Into the bibliography jungle:

using random forests to predict dissertations' reference section

Silvia E. Gutiérrez De la Torre*, Julián Equihua**,
Andreas Niekler*, and Manuel Burghardt*
*University of Leipzig | **Helmholtz-Centre for Environmental Research

TABLE OF CONTENTS

01
02
03
04

01

02

03

04

01.

INTRODUCTION

Citation Analysis of Dissertations

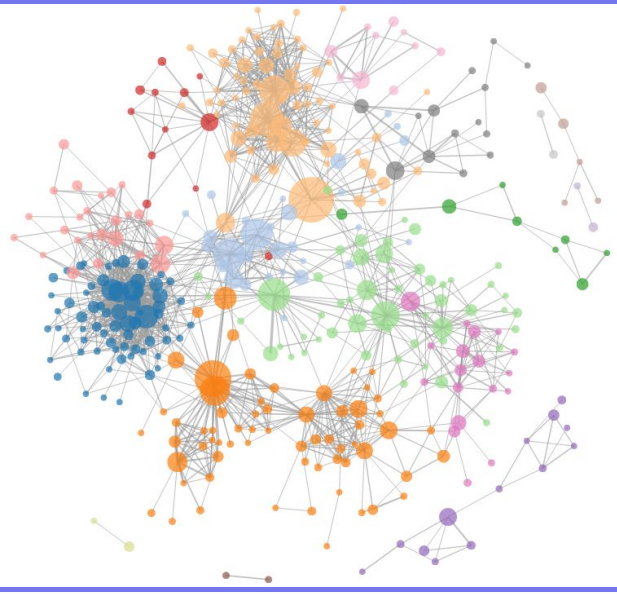
01

02

03

04

Citation Analysis (CA) 🎓



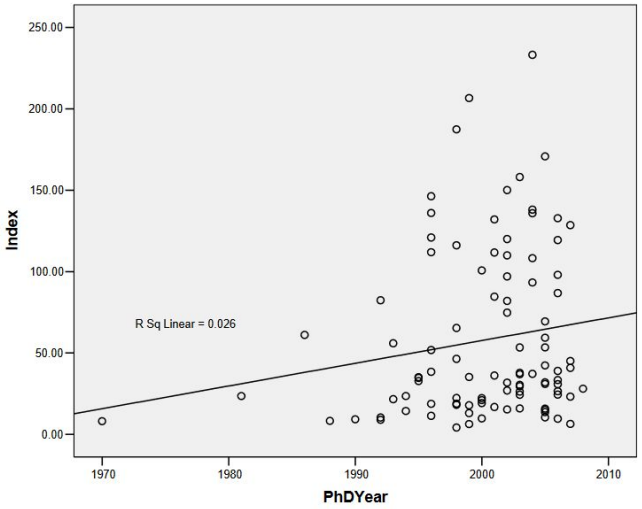


Figure 11. Regression analysis on interdisciplinarity index and year of graduation

Table 40. Descriptive statistics for the interdisciplinarity index

Statistic	Value
Mean	58.04
Median	35.23
Mode	N/A
Min.	4.19
Max	233.22
Standard Deviation	51.66

Figure 10 depicts how the dissertations were grouped according to their score on the interdisciplinarity index.

Interdisciplinarity

01

02

03

04

TABLE 1
Top 33 Core Journals Cited

Rank Core Journal	Citations (%)	Accum. Citations (%)	Journal Title	Impact Factor 2003	Rank Impact Factor 2003	Cost per Citation (€)
1	6.87	6.87	Journal of the American Chemical Society	6.516	4	0.20
2	5.28	12.15	Journal of Organic Chemistry	3.297	9	0.18
3	4.60	16.75	Tetrahedron Letters	2.326	15	1.09
4	2.52	19.27	Tetrahedron	2.641	12	2.27
5	2.03	21.30	Angewandte Chemie. International Edition in English	8.427	3	0.85
6	1.78	23.08	Chemosphere	1.904	22	1.08
7	1.47	24.55	Biochemistry	3.922	7	1.05
8	1.43	25.98	Journal of Medicinal Chemistry	4.820	5	0.55
9	1.36	27.34	Journal of the Electrochemical Society	2.361	13	0.26
10	1.31	28.65	Afinidad	0.157	31	0.01

Citation Analysis of Ph.D. Dissertation References as a Tool for Collection Management in an Academic Chemistry Library

Núria Vallmitjana and L. G. Sabaté

A bibliometric study was carried out on the citations within the chemistry field Ph.D. dissertations to ascertain what types of documents are the most frequently used in the research process, the most frequently consulted journals and obsolescence rate of the journals. The analysis covered 46 doctoral theses presented at the Institut Químic de Sarriá (IQS) from 1995 to 2003. The results obtained from the 4,203 citations revealed that the most frequently used documents were scientific papers, which accounted for 79 percent of the total; 33 journals met 50 percent of the informational needs; and the age of 50 percent of the citations was no older than 9 years. Finally, the results can be used as a tool for the collection management of the library.

