digital approaches ("we would just hire 20 interns to do this by hand").

4.2 USE CASES

The following use cases were submitted to the workshop:

- Co-occurrence of named entities in newspaper articles
- SHEBANQ
- Word frequency patterns over time
- Polimedia
- Location extraction and visualisation
- contaWords
- Quantifying historical perspectives

From this list, the participants chose to discuss the first 4 use cases in the breakout sessions. The participants were asked to form groups with at least one researcher from (Digital) Humanities as well as Computer Science.

4.2.1 *Constructing social networks with co-occurrence*

This use case was submitted by Jacqueline Hicks (KITLV) under the original title "Co-occurrence of Named Entities in Newspaper Articles".

Use case description

The computational strategy is to use the co-occurrence of named entities in newspaper articles to represent a real-world relationship between those entities.

Main discussion points[4]

The discussion started with explaining the purpose of the tool: As well as locating names of people appearing together in one sentence in a newspaper article, it was also used in the project to help disambiguate entities.

The tool makes use of the widely known and used Stanford NER, its performance is documented on CoNLL 2002 and 2003 NER data[5]. This data is not similar to the data used in the example use case. To be able to evaluate the performance of the Stanford NER in the new domain, the researcher would need a corresponding "ground truth" data set, that is, manually constructed reference data that can be used to check the results of the automatic NER process. Developing a ground truth for a new domain is a very time consuming operation.

The research task is to find out whether the tool can help detect changes in communities of elite that changed over regime transitions when the Indonesian authoritarian government fell after 30 years in power. However, the task turned out to be difficult to solve as insufficient data was available for the time before 1998. More time is needed to add linguistic context to the co-occurrences to find what sort of relationships ties the entities together in a sentence. A co-occurrence of two entities can mean that they participated in the same event, that one person commented on the other or that they were in competition with each other. With such diverse relations, it is difficult to draw conclusions from the automatically generated graph.

BIASES OF THE SOURCE SELECTION The data was collected from several listserves of news articles on Indonesian politics. The articles on these listserves were handpicked by those running them and so could not be considered free from bias. They include, for example, the articles in English language, chosen for the interest of foreign and Indonesian readers generally interested in political reform, as it was originally started to share information among activists under the authoritarian government. Since these biases are known, they are easily dealt with as limitations of the study in the same way that research limitations are usually explained when writing in the social sciences. This is in contrast to the computational filtering which introduces biases which are not known to the social scientist.

PROVENANCE OF THE DATA All articles had date and newspaper source on them.

[4] The summary of the discussion is based on notes kindly made available to us by Lynda Hardman.

[5] http://nlp.stanford.edu/ner/

UTILITY OF THE TOOL Utility is limited and only good for some initial explorations. The idea of the session was to find ways to integrate qualitative information from interviews with Indonesian elites about their network with the computational techniques. The group discussed the idea to investigate the changes in the political system by creating two networks for the time before and another two for the time after the transition. The networks could then be compared and in case the networks of the same period coincide, but the networks across periods do not, they may be used to reveal interesting differences as basis for further research. Jacky would have to explain the differences by investigating through political sciences literature how a military group fragmented in this way around person X and/or came together (again). Jacky already wrote a paper on how social scientists have identified populations of elites in the past and how this can be done differently with computational tools [26].

Summary

In general, to methodically evaluate tools is extremely time intensive. It requires intensive exchange between the user and developer. As publishing papers is an incentive to work in academia, the lack of forums to publish about tool criticism is a problem.

4.2.2 *SHEBANQ*

This use case was submitted by Dirk Roorda (DANS).

Use case description

SHEBANQ[6] allows users to query the Hebrew text database created over the years by the ETCBC group at the VU University Amsterdam. There is an associated, offline tool, LAF-Fabric for more refined and intense processing of the data. The data is encoded in Linguistic Annotation Framework, an ISO standard. LAF-Fabric is a python tool to deal with big LAF resources efficiently. There are several modules on top of it that exploit the structure of this particular research.

Main discussion points[7]

THE TOOL At the beginning of the breakout session, Dirk Roorda introduced the participants to some of the functionalities. SHEBANQ should actually be seen as a collection of tools to annotate and query the Hebrew Bible. It is developed at the Eep Talstra Centre for Bible and Computing which is located at VU University. Not only is the

[6] http://shebanq.ancient-data.org/

[7] The summary is based on the notes taken collaboratively by Dirk Roorda (session leader), Michiel Cock, Arjen de Vries, Liliana Melgar and Myriam Traub and the personal notes of Myriam Traub.

tool freely available through the *ETCBC's organizational github account*[8], a user can also download the documentation of the tool as well as executable *IPython Notebooks* that demonstrate some uses of the tool.

THE DATA SHEBANQ is designed to support analysis of a specific version of the Hebrew Bible and is therefore tailored to cater to the specific requirements of the data set. Using the tool on a different data set therefore does not seem reasonable. The data is encoded in the Linguistic Annotation Framework (LAF), an ISO standard [31].

THE USER SHEBANQ encourages a community of people to come forward with their attempts to answer research questions by means of formalizing questions into tasks that can be run on the data. A unique feature of SHEBANQ is the possibility to share and publish queries:

> "If you want to cite your shared query in a publication, you can also publish it. Your query and its results on a particular version om the database will be frozen, so that others will see exactly the same results later on. When newer versions of the database arrive in SHEBANQ, you can run the same query on the new data. You can modify that version of your query and publish it separately from earlier versions."[9]

This is seen as an important and novel feature that can facilitate the discussion among users on the fitness of a query for a given research task.

Summary

It is vitally important to make explicit how the data in the ETCBC database has been encoded. Who has done it by what methods? Especially when the same researcher draws conclusions from the database as the one who has contributed relevant parts of the encoding. That is not necessarily bad, as long as his/her method of encoding is well described and can be subject to criticism. Another matter is whether other researchers are willing to contribute data to SHEBANQ. That will only happen if others can identify with the way of encoding and trust that SHEBANQ is impartial. Maybe SHEBANQ should allow multiple encoding styles and give other researchers partial ownership.

4.2.3 *Word frequency patterns over time*

This use case was submitted by Marijn Koolen (UvA).

[8] https://github.com/ETCBC

[9] http://shebanq.ancient-data.org/

Use case description

The use case aims at looking into tools that chart word frequencies using timebased counts of n-grams found in digital sources. Examples of such tools are the Google Ngram Viewer[10] and the Ngram Viewer bases on historic newspapers which was developed by KB[11].

Main discussion points[12]

Criticism is not only a playing field for Computer Sciences and Humanities but also for libraries and social sciences. It is, however, sometimes difficult to distinguish tool criticism from data criticism since tools have been used to create the data. These tools may not be available for criticism, which needs to be explicitly accounted for.

THE TOOL The chosen tools are designed to visualize word counts on a time line. In the experience of the researchers, this task is not as simple as it may appear: three different programs give three different counts for the total word count. In linguistic annotation, when different people annotate the same text, different schemes are used. The resulting conflicts need to be resolved by writing down the choices and agreements. This could be done similarly for coding / tools. The different results show that a tool does interpretation, too (in the sense of defining what a "word" is). Humanists are often put off if such counts are off by 1, because they tend to have precise ideas about text length. Without statistics, it is hard to say how much difference/variance is 'allowable' for a humanities researcher. This also applies for search engines. One participant recalls different answers from different search engines on the same query. She concluded that tools are not neutral, and that accuracy/concreteness are an illusion.

Interestingly, though, textual scholars seem to cope with this lack of precision very well *until* they start using a technical tool? We should remind people that tools and code are human engineered creative contraptions that have all queer human decisions and ambiguities built in. History depends on who writes it, code depends on who writes it.

THE USER The group further discussed the skills required from a humanities scholar to perform the tool criticism. They agreed that to understand a program, to some extent programming skills are required and that it should be part of the education in DH. A better documentation of the program code could make it easier to understand it and/or code in a way that is easier understandable. The readability of the code, however, may affect the efficiency. Understanding the im-

[10] https://books.google.com/ngrams

[11] http://lab.kbresearch.nl/find/Ngrams#

[12] The summary is based on the notes kindly made available to us by Joris van Zundert.

plications of program code requires deep inspection and knowledge of the code. It can be very complicated for good reasons. If scholars cannot invest a considerable amount of time in understanding the code, they will have to trust the experts. In that case, the developers need to explain the tool, for example how it counts words and why it may come up with different totals for word counts than another tool.

The scholars may also need the programming skills to understand the methodology that a certain tool may force on them. This is foremost a task for humanities researchers to experiment and judge the methodology. If the source code is not (freely) available, it requires the scholar to experiment in order to to find major shortcomings or bias [1]. The use of tools should be embedded in a research process that iterates between distant reading and close reading to foster understanding. Tool support for this could be provided with "Sub Corpus Modeling" [46]. Ideally, there are multiple tools available that a scholar can choose from. In order to make an informed decision for choosing one above the other, the criteria need to be clear.

An important aspect of criticism is seen in its publication. Results of tool criticism should be reported to other users. This, in turn, raises the question of trust. Criticism cannot be considered as neutral, as it depends on the persona, background, status, etc. of the critic.

TOOL BUILDERS Tool builders and computer scientists could learn from the humanities that there are more perspectives / more possible choices what the 'data' are. However, computer scientists are *not* interested in what DH does. CS studies process and abstraction. We should NOT suggest that DH is the field where Humanities and CS meet. It is maybe where AI and Humanities meet.

Summary

CS/AI need to evaluate a tool in a way that is tailored to a humanities researcher. The commonly used computational metrics usually do not answer that question. The DH, however, are in a 'it's all up in the air' period (as opposed to times in science where things seem to be clear and generalized); and scholars are not even sure about the standards against which they should be evaluated. Therefore, in order to define the requirements of humanities scholars, more discussions between the two disciplines is needed.

4.2.4 *Polimedia*

This use case was submitted by Laura Hollink (CWI).

Use case description

PoliMedia[13] is designed for specific humanities research tasks that require the possibility to do a cross-media analysis [33]. An example use case might be: studying several events related to “the resignation of Aantjes” by comparing information from different media. With PoliMedia, researchers can search among the debates in the Dutch Parliament (Dutch Hansard), Dutch historical newspapers archive and ANP radio bulletins, in a uniform search interface. The functionality is proven useful and the system design is highly valued. However, there are still obvious limitations.

Main discussion points[14]

During the discussion, the PoliMedia group particularly wrote down a list of deficiencies of resource bias, then brainstormed about the solutions from the “tool side”.

BIASES OF THE SOURCE SELECTION Some bias issues of the dataset are known and might be quantified or circumvented. One problem is the coverage and selection of the resources: PoliMedia does not make use of data from television programs and news (but it does have ANP as a data source); the dataset covers only one radio station, so opinions might be limited; the selection of KB newspaper items for PoliMedia are significantly different in amount related to different newspaper brands. Additionally, there are technical issues such as OCR errors in the database, hindering users in retrieving the complete results. There are also biases that the creators of the system cannot circumvent or quantify. On one hand, some of the links/search results are definitely lost due to the bias in the phase collecting the resource and system’s technical issues. We do not know what we are missing in the database and how those missing files would influence researchers’ conclusions. On the other hand, bias can be caused by a chain of uncertainty: we don’t know what the bias is of the off-the-shelf topic extraction tool.

DATA PROVENANCE Provenance of the data is clear and all search results link to original sources where a user could check if the digital versions are correct. However, the provenance of the algorithm is unclear: e.g. the system limits links to articles written within 7 days of the debate. This would be a limitation if the user needs more information. Such issues could be solved by a collaboration/dialog between tool makers and users, to explain and point out the impact of the al-

[13] http://www.polimedia.nl/

[14] The summary of the discussion is based on notes kindly made available to us by Laura Hollink.

gorithms on specific research questions. It is also possible to change the tool so the user can define a time period.

SOLUTION BRAINSTORM In regard to to the limitations discussed, the group wrote down some questions and brainstormed solutions for them:

How to convince a reviewer that dataset and tool are good enough to draw quantitative conclusions from it?

Solution: *Sandbox:* on the spot evaluation of that particular query: The general goal is to provide the user the means on the spot get a feeling or even a measure of the bias. For dealing with the bias of data selection, practically user can always manually go to KB archive for more complete files that should be in there. The system can also compare the results with present links to on the spot, evaluating for that particular query. Till here users might still miss some links, but at least they cannot systematically miss out on things. For dealing with OCR issues, the system can provide relevance feedback, and does a query expansion to help users finding miss-OCRed versions of their query.

Quality can vary per query (e.g. simple/complex query, OCR errors, etc.), how to deal with the specific quality issues?

Solution: *Queries Sharing:* If a user took time making queries for a particular topic and find meaningful results, other users may also need the "accurate queries" when searching for similar topics. Solution: Triangulation: could be possible if we had multiple versions of the linking algorithm.

Solution: *Sharing the research process:* validated subsets that could be reused and criticized.

FUTURE QUESTIONS AND RESEARCH DIRECTION Given what we know about the quality of the tool/data, what can we do in our research:

- What can prove that something is there: e.g. media said x, this debate is discussed by x
- We can never prove that something is not there: e.g. nobody said x, this debate is never discussed.
- We can find preliminary results for quantitative questions: e.g. this debates is discussed more in the media than another one. Further research would be needed for definite conclusions about these kinds of questions

Summary

The group discussed different biases that influence research tasks. Some biases were found easy to assess and circumvent (limited number of sources included), others were more difficult (missing links and cascading of biases from tools used for preprocessing). The question was raised, how a reviewer could be convinced that tool and data are suitable to perform the task. Solutions could be a "sandbox" approach (on the spot evaluation of a query, ask user to give some manual results and check), a community approach (share queries, quality of queries, validate queries and answers) or cross validation with other tools.

4.3 RESULTS

At the end of the workshop, the participants agreed that the idea of *Tool Criticism* as part of the Digital Humanities' research practices should be fostered. This could be achieved in different ways. A traditional way that would reach a large audience could be a journal article (Digital Scholarship in the Humanities[15], Digital Humanities Quarterly[16] or a conference contribution (DH 2016, DH BeNeLux 2016).

Complementary to this, a more "interactive" approach in the form of a website[17] or a forum was suggested. This could be used to obtain feedback from users on a selected set of powerful tools. It would be interesting to be able to collect use cases and to compare evaluations of different tools that were designed to support similar tasks (such as named entity extraction). The insights gained from these examples could be used to create checklists and guidelines for both, tool builders and users. The checklists should, however, not only focus on general tasks, but also on very specific ones.

In order to encourage the direct exchange of ideas between tool builders and humanities scholars and to complement creation and evaluation of tools, hackathons could be organized. This could be done in one-day events, such as a follow-up workshop or at larger scale as part of a Dagstuhl or Lorentz Center seminar. Ideally, those activities should result in the establishment of a European network for tool criticism.

4.4 APPENDIX: QUESTIONS

These questions are intended to provide starting points for the breakout sessions and to stimulate the discussion, in case is comes to a standstill. You do not need to answer all questions and it may well be that the splitting into two separate categories does not work well for your use case. If so, please feel free to add, remove, merge, move, or reword the questions in a way that they fit your needs.

[15] http://dsh.oxfordjournals.org/

[16] http://www.digitalhumanities.org/dhq/

[17] see for example http://programminghistorian.org/

Data set criticism

1. What type of data does the tool make use of?
    a) Is the tool able to cope with multiple data sets (of different types)?
    b) What is the relation between data set and tool?
    c) How does the tool deal with anomalies and outliers?

2. Is documentation on the curation, representativity, biases and pitfalls of the data set available?

3. Is provenance data on the data set available?

4. Who created the data set?
    a) Who was involved? What is the reputation / scientific impact / qualification of the people involved?
    b) What institutions were involved? What is the reputation / scientific impact of the institutions involved?

5. When and how was the data set published?

6. Was the data collected for a specific task / research question?
    a) How does this differ from your intentions?
    b) Is the data set credible and objective?

7. Do other versions of the data set exist?
    a) Are there older / more recent versions of the data set?
    b) How do the versions differ?

8. Does the data show a particular political or cultural bias?
    a) Is this bias of importance for your research question?

9. Do similar data sets from other sources exist?
    a) Can you use the other data set(s) to answer the same research question?
    b) Can you use the other data set(s) to detect / quantify biases in your data set (triangulation)?

Tool criticism

1. Was the tool developed to perform a specific task?
    a) How does this task differ from yours?
    b) For which part of your research cycle do you think the tool is suited (exploration, hypothesis generation, ...)?

2. Is documentation on the precision, recall, biases and pitfalls of the tool available?

3. Is provenance data available on the way the tool manipulates the data set? (i.e. algorithms, choices when selecting, NLP pipeline)
    a) What would it take to make the tool suitable for drawing quantitative conclusions?

4. Which versions of the tool are available?
    a) What are the differences between the versions?
    b) Which version caters best to the requirements of your research task?

5. Who are the developers behind the tool?
    a) Who was involved? What is the reputation / scientific impact / qualification of the people involved?
    b) What institutions were involved? What is the reputation / scientific impact of the institutions involved?
    c) Do you know which scientific discipline the tool was built for? Does this matter for your research task?

6. Do you know similar tools?
    a) Can you use other tools to answer the same research question?
    b) Can you use the other tools to detect / quantify biases in your data set (triangulation)?

5 QUERYLOG-BASED ASSESSMENT OF RETRIEVABILITY BIAS IN A LARGE NEWSPAPER CORPUS

Bias in the retrieval of documents can directly influence the information access of a digital library. In the worst case, systematic favoritism for a certain type of document can render other parts of the collection invisible to users. This potential bias can be evaluated by measuring the *retrievability* for all documents in a collection. Previous evaluations have been performed on TREC collections using simulated query sets. The question remains, however, how representative this approach is of more realistic settings. To address this question, we investigate the effectiveness of the retrievability measure using a large digitized newspaper corpus, featuring two characteristics that distinguishes our experiments from previous studies:

(1) Compared to TREC collections, our document collection contains noise originating from OCR processing, historical spelling and use of language.

(2) Instead of simulated queries, the collection comes with real user query logs including click data.

First, we assess the retrievability bias imposed on the newspaper collection by different IR models. We assess the retrievability measure and confirm its ability to capture the retrievability bias in our setup. Second, we show how simulated queries differ from real user queries regarding term frequency and prevalence of named entities, and how this affects the retrievability results.

5.1 INTRODUCTION

For many digital libraries and archives, users are limited to the retrieval system offered by the data custodian. It is important for users that all relevant documents are equally likely to be retrieved, i.e. that retrieved results are not biased by hidden technological artefacts. If, however, the bias in the search technology impacts the findings of research tasks in a way that it renders relevant documents inaccessible or over-represents specific types of documents, this can lead to a skewed perception of the archive's contents. It is therefore important to provide data custodians and users with a measure to quantify the degree to which the retrieval system provides a neutral way of giving access to a document collection.

In the domain of Information Retrieval (IR), Azzopardi and Vinay introduced a way to measure how retrieval systems influence the accessibility of documents in a collection [3]. The *retrievability score* of a document d, $r(d)$, measures how *accessible* a document is. It is determined by several factors, including the matching function of the retrieval system and the number of documents a user is willing to evaluate. The retrievability score is the result of a cumulative scoring function, defined as:

$$r(d) = \sum_{q \in Q} o_q \cdot f(k_{dq}, c),$$

where c defines the number of documents a user is willing to examine in a ranked list. The coefficient o_q weights the importance of a query. The function $f(k_{dq}, c)$ is a generalized utility/cost function, where k_{dq} is the rank of d in the result list for q. f is defined to return a value of 1 if the document is successfully retrieved below rank c, and 0 otherwise. In summary, $r(d)$ counts for how many queries $q \in Q$ a document d is retrieved at a rank lower than a chosen cutoff c.

Using TREC collections and simulated queries, Azzopardi and Vinay demonstrated the effectiveness of retrievability as a measure for bias, and how retrievability can be used to compare the bias of different retrieval models [3]. We add to their findings by examining the effectiveness of the retrievability measure, and the query simulation procedure in a more realistic setting and we answer the following research questions:

- *RQ1: Is the access to the digitized newspaper collection influenced by a retrievability bias?*

We use the retrievability measure following a similar experimental setup as described in [3] to the digitized historic newspaper archive of the National Library of the Netherlands. This allows us to investigate the retrievability inequality of documents on a digitized – and therefore error-prone – corpus.