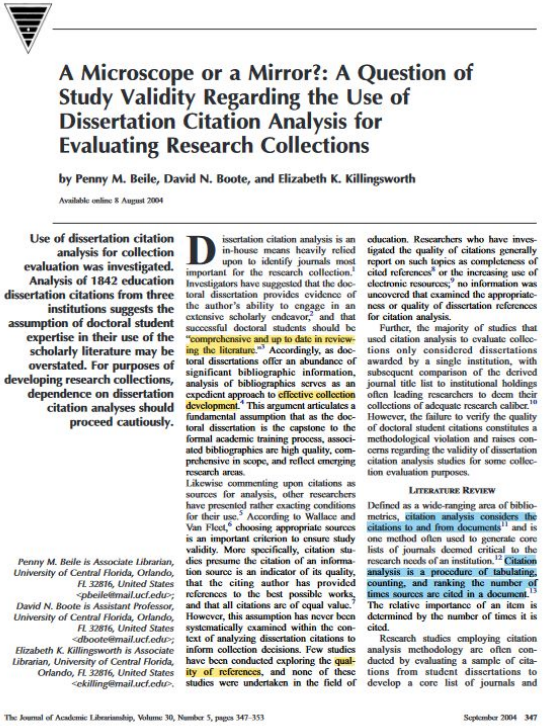
Subject area / Source freq.

01

02

03

04



A Microscope or a Mirror?: A Question of Study Validity Regarding the Use of Dissertation Citation Analysis for Evaluating Research Collections

by Penny M. Beile, David N. Boote, and Elizabeth K. Killingsworth

Available online 8 August 2004

Use of dissertation citation analysis for collection evaluation was investigated. Analysis of 1842 education dissertation citations from three institutions suggests the assumption of doctoral student expertise in their use of the scholarly literature may be overstated. For purposes of developing research collections, dependence on dissertation citation analyses should proceed cautiously.

Dissertation citation analysis is an in-house means heavily relied upon to identify journals most important for the research collection.[1] Investigators have suggested that the doctoral dissertation provides evidence of the author's ability to engage in an extensive scholarly endeavor,[2] and that successful doctoral students should be "comprehensive and up to date in reviewing the literature."[3] Accordingly, as doctoral dissertations offer an abundance of significant bibliographic information, analysis of bibliographics serves as an expedient approach to effective collection development.[4] This argument articulates a fundamental assumption that as the doctoral dissertation is the capstone to the formal academic training process, associated bibliographies are high quality, comprehensive in scope, and reflect emerging research areas.

Likewise commenting upon citations as sources for analysis, other researchers have presented rather exacting conditions for their use.[5] According to Wallace and Van Fleet,[6] choosing appropriate sources is an important criterion to ensure study validity. More specifically, citation studies presume the citation of an information source is an indicator of its quality, that the citing author has provided references to the best possible works, and that all citations are of equal value.[7] However, this assumption has never been systematically examined within the context of analyzing dissertation citations to inform collection decisions. Few studies have been conducted exploring the quality of references, and none of these studies were undertaken in the field of education. Researchers who have investigated the quality of citations generally report on such topics as completeness of cited references[8] or the increasing use of electronic resources;[9] no information was uncovered that examined the appropriateness or quality of dissertation references for citation analysis.

Further, the majority of studies that used citation analysis to evaluate collections only considered dissertations awarded by a single institution, with subsequent comparison of the derived journal title list to institutional holdings often leading researchers to deem their collections of adequate research caliber.[10] However, the failure to verify the quality of doctoral student citations constitutes a methodological violation and raises concerns regarding the validity of dissertation citation analysis studies for some collection evaluation purposes.