- *RQ2: Can we correlate features of a document (such as document length, time of publishing, and type of document) with its retrievability score?*

We investigate whether documents with specific features are particularly susceptible or resistant towards retrievability bias. This allows to better understand the origin of retrievability bias.

- *RQ3: To what extent are retrievability experiments using simulated queries representative of the search behavior of real users of a digital newspaper archive?*

The availability of user logs allows us to compare retrievability patterns of simulated queries to those generated with real user queries. We investigate how the results differ, for example, what types of documents the queries favor most. Finally, we compare the retrieved document sets with the documents viewed by users to explore how well the results match with users' interests.

Our study investigates the applicability of the retrievability concept to a digitized newspaper collection and the representativeness of simulated query sets of user queries.

5.2 RELATED WORK

The *Gini coefficient* and the *Lorenz curve* were introduced by Azzopardi and Vinay as means to assess and express potential bias in the accessibility of documents in a collection [3]. Both indicators were originally developed to measure and visualize a degree of inequality in societies [21], such as deprivation and satisfaction [60]. A "perfect tyranny", where one "tyrant" owns the entire fortune, is represented by a Gini coefficient of $G = 1$, whereas for the "perfect communist" scenario $G = 0$. Both have been used in several studies to facilitate the comparison of retrievability inequality of different IR models, subsets of the document collection, parameter sets and cutoff values [3, 4, 54, 53]. We follow these examples and use Lorenz curves and Gini coefficients to assess the retrievability inequality in a digitized newspaper archive, but we also show what other indicators could be used to better understand the *source* of the inequality.

Several additional studies investigated different aspects of retrievability. Most of these studies largely followed the approach introduced in [3], as well as its metrics. Subdomains of IR that are very sensitive to recall are legal and patent retrieval. An IR model that performs poorly on a specific patent collection can therefore have a devastating effect on the result of the search task. A study by Bache and Azzopardi comparing the retrievability of documents in the MAREC[1] collection through different retrieval models [4] adapted the process used in [3] to generate queries to better simulate the search behavior of patent searchers. They included only bi-term queries as it allowed them to use Boolean operators. Our study shows that even more improvements to the query simulation process are necessary.

To facilitate comparisons across corpora, Bache and Azzopardi suggest that the document to query ratio (DQR) should be kept constant [4]. A high DQR, meaning that a relatively small number of queries is applied to a large data set, may lead to an unrealistically high *Gini* coefficient as a large fraction of documents is never retrieved. Low DQR values are very difficult for experiments with large corpora and real queries. None of the studies we found addresses this problem. The main reason for this being that most studies on retrievability make use of TREC document collections [5, 7, 8, 53, 54, 56], or a freely available corpus of patents from the US patent and trademark office [6]. As these data collections are not provided with query logs from real users, the queries for these studies were generated from the terms in the collection, which allows the researchers to create

[1] www.ir-facility.org/prototypes/marec

any number of queries to meet a predefined DQR. We show how a high DQR influences the results of a retrievability study with queries based on user logs and suggest compensation strategies.

5.3 APPROACH

To answer *RQ1*, we explore whether we can identify a retrievability bias with an approach similar to that reported by Azzopardi and Vinay in [3]. We assess the bias by calculating retrievability scores for every document in the collection for three different IR models, two different query sets (real and simulated), and several cutoff values c. For all of these conditions, we calculate the Gini coefficient. Additionally, we visualize the bias in the retrievability results using Lorenz curves.

To verify that the retrievability scores we generated are meaningful, we test in a known-item-search setup, whether documents with a lower $r(d)$ score are actually harder to find than documents with a higher $r(d)$ score. This is achieved by comparing the mean reciprocal ranks (MRR) of target documents of low scoring and high scoring documents for significant differences.

Understanding how specific document features contribute to a potential retrievability bias would allow a data custodian or a user to make a prediction of how likely they would be able to find documents with this feature in a specific retrieval task. We analyze whether features, such as time of publishing, estimated OCR quality or the newspaper title a document originates from, correlates with a higher or lower retrievability of a document (*RQ2*). Furthermore, we investigate the influence of different parameters (specifically stemming, use of Boolean operators and stopwords) on the retrievability of documents.

As queries play an essential role in any retrieval task, we compare how representative simulated queries are for real user queries. We analyze and compare the composition and length of simulated and real queries and how their result sets differ (*RQ3*). To find out which setup best caters to the users' interests, we compare how well the result sets we obtained in our previous experiments overlap with the documents that were actually viewed.

5.4 EXPERIMENTAL SETUP

We describe the collection of historic newspapers, the query sets and the parameters we used. To obtain comparable results, we followed the experimental setup of [3] as closely as possible, namely to assess the retrievability of documents through a cumulative scoring model. This means that a document score is given for each query for which a document ranks above a pre-specified cutoff rank (c). We quantified the extent to which the retrievability scores of different retrieval models vary using Lorenz curves and Gini coefficients. To verify the meaningfulness of the retrievability scores, we measure the effectiveness of queries designed to retrieve previously selected documents. An analysis of document features and their correlation with retrievability scores concludes our exploration of the bias in our document collection. The second part of our research investigates the representativeness of retrievability results by comparing the results with view data from the user logs.

5.4.1 *Data Sets*

We used three different data sets. The National Library of the Netherlands[2] (KB) provided us with the data of their entire digitized newspaper archive along with server logs from which we could extract the queries users issued via the library's webinterface, Delpher[3]. Additionally, we generated a set of simulated queries from the body text of the documents.

5.4.1.1 *Historic Newspaper Collection*

The newspaper data set made available to us ranges from 1618 to 1995[4] and consists of more than 102 million OCRed newspaper items. This comprises articles, advertisements, official notifications, and the captions of illustrations (see Table 4 for details).

As the archive spans almost four centuries, the newspaper pages vary strongly in visual appearance which is known to influence the performance of OCR software [28, 34]. The very high vocabulary size (see Table 4) indicates that the corpus might contain a high number of OCR errors, which can impact retrieval tasks [48]. The OCR quality has not been evaluated, therefore the actual error rates for the documents in this collection are unknown. An estimation of the quality by the OCR engine, however, is included in the metadata in the form of page confidence values.

From the KB data, we extracted and tokenized the body text of the newspaper items, which excludes the headings and meta data. We removed all stopwords and terms with fewer than three characters and kept only numbers with four digits, as these are likely to represent years and can therefore be used as query terms by users. The large majority of items (98%) are written in Dutch. As a stemmer for Dutch text was not available in the Indri[5] search engine, we created

[2] www.kb.nl

[3] www.delpher.nl

[4] A small number of documents from the 20th century is incorrectly dated to 2011 in the metadata.

[5] http://www.lemurproject.org/indri.php

Newspaper Collection		**1618 - 1995**
Total Size		102,718,528
Vocabulary Size (no stopwords, min. 2 characters)[7]		353,086,358
Articles	67%	69,237,655
Advertisements	29%	29,591,599
Official Notifications	2%	1,918,375
Captions	2%	1,970,899

User Logs	**March - July 2015**
Log Size (Number of HTTP Requests)	107,684,434
Number of Queries	4,169,379
Number of Unique Queries	1,051,676
Number of Unique IPs	162,536
Number of Document Views	3,328,090
Number of Unique Documents Viewed	2,732,139

Table 4: Data sets used based on the historic newspaper collection from KB.

a stemmed version during preprocessing. We used the default Snowball stemmer for Dutch[6].

5.4.1.2 *Real Queries*

Under conditions of strict confidentiality, the KB made user logs available to us that were collected between March 2015 and July 2015. In order to protect the privacy of the users, the logs had been anonymized by hashing the IP addresses, which enabled us to trace queries that originated from the same address without identifying the user. Delpher provides an advanced search interface, which allows users to apply boolean operators and facets based on metadata to their search queries. We processed the query logs the same way as the document collection by removing operators and stopwords, and stemming. For the latter, we again used the Snowball stemmer[8] (see Table 5 for details).

5.4.1.3 *Simulated Queries*

To be able to compare our results with those reported in [3], we created a simulated query set. For this, we counted the unique terms and bigrams in the preprocessed documents and extracted the top 2

[6] https://lucene.apache.org/core/4_0_0/analyzers-common/org/apache/lucene/analysis/nl/DutchAnalyzer.html

[7] Stopwords removed, length of term at least 2 characters

[8] https://pypi.python.org/pypi/PyStemmer/1.3.0

Query Set	Composition	Size	DQR
Sim. Queries	single term	2,000,000	
	bi-term	2,000,000	
	total	4,000,000	26
Real Queries	no op., no stopw., st.	957,239	107

Table 5: Sizes and document to query ratios (DQR) of the query sets.

million terms as single term queries and the top 2 million bigrams as bi-term queries (see Table 5). The frequencies for the two query sets ranged from more than 180 million to 5 for the single term queries and from more than 10 million to 20 for the bi-term queries. We did not filter for OCR errors, therefore frequently occurring misspellings can still be found in the simulated queries.

5.4.1.4 *Document Query Ratio*

Azzopardi and Vinay use query sets of which the size are comparable to the size of the corpus [3]. In this setting all documents have a fair chance to be retrieved. As we used real user queries in a very large corpus, it was not possible for us to influence the DQR. Consequently, the DQR values in our experiments vary greatly for the different query sets (see Table 5), as opposed to the study reported in [3], where the DQRs were 0.57 (AQUAINT) and 0.43 (.GOV). This issue has not been addressed in previous studies investigating retrievability of large document collections.

5.4.2 *Setup for Retrievability Analysis*

We compute retrievability scores based on three of the retrieval models used in [3]: TFIDF, Language Model using Bayes Smoothing with $\mu = 1{,}000$ (LM1000), and BM25.

Azzopardi et al. chose to report their results for $c = 100$ [3], therefore we also included these values for comparison. Additionally, we report on a cutoff value of $c = 10$ as it best represents the behavior of our users. The default number of results per page the Delpher interface shows is 10 and an analysis of the user logs showed that only a small fraction of users go beyond this. For the results based on the real queries, we also report on $c = 1000$, as this result set was of comparable size to the $c = 100$ results for the simulated queries.

We did not apply the query weights o_q as the by far largest fraction of real queries were issued only once.

5.4.3 *Setup for Retrievability Validation*

We validated the effectiveness of the retrievability scores for the newspaper collection. We examined whether documents with a low retrievability score are harder to retrieve than documents with a high score when a query is *specifically* designed to return the targeted document. We performed one experiment per query set. For simulated queries we follow Azzopardi and Vinay and use BM25 at $c = 100$ (stemmed, stopwords and operators removed) [3]. For the smaller set of real queries, we chose the same parameters but with a cutoff at $c = 1,000$, as the result set is more similar in size to the chosen set for the real queries. We included the documents with $r(d) = 0$, as they represent the group of documents that is supposedly the least accessible one.

For both result sets we generate queries from the target documents which contain OCR misspellings. In the experiment described in Subsection 5.4.2 the impact of these misspellings was lowered as a side effect of selecting the most frequent terms in the large corpus. Here, we select terms from a single document, which required us to apply filters as very rare misspellings being part of queries led to very high mean reciprocal rank (MRR) values, but are very unlikely to be used as queries by users. First, we created a dictionary of terms that occurred in more than one document, but in fewer than 25% of all documents and for which the document frequency was *not* equal to collection frequency. This allowed us to exclude extremely rare misspellings that occur in only one document or only once in multiple documents, and very generic terms. The dictionary we created from these terms was used to determine a list of suitable documents. We removed all words from the documents that did not appear in the dictionary or appeared only once in the document. All documents with fewer than four unique words were discarded for the experiment. By applying these filters, we removed 38,026,541 documents from the collection, leaving 64,691,987.

We divided the remaining documents into four bins, the same number of bins as used in [3]. For the division into bins, however, we diverged from the description given in [3] (where documents were ordered by retrievability and then divided into quartiles) because due to a different distribution of $r(d)$ values, the lower scores would have dominated the lower quartiles. Instead of binning on $r(d)$, we used a strategy that is inspired by the distribution of wealth measurements in economics. In our case, wealth is represented as the number of data points per $r(d)$. It is calculated for each $r(d)$ score by multiplying the score with its number of documents. Then we successively merged the $r(d)$ bins, until their summed up wealth reached the threshold of 25% of the total wealth. This led to four bins that roughly correspond to quartiles.

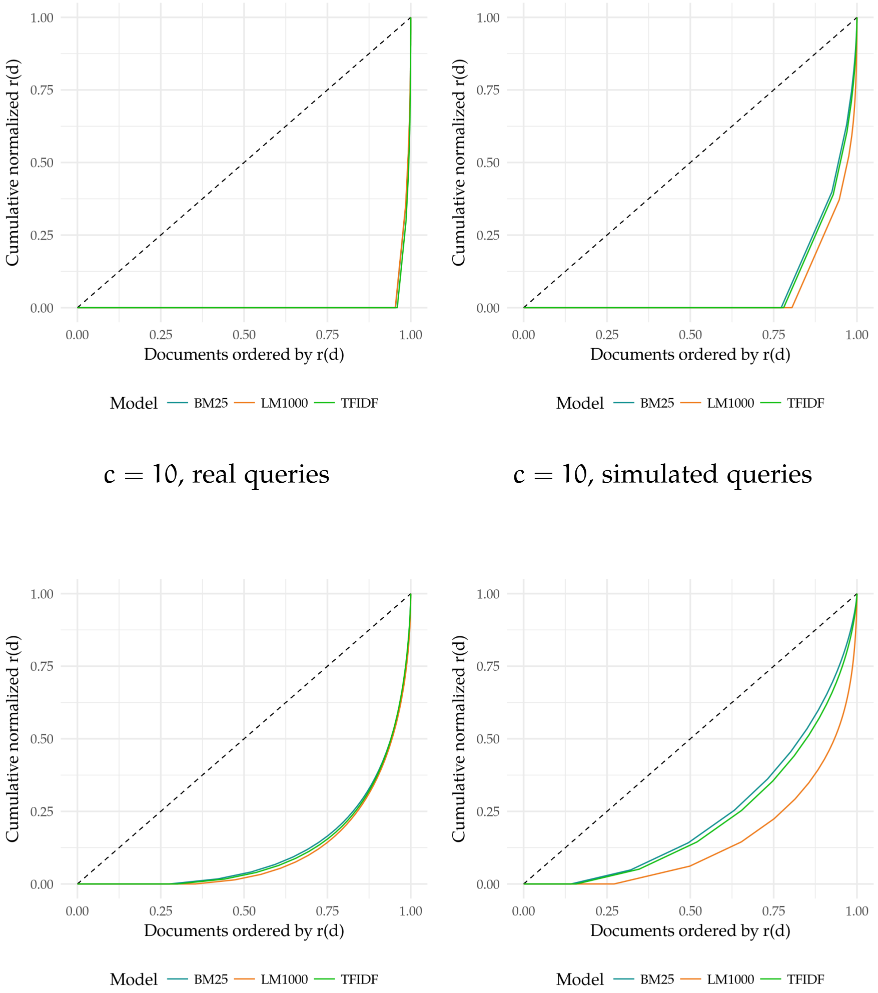
From each bin, we picked a random sample of 1,000 documents. We randomly selected 2 to 3 of the most frequent terms of each document to use as a query, as the mean number of terms issued by users was 2.32. The 1,000 queries we created this way were issued against the collection using the same IR model as before, BM25. We determined the rank of the target documents in the result lists and calculated the MRR for each bin as a measure of its retrieval performance.

5.5 RETRIEVABILITY ASSESSMENT

The high DQR value for our setup suggested that the fraction of documents with $r(d) = 0$ will be relatively high, especially for low cutoff values. Therefore, a large inequality in the retrievability scores was to be expected (*RQ1*). We describe the measured retrievability bias in different result sets and explore how to deal with the non-retrieved documents.

5.5.1 *Assessment of Retrievability Inequality*

We first look at the retrievability bias for both query sets at $c = 10$, which is the most realistic representation for the bias users of the archive are confronted with. The *Lorenz* curves depict a high inequality in the retrievability scores (see Fig. 11), with almost identical curves for the TFDIF, BM25 and LM1000 models. This is also reflected in the high *Gini* coefficients ranging from 0.97 to 0.98 for the real and from 0.85 to 0.89 for the simulated queries (see Table 6). The largest part of both curves consists of a flat line, which represents documents that were not retrieved. The setup with the highest *Gini* coefficient (TFIDF at $c = 10$, real queries) also contains the highest fraction of non-retrieved documents (96%).



c = 10, real queries

c = 10, simulated queries

c = 1000, real queries

c = 100, simulated queries

Figure 11: Lorenz curves visualize the inequality of retrievability scores for the *real queries* (left) and the simulated queries (right) at different cutoff values c.

By contrast, the *Lorenz* curves for the higher cutoff values depicted in Fig. 11 indicate a more balanced distribution of $r(d)$ values. The curves for all models show a smaller deviation from the equality diagonal and both the *Gini* coefficient, as well as the fractions of documents with $r(d) = 0$, are lower. This suggests that the large number of documents with $r(d) = 0$ has a strong influence on both the shape of the *Lorenz* curve and the *Gini* coefficient. As never-retrieved documents are inevitable in a realistic scenario such as ours, it is important to find a way to address this problem.

To further explore the influence of the $r(d) = 0$ values, we created a Union_c result set, that contains only documents retrieved by at least one of the models. While this removed most of the documents with $r(d) = 0$, a surprisingly large number of zeros still remained in the

	Model	C					
		10		100		1000	
		G	Z	G	Z	G	Z
Real	TFIDF	0.98	96%	0.91	78%	0.77	30%
	BM25	0.97	95%	0.89	75%	0.76	28%
	LM1000	0.97	95%	0.90	77%	0.78	35%
Sim.	TFIDF	0.86	78%	0.55	16%	-	-
	BM25	0.85	77%	0.52	14%	-	-
	LM1000	0.89	80%	0.71	27%	-	-

Table 6: Gini coefficients (G) and fractions of documents with $r(d) = 0$ (Z) for the complete data set.

	Model	C					
		10		100		1000	
		G	Z	G	Z	G	Z
Real	TFIDF	0.71	47%	0.74	36%	0.71	13%
	BM25	0.64	40%	0.69	29%	0.70	10%
	LM1000	0.63	39%	0.71	33%	0.73	20%
Sim.	TFIDF	0.52	26%	0.50	5%	-	-
	BM25	0.48	24%	0.46	3%	-	-
	LM1000	0.63	34%	0.67	18%	-	-

Table 7: Gini coefficients (G) and fractions of documents with $r(d) = 0$ (Z) for the Union_c data set.

subset. The number of zero-scoring documents for TFIDF at $c = 10$, for example, was only reduced from 96% to 47%. Even with never-retrieved documents removed, the inequality in the Union_c data set remains quite high for $c = 10$ with *Gini* coefficients ranging from 0.48 (BM25) to 0.63 (LM1000) (see Table 7). The remaining zero-scoring documents are a first indication that, while their Lorenz curves and Gini coefficients are similar, the models actually retrieve very different sets of documents.

We finally removed *all* documents with $r(d) = 0$ to measure the inequality among the retrieved documents. This caused the *Gini* coefficients to drop to values between 0.40 and 0.46 (real queries at $c = 10$). This again shows the large influence of a high fraction of zeros on the overall *Gini* score.

The similarity of the different models' *Lorenz* curves indicates a similar degree of bias in the $r(d)$ scores, but it does not allow further insights into the type of bias, i.e. whether it originates from the high

	Model	C		
		10	**100**	**1000**
		G	G	G
Real	TFIDF	0.46	0.59	0.67
	BM25	0.40	0.56	0.67
	LM1000	0.40	0.56	0.66
Sim.	TFIDF	0.35	0.47	-
	BM25	0.32	0.44	-
	LM1000	0.43	0.60	-

Table 8: Gini coefficients (G) for the *Non Zero* data set from which all documents with $r(d) = 0$ were removed.

Query Set		**Bin**			
		1st	**2nd**	**3rd**	**4th**
Simulated	MRR	0.19	0.28	0.36	0.45
	D	0.20	0.12	0.08	-
Real	MRR	0.17	0.26	0.34	0.38
	D	0.20	0.11	0.05*	-

Table 9: MRR values are higher for items in the quartiles with higher $r(d)$ scores. An * indicates that the Kolmogorov-Smirnov test did not confirm a significant difference ($p > 0.05$) between the indicated bin and the fourth bin. D is the maximum vertical deviation as computed by the KS test.

DQR, from the users' interest, or from a technological bias towards particular document features.

Fig. 12 shows the frequencies of $r(d)$ values (log scale), with a long tail distribution for both query sets. The maximum $r(d)$ value for the real queries is $r(d) = 4319$, while for the simulated queries this is much smaller (max $r(d) = 807$). This shows one possible cause for the bias towards higher fractions of documents with $r(d) = 0$ within the real queries: they tend to retrieve the same documents more often, leading to a smaller number of unique retrieved documents. This indicates that the query sets themselves may be biased, the real query set towards the users' interest and the simulated query set towards the language use in the document collection.

5.5.2 *Validation of the Retrievability Scores*

We validated our results using a known-item-search experiment (see Subsection 5.4.3) to confirm that documents with low $r(d)$ scores are indeed harder to find.

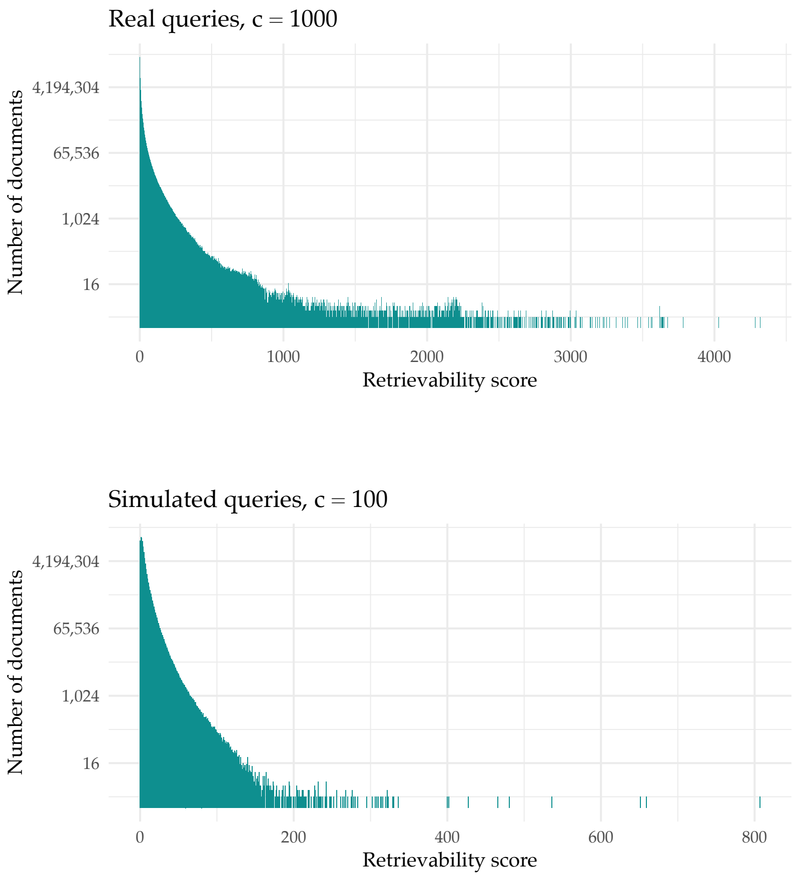


Figure 12: Log scale representation of the distribution of retrievability scores $r(d)$ for BM25 based on the complete KB dataset.

The results show that the MRR values indeed increase for the bins containing the documents with the higher $r(d)$ values (see Table 9). With one exception the differences in the ranks between the bins proved to be significant in a Kolmogorov-Smirnov test. This suggests that documents in the first bin are significantly more difficult to retrieve than documents in the fourth bin.

This pattern is similar to the findings in [3] and confirms that a document's retrievability score is a good indication of how hard it is to retrieve the document by a user.

5.5.3 *Document Features' Influence on Retrievability*

To better understand the inequality in our document collection, we explored whether we can identify subsets within the archive that are particularly susceptible or resistant towards retrievability bias (*RQ2*).

• The *time of publishing* of the newspapers in our collection spans a period of nearly 400 years. Newspapers that belong to the early issues are very different from today's newspapers in terms of content as well as visual appearance. This affects the performance of OCR soft-

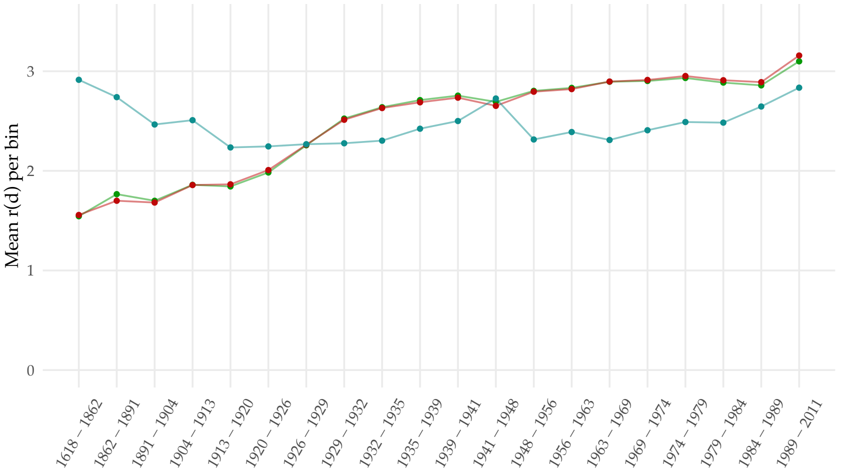
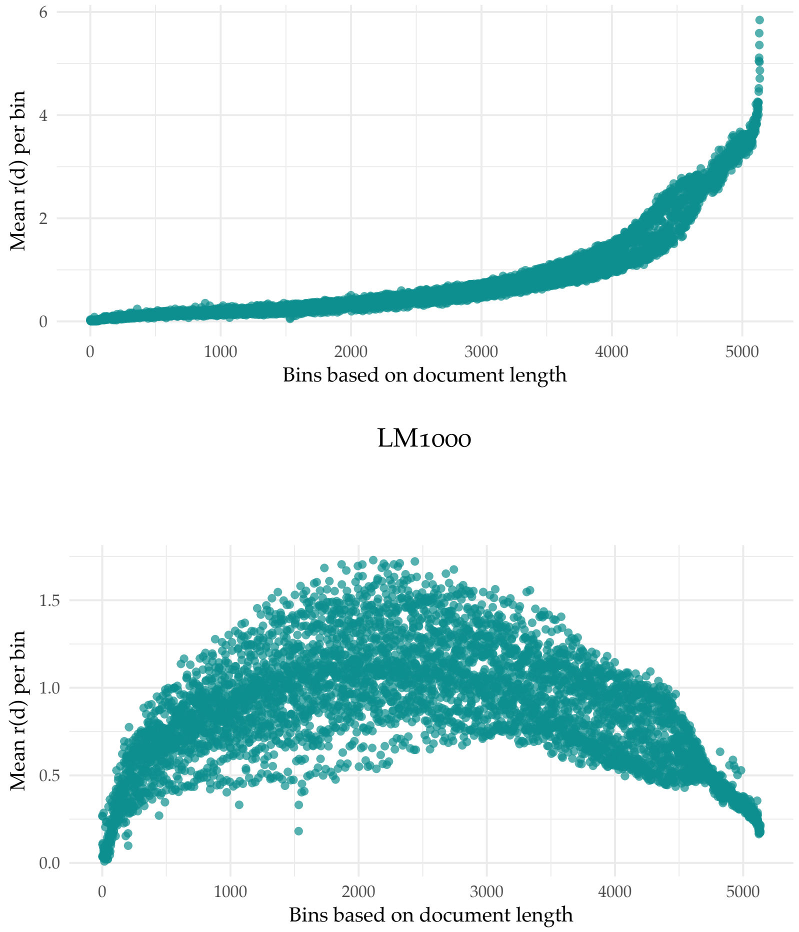


Figure 13: The mean $r(d)$ scores (20 equally sized bins, based on $\mathtt{Union}_c$ data, real queries for $c = 100$) for BM25 (**green**) and TFIDF (**red**) are nearly identical and double in value over time. LM1000 (**blue**) does not show this upward trend.

ware, which results in high OCR error rates in older newspapers. We are therefore interested if this is reflected in the $r(d)$ values. For the analysis, we ordered the newspaper items in the $\mathtt{Union}_c$ set by publishing date, divided them into 20 equally sized bins (1,7M items per bin) and calculated the mean retrievability score for each bin. Note that due to the much lower number of documents in the early periods of the archive, the 20th century occupies by far the most bins. The results for BM25 and TFIDF show a very small upward trend for later documents (see Fig. 13). This trend is, however, not visible for LM1000 and could also not be confirmed in an analysis of the raw data.

- The *document length* in our collection varies from 33 to 381,563 words with a mean length of 362 words. As Azzopardi and Vinay found that longer documents in their collections were more retrievable than short ones [3], we were interested in finding out whether the same holds for our collection. We sorted all items in the collection according to their length and divided them into bins of 20,000 documents, leading to 5,135 bins in total. For each bin, we calculated the *mean* $r(d)$. While the pattern we obtained for LM1000 shows an upwards trend for longer documents and thereby confirms this assumption (see Fig. 14), the results for BM25 and TFIDF[9] indicate that documents of medium length are most retrievable, whereas documents at both extremes are less retrievable. We can see a bias in both patterns, while LM1000 clearly favors longer documents, BM25 and TFIDF overcompensate for long documents, while they seem to fail to compensate for short ones.

[9] The pattern for TFIDF looks very similar to BM25, therefore we did not include the plot.



LM1000

BM25

Figure 14: Document length vs. r(d) for c=100, bins of 20,000 documents

• The library's OCR engine assigns *confidence scores* to each page (*PC*), word (*WC*) and character (*CC*) in the corpus. This is intended to give an indication of the quality of the OCR processing. From our contacts with the KB we learned that, during the post-processing, the scores were adapted based on the occurrence of a term in a Dutch word list. A formal evaluation of error rates in the KB data has not yet been performed, therefore we do not know to what extent these PC values are realistic. We divided the collection into bins of 20,000 documents based on their PC value and plotted the mean $r(d)$ score for each bin. The resulting plot shows an upward trend for increasing confidence values (see Fig. 15). Documents with an PC score very close to 1.0, however, seem to be less retrievable. A closer look revealed that these documents often contain only very short texts, which makes them harder to find.

• *Newspaper titles* do not only vary with respect to their political orientation, but also concerning the content they provide to their readers. The mean number of articles per newspaper title in the archive

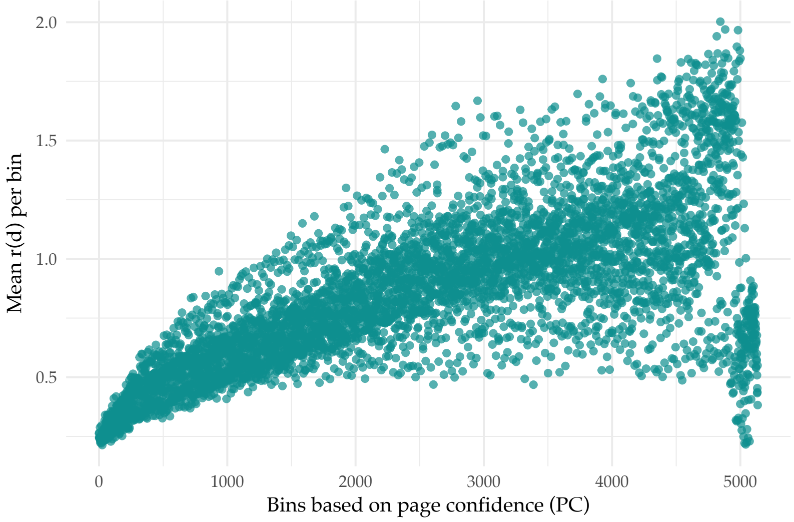


Figure 15: Mean $r(d)$ scores versus page confidence (PC) scores for bins of 20,000 documents

is 82,638, with a median of 127 and a range from one to 16,348,557 documents. We list differences in retrievability scores of the 10 most prevalent newspaper titles in our collection (see Table 10). While the differences seem small, three regional titles have a higher mean $r(d)$ than the seven national titles. Again, this may be caused by a bias in user preferences.

• We computed the mean $r(d)$ scores of the four *types of documents* in the archive for the two query sets. The means resulting from simulated queries show relatively small differences (see Table 11), whereas the mean scores obtained through real queries show a much higher score for official notifications. This again shows the large difference in the document sets retrieved by the two query sets.

From these results we can conclude that the large fraction of never retrieved documents is inevitable in realistic setups and needs to be addressed when assessing retrievability bias. We found evidence for a relation between low OCR confidence values, and short document length and a lower retrievability of documents. When comparing the degree of bias among the three IR models, we found LM1000 to show a greater bias for simulated queries. A comparison of the distributions of retrievability scores indicated a higher variety in $r(d)$ scores for real queries, and a bias towards official notifications for real queries which is not present in the simulated queries.

5.6 REPRESENTATIVENESS OF THE RETRIEVABILITY EXPERIMENT

We explore to what extent the different types of bias we see in the retrievability experiments are representative for bias in the documents actually viewed on the library's website (*RQ3*). For this purpose we

Top 10 Newspaper Titles	**Mean r(d)**
Rotterdamsch nieuwsblad*	0.05
Algemeen Handelsblad	0.06
De Telegraaf	0.06
Het Vaderland: staat- en letterkundig nieuwsblad	0.07
Leeuwarder courant*	0.07
De Tijd: godsdienstig-staatkundig dagblad	0.08
Het vrije volk: democratisch-socialistisch dagblad	0.10
Limburgsch dagblad*	0.12
Nieuwsblad van het Noorden*	0.14
Leeuwarder courant: hoofdblad van Friesland*	0.15

Table 10: Mean $r(d)$ values for the most prevalent newspaper titles for BM25 at $c = 10$, real queries. An * indicates a regional newpaper title.

	Real	**Simulated**
Article	0.90	3.89
Advertisement	0.51	3.32
Official notification	4.80	3.22
Caption	0.84	3.06

Table 11: Mean $r(d)$ for different types of articles (BM25, c=100).

compare the reported results with click data from the user logs, and revisit the use of simulated queries versus real queries.

5.6.1 *Retrieved versus Viewed*

The *Lorenz* curve in Fig. 16 (left) shows the inequality in the corpus with respect to the number of views. With only 2.7M out of 102M documents that are viewed, the fraction of documents that is never viewed by users is even larger than the fraction of never retrieved documents in our $c = 10$ experiments. This confirms that a large fraction of not-accessed documents is not only an artifact in our retrievability experiments caused by a relatively small query set: it also reflects the fact that in most large digital libraries, the number of views in any reasonable observation period will be small in comparison to the number of documents in the collection. Since the retrievability and the viewing scores are dominated by the large number of never accessed documents, neither the *Lorenz* curves nor the *Gini* coefficients are very informative measures of bias.

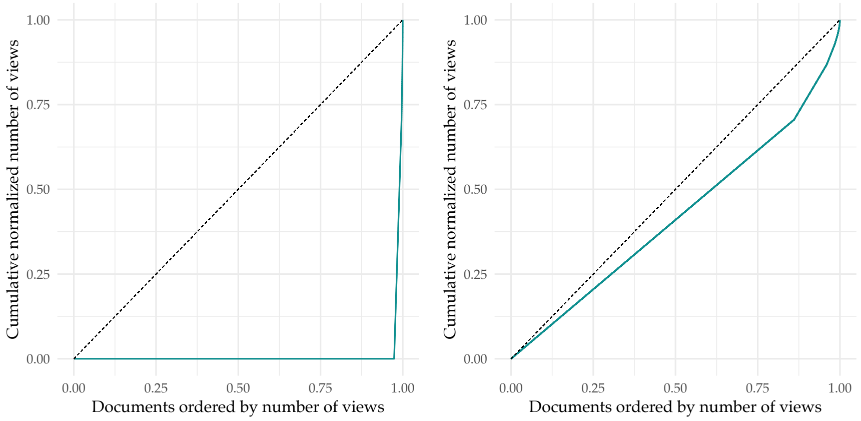


Figure 16: The *Lorenz* curve of viewed documents shows that only a small fraction of the collection was accessed (left) (Gini = 0.98). When non-viewed documents are removed, the inequality largely disappears, because most documents that are viewed, are viewed only once (right) (Gini = 0.16).

DISTRIBUTION OF R(D) SCORES AND VIEW FREQUENCIES For documents that are never accessed, it is hard to classify whether this is indeed the result of the small number of user views, the result of bias in user interest, or the result of technical bias in the retrieval system. Focussing only on the accessed documents would ignore the latter type of bias. However, even if we discard the non-accessed documents, the *Lorenz* curve of only the 2.7M viewed documents (see Fig. 16 (right)) is not much more informative. Here we see the opposite: extremely low inequality, which results from the fact that the large majority (86%) of the viewed documents is only viewed once.

A log scale bar chart of the (non-zero) viewing frequencies (as in Fig. 17) provides more insight than the Lorenz curves. While the viewed documents dataset is smaller, the shape of the view frequency distribution is very similar to that of the retrievability score of the real queries in Fig. 12, and even more similar than the scores of the simulated queries. Again, this suggests that simulated queries do not necessarily represent real user behavior.

VIEWED BUT NOT RETRIEVED To explore if the unique documents retrieved in our experiment using real queries are representative for the 2.7M unique documents actually viewed by the users, we investigated the overlap between the two. Given that most users only look at the first page with 10 results, we looked at the overlap for BM25 at $c = 10$, where we have 4.7M unique documents that are retrieved at least once. Less than 0.6M of these were also viewed, leaving 2.1M documents that were viewed but not retrieved in our top 10.

To find out what the reasons for the small overlap were, we performed a preliminary manual assessment of the top viewed documents that had *not* been retrieved by BM25 at $c = 10$. The most

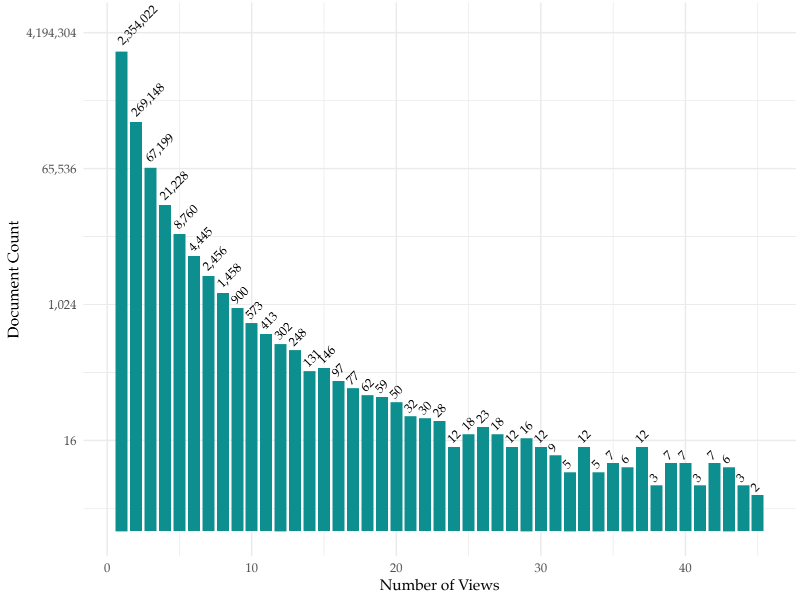


Figure 17: Log scale representation of the frequencies of document views based on the query logs (cut off at 45 views for better readability).

viewed document in this subcollection is a very short article describing an incident, in which a cow accidentally "caught" a rabbit ("Men kan niet weten hoe een koe een haas vangt."[10]). From the user logs, we learned that this was caused by deep linking: the article was accessed in response to a hyperlink in a newsletter, not in response to a direct search action. The second most viewed article[11] was retrieved in response to a direct search action, but by making use of the search interface's time facet which allows users to narrow down the search results to specific time periods.