Literature Review

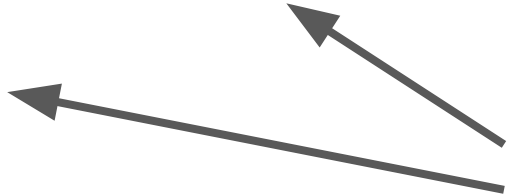
Defined as a wide-ranging area of bibliometrics, citation analysis considers the citations to and from documents[11] and is one method often used to generate core lists of journals deemed critical to the research needs of an institution.[12] Citation analysis is a procedure of tabulating, counting, and ranking the number of times sources are cited in a document.[13] The relative importance of an item is determined by the number of times it is cited.

Research studies employing citation analysis methodology are often conducted by evaluating a sample of citations from student dissertations to develop a core list of journals and

Penny M. Beile is Associate Librarian, University of Central Florida, Orlando, Fl. 32816, United States <pbeile@mail.ucf.edu>; David N. Boote is Assistant Professor, University of Central Florida, Orlando, Fl. 32816, United States <dboote@mail.ucf.edu>; Elizabeth K. Killingsworth is Associate Librarian, University of Central Florida, Orlando, Fl. 32816, United States <ekilling@mail.ucf.edu>.

study. As evidenced by student reliance on items of questionable value, the presumed quality of dissertation citations was not substantiated by this study.

"...the presumed quality of dissertation citations was not substantiated by this study."

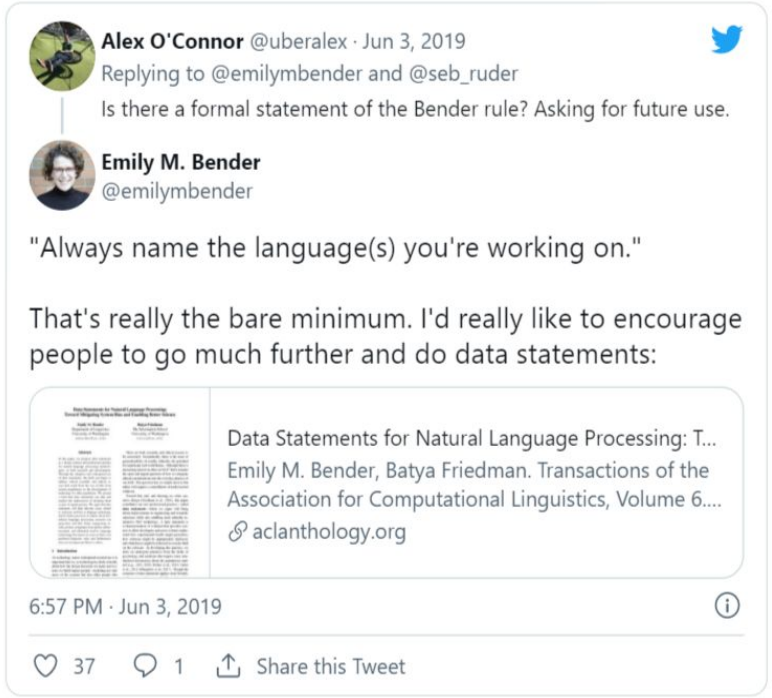


Collection Development

State of the art of RM tools

tool	update	repository	method	language	training_set	elds
Neural ParsCit	2019	https://github.c	LSTM	Python 2.7	English journal papers	uthors, booktitle, date, edi
GROBID	2021	https://github.c	CRF/LSTM	Java	Mostly English journal p	uthors, date, editor, issue,
ParsCit	2018	https://github.c	CRF	Perl	English journal papers	uthors, booktitle, date, edi
CERMINE	2018	https://github.c	CRF	Java	English journal papers	uthors, DOI, issue, pages,
Science Parse	2019	https://github.c	CRF	Java	English journal papers	uthors, title, volume, year,
AnyStyle Parser	2021	https://github.c	CRF	Ruby	English journal papers	uthors, booktitle, date, DC
Biblio	2004	http://search.c	regular expres	Perl	Mostly English journal p	uthors, date, editor, genre
citation	2016	https://github.c	regular expres	Ruby	English journal papers	uthors, title, URL, year
PDFSSA4MET	2013	https://github.c	regular expres	Python	English journal papers	ages, title, volume, year
Citation-Parser	2017	https://github.c	rules	Python 2.7	English journal papers	uthors, booktitle, issue, jo
BibPro	2012	https://github.c	template mat	Java	English journal papers	uthors, editor, institution, ssue, journal, pages, volun
free_cite	2008	https://github.c	CRF	Ruby	English journal papers	uthors, booktitle, date, edi
Reference Tagger	2018	https://github.c	CRF	Python 3.4	English journal papers	uthors, issue, journal, pag
SPV2	2019	https://github.c	LSTM	Python	English journal papers	uthors, title, venue, year