Other often viewed documents were retrieved in our experiment, but with a ranking slightly above the $c = 10$ cut-off. That this is not just anecdotical but a larger issue is confirmed by the much larger overlap for the higher cut-off values. $c = 100$ retrieved 1.5M viewed documents, and $c = 1000$ retrieved more than 2.4M of the 2.7M viewed documents.

These results can be interpreted in two ways. First, small differences in the ranking scheme can have quite dramatic effects due to the all-or-nothing scoring function. This suggests that a smoother cost function based on the ranking might be worthwhile. Another potential interpretation is that the experimental setup needs to reflect the real search engine better, and also take the faceted search parameters, pagination, search operators and other more complex search settings into account.

[10] http://resolver.kb.nl/resolve?urn=ddd:110540686:mpeg21:a0015

[11] http://resolver.kb.nl/resolve?urn=ddd:000011882:mpeg21:a0004

Model	C	Real	Simulated
BM25	10	56.19 %	91.19 %
	100	7.94 %	73.51 %
TFIDF	10	53.48 %	91.44 %
	100	8.19 %	75.53 %
LM1000	10	54.74 %	89.24 %
	100	8.75 %	70.62 %

Table 12: The percentages of results from query logs and simulated queries that are *not* found by the other query set show that for small values of c the results vary strongly.

5.6.2 *Real versus Simulated Queries*

Since real query logs for large document collections are hard to obtain, most retrievability experiments reported in the literature use simulated queries, typically based on sampling the most popular n-grams. However, our results seem to suggest such queries might not be representative of real user queries.

QUALITATIVE COMPARISON OF OFTEN RETRIEVED DOCUMENTS To get a better intuition of the type of documents retrieved, we manually explored the top 10 articles for both query sets (for BM25 at $c = 10$). The top results for the real queries completely consisted of articles that contained lists of names [12]. This is because the logs from the KB contain a large number of queries with names and locations.

We compared this finding to the top results set retrieved by the simulated queries. Here, the top scoring documents either contain a very repetitive text pattern (e.g. repetitive poems[13]), or the documents themselves are near duplicates of other documents (e.g. chain letters, advertisements with identical text, or other documents that were published multiple times[14]). This finding might indicate another drawback of the way the simulated queries are traditionally sampled: frequently occurring terms are more likely to be included in the query set.

OVERLAP IN RETRIEVED DOCUMENTS The variety of $r(d)$ values is much larger on the real queries, indicating that the two query sets might retrieve very different documents (see Fig. 12). We explored the overlap of documents that were retrieved by the real queries and the (larger) set of simulated queries. For all three models, at $c =$

[12] see for example http://resolver.kb.nl/resolve?urn=ddd:010179873:mpeg21:a0001

[13] http://resolver.kb.nl/resolve?urn=ddd:010210514:mpeg21:a0150

[14] see http://resolver.kb.nl/resolve?urn=ddd:010691557:mpeg21:a0069

10, more than half of the documents retrieved by the real queries are *not* found in the results from the simulated queries (see Table 12). This again suggests that we should improve the construction of our simulated query set to better represent real queries. Note that the fraction of documents that are retrieved by both approaches is considerably higher for $c = 100$, where less than 9% of the documents in the result set of the real queries are not found in the results of the simulated queries.

DIFFERENCES BETWEEN QUERY SETS In addition to the difference between the documents retrieved by both types of queries we also looked at the characteristics of the query sets themselves. The two query sets differ not only in size (as indicated in Table 5). The mean length of the real queries is 2.32 and all queries use a total of 253,637 unique terms. As we followed Azzopardi and Vinay and only used single and bi-term queries for the simulated query set [3], its mean query length is much smaller (1.5). The number of unique terms (2,028,617) is, however, much higher. This suggests that even by sampling only the most popular (bi)terms, we would over estimate the vocabulary used by users to formulate their queries.

We manually assessed the number of terms that refer to named entities in the 100 most frequent terms in both query sets. For the simulated queries, we found only 5 mentions of persons or locations, as opposed to 56 named entities in the real queries, confirming again the large differences in this aspect between the two sets.

Table 11 shows a higher retrievability of *official notifications* for the real queries. We compare this finding with the fractions of viewed documents for each type. While these fractions are very low for articles (only 2.61% viewed), advertisements (2.07%) and captions (4.01%), a much higher fraction of the official notifications was viewed (40.10%). This again shows that retrievability measured by real queries are more representative than synthesized queries.

5.6.3 *Representativeness of Parameters used*

Apart from queries and document features, retrievability can also be influenced by the parameters used in the retrieval setup, namely the inclusion or exclusion of stopwords and operators, and stemming. While we followed the parameter settings used by Azzopardi and Vinay so far (PS1) [3], we compare the results obtained with the real queries using two alternative parameter settings (PS2 and PS3):

PS1: operators removed, stopwords removed, stemmed (used by [3])

PS2: operators removed, stopwords kept, unstemmed

	PS1	Shared	PS2	C
BM25	1,939,710	2,758,599	1,971,087	10
TFIDF	1,667,374	2,485,412	1,689,125	
LM1000	2,141,563	2,620,988	1,317,420	
BM25	7,436,058	17,923,267	7,232,087	100
TFIDF	6,672,656	16,385,354	6,381,519	
LM1000	7,384,854	16,711,774	4,804,696	

Table 13: Numbers of documents retrieved only by one parameter set (*PS*) and number of documents retrieved by both sets.

PS3: operators, stopwords removed, stemmed[15]

Parameter sets *PS2* and *PS3* resulted in nearly identical *Gini* coefficients to those we reported in Table 6 for *PS1*. This suggests that the removal of stopwords, or the use of stemming and operators, has no influence on the extent of inequality in the document retrieval. The question remains, however, whether and how the underlying retrieved document sets differ and how this relates to the documents the users found sufficiently relevant to view.

DIFFERENCES IN RETRIEVED DOCUMENT SETS We compared the retrieved document sets from *PS1* and *PS2* for their overlap and found that while the majority of documents retrieved in one setting is also retrieved in the other, still a large fraction is only found in one setting (see Table 13). Note that even though this difference is not reflected in the *Gini* coefficient, *Lorenz* curves or r(d) distribution plots, it is a form of retrieval bias that may have a huge impact on the user's task.

Again, as c increases, the fraction of shared documents between the parameter sets increases as well. To judge which of the document sets is the more favorable for our use case, we compare the overlaps of the result sets with documents that were viewed by users (e.g using views as a proxy for relevance judgements).

The combinations of IR model and parameter set vary strongly with respect to their ability to retrieve the viewed documents (see Table 14). BM25 and TFIDF achieved better results with *PS2* than with *PS1*, but both are outperformed by LM1000 in all settings. The best result is achieved by using LM1000 with *PS3* with 29% of the viewed documents retrieved, so that in this case, the retrieval model with the most bias also performs better. This is in contrast to results reported by [3], where better performing models typically also show less bias.

[15] As restrictions of the Indri toolkit (`http://www.lemurproject.org/`) did not allow us to run this set of parameters for BM25 and TFIDF, these results are available only for LM1000.

	PS1	PS2	PS3	C
BM25	504,022	598,969	-	10
TFIDF	435,413	527,461	-	
LM1000	742,548	706,425	781,908	
BM25	1,422,231	1,511,973	-	100
TFIDF	1,323,284	1,423,589	-	
LM1000	1,788,719	1,741,290	1,840,285	

Table 14: Viewed documents that were retrieved by each model for the different parameter sets (*PS*) for a total of 2,732,139 viewed documents.

5.7 conclusions and outlook

Measuring the variation in the retrievability of documents in a collection complements standard IR evaluations that focus on efficiency and effectivity. No previous study has investigated how well retrievability studies represent the search behavior of real users and how they could be applied to a large collection of digitized documents that contain an unknown number of misspellings due to OCR processing. Our focus was on the exploration of the applicability of retrievability studies to a large digitized document collection and an evaluation of the representativeness of simulated queries for real users' search behavior.

While *Gini* coefficients and *Lorenz* curves allowed us to detect and quantify a retrievability bias in the document collection for three standard IR models, they were not sufficiently expressive to help us understand the source of it. We looked at the differences among the documents retrieved, and showed that large differences are common even for models with similar *Gini* coefficients and *Lorenz* curves.

In addition, we explored several influencing factors: the document to query ratio, document features, characteristics of query sets and the use of different parameter sets.

When comparing the characteristics of simulated queries to those of real users' queries we found substantial differences with respect to composition of the query sets, number of (unique) terms used, and use of named entities. Real users' queries contained a much higher fraction of named entities than we found in the simulated query set.

Finally, we compared how effectively combinations of specific parameter settings could retrieve the documents users viewed. Based on the results from this study, the setup that best covers the users' information needs is the combination of real queries with operators on LM1000. Note that according to the inequality assessment, the least biased model is BM25. This shows, that switch to a model with a lower retrievability bias might hurt the system's performance in terms of retrieving the most relevant documents.

Simulated queries that are representative for the search behavior of real users are a key ingredient for a realistic assessment of retrievability bias. Future work should therefore focus on how the generation of simulated queries can be adapted in a way that they better represent the type of queries real users issue on a specific collection.

6 IMPACT OF CROWDSOURCING OCR IMPROVEMENTS ON RETRIEVABILITY BIAS

Digitized document collections often suffer from OCR errors that may impact a document's readability and retrievability. We studied the effects of correcting OCR errors on the retrievability of documents in a historic newspaper corpus of a digital library. We computed retrievability scores for the uncorrected documents using queries from the library's search log, and found that the document OCR character error rate and retrievability score are strongly correlated. We computed retrievability scores for manually corrected versions of the same documents, and report on differences in their total sum, the overall retrievability bias, and the distribution of these changes over the documents, queries and query terms. For large collections, often only a fraction of the corpus is manually corrected. Using a mixed corpus, we assess how this mix affects the retrievability of the corrected and uncorrected documents. The correction of OCR errors increased the number of documents retrieved in all conditions. The increase contributed to a less biased retrieval, even when taking the potential lower ranking of uncorrected documents into account.

6.1 INTRODUCTION

Digitized collections are the foundation for services and research tasks that would be much more difficult (if not impossible) to perform on collections of physical items. Examples of such tasks are full-text search and quantification of changes in textual features over long time periods. Most of these services, however, rely on the use of retrieval systems.

How well these systems perform has been investigated with regard to many different aspects, such as precision and recall, and based on many different types of corpora, such as community-created TREC collections, digital libraries or Web archives. The *retrievability measure* as introduced by Azzopardi and Vinay extends these evaluation measures by means to detect and assess bias when retrieving documents [3].

In a previous study, we used retrievability to investigate whether a retrievability bias influences access to a digitized collection of historic newspapers and to measure the extent of this bias [49]. We found a relation between document features, such as document length, and retrievability. In this study, we focus on the effects of OCR quality on retrievability and how a (partial) manual correction of the OCR errors impacts the accessibility of document. We investigate the following research questions.

- *RQ1: What is the relation between a document's OCR character error rate and its retrievability score?* By relating the retrievability scores of documents with the character error rates of their content, we investigate how the quality of OCR processing impacts a document's retrievability.
- *RQ2: How does the correction of OCR errors impact the retrievability bias of the corrected documents (*direct impact*)?* Assuming that the complete set of documents has been corrected, we investigate if the correction makes retrieval more or less biased in terms of retrievability, and how differences in retrievability scores are distributed over documents, queries and query terms.
- *RQ3: How does the correction of a fraction of error-prone documents influence the retrievability of non-corrected ones (*indirect impact*)?* Typically, only small fractions of a collection are corrected. We investigate how this affects the other documents in the collection by comparing the retrievability scores in a mixed collection where 50% of the collection has been corrected with those of an uncorrected only collection.

6.2 APPROACH

To investigate whether and how errors in OCRed documents influence their retrievability, we performed a series of experiments that make use of the concept of *retrievability* as introduced by Azzopardi and Vinay [3]. For this, we used different subsets of a digitized newspaper collection and search queries that were collected from users of the online access portal of the archive.

The National Library of the Netherlands (KB)[1] made a ground truth data set available that contains the manually corrected versions of 100 newspaper issues. By comparing these documents with their original versions, we were able to assess the number of incorrect characters and compute the character error rates (*CER*) for each document. This allowed us to investigate a relation between the documents' quality and their retrievability scores (*RQ1*).

The manual correction of OCR errors *directly* impacts the retrievability of these documents. We investigated this effect with two retrievability experiments based on a small document collection and two versions of query sets that were originally collected from users of the digital archive. By comparing the *r(d)* scores, we investigate which documents and queries gained or lost *r(d)* scores through the correction and how this influences the total number of retrieved documents (*wealth*) and retrieval bias (*inequality*) of the results (*RQ2*).

Since correction of OCR errors is often performed manually, it is a costly process. As a consequence only relatively small fractions of

[1] www.kb.nl

a collection are corrected. The same document may score lower in a corpus consisting of only highly findable documents than it would as part of a collection of documents that are difficult to find. Therefore, we explored how the correction of only a part of the collection *indirectly* impacts the retrievability of documents that remain uncorrected (*RQ3*).

6.3 RELATED WORK

6.3.1 *OCR Quality and Retrieval*

In 2015, we conducted a series of interviews with digital humanities scholars on their use of digital archives for their research. All agreed that the (OCR) quality of digitized documents makes digital libraries unsuited for "distant reading" and other computational approaches [48]. Several studies investigated the applicability of crowdsourcing tasks to transcribe documents [28, 32] or the use of a tool that combines the search in a digitized corpus with correction of OCR errors [37]. While the results from these studies can help improve data quality more efficiently, it remains unclear how this correction affects a scholar's research.

Mittendorf and Schäuble investigated how robust IR systems are toward OCR errors in digitized documents [36]. They found that longer documents describing a single topic redundantly have a better chance of retrieval than documents that are either short or discuss different topics.

Taghva et al. investigated the performance of the vector space model on OCRed documents [44]. They found that for their full text collection neither average precision, nor recall of the documents is affected by OCR errors. 674 documents were used in a OCR processed version and a manually corrected ground truth version. The character error rate was estimated to be around $10 - 20\%$ and the average length of the documents is reported to be around 40 pages. This confirms the findings of Mittendorf et al. that the effect of OCR errors on long documents can be expected to be very low. Since our corpus is characterized by relatively short documents with a high estimated error rate, we expect a higher impact of OCR errors than in the studies of Mittendorf et al. and Taghva et al.

Ohta et al. studied whether the effect of OCR errors on document retrieval can be compensated by generating additional search terms based on a character confusion matrix [39]. They based their study on two collections of documents obtained from the *Elsevier Electronic Subscriptions* service and published between 1995 and 1996. The document collection in this case can therefore be expected to be very homogenous in terms of layout, fonts, document length, quality of the physical copy and as a consequence cause little variation in error rates and error types. In our case, documents vary strongly in all of these aspects and therefore errors are less systematic as in the documents of Ohta et al. A statistical approach would be difficult, as it could only be applied to subsets of very similar documents.

Chiron et al. investigated for the AmeliOCR project how OCR errors are distributed in a large and diverse digitized corpus [15]. They found that about 15% of the misspelled terms represent named entities and that even 80% of the top 500 queries contain at least a mention of one. In a manual inspection of the 100 most frequent terms in the query set we used for [49], we found that 56 were named entities. The frequency of named entities in the document collection, however, can be very low and they may not even be found in common dictionaries. This makes them particularly susceptible to OCR errors. This combination, i.e. terms that occur very frequently in queries, but very infrequently in documents, is the reason that OCR errors in these terms can have a disproportional effect on the retrieval results [15].

6.3.2 *Retrievability Assessment*

The foundation for the assessment of retrievability bias in document collections is the work by Azzopardi and Vinay [3]. They introduced *retrievability* as an extension to traditional IR measures, one that does not require the availability of relevance judgments. It considers the number of results that a user is willing to examine (c). If the rank k_{dq} of a document d is retrieved within the cutoff value c, the utility/cost function f returns a score of 1, otherwise 0.

$$r(d) = \sum_{q \in Q} o_q \cdot f(k_{dq}, c)$$

o_q allows different weighting of queries according to their importance. We use $o_q = 1$ for all queries. To measure a potential bias among *r(d)* scores, [3] suggested to use the Gini coefficient, which was introduced to measure inequalities in societies [21]. Wilkie and Azzopardi later compared it to other inequality measures and confirmed its aptitude for retrievability analyses [56].

Follow-up studies confirmed the applicability of the retrievability measure to assess bias in retrieval models [52] and its relatedness to retrieval effectiveness [8, 9]. Several studies found that Okapi BM25 induces the least bias and can therefore be considered to be the fairest retrieval model [49, 52, 55]. While Azzopardi and Vinay and most subsequent retrievability studies (e.g. [5, 9, 41, 56]) made use of *simulated user queries* [3], we follow the line of our previous study and use queries collected from real users of the digital library [49]. In [49], we investigated the applicability of the retrievability metric on a digitized newspaper collection and questioned the representativeness of simulated queries for the search behavior of real users. Our findings revealed significant differences in number of query terms used and the frequency of named entity queries.

The current study extends the findings of [49] in several aspects. Our first study was based on the complete archive which comprises more than 102 million documents. The relatively high *document - query ratio* (DQR) had a large impact on the inequality in the *r(d)* scores because a large fraction of the documents was never retrieved. By focusing on a small subset of the newspaper collection in this study, we prevented a high DQR rate, and analyze an inversed scenario where the number of queries exceeds by far the number of documents. Finally, the availability of a ground truth data set enables us to investigate retrieval results on a corrected document collection, a collection containing errors, and a mixed collection.

6.4 EXPERIMENTAL SETUP

The setup we used for our experiments follows the setup used in [3, 49], modifications are explicitly described in this section.

6.4.1 *Document Collections*

We use different subsets of the historic newspaper archive, a manually corrected ground truth subset and queries collected from the online users of the archive.

OCR Ground Truth Corpus ($822_{\mathtt{GTcor}}$) For a small subset of the newspaper collection of the National Library of the Netherlands, the OCR text has been manually corrected. This subset covers 100 newspaper issues published between 14-06-1618 and 26-10-1624 ($17^{th}\mathtt{century}$ subset) and between 04-10-1940 and 29-09-1944 (*WWII* subset). The $17^{th}\mathtt{century}$ sub-collection constitutes the part of the archive with the oldest documents. It is prone to OCR errors as the decay of the physical material, the layout and the (gothic) fonts make character recognition very difficult. The *WWII* collection includes illegal newspapers, printed secretly, often in non-professional settings. Some of these articles therefore have a lower OCR quality than pro-German papers of the same period with better print quality. Combined, they include a total of 822 newspaper items. Note that this corpus is very small compared to 100M item corpus used in our first study [49].

OCRed Corpus containing Errors ($822_{\mathtt{GTerr}}$) We used the uncorrected versions of the articles in $822_{\mathtt{GTcor}}$ to build the $822_{\mathtt{GTerr}}$ corpus.

Mixed Documents Corpus ($1644_{\mathtt{mix}}$ and $1644_{\mathtt{err}}$) We extended $822_{\mathtt{GTcor}}$ and $822_{\mathtt{GTerr}}$ with an equal number of articles that originate from the same newspaper titles as in the $822_{\mathtt{GTerr}}$ collection. We selected the 503 earliest articles from the KB collection and a random sample of 319 articles from the WWII period ($822_{\mathtt{mixin}}$). These documents added to 822_{GTcor} yields the 1644_{mix} corpus, and added to 822_{GTerr} yields the 1644_{err} corpus.

6.4.2 *Query Set*

The queries we used were collected from the users of the library's Web interface (Delpher.nl) between March and July 2015. The data set comprises a total number of $1,008,915$ queries from $162,536$ unique users with an average length of three terms. We removed stopwords[2] and terms shorter than three characters from the queries. The final, deduplicated, query set comprises $859,716$ *multi-term queries*. Additionally, we created a *single-term query set* by extracting all $259,091$ unique terms.

6.4.3 *OCR Quality Assessment*

We measured the OCR quality of 822_{GTerr} set using the OCRevalUAtion tool[3] developed by the IMPACT project. It allowed us to compute the character error rates (*CER*) for each article in 822_{GTerr}.

6.4.4 *Setup for Retrievability Analysis*

We investigated whether and how OCR quality impacts retrievability by comparing how retrievability scores ($r(d)$) differ between documents containing errors and their corrected versions. To compute the $r(d)$ score for each document, we issued all queries against the document collections using the Indri search engine[4] and BM25 as retrieval model. For each document we calculated how often it was retrieved in the top c results (for cut-off values of $c = 1, 10$ and 100) and how often it was retrieved at all ($c = \infty$).

The *wealth*, or the total sum of all *r(d)* scores, depends on the number of queries issued and the number of results taken into account (c). To assess differences between the results obtained from the different corpora we calculated the wealth for each corpus for all values of c. An increase or decline in retrieval bias is determined using the *Gini coefficient*, which is a measure developed to express inequalities in societies [21].

6.4.5 *Impact Analysis*

Assessment of Query Impact We investigated the *impact* each unique query *term* has on the total wealth of a document collection. For this,

[2] http://snowball.tartarus.org/algorithms/dutch/stop.txt

[3] http://www.digitisation.eu/training/succeed-training-materials/ocr-evaluation/ocrevaluation/the-ocrevaluation-tool/

[4] http://www.lemurproject.org/indri.php

we issued all unique single query terms against the document collections and recorded the matching query - document pairs. We used these to assess, for every multi-term query - document pair, which of the terms in the query was responsible for retrieving the documents that appeared on the result list for said multi-term query. We then assigned each successful term a score of $\frac{1}{n_t}$ where n_t is the number of successful query terms t for a document - multi-term query pair. The sum of all of these scores for all occurrences of a query term is its *impact score* and the sum of all impact scores equals the total wealth of all $r(d)$ scores for a corpus.

Assessment of Direct Impact We investigated the differences in the retrievability of documents before (822_{GTerr}) and after (822_{GTcor}) error correction. For this, we evaluate the total number of documents retrieved (*wealth*), the equality of the *r(d)* scores' distribution, and we analyze qualitatively the documents and queries for which the differences between the experimental conditions are the largest. We measure the difference in inequality among the *r(d)* scores for the two versions of the document collection using the Gini coefficient. A high Gini coefficient indicates a large inequality in the distribution, a low Gini coefficient indicates a more equal and therefore less biased distribution. Then we investigate the difference in *r(d)* scores for each document in both versions. A gain in *r(d)* scores indicates that the document benefited from the correction of its content. A decrease in *r(d)* scores shows that its corrected version was retrieved by fewer queries than the original version. We manually assessed the documents with the largest differences and the queries that retrieved those documents to find out what caused the drop or increase in *r(d)* scores.

Assessment of Indirect Impact For this experiment we used the 1644_{mix} and the 1644_{err} data sets. Again, we evaluated differences in the overall wealth of distributed *r(d)* scores, the inequality between documents in terms of *r(d)* scores and the differences between documents in direct comparison. Differences in the results are caused by the interlace between the rankings of the corrected and unchanged documents. The analyses we perform for this section are similar to those of the *direct impact* experiments.

6.4.6 *Limitations*

Since relevance judgments are not available for this document collection, we were not able to explore how OCR errors correlate with precision and recall. In our mixed experiment, we only evaluated a correct/incorrect ratio of 50:50, other ratios are planned for future work.

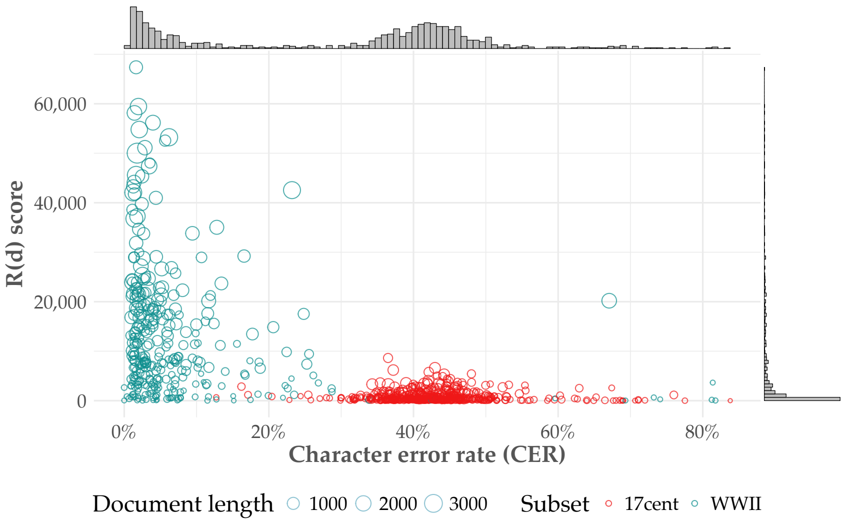


Figure 18: The 17thcentury collection has a higher character error rate (*CER*) than the *WWII* collection. The *r(d)* scores and *CER* for $c = \infty$ are strongly correlated: the higher the error rate, the less retrievable is a document.

6.5 RESULTS

6.5.1 *OCR Quality versus Retrievability*

First, we studied to what extent a document's OCR error rate and its *r(d)* score are related (*RQ1*).

QCR Quality We evaluated the OCR quality using the OCRevalUAtion tool[5]. The results showed that the mean character error rate (*CER*) of the collection is high: 29% (with a median CER of even 37%). We found a clear difference in the *CER* distributions of the two sub-collections (see Fig. 18). As expected, the more recent documents from *WWII* suffer from far fewer mis-recognized characters (median CER = 3.97%) than the documents from the 17thcentury (median CER = 42.00%).

Retrievability in 822_{GTerr} An analysis of the *r(d)* scores showed that we retrieved $4,521,030$ documents from 822_{GTerr} ($c = \infty$) in total. The scores ranged from $r(d) = 0$ (16 documents, of which two are part of the *WWII* sub-collection and 14 are part of the 17thcentury sub-collection) to $r(d) = 65,347$. Most documents are in the lowest bin ($r(d) < 674$), as shown in the margin histogram on the right of Fig. 18. The median scores are

- $r(d) = 991$ for 822_{GTerr},
- $r(d) = 447$ for 17thcentury, and
- $r(d) = 8,237$ for the *WWII* sub-collection.

[5] https://github.com/impactcentre/ocrevalUAtion

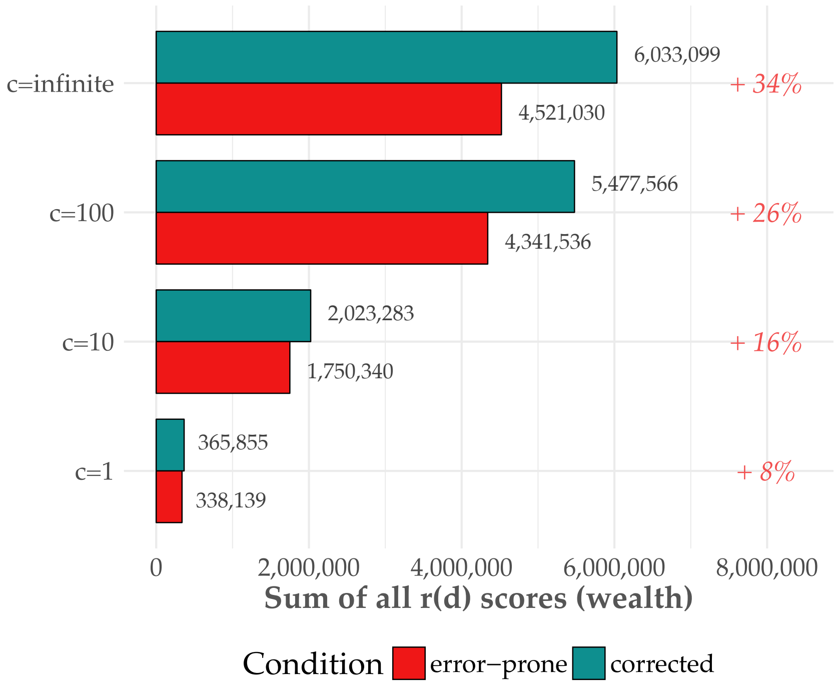


Figure 19: Difference in distributed wealth between the uncorrected and corrected corpus.

This confirms the hypothesis that the *WWII* documents are easier to retrieve due to their better OCR quality.

We found a strong correlation between OCR quality and retrievability of a document for results with $c = \infty$. Documents with a low *CER* generally obtained higher *r(d)* scores (see Figure 18). The correlations of -0.57 (Pearson) and -0.61 (Spearman) were both strong and significant with $p < 0.001$. While this correlation may suggest that low $r(d)$ scores are *caused* by high OCR error rates, other explanations could be that our modern query set just better matches the *WWII* sub-collection, or that $17^{th}century$ documents are harder to retrieve in general. To establish a causal relation, we study the direct impact of the crowd-sourced improvements on the *r(d)* scores in the next section.

6.5.2 *Direct Impact Assessment*

Next, we studied how the correction of OCR errors influences retrievability bias (*RQ2*). For this, we measure the direct impact of correcting OCR errors by comparing the $r(d)$ scores over 822_{GTcor} with the corresponding scores in 822_{GTerr}.

Wealth We found that more documents were retrieved from 822_{GTcor} than from 822_{GTerr} and that the relative difference increases for larger values of c (see Fig. 19). The total wealth at $c = 1$ indicates how many queries could be matched with at least one document. For $c = 1$, 8%

Corpus	c = 1	c = 10	c = 100	c = ∞
822_{GTerr}	0.75	0.72	0.74	0.74
822_{GTcor}	0.68	0.59	0.61	0.61
1644_{err}	0.78	0.70	0.73	0.73
1644_{mix}	0.73	0.63	0.66	0.66

Table 15: Gini coefficients indicating to which extent the distribution of *r(d)* scores among documents for different c's is biased (higher values indicate more bias).

c	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	-5,039	-56	8	34	99	7,160
10	-7,124	-153	177	332	647	8,408
100	-7,040	24	652	1,382	1,941	25,647
∞	-7,019	275	912	1,840	2,292	27,926

Table 16: Summary statistics of differences in *r(d)* scores between the two corpora.

more documents are retrieved from 822_{GTcor} than from 822_{GTerr}, which means that fewer queries retrieved no documents at all. For $c = \infty$ the total wealth increases by 34% (see Fig. 19). This suggests that for users willing to examine *all* search results (which is not uncommon in a research library) the impact of the error-correction is much larger. Correcting the OCR errors thus indeed leads to higher numbers of documents retrieved, even for small c's, and the effect increases when more results are taken into account.

Equality We computed and compared Gini coefficients for 822_{GTerr} and 822_{GTcor} to find out whether the increase in wealth contributed to a more equal or more biased distribution of *r(d)* scores (see Table 15). Gini coefficients for 822_{GTcor} are consistently lower than for 822_{GTerr} for all c's. The correction of the documents thus contributed to *less biased* retrieval for all c's. In contrast to other studies [3, 5, 8] and our earlier findings in [49], Gini coefficients do *not* show a clearly decreasing trend for larger cutoff values c. This suggests that in this experiment, ranking does not contribute much additional bias. This may be caused by the relatively small corpus size.

Increased retrieval per document We investigated how the changes of *r(d)* scores were distributed among documents, i.e. whether many documents gained a little or whether very few documents gained a lot. For Fig. 20 we ordered documents according to their difference in *r(d)* scores between 822_{GTerr} and 822_{GTcor}. We see a few documents on the left of the 0-axis, these documents had a higher *r(d)* score in the uncorrected corpus. Closer inspection indicated that these were false positive matches *caused by* OCR errors. Their decreasing scores can

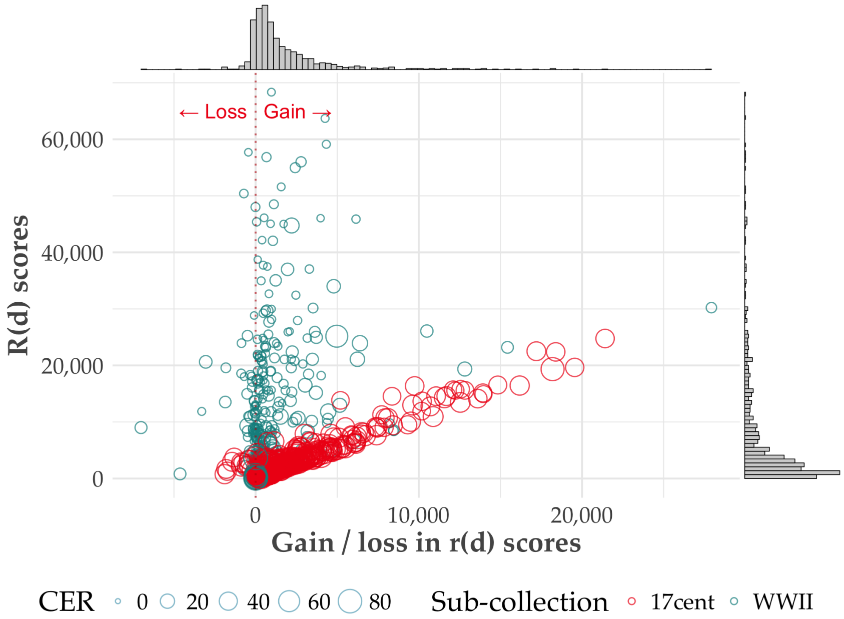


Figure 20: Documents ordered by their gain/loss in *r(d)* scores ($c = \infty$). The position on the y-axis represents their *r(d)* scores for 822_{GTcor}.

therefore be interpreted as a potential improvement in precision. For most documents, OCR correction increased their $r(d)$ score, and they are therefore found on the right of the 0-axis. This can be interpreted as a potential improvement in recall. We see clearly different patterns for the two corpora, with many 17^{th}century documents improving more but scoring overall lower than the *WWII* documents. Several documents scored very low in 822_{GTerr}, but gained a lot from the correction. This is one explanation for why Gini coefficients for 822_{GTcor} show less bias than for 822_{GTerr}. Most documents, however, have a modest *r(d)* score and gained a modest amount, as shown in the margin histograms. The distributions of the differences in *r(d)* scores in Table 16, show that for all cutoff values, the median of the differences is positive, and increases from 8 ($c = 1$) to 912 ($c = \infty$). The maximum loss and the maximum gain in *r(d)* scores increase for larger cutoff values c, the latter to a much larger extent. Note that for $c = 1$ and $c = 10$ the entire first quartile is filled with documents that scored *worse* in the corrected version. This shows that the competition in the top results makes the gain of some documents the loss of others.

Increased retrieval per query In a final step, we investigated how the changes of *r(d)* scores were distributed among the queries, i.e. if many queries contributed a little or if only a few queries that contributed a lot to the change in wealth. The large majority of queries does not match with any of the documents in our collection. Only 384,486 out of 859,716 queries retrieved at least one document from either of the document collections. This is due to misspellings from users, invalid words, numbers, words in foreign languages or sim-

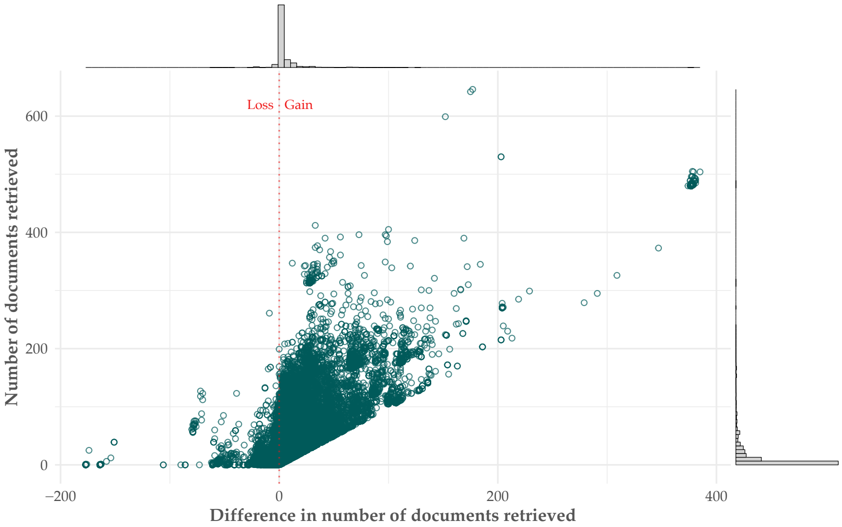
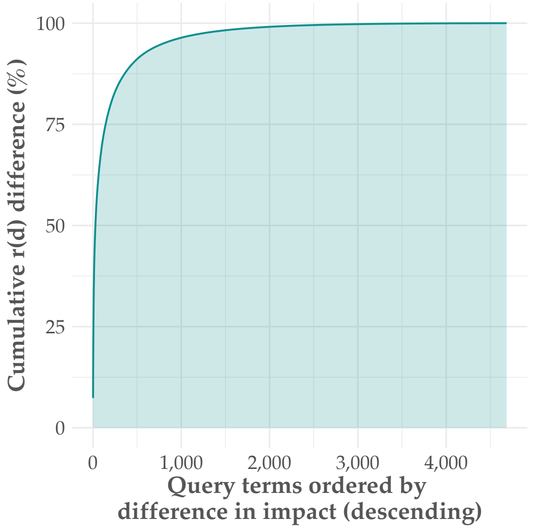


Figure 21: Queries ordered by their gain/loss in number of retrieved documents. The position on the y-axis represents the number of documents retrieved from 822_{GTcor}.

ply queries that are unrelated to our (small) corpus. In Figure 21 we ordered these queries by how many more (or less) documents they retrieved in 822_{GTcor}. Note that despite the small corpus size, we still see outliers with very large gains (to over 400 documents more retrieved for some queries). Also note that some queries have a negative gain, which means that for these queries, the OCR errors caused more false positive matches than false negatives.

Finally, we were interested in finding out which query *terms* are responsible for most of the increase in wealth. Figure 22 shows that most of the increase can be attributed to *very few* query terms. The top ten queries[6] (see table adjacent to Fig 22) contribute 35% of the increase. This disproportionately large impact originates from a combination of the terms' high frequency in the users' queries and the large extent to which they are susceptible to errors in OCR processing.

[6] Translations: new, Amsterdam, end, Mister, died/dead, grand/large, Willem (name), two, three, old



Query	Frequency in			Cum.
Term	Queries	822_{GTerr}	822_{GTcor}	Impact
nieuwe	1,903	99	166	7.36%
amsterdam	7,885	41	57	14.65%
ende	185	103	480	18.69%
heer	826	20	89	21.99%
overleden	3,698	5	18	24.78%
groot	1,573	125	153	27.33%
willem	5,375	5	13	29.81%
twee	319	64	175	31.83%
drie	401	34	120	33.81%
oude	991	50	78	35.41%

Figure 22: The accumulated impact scores of single-term queries show that very few query term contribute a large fraction of the overall wealth. The top ten query terms account for more than a third of the increase (see Table).

6.5.3 *Results of Indirect Impact Assessment*

Finally, we investigated the influence of OCR error correction on the retrieval of documents that remain uncorrected (*RQ3*). We investigate for the typical case of a partial error-correction how the improved retrievability of the corrected documents impacts the *r(d)* scores of the documents that have not (yet) been corrected.