Bender Rule



Content credit: Emily M. Bender

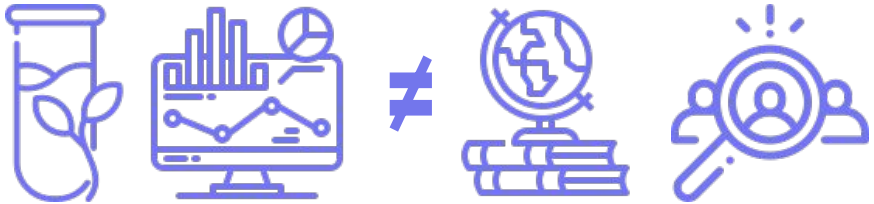
01

02

03

04

Domain



Gap icons created by Freepik & xnimrodx- & monkik Flaticoncc

01
02
03
04

Gold Standard

20 citation pages
~ 20 references
~ 400 ref. /dissertation

Gold icon created by Becris- Flaticon

non-English
humanities
book-length texts (PhD)

Gap icons created by Freepik - Flaticoncc

01
02
03
04

02.
Related Work

Which research questions am I trying to answer?

EXCITE

01

02

03

04

✔ Social Sciences

✔ non-English (German)

< 5 dissertations in GS

Opening Books and the National Corpus of Graduate Research

✔ Multidisciplinary

✔ PhD (✗ mostly English)

✗ no code release for RM

Dissertation icon created by eucalyp- Flaticon

Körner et al (2017)

✔ line-based conditional random field

✔ reduces model complexity

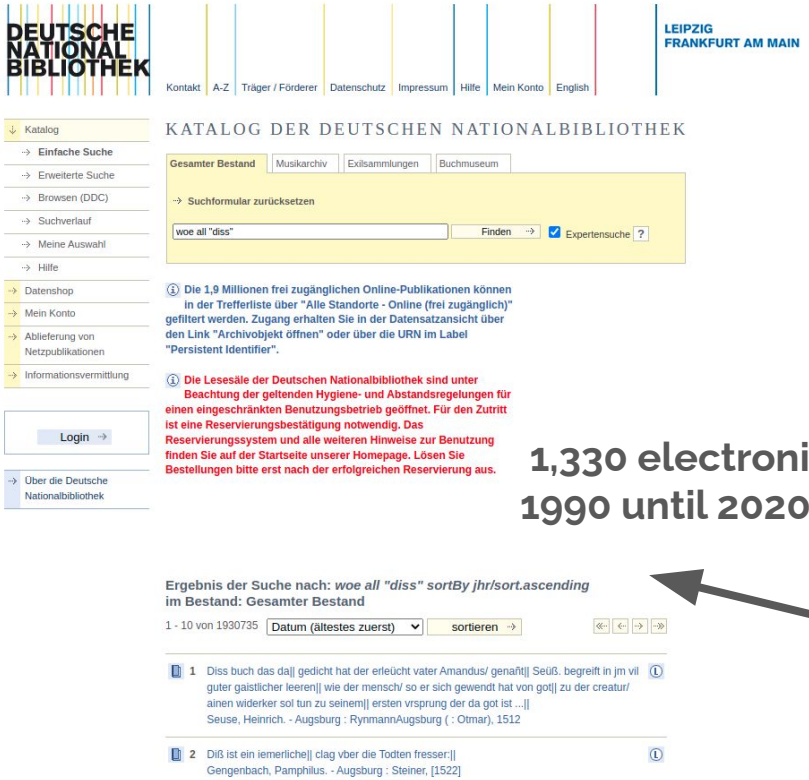
✗ hardly scalable for the ca. 12,000 lines long dissertations

Dissertation icon created by icongeek26- Flaticon

01
02
03
04

03. Methodology

How will I address this issue (computationally)?