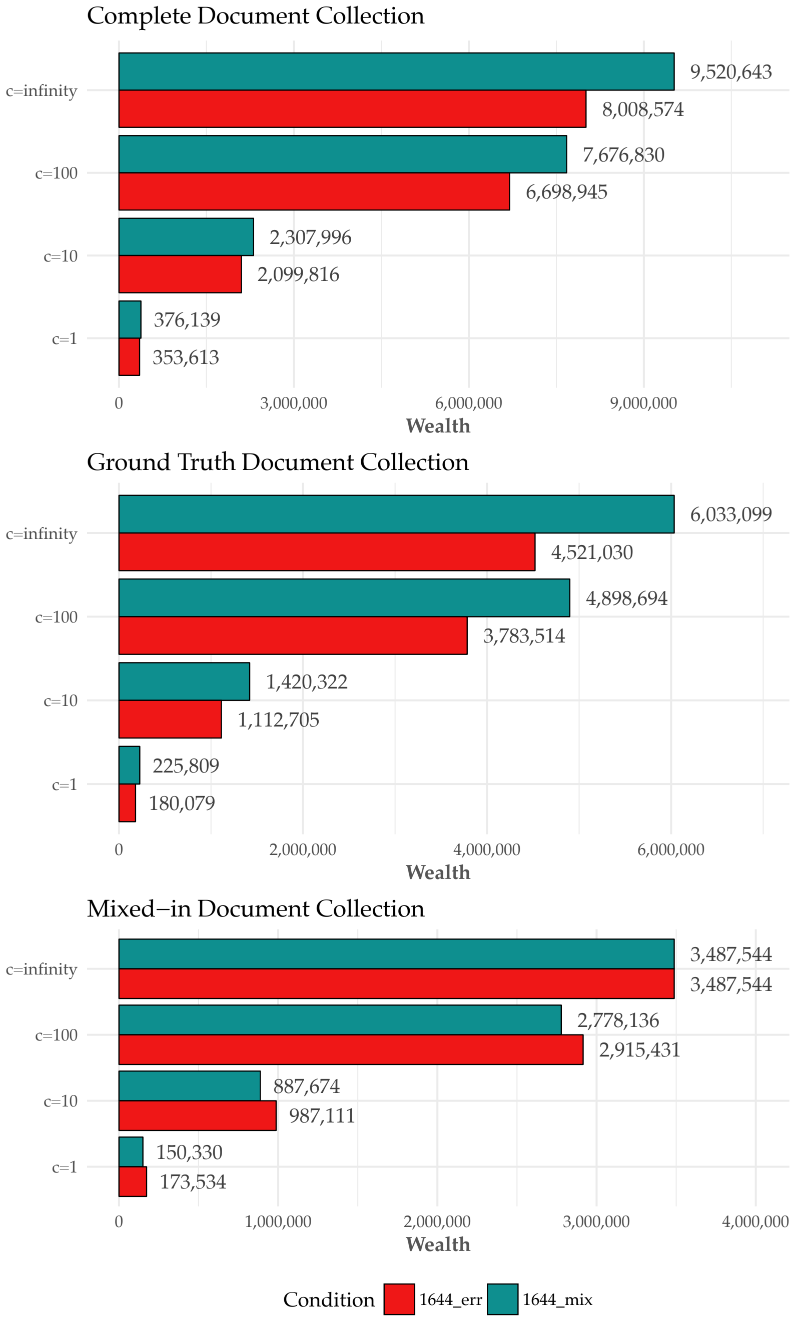


Figure 23: Wealth in *r(d)* scores for the complete collection (top), the 822_{GT} documents (middle) and the mixed in documents, 822_{mixin} (bottom).

Wealth When looking at the *r(d)* scores of the mixed collection, we see that the correction of half the documents still leads to an increase in wealth for the complete corpus for all values of c (see Fig. 23). We first focused on the 822_{GTcor} documents within the mixed corpus. These are retrieved for the same queries as in the previous section. The mixed-in documents only cause differences in ranking. For $c = \infty$, we thus see identical *r(d)* scores and total wealth as in Section 6.5.2.

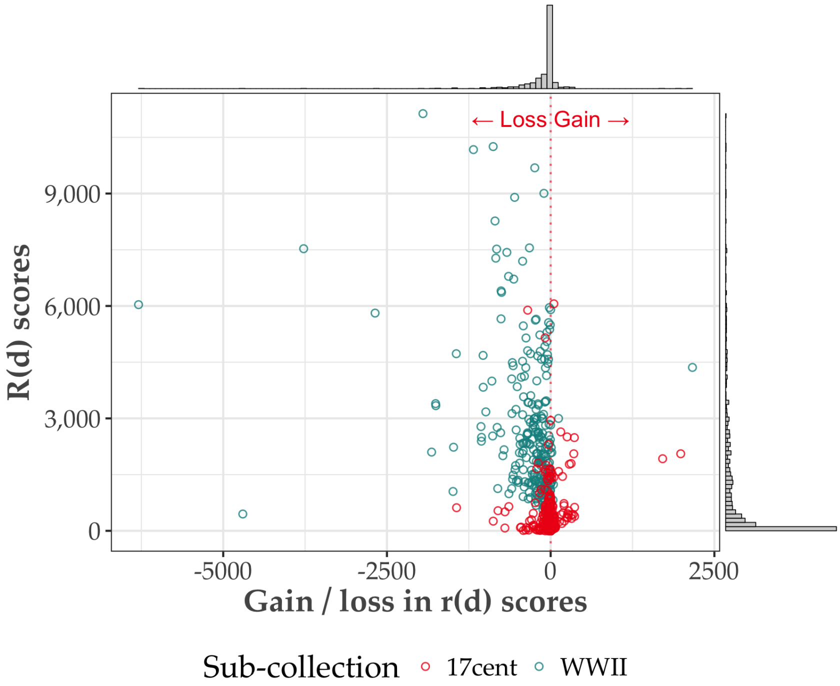


Figure 24: Documents ordered according to their difference in *r(d)* scores (non-GT documents at c = 10). Position on the y-axis indicated the *r(d)* score in the mixed condition. Documents in the left part of the graph lost *r(d)* scores.

For the lower c values, we see lower wealth due to competition in the ranking with the unaltered documents, but also large gains caused by the manual OCR correction.

In the remainder of this section, we focus solely on the documents that remain uncorrected, $822_{\mathtt{mixin}}$. In terms of distributed *r(d)* scores we found a decrease in wealth for the mixed-in documents for values of c from c = 1 to c = 100. This is because the corrected versions push many mixed-in documents to higher ranks that exceed the number of documents we take into account (c). This difference in wealth is largest for c = 1 (−13%), followed by c = 10 (−10%) and c = 100 (−5%). For larger values the ranking does not take effect and the wealth remains the same.

Equality When we compare the Gini coefficients we obtained for different values of c, we see that they are lower for the corpus that was partially corrected, $1644_{\mathtt{mix}}$. Again, the correction of a part of the collection has reduced retrievability bias (see Table 15).

Retrieval per document The *r(d)* scores of most mixed-in documents changed very little after the correction of the other documents. Most documents' *r(d)* scores drop slightly (see Fig. 24), which could be expected as they now compete with corrected documents for low ranks. In total, 522 documents have lost in r(d) scores, of which 266 are from the *WWII* sub-collection and 256 from $17^{\text{th}}\mathtt{century}$. We also see that 171 documents gain in *r(d)* scores, of which 8 are from *WWII* and 163 from 17^{th}*century* (see Fig. 24). These documents profit from false positive matches that disappeared through the correction.

Overall, we found that even in this mixed condition, the overall positive effect in improved retrievability for the corrected documents by far outweighs the slightly reduced retrievability of the unchanged documents. The net effect of the correction is still an overall *reduction of retrievability bias*.

6.6 CONCLUSIONS

Many text documents in digital libraries are affected by errors caused by OCR engines. It is therefore vital to understand how these errors and their (partial) correction impact retrieval tasks of digital library users. We investigated the relation between OCR quality of digitized newspaper articles and their retrievability and found a strong correlation: high error rates correlate with low retrievability scores. We compared the overall retrievability of a manually corrected ground truth document collection with the results obtained from the same documents but in their original, uncorrected version. Our analyses showed that error correction leads to both higher and more equally distributed retrievability scores.

The higher scores are mainly caused by a disproportionately small set of query terms, that are both very frequent in the query set and highly susceptible to OCR errors. This shows that for retrievability studies with real user queries, understanding the impact of a (biased) query set on the retrievability bias is important, while this is typically not considered in the literature, where synthetic query sets are more prevalent.

Our findings could be used for improving and evaluating automatic OCR-error correction techniques, or to improve query expansion techniques designed to deal with OCR-errors in uncorrected texts.

Furthermore, we looked at interference effects that the correction of a subset may have on documents that are excluded from the correction. We found that the reduced scores for the excluded documents do not outweigh the improved scores of the corrected version. The overall outcome is still a less biased retrieval result. Because we lack relevance judgments for this corpus, we cannot measure the improvement of the correction in terms of precision and recall. We can, however, conclude that the error correction has led to more documents being retrieved overall while reducing the retrievability bias in all experimental setups.

7 CONCLUSIONS AND DISCUSSION

Cultural heritage institutions face huge challenges when trying to ensure high data quality in their digital collections. With increasing numbers of documents becoming digitally available, it becomes more and more difficult for data custodians to define and maintain data quality standards that meet the requirements of users. Quality issues in data sets can originate from different sources (such as digitization process, collection and digitization policy, legal conditions) and therefore vary in types and impact. In order to establish the desired level of data quality, an institution needs to understand which data quality criteria matter the most, how to develop meaningful metrics and how to measure the impact of increased quality.

7.1 SUMMARY

For expert users, such as Humanities scholars, it is important to understand whether results obtained from digital archives are sufficiently reliable to be used in publications. For this reason, data custodians need to make sure that the level of quality of the data and infrastructure they provide is sufficient to support the research tasks of their users. This, however, requires an understanding of users' tasks, potential issues in data quality, and potential biases in tools.

In some cases, a discrepancy exists between the data (lay) users require to search an archive, and the data that is deemed useful by experts. An example is the annotation of subject types (e.g. *portrait*, *landscape*) in the metadata of artworks, which from the perspective of an art historian can be subjective. Such additions to the metadata records should ideally be made by domain experts, but since they are scarce and expensive, other approaches need to be considered. We explored how reliable annotations can be obtained for artwork classification through crowdsourcing (Chapter 2). For this, we conducted several experiments where we presented online users a classification task set up as a multiple choice guessing game. We found that non-experts became better at classifying a painting into the categories of a professional vocabulary while playing the game. After aggregating the votes we collected from users, we found that the deviation from our groundtruth experts' classification is greatly reduced. From this study we conclude that crowdsourcing can not only be used to enrich cultural heritage data with common knowledge, but also with high-quality data based on knowledge that is typically not found in untrained crowd workers. This raised the questions, which data should be collected in the first place in order to be useful for Humanities researchers, and for which kinds of research tasks would this data be used. From the insights we gained in the later studies, we would suggest to also assess whether this input is *fit for use* for user tasks. This should be evaluated both, for content and level of quality. For this, it would be necessary to first investigate search behavior of users. It may be interesting to know if the query terms of visitors of the RMA website overlap with the terms defined in the Art & Architecture Thesaurus[1]. Another limitation of the suggested approach is the dependence on the availability of expert-approved labels that can be used as feedback for users. As He et al. showed, users would not be able to learn correct categories for the items without receiving feedback for their choices. Using labels that were contributed by users and that have not been approved by experts bears the risk of introducing a systematic error. If users are taught a wrong label for a category, this misconception may cascade through the game and lead to incorrect labels for an entire category. We have not found a solution for this problem.

In the interviews we conducted with humanities scholars, we learned that the access to digital sources in their current state is seen critically by many humanities scholars (Chapter 3). When working with digital archives, they perceive the data quality of digitized documents to be very low. Knowing that the same data is used as a basis for retrieval systems, the scholars raised questions about the trustworthiness of results that can be obtained from search functions and whether they can actually be used for publications. We concluded from our findings that these issues will not be solved simply by *improving* the digitization infrastructure, but it requires a better *understanding of the impact of data quality* on user tasks. For this, all three parties – the computer scientists developing the tools, the scholars designing the research tasks, and the data custodians providing the infrastructure and data – need to tackle the challenges jointly. One limitation of our study is that it does not include more (complex) use cases to further investigate the requirements of *distant reading* tasks. While the interviews with historians gave us a good view of the challenges scholars from this discipline face, it would be interesting to extend the investigations to scholars from other disciplines within the humanities, i.e. linguists, to gain more information about the types of tasks they (would like to be able to) perform. Bulger et al., for example, mention *word lists*, *fre-*

[1] http://www.getty.edu/research/tools/vocabularies/aat/index.html

quency analysis, *collocation analysis*[2] and *concordance analysis*[3] as some of the main techniques of corpus linguists to analyze text.

In order to foster discussion between humanities researchers and computer scientists regarding the pitfalls of using computational tools in humanities research, we organized the *"Tool Criticism for Digital Humanities"* workshop (Chapter 4). During the workshop, participants from different disciplines presented use cases that were closely examined for potential bias introduced by digital tools by mixed groups with backgrounds in Computer Science and (Digital) Humanities. At the end of the workshop, participants agreed that the idea of *Tool Criticism* as part of Digital Humanities research practices should be fostered. Ideally, more use cases and research tasks (generic as well as specific ones) should be collected as a basis to develop guidelines for tool developers as well as scholars.

The basis for many research tasks is the retrieval of (all) relevant documents. To find out how biased the discoverability of documents in a large historic newspaper archive is, we conducted a large-scale analysis of retrievability bias (Chapter 5). In this study, we investigated whether access to the documents within the KB's newspaper archive through standard retrieval models is biased and whether we can link this bias to features in the documents. Additionally, we evaluated how well the results gained through the typical setup of retrievability studies represents the actual bias as experienced by real users of document collections. We found that the frequently used approach of using artificial queries generated from the corpus' terms does not reflect the bias experienced by users adequately. The findings of this study can be used to design retrievability studies in a way that they better represent the types of bias users face.

Data quality was mentioned multiple times by the humanities scholars we interviewed for Chapter 3. We therefore decided to study the impact of (low) data quality on retrievability and – as a consequence – on search tasks of users (Chapter 6). The low data quality in documents we investigated partially originates from the imperfect digitization of historic newspapers. As expected, we found that high error rates in the digitized documents correlate with low retrievability. We then investigated how the correction of the entire document collection influences the retrievability scores. We found that the retrievability scores for corrected documents are both higher and more equally distributed. The changes in retrievability scores between the uncorrected and the corrected versions of the documents are to a large

[2] "examines high-frequency keyword combinations; either adjacent (e.g. strong tea/powerful tea), or non-adjacent (i.e. within 4-5 words to the left and/or right of the word investigated). Words or terms that co-occur more often than would be expected by chance are examined." [13]

[3] "list of specific keywords or collocates displayed within the context for which they were used. The keyword is usually listed within the context of the five words that precede and succeed it." [13]

extent caused by a relatively small set of query terms. These terms, however, are very frequent in user queries and susceptible to OCR errors. As the (manual) correction of a document is costly, institutions often improve only a fraction of the whole collection. Therefore, we investigated what the interference effects of such a partial correction are. Despite the "unfair" preferential treatment of a part of the document collection, the overall retrievability bias decreased.

Ideally, we would have evaluated the impact of correcting OCR errors on retrieval bias (Chapter 5 and 6) not only using the retrievability measure, but also by measuring differences in precision and recall. For this, however, we would have needed relevance judgements. Unfortunately, we did not have any judgments for the KB data available, therefore, this was not possible. Another limitation for the study we describe in Chapter 5 is the high document to query ratio. This is a problem any retrievability study using a very large document collection will have to deal with, but the problem is more prevalent for studies based on real user queries.

Finally, it would be good to study the effects of OCR errors on retrievability bias (Chapter 6) on a larger dataset to confirm that the reduction of retrievability bias is not just caused by the small document collection.

7.2 DISCUSSION AND FUTURE WORK

The findings of the work presented in this thesis help us better understand what requirements have to be met in order to provide better support to Humanities scholars wishing to make use of digital archives.

As it currently stands, scholars are hesitant to publish work based on results from digital archives as they doubt their validity. They perceive the quality of digitized documents as very low and are only given very little or no information about how this may affect (search) results. Given the size of many digital archives and the complexity of the task, it is unrealistic to expect that data quality will be improved to 100% accuracy.

We believe scholars should be given the possibility to better understand the inner workings of digital archives and make *tool criticism* part of their research routines. The idea of *tool criticism* is derived from the practice of *source criticism*, which is used in many sub-disciplines of the humanities. Source criticism describes a set of principles and methods to evaluate a (historic) source of information for its authenticity, reliability and relevance.

Performing tool criticism, however, requires the scholar to be aware of sources for potential fallacies. Our work contributes to this to the extent that we have identified and quantified some sources of bias and how they affect user tasks. Bias can be introduced into a data set through different sources.

COLLECTION of documents is typically determined by the collection policy of the data custodian and often focuses on particular types of items, periods or topics (Section 4.1.1). Additionally, survivorship bias may bias which documents *can* be collected by the institution in the first place.

PHYSICAL DOCUMENT FEATURES, such as low quality print or decay may negatively impact the digitization process and as a consequence decrease the accessibility of documents (Chapter 3).

CONTENT characteristics of documents (e.g. length and repetition of specific terms) influence their retrievability as we have shown in Chapters 5 and 6.

DIGITIZATION POLICY of the data custodian determines which documents are prioritized for digitization. More weight can be assigned to documents that are popular with users, rare or close to decay.

DIGITIZATION of documents can be a complex process and its success depends the use of apt technology for digitization and post-processing, e.g. OCR engines that were trained on historic data sets (Chapters 3 and 6).

QUALITY IMPROVEMENT is sometimes applied to digitized documents by (manually) correcting errors of the OCR engine. In Chapter 6 we showed how this impacts the retrievability of the corrected, as well as the uncorrected documents.

SELECTION of relevant documents from the documents available in a collection is often done manually by scholars (Section 4.2.1). Which documents are deemed relevant to the research tasks and added to the selection is influenced by the (search) skills of the users, and their knowledge of the domain (use of historic language, terminology) and the purpose for which the documents are collected.

ACCESS to documents is usually granted through an (online) search interface. Functionalities such as facets and logic operators, as well as the retrieval model used, can influence which documents are easier or harder to find than others (Chapter 5).

PRESENTATION of search results influences how (much) information about the retrieved documents is presented and how many search results a user sees without scrolling (Chapters 5 and 6).

USERS themselves have a strong influence on the outcome of their search tasks. They should ideally have the (technical) knowledge to understand the way a search engine works (facets and operators), patience to explore a large number of documents and enough knowledge about the targeted documents to be able to use suitable terms when formulating queries (Chapters 5 and 6).

Additionally, combinations of these sources may increase bias in results and create new biases. As it now stands, it is very difficult for scholars to understand the impact of the sources of bias and their combined effect on their search results.

Changing this is challenging and not only requires effort from the scholars, but also from the developers of the software tools and data custodians. Only if these three parties collaborate can the foundations be laid for a comprehensive *tool* and *data criticism*.

HUMANITIES SCHOLARS need to familiarize themselves with the implications of data processing by different software tools. This should include an understanding of tools that are used along the entire processing pipeline of a digitization setup, as well as the tools used in the data provision infrastructure of the digital archive, such as search technologies.

On top of this, we believe scholars should develop a standard approach to investigating their digital research environment and the results it produces. For historic sources, the concept of *source criticism* already exists, an analog concept should be developed to criticize tools and the data they produce.

Based on our experiences from the workshop we organized, we believe that a good approach to this would be to collect typical research tasks and subject them to a discussion with tool developers to find potential sources of bias. The results of these discussions should then be used to formulate guidelines for a common understanding of *tool criticism.*

TOOL DEVELOPERS should provide as much information about the assumptions their tools are based on and the effect they can potentially have on data in terms of bias. They should conduct evaluations on bias and communicate their findings to humanities scholars and data custodians.

In discussions with the Humanities scholars, they should try to gain a better understanding of their research tasks and the requirements of users. Understanding how their tools are actually used can help them find out what potential biases to look for and how to communicate them effectively to the users.

DATA CUSTODIANS should extend their already extensive efforts of providing quality data to users by adding as much information about their collections as possible. As they have the role of an intermediary between Humanities scholars and the software tool developers, they should engage in guiding discussions between the parties involved and acquiring a close understanding of both sides' interests and requirements. Given their role as designers and maintainers of the digital archive infrastructure, they should oversee its *Quality of Design* (denoting how well the system specifications meet users' requirements) and *Quality of Conformance* (denoting how well the implementation corresponds to the specifications).

As much information as possible should be provided to the users. Most data custodians focus on the scope of the collection, but this information should also contain tools used in the digitization process and for data access, and potential limitations of the technology used.

The recommended responsibilities and tasks for each party are complex in their own right and their dependencies on other parties input require close interaction. Ways of fostering this interaction, such as interdisciplinary hackathons, Lorentz Center or Dagstuhl seminars, should be organized within the communities. Insights gained should be distilled into guidelines and incorporated into digital humanities and tool providers' education in order to reach the goal of providing measures of tool bias and data quality in research environments for the field.

SUMMARY

Cultural heritage institutions, such as galleries, libraries, archives and museums increasingly make their collections digitally available. These provide many ways for users of these digital platforms to retrieve, aggregate, analyze and visualize the data. Users need to familiarize themselves with new digital tools of different kinds and be aware of the implications on the results. This is particularly true for humanities scholars who want to include results of their analyses in their publications.

Judging whether insights derived from these analyses constitute a real trend or whether a potential conclusion is just an artifact of the ensemble of tools used, can be difficult and requires an understanding of the processing chain of the tools used.

In order to detect and correct errors, however, human input is in many cases still indispensable. Since experts are expensive and scarce, we conducted a study showing how crowdsourcing tasks can be designed to allow lay users to contribute information at the expert level in order to increase the number and quality of descriptions of cultural heritage items.

In order to improve the quality of the data they provide, data custodians need to understand the (search) tasks their users perform and the level of trustworthiness they expect from the results. Through interviews with historians, we studied their use of digital archives for research purposes and classified typical research tasks and their requirements for data quality.

Most cultural heritage archives provide, at best, very generic information about the quality of their digitized collections. Humanities scholars performing research tasks, however, need to be able to assess how *data quality* of the archive and *inherent bias in tools* involved in the creation, retrieval, aggregation, analysis and visualization of digital items influences their research tasks. Therefore, they need specific information on the data quality of the used subcollection and the biases the tools provided may have introduced into the analyses.

We studied whether access to a historic newspaper archive is biased and which types of documents benefit from or are disadvantaged by the bias. We used an existing retrievability measure we applied on artificially generated queries. Since we also had access to real user queries and page view data, we were able to investigate how well the typical setup of these studies reflects the real users' experience. We found large differences in the characteristics of the query sets and in the results for different parameter settings of the experiments.

In archives of digital historic documents, a prevalent data quality issue is errors caused by Optical Character Recognition (OCR). Since these are relatively easy to spot for users closely examining an item, it has caused some concern by the humanities researchers about the trustworthiness of results based on digitized data. We evaluated the impact of OCR quality on retrieval tasks, and studied the effect of manually improving (parts of) a collection on the retrievability bias.

The insights we gained helped us understanding researchers' needs better and identifying and measuring biases in accessing collections. Our work provides a small number of examples that demonstrate that data quality and tool bias are real concerns to the DH community. Further studies such as those we carried out, while essential, are insufficient to address the challenges to the community. In addition, intense multidisciplinary exchange between data custodians, tool developers and humanities scholars and efforts from each side is required:

HUMANITIES SCHOLARS need to enhance the *awareness* in their field that software tools and (big) data sets are not free of bias and develop the *skills* to detect and evaluate potential types of biases and their impact on research tasks. Based on these skills, guidelines should be developed that help the individual scholar to perform such *tool criticism*.

TOOL DEVELOPERS need to be more *transparent* and provide sufficient information about their tools to allow the task-based evaluation of their tools' performance.

DATA CUSTODIANS need to make as much *information about their collection* available as possible. This should not only include the scope of the items in the collection, but also which tools were used in the digitization process, and what the limitations of the provided data and infrastructure are.

The goal should be a mutual understanding of each others' assumptions, approaches and requirements and more transparency concerning the use of tools in the preprocessing of data. This will help scholars to develop effective methods of *digital tool criticism* to critically assess the impact of existing tools on their (re-)search results and to communicate on an equal footing with tool developers on how to develop future versions that better suit their needs.

SAMENVATTING

Erfgoedinstellingen, zoals galeries, bibliotheken, archieven en musea, stellen hun collecties steeds vaker digitaal beschikbaar. Dit biedt gebruikers van deze digitale platforms veel mogelijkheden om gegevens op te halen, te aggregeren, te analyseren en te visualiseren. Gebruikers moeten zich vertrouwd maken met verschillende soorten nieuwe digitale tools en zich bewust zijn van de gevolgen voor de resultaten. Dit geldt in het bijzonder voor geesteswetenschappers die de resultaten van hun analyses in hun publicaties willen opnemen.

Het kan moeilijk zijn om te beoordelen of de inzichten uit deze analyses een echte trend vormen, of dat een mogelijke conclusie slechts een artefact is van het ensemble van gebruikte tools. Dit vereist begrip in de verwerkingsketen van de gebruikte tools.

Menselijke input is echter in veel gevallen nog steeds onmisbaar om fouten te kunnen opsporen en corrigeren. Omdat experts duur en schaars zijn, hebben we een studie gedaan die laat zien hoe crowdsourcing-taken zo kunnen worden opgezet dat leken informatie op expertniveau kunnen toevoegen en zo bijdragen aan het verhogen van het aantal en de kwaliteit van beschrijvingen van cultuurhistorische objecten.

Om de kwaliteit van de gegevens die zij aanleveren te verbeteren, moeten de gegevensbeheerders inzicht hebben in de (zoek)taken die hun gebruikers uitvoeren en de mate van betrouwbaarheid die zij van de resultaten verwachten. Door middel van interviews met historici hebben we het gebruik van digitale archieven voor onderzoeksdoeleinden bestudeerd en hebben we typische onderzoekstaken en hun eisen voor datakwaliteit geclassificeerd.

De meeste erfgoedinstellingen leveren, in het beste geval, zeer generieke informatie over de kwaliteit van hun gedigitaliseerde collecties. Geesteswetenschappers moeten echter kunnen beoordelen hoe hun onderzoekstaken beïnvloed worden door de *datakwaliteit* van het archief en door de *inherente bias van tools* voor het creëren, terugvinden, aggregeren, analyseren en visualiseren van digitale items. Daarom hebben ze specifieke informatie nodig over de datakwaliteit van de gebruikte deelcollectie en mogelijke vertekeningen die gebruikte tools in de analyses hebben geïntroduceerd.

We onderzochten in hoeverre er bias bestaat bij de toegang tot een historisch krantenarchief en welke soorten documenten hoger of lager in de resultatenlijst terechtkomen door de bias. We gebruikten een bestaande *retrievability* maatstaf die we toepasten op kunstmatig gegenereerde zoekopdrachten. Aangezien we ook toegang hadden tot echte zoekvragen van gebruikers en hoe vaak pagina's werden

weergegeven, konden we onderzoeken hoe goed de typische opzet van deze studies de echte gebruikerservaring weerspiegelt. We vonden grote verschillen in de karakteristieken van de query sets en in de resultaten voor de verschillende parameterinstellingen van de experimenten.

Fouten veroorzaakt door OCR (Optical Character Recognition) komen veelvuldig voor in archieven met digitale historische documenten. Aangezien deze fouten relatief gemakkelijk te herkennen zijn voor gebruikers die een item nauwkeurig onderzoeken, heeft dit bij de geesteswetenschappelijke onderzoekers enige bezorgdheid gewekt over de betrouwbaarheid van bevindingen op basis van gedigitaliseerde gegevens. We evalueerden de impact van OCR-kwaliteit op de retrievaltaken, en onderzochten het effect van handmatige verbetering van (delen van) een collectie op de retrievability bias.

De inzichten die we hebben opgedaan, hebben ons geholpen om de behoeften van onderzoekers beter te begrijpen en om bias bij de toegang tot collecties te identificeren en te meten. Ons werk geeft een klein aantal voorbeelden die aantonen dat datakwaliteit en de bias van tools mogelijk zorgwekkend zijn voor de DH-gemeenschap. Nader onderzoek zoals we dat hebben gedaan is weliswaar essentieel, maar niet voldoende om de uitdagingen voor de gemeenschap aan te pakken. Daarnaast is een intensieve multidisciplinaire uitwisseling tussen gegevensbeheerders, toolontwikkelaars en geesteswetenschappers benodigd, die inspanningen van verschillende zijden vereist:

GEESTESWETENSCHAPPERS moeten in hun vakgebied het *bewustzijn* vergroten dat softwaretools en (grote) datasets niet vrij zijn van bias. Tevens moeten zij *vaardigheden* ontwikkelen om mogelijke vooroordelen en hun impact op onderzoekstaken op te sporen en te evalueren. Op basis van deze vaardigheden moeten richtlijnen worden ontwikkeld die de individuele wetenschapper kunnen helpen bij het uitvoeren van dergelijke *toolkritiek*.

ONTWIKKELAARS van tools moeten *transparanter* zijn en voldoende informatie verstrekken over hun tools om een taakgerichte evaluatie van de prestaties van hun tools mogelijk te maken.

GEGEVENSBEHEERDERS moeten zoveel mogelijk *informatie over hun collectie* beschikbaar stellen. Daarbij moet niet alleen worden gekeken naar de omvang van de items in de collectie, maar ook welke tools zijn gebruikt bij het digitaliseringsproces en wat de beperkingen zijn van de geleverde data en infrastructuur.

Het doel moet een wederzijds begrip zijn van elkaars aannames, benaderingen en eisen, maar ook meer transparantie over het gebruik van tools bij de preprocessing van data. Dit zal wetenschappers helpen om effectieve methoden van *digitale toolkritiek* te ontwikkelen om de impact van bestaande tools op hun zoek- en onderzoeksresultaten kritisch te beoordelen en tevens om op gelijke voet met toolontwikkelaars te communiceren over hoe zij toekomstige versies kunnen ontwikkelen die beter aansluiten bij hun behoeften.

PUBLICATIONS BY THE AUTHOR

JOURNAL PUBLICATIONS

- Samar, T., **Traub, M. C.**, van Ossenbruggen, J.R., Hardman, L. (Lynda) and de Vries, A.P., *Quantifying retrieval bias in Web archive search*, International Journal on Digital Libraries Volume 19, Issue 1, p. 57- 75, 2018.
- **Traub, M. C.**, Lamers, M.H. and Walter, W., *A Semantic Centrality Measure for Finding the Most Trustworthy Account*, IADIS International Journal on Computer Science and Information Systems, Vol. 6 No.1, pp 45-57, 2011.

BOOK CHAPTERS

- Hannemann, J., **Traub, M.C.**, Böhme, C., *CONTENTUS: Next Generation Multimedia Libraries*, pages 264-269, In: Towards the Internet of Services: The THESEUS Research Program. Cognitive Technologies, 2014.
- Böhme, C., **Traub, M.C.**, Bergholz, A., Hannemann, J., Svensson, L., *Semantic Linking in Contentus*, pages 329-341, In: Towards the Internet of Services: The THESEUS Research Program. Cognitive Technologies, 2014.

CONFERENCE PUBLICATIONS

- **Traub, M.C.**, Samar, T., Ossenbruggen, J., Hardman, L., *Impact of Crowdsourcing OCR Improvements on Retrievability Bias*, Joint Conference on Digital Libraries (JCDL) 2018, Fort Worth, USA, 2018. (full paper)

  Best Student Paper Award & **Vannevar Bush Best Paper Award**
- **Traub, M.C.**, Samar, T., Ossenbruggen, J., de Vries, A., Hardman, L., *Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus*, Joint Conference on Digital Libraries (JCDL) 2016, Newark, USA (full paper)
- **Traub, M.C.**, Ossenbruggen, J., Hardman, L., *Impact Analysis of OCR Quality on Research Tasks in Digital Archives*. 19th International Conference on Theory and Practice of Digital Libraries (TPDL), 2015. (full paper)

- **Traub, M.C.**, *Measuring and Improving Data Quality of Media Collections for Professional Tasks*. In: Proceedings of Information Interaction in Context 2014 (IIiX 2014), August 26 - August 29, Regensburg, Germany. (doctoral consortium paper)

- **Traub, M.C.**, van Ossenbruggen, J., He, J., Hardman, L., *Gamesourcing Expert Painting Annotations*. CHI Sparks 2014, April 3, 2014 Den Haag, The Netherlands. (demo track)

- **Traub, M.C.**, Ossenbruggen, J., He, J., Hardman, L., *Measuring the Effectiveness of Gamesourcing Expert Oil Painting Annotations*. In: Advances in Information Retrieval, European Conference on Information Retrieval (ECIR), 2014, April 13 -16, Amsterdam, The Netherlands. (full paper)

- Perez Romero, L., **Traub, M.C.**, Leyssen, M. H. R and Hardman, L., *Second Screen Interactions for Automatically Web-enriched Broadcast Video*. 4th International Workshop on Future Television: Linking Television and Web across Screens, June 24 - 26 2013, Como, Italy. (workshop paper)

- Dijkshoorn, C., Leyssen, M.H.R., Nottamkandath, A. , Oosterman, J. , **Traub, M.C.**, Aroyo, L., Bozzon, A., Fokkink, W.J., Houben, G.J., Hovelman, H., Jongma, L., van Ossenbruggen, J.R., Schreiber, G. and Wielemaker, J., *Personalized Nichesourcing: Acquisition of Qualitative Annotations from Niche Communities*, In: S Berkovsky, E Herder, P Lops, & O.C Santos (Eds.), Workshop on Personalized Access to Cultural Heritage, 2013. (workshop paper)

- **Traub, M.C.**, Lamers, M.H. and Walter, W., *A Semantic Centrality Measure for Finding the Most Trustworthy Account*. In:Proceedings of the IADIS International Conference Informatics 2010 (IADIS 2010), July 16 - 31, 2010. (full paper)

REPORTS

- **Traub, M.C.** and Ossenbruggen, J.R., *Workshop on Tool Criticism in the Digital Humanities*, July 2015. (workshop report)

- Leyssen, M.H.R., **Traub, M.C.**, van Ossenbruggen, J.R., Hardman, L., *Is it a bird or is it a crow? The influence of presented tags on image tagging by non-expert users*, 2012. (technical report)

BLOG POSTS

- **Traub, M.C.**, *Assessing bias in search results*, codecentric Blog, https://blog.codecentric.de/en/2019/05/assessing-bias-search-results/, May 2019

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

[1] Alberto Acerbi, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. The expression of emotions in 20th century books. *PLoS ONE*, 8(3):e59030, 03 2013. doi: 10.1371/journal.pone.0059030. URL http://dx.doi.org/10.1371%2Fjournal.pone.0059030.

[2] Bea Alex, Claire Grover, Ewan Klein, and Richard Tobin. Digitised historical text: Does it have to be mediOCRe? In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 401–409. ÖGAI, September 2012. URL http://www.oegai.at/konvens2012/proceedings/59_alex12w/. LThist 2012 workshop.

[3] Leif Azzopardi and Vishwa Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 561–570, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458157. URL http://doi.acm.org/10.1145/1458082.1458157.

[4] Richard Bache and Leif Azzopardi. Improving access to large patent corpora. In Abdelkader Hameurlain, Josef Küng, Roland Wagner, Torben Bach Pedersen, and A.Min Tjoa, editors, *Transactions on Large-Scale Data- and Knowledge-Centered Systems II*, volume 6380 of *Lecture Notes in Computer Science*, pages 103–121. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-16174-2. doi: 10.1007/978-3-642-16175-9_4. URL http://dx.doi.org/10.1007/978-3-642-16175-9_4.

[5] Shariq Bashir. Estimating retrievability ranks of documents using document features. *Neurocomputing*, 123(0):216 – 232, 2014. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2013.07.011. URL http://www.sciencedirect.com/science/article/pii/S0925231213007455. Contains Special issue articles: Advances in Pattern Recognition Applications and Methods.

[6] Shariq Bashir and Andreas Rauber. Improving retrievability of patents in prior-art search. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, and Keith van Rijsbergen, editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 457–470. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-12274-3. doi: 10.1007/978-3-642-12275-0_40. URL http://dx.doi.org/10.1007/978-3-642-12275-0_40.

[7] Shariq Bashir and Andreas Rauber. Improving retrievability and recall by automatic corpus partitioning. In Abdelkader Hameurlain, Josef Küng, Roland Wagner, Torben Bach Pedersen, and A.Min Tjoa, editors, *Transactions on Large-Scale Data- and Knowledge-Centered Systems II*, volume 6380 of *Lecture Notes in Computer Science*, pages 122–140. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-16174-2. doi: 10.1007/978-3-642-16175-9_5. URL http://dx.doi.org/10.1007/978-3-642-16175-9_5.

[8] Shariq Bashir and Andreas Rauber. Automatic ranking of retrieval models using retrievability measure. *Knowledge and Information Systems*, 41(1):189–221, 2014. ISSN 0219-1377. doi: 10.1007/s10115-014-0759-6. URL http://dx.doi.org/10.1007/s10115-014-0759-6.

[9] Shariq Bashir and Andreas Rauber. Retrieval models versus retrievability. In Mihai Lupu, Katja Mayer, Noriko Kando, and Anthony J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, pages 185–212. Springer Berlin Heidelberg, 2017. ISBN 978-3-662-53817-3. doi: 10.1007/978-3-662-53817-3_7. URL https://doi.org/10.1007/978-3-662-53817-3_7.

[10] Adrian Bingham. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–231, 2010. doi: 10.1093/tcbh/hwq007. URL http://tcbh.oxfordjournals.org/content/21/2/225.short.

[11] Marc Bron. *Exploration and Contextualization through Interaction and Concepts*. PhD thesis, Universiteit van Amsterdam, 2013. URL dare.uva.nl/document/502718.

[12] Christine D Brown. Straddling the humanities and social sciences: The research process of music scholars. *Library & Information Science Research*, 24(1):73 – 94, 2002. ISSN 0740-8188. doi: http://dx.doi.org/10.1016/S0740-8188(01)00105-0. URL http://www.sciencedirect.com/science/article/pii/S0740818801001050.

[13] Monica E. Bulger, Eric T. Meyer, Grace De la Flor, Melissa Terras, Sally Wyatt, Marina Jirotka, Katherine Eccles, and Christine McCarthy Madsen. Reinventing research? information practices in the humanities. *SSRN eLibrary*, pages 1–83, March 2011. URL http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities.

[14] Laura Carletti, Gabriella Giannachi, and Derek McAuley. Digital humanities and crowdsourcing: An exploration. In *MW2013: Museums and the Web 2013*, 2013. URL

http://mw2013.museumsandtheweb.com/paper/digital-humanities-and-crowdsourcing-an-exploration-4/.

[15] G. Chiron, A. Doucet, M. Coustaty, M. Visani, and J. P. Moreux. Impact of OCR errors on the use of digital libraries: Towards a better access to information. In *JCDL 2017*, pages 1–4, June 2017. doi: 10.1109/JCDL.2017.7991582.

[16] Daniel Jared Cohen and Roy Rosenzweig. *Digital history: A guide to gathering, preserving, and presenting the past on the web*, volume 28. University of Pennsylvania Press Philadelphia, 2006.

[17] W. B. Croft, S. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. Technical report, University of Massachusetts, Amherst, MA, USA, 1993.

[18] Chris Dijkshoorn, Mieke H. R. Leyssen, Archana Nottamkandath, Jasper Oosterman, Myriam C. Traub, Lora Aroyo, Alessandro Bozzon, Wan Fokkink, Geert-Jan Houben, Henrike Hovelmann, Lizzy Jongma, Jacco van Ossenbruggen, Guus Schreiber, and Jan Wielemaker. Personalized nichesourcing: Acquisition of qualitative annotations from niche communities. In *6th International Workshop on Personalized Access to Cultural Heritage (PATCH 2013)*, pages 108–111, 2013. URL http://oai.cwi.nl/oai/asset/21391/21391B.pdf.

[19] Norbert Fuhr, Preben Hansen, Michael Mabe, Andras Micsik, and Ingeborg Sølvberg. Digital libraries: A generic classification and evaluation scheme. In Panos Constantopoulos and IngeborgT. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries*, volume 2163 of *Lecture Notes in Computer Science*, pages 187–199. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-42537-3. doi: 10.1007/3-540-44796-2_17. URL http://dx.doi.org/10.1007/3-540-44796-2_17.

[20] F. Galton. Vox populi. *Nature*, 75(1949):7, 1907. doi: 10.1038/075450a0.

[21] George Garvy. Inequality of income: Causes and measurement. In *Studies in Income and Wealth, Volume 15*, pages 25–48. NBER, 1952.

[22] Jennifer Golbeck, Jes Koepfler, and Beth Emmerling. An experimental study of social tagging behavior and image content. *Journal of the American Society for Information Science and Technology*, 62(9):1750–1760, 2011. ISSN 1532-2890. doi: 10.1002/asi.21522. URL http://dx.doi.org/10.1002/asi.21522.

[23] Jiyin He, Jacco van Ossenbruggen, and Arjen P. de Vries. Do you need experts in the crowd? a case study in image annotation for marine biology. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 57–60, Paris, France, France, 2013. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. URL http://dl.acm.org/citation.cfm?id=2491748.2491763.

[24] Kurtis Heimerl, Brian Gawalt, Kuang Chen, Tapan Parikh, and Björn Hartmann. Communitysourcing: engaging local crowds to perform expert work via physical kiosks. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pages 1539–1548, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi: 10.1145/2208516.2208619. URL http://doi.acm.org/10.1145/2208516.2208619.

[25] Bernd Heinrich, Marcus Kaiser, and Mathias Klier. How to measure data quality? a metric-based approach. In S. Rivard and J. Webster, editors, *Proceedings of the 28th International Conference on Information Systems (ICIS). Montreal, Queen's University*, 2007. URL http://epub.uni-regensburg.de/23633/.

[26] Jacqueline Hicks, Ridho Reinanda, and Vincent Traag. Old questions, new techniques: A research note on the computational identification of political elites. *Comparative Sociology*, 2015.

[27] Michiel Hildebrand, Jacco van Ossenbruggen, Lynda Hardman, and Geertje Jacobs. Supporting subject matter annotation using heterogeneous thesauri: A user study in web data reuse. *International Journal of Human-Computer Studies*, 67(10):887 – 902, 2009. ISSN 1071-5819. doi: http://dx.doi.org/10.1016/j.ijhcs.2009.07.008. URL http://www.sciencedirect.com/science/article/pii/S1071581909000950.

[28] Rose Holley. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4), 2009. doi: 10.1045/march2009-holley. URL http://dx.doi.org/10.1045/march2009-holley.

[29] Rose Holley. Many hands make light work: Public collaborative OCR text correction in Australian Historic Newspapers. Technical report, National Library of Australia, March 2009. URL http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.

[30] Mehdi Hosseini, IngemarJ. Cox, Natasa Milic-Frayling, Gabriella Kazai, and Vishwa Vinay. On aggregating labels from multiple crowd workers to infer relevance of documents. In Ricardo Baeza-Yates, ArjenP. Vries, Hugo Zaragoza, B.Barla Cambazoglu,

Vanessa Murdock, Ronny Lempel, and Fabrizio Silvestri, editors, *Advances in Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 182–194. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28996-5. doi: 10.1007/978-3-642-28997-2_16. URL http://dx.doi.org/10.1007/978-3-642-28997-2_16.

[31] Nancy Ide and Keith Suderman. The linguistic annotation framework: a standard for annotation interchange and merging. *Language Resources and Evaluation*, 48(3):395–418, 2014.

[32] Kimmo Kettunen, Timo Honkela, Krister Lindén, Pekka Kauppinen, Tuula Pääkkönen, Jukka Kervinen, et al. Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods. In *IFLA World Library and Information Congress Proceedings 80th IFLA General Conference and Assembly*, 2014.

[33] Martijn Kleppe, Laura Hollink, Max Kemman, Damir Juric, Henri Beunders, Jaap Blom, Johan Oomen, and Geert-Jan Houben. Polimedia-analysing media coverage of political debates by automatically generated links to radio newspaper items. In *LinkedUp Veni Competition on Linked and Open Data for Education*. CEUR-WS, 2014.

[34] Edwin Klijn. The current state-of-art in newspaper digitization a market perspective, January 2008. URL http://www.dlib.org/dlib/january08/klijn/01klijn.html.

[35] David Mimno. Computational historiography: Data mining in a century of classics journals. *J. Comput. Cult. Herit.*, 5(1):3:1–3:19, apr 2012. ISSN 1556-4673. doi: 10.1145/2160165.2160168. URL http://doi.acm.org/10.1145/2160165.2160168.

[36] Elke Mittendorf and Peter Schäuble. Information retrieval can cope with many errors. *Inf. Retr.*, 3(3):189–216, oct 2000. ISSN 1386-4564. doi: 10.1023/A:1026564708926. URL http://dx.doi.org/10.1023/A:1026564708926.

[37] Günter Mühlberger, Johannes Zelger, and David Sagmeister. User-driven correction of OCR errors: Combining crowdsourcing and information retrieval technology. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 53–56, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2588-2. doi: 10.1145/2595188.2595212. URL http://doi.acm.org/10.1145/2595188.2595212.

[38] Bob Nicholson. Counting culture; or, how to read Victorian newspapers from a distance. *Journal of Victorian Culture*, 17(2):238–246, 2012. doi: 10.1080/13555502.2012.683331. URL http://dx.doi.org/10.1080/13555502.2012.683331.

[39] M. Ohta, A. Takasu, and J. Adachi. Retrieval methods for english-text with missrecognized ocr characters. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, volume 2, pages 950–956 vol.2, Aug 1997. doi: 10.1109/ICDAR.1997.620651.

[40] Ken Orr. Data quality and systems theory. *Commun. ACM*, 41 (2):66–71, feb 1998. ISSN 0001-0782. doi: 10.1145/269012.269023. URL http://doi.acm.org/10.1145/269012.269023.

[41] Thaer Samar, Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman, and Arjen P. de Vries. Quantifying retrieval bias in web archive search. *International Journal on Digital Libraries*, Apr 2017. ISSN 1432-1300. doi: 10.1007/s00799-017-0215-9. URL https://doi.org/10.1007/s00799-017-0215-9.

[42] Carolyn Strange, Daniel McNamara, Josh Wodak, and Ian Wood. Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8(1), 2014.

[43] Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva. The effects of noisy data on text retrieval. *J. Am. Soc. Inf. Sci.*, 45(1):50–58, jan 1994. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(199401)45:1<50::AID-ASI6>3.0.CO;2-B. URL http://dx.doi.org/10.1002/(SICI)1097-4571(199401)45:1<50::AID-ASI6>3.0.CO;2-B.

[44] Kazem Taghva, Julie Borsack, and Allen Condit. Effects of ocr errors on ranking and feedback using the vector space model. *Information Processing & Management*, 32(3):317 – 327, 1996. ISSN 0306-4573. doi: https://doi.org/10.1016/0306-4573(95)00058-5. URL http://www.sciencedirect.com/science/article/pii/0306457395000585.

[45] Kazem Taghva, Russell Beckley, and Jeffrey Coombs. The effects of OCR error on the extraction of private information. In Horst Bunke and A.Lawrence Spitz, editors, *Document Analysis Systems VII*, volume 3872 of *Lecture Notes in Computer Science*, pages 348–357. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-32140-8. doi: 10.1007/11669487_31. URL http://dx.doi.org/10.1007/11669487_31.

[46] Timothy R. Tangherlini and Peter Leonard. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, 41(6):725 – 749, 2013. ISSN 0304-422X. doi: http://dx.doi.org/10.1016/j.poetic.2013.08.002. URL http://www.sciencedirect.com/science/article/pii/S0304422X13000648. Topic Models and the Cultural Sciences.

[47] Simon Tanner, Trevor Muñoz, and Pich Hemy Ros. Measuring mass text digitization quality and usefulness. *D-Lib Magazine*, 15 (7/8):1082–9873, 2009.

[48] Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. Impact analysis of ocr quality on research tasks in digital archives. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries*, volume 9316 of *Lecture Notes in Computer Science*, pages 252–263. Springer International Publishing, 2015. ISBN 978-3-319-24591-1. doi: 10.1007/978-3-319-24592-8_19. URL http://dx.doi.org/10.1007/978-3-319-24592-8_19.

[49] Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. Querylog-based assessment of retrievability bias in a large newspaper corpus. In *Proceedings of the Joint Conference on Digital Libraries 2016*, 2016.

[50] Luis von Ahn and Laura Dabbish. ESP: Labeling images with a computer game. In *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*, pages 91–98. AAAI, 2005. URL http://dblp.uni-trier.de/db/conf/aaaiss/aaaiss2005-3.html#AhnD05.

[51] Anna Weymann, Rodrigo A. Luna Orozco, Christoph Mueller, Bertram Nickolay, Jan Schneider, and Kathrin Barzik. *Einführung in die Digitalisierung von gedrucktem Kulturgut - Ein Handbuch für Einsteiger*. Ibero-American Institute (Berlin), 2010. ISBN ISBN 978-3-935656-40-8. URL http://www.iai.spk-berlin.de/fileadmin/dokumentenbibliothek/handbuch/Handbuch_Digitalisierung_IAI_IPK_Online_druck.pdf.

[52] Colin Wilkie and Leif Azzopardi. A retrievability analysis: Exploring the relationship between retrieval bias and retrieval performance. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 81–90, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. doi: 10.1145/2661829.2661948. URL http://doi.acm.org/10.1145/2661829.2661948.

[53] Colin Wilkie and Leif Azzopardi. Best and fairest: An empirical analysis of retrieval system bias. In Maarten de Rijke, Tom Kenter, ArjenP. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 13–25. Springer International Publishing, 2014. ISBN 978-3-319-06027-9. doi: 10.1007/978-3-319-06028-6_2. URL http://dx.doi.org/10.1007/978-3-319-06028-6_2.

[54] Colin Wilkie and Leif Azzopardi. Efficiently estimating retrievability bias. In Maarten de Rijke, Tom Kenter, ArjenP. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 720–726. Springer International Publishing, 2014. ISBN 978-3-319-06027-9. doi: 10.1007/978-3-319-06028-6_82. URL http://dx.doi.org/10.1007/978-3-319-06028-6_82.

[55] Colin Wilkie and Leif Azzopardi. Efficiently estimating retrievability bias. In Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 720–726. Springer International Publishing, Cham, 2014. ISBN 978-3-319-06028-6. doi: 10.1007/978-3-319-06028-6_82. URL https://doi.org/10.1007/978-3-319-06028-6_82.

[56] Colin Wilkie and Leif Azzopardi. Retrievability and retrieval bias: A comparison of inequality measures. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*, pages 209–214. Springer International Publishing, 2015. ISBN 978-3-319-16353-6. doi: 10.1007/978-3-319-16354-3_22. URL http://dx.doi.org/10.1007/978-3-319-16354-3_22.

[57] Sjaak Wouters. Semi-automatic annotation of artworks using crowdsourcing. Master's thesis, Vrije Universiteit Amsterdam, The Netherlands, 2012.

[58] Hong (Iris) Xie. Evaluation of digital libraries: Criteria and problems from users' perspectives. *Library & Information Science Research*, 28(3):433 – 452, 2006. ISSN 0740-8188. doi: http://dx.doi.org/10.1016/j.lisr.2006.06.002. URL http://www.sciencedirect.com/science/article/pii/S0740818806000697.

[59] Hong Iris Xie. Users' evaluation of digital libraries (dls): Their uses, their criteria, and their assessment. *Inf. Process. Manage.*, 44(3):1346–1373, may 2008. ISSN 0306-4573. doi: 10.1016/j.ipm.2007.10.003. URL http://dx.doi.org/10.1016/j.ipm.2007.10.003.

[60] Shlomo Yitzhaki. Relative deprivation and the Gini coefficient. *The Quarterly Journal of Economics*, 93(2):pp. 321–324, 1979. ISSN 00335533. URL http://www.jstor.org/stable/1883197.

ACKNOWLEDGEMENTS

"Kokoro wa hanatan koto wo yosu."
(Be ready to free your mind.)

— Gichin Funakoshi

This work would not have been possible without the support of many people. I would like to thank

Maarten for your encouragement to do a PhD, your friendship and your mentoring.

CWI and all its employees for making it an inspiring place to do research, learn and grow.

The Information Access group at CWI, and in particular Thaer, Gebre, Emma, Alex, Desmond, Lilia, Martine and Mieke for the chocolate breaks, the mutual support and the numerous table football matches.

The SEALINCMedia / LinkedTV project teams, and in particular Archana and Chris, for the stimulating collaboration.

The National Library of the Netherlands, the Rijksmuseum Amsterdam and the Netherlands Institute for Sound and Vision for supporting my research and giving me the opportunity to share and discuss my results.

My co-authors for your valuable contributions to our joint publications.

My employers bd4travel and codecentric AG for giving me the time and space to finish the last bits of this project.

Antoinette for your time and advice.

My assessment committee consisting of Katriina Byström, Sally Wyatt, Joris van Eijnatten, Toine Pieters, and Arno Siebes.

My supervisors Lynda and Jacco for your scientific guidance and the inspiring supervision meetings.

My friends Irini, Liliana, Hugo, Astrid, Claudia, Claudia, Corinna, Barbara and Peter for staying in touch despite long stretches of silence from my side, for lending an ear when I needed it, and of course for the countless visits to museums. Special thanks goes to Hugo for translating the Samenvatting, to Peter for proofreading and for the impact therapy trainings, and to Liliana and Astrid for supporting me as paranimfs.

My parents Irmgard and Ewald for always believing in me and for your unfailing support during difficult times.

Joachim for your patience, humour and love. I could not have done this without you.

Apologies if I have missed out anyone.

As I am making the final adjustments to this thesis, the entire world is in an exceptional state and it is uncertain when we will be able to meet and celebrate together. I hope this day will come soon and I will do my best to make the party worth the wait.

SIKS DISSERTATIONS

SIKS dissertations published since 2011:

2011

01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

09 Tim de Jong (OU), Contextualised Mobile Media for Learning

10 Bart Bogaert (UvT), Cloud Content Contention

11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective

12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval

16 Maarten Schadd (UM), Selective Search in Games of Different Complexity

17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness

18 Mark Ponsen (UM), Strategic Decision-Making in complex games

19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles

20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

22 Junte Zhang (UVA), System Evaluation of Archival Description and Access

23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media

24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics

26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

29 Faisal Kamiran (TUE), Discrimination-aware Classification

30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

33 Tom van der Weide (UU), Arguing to Motivate Decisions

34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization

39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications

05 Marijn Plomp (UU), Maturing Interorganisational Information Systems

06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms

10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems

13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems

15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.

16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance

18 Eltjo Poort (VU), Improving Solution Architecting Practices

19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution

20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval

22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval

25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

26 Emile de Maat (UVA), Making Sense of Legal Text

27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games

28 Nancy Pascall (UvT), Engendering Technology Empowering Women

29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval

30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making

31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure

32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning

33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)

34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications

35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics

36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes

37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation

38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms

39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks

40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia

41 Sebastian Kelle (OU), Game Design Patterns for Learning

42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning

43 Withdrawn

44 Anna Tordai (VU), On Combining Alignment Techniques

45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions

46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation

47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior

48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data

49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions

50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering

51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013

01 Viorel Milea (EUR), News Analytics for Financial Decision Support

02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing

03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics

04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling

05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns

06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience

07 Giel van Lankveld (UvT), Quantifying Individual Player Differences

08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators

09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications

10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.

11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services

12 Marian Razavian (VU), Knowledge-driven Migration to Services

13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly

14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning

15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications

16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation

17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid

18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification

19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling

20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval

21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation

22 Tom Claassen (RUN), Causal Discovery and Logic

23 Patricio de Alencar Silva (UvT), Value Activity Monitoring

24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning

25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System

26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning

27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance

28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience

29 Iwan de Kok (UT), Listening Heads

30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support

31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications

32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere

34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search

35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction

36 Than Lam Hoang (TUe), Pattern Mining in Data Streams

37 Dirk Börner (OUN), Ambient Learning Displays

38 Eelco den Heijer (VU), Autonomous Evolutionary Art

39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems

40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games

41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning

43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014

01 Nicola Barile (UU), Studies in Learning Monotone Models from Data

02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method

03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions

04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability

06 Damian Tamburri (VU), Supporting Networked Software Development

07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints

09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

10 Ivan Salvador Razo Zapata (VU), Service Value Networks

11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support

12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control

13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

14 Yangyang Shi (TUD), Language Models With Meta-information

15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations

19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support

20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link

21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments

22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training

23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era

24 Davide Ceolin (VU), Trusting Semi-structured Web Data

25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction

26 Tim Baarslag (TUD), What to Bid and When to Stop

27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty

28 Anna Chmielowiec (VU), Decentralized k-Clique Matching

29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software

30 Peter de Cock (UvT), Anticipating Criminal Behaviour

31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support

32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data

33 Tesfa Tegegne (RUN), Service Discovery in eHealth

34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.

35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach

36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models

37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying

38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.

39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital

40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education

41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text

42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models

43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments

44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.

45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach

46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity

47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015

01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response

02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls

03 Twan van Laarhoven (RUN), Machine learning for network data

04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments

05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding

06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes

07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis

08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions

09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems

10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning

11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins

12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks

13 Giuseppe Procaccianti (VU), Energy-Efficient Software

14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations

15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation

16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork

17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs

18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories

19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners

20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination

21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning

22 Zhemin Zhu (UT), Co-occurrence Rate Networks

23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage

24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation

25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection

26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure

27 Sándor Héman (CWI), Updating compressed colomn stores

28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO

29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains

30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning

31 Yakup Koç (TUD), On the robustness of Power Grids

32 Jerome Gard (UL), Corporate Venture Management in SMEs

33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources

34 Victor de Graaf (UT), Gesocial Recommender Systems

35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016

01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UVA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web

19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UVA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UVA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (UvT), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UVA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TUE), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UVA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VU), Logics for causal inference under uncertainty

23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People’s Preferences, Perceptions and Behaviors

28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (UvT), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018

01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TUE), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TUE), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech

30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data

06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VU), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming

36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TUE), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam C. Traub (CWI, UU), Measuring Tool Bias & Improving Data Quality for Digital Humanities Research

“I’ve traveled many roads
And not all of them were good
The foolish ones taught more to me
Than the wise ones ever could”

— Calvin Russell, Crossroads

Contents

Cultural heritage institutions increasingly make their collections digitally available. Consequently, users of digital archives need to familiarize themselves with new kinds of different digital tools. This is particularly true for humanities scholars who include results of their analyses in their publications. Judging whether insights derived from these analyses constitute a real trend or whether a potential conclusion is just an artifact of the tools used, can be difficult.

To correct errors in data, human input is in many cases still indispensable. Since experts are expensive, we conducted a study showing how crowdsourcing tasks can be designed to allow lay users to contribute information at the expert level to increase the number and quality of descriptions of collection items. However, to improve the quality of their data effectively, data custodians need to understand the (search) tasks their users perform and the level of trustworthiness they expect from the results. Through interviews with historians, we studied their use of digital archives and classified typical research tasks and their requirements for data quality.

Most archives provide, at best, very generic information about the data quality of their digitized collections. Humanities scholars, however, need to be able to assess how data quality and inherent bias within tools influence their research tasks. Therefore, they need specific information on the data quality of the subcollection used and the biases the tools provided may have introduced into the analyses.

We studied whether access to a historic newspaper archive is biased, and which types of documents benefit from, or are disadvantaged, by the bias. Using real and simulated search queries and page view data of real users, we investigated how well typical retrievability studies reflect the users' experience. We discovered large differences in the characteristics of the query sets and in the results for different parameter settings of the experiments.

Within digital archives, OCR errors are a prevalent data quality issue. Since these are relatively easy to spot, it has caused some concern about the trustworthiness of results based on digitized documents. We evaluated the impact of OCR quality on retrieval tasks, and studied the effect of manually improving (parts of) a collection on retrievability bias.

The insights we gained helped us understanding researchers' needs better. Our work provides a small number of examples, which demonstrate that data quality and tool bias are real concerns to the Digital Humanities community. To address these challenges, intense multidisciplinary exchange is required:

- **Humanities scholars** need to enhance the awareness that software tools and data sets are not free of bias and develop skills to detect and evaluate biases and their impact on research tasks. Guidelines should be developed that help scholars to perform tool criticism.
- **Tool developers** need to be more transparent and provide sufficient information about their tools to allow the task-based evaluation of their tools' performance.
- **Data custodians** need to make as much information about their collection available as possible. This should include which tools were used in the digitization process, in addition to both the limitations of the provided data and infrastructure used.

The goal should be a mutual understanding of each others' assumptions, approaches and requirements and more transparency concerning the use of tools in the processing of data. This will help scholars to develop effective methods of digital tool criticism to critically assess the impact of existing tools on their (re-)search results and to communicate on an equal footing with tool developers on how to develop future versions, which better suit their needs.

ISBN 978-3-00-065364-3