1,330 electronic theses + metadata
1990 until 2020

Corpus

01
02
03
04

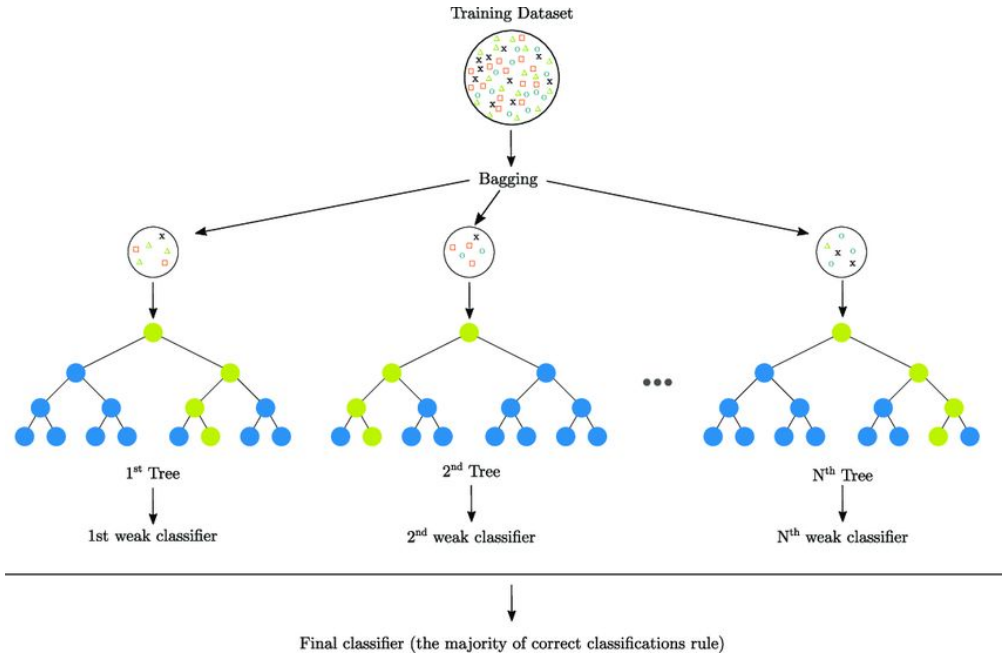
01

02

03

04

Random Forest



01

02

03

04

ɔgraphie

id (1930): *Meios de Expressão e Alterações Semânticas*. Ri
neiro

ter (1986): „Kontextualisierung“, in: *Studium Linguistik* 19, S. 2

do Nascimento, Maria Fernanda (1989): „Língua falada, Lí
crita – na investigação e na prática pedagógica“, in: *Congresso s*
investigação e ensino do português 1987. Actas. Lissabon: IC
107–111

Mrotzek (1992): *Diskursforschung und Kommunikation in Ins*
men. [Studienbibliographien Sprachwissenschaft] Heidelberg: Gr

n, Christiane (1991): *Modalpartikeln als Übersetzungsprob*
[eidelberger Beiträge zur Romanistik; 26] Frankfurt/M. u. a.: Lan

nn, J. (1981): „Ethnomethodologische Konversationsanalyse“
hröder/Steger (eds.), *Dialogforschung*. Jahrbuch 1980 des Inst
r deutsche Sprache. Düsseldorf, S. 9–51

ch, Manfred (1980): „Semantic Stucture and Illocutionary Force“
arle/Kiefer/Bierwisch (eds.), *Speech Act Theory and Pragma*
ordrecht, S. 1–35

Predict if page is part of the bibliography

Feature extraction

```
#List files
pdf_files <- list.files(path = "pdfs/", pattern = ".pdf$",
full.names = T)
# Read files
pdf_text <- purrr::map(pdf_files, .f = pdftools::pdf_text)
#Extract dates using regex
pdf_titles <- stringr::str_match(pdf_files,
"pdfs//(.*)\\..*$")[,2]
#Create dataframe with filenames and total occurrences of dates
thesis_dates <- purrr::map(pdf_text, ~
lapply(stringr::str_extract_all(string = .x, "([1-2][0-9]{3})"),
length))
df <- tibble(
  pdf_name = pdf_titles,
  dates = purrr::map(thesis_dates, .f=tibble))
#Create dataframe with proportional occurrences
df2 <- tibble(
  pdf_name = unlist(df["pdf_name"]),
  dates = unlist(df["dates"])/unlist(df["words"]))
```

Literaturverzeichnis (section headers)
Hrsg (abbreviations)
Frankfurt (publication places)
1993 (publication years)
ebd (negative feat., i.e. footnotes)
position (n_pag/total_pages)

Tagged dataset

dates_sw	pubplaces	page_n	bibtest
0.06211	0.01527	152	0
0.06338	0.00352	153	0
0.06609	0.00575	154	1
0.08397	0	155	1

Table 1
Example of tagged dataset to train RF model, each line is one page of the same PDF file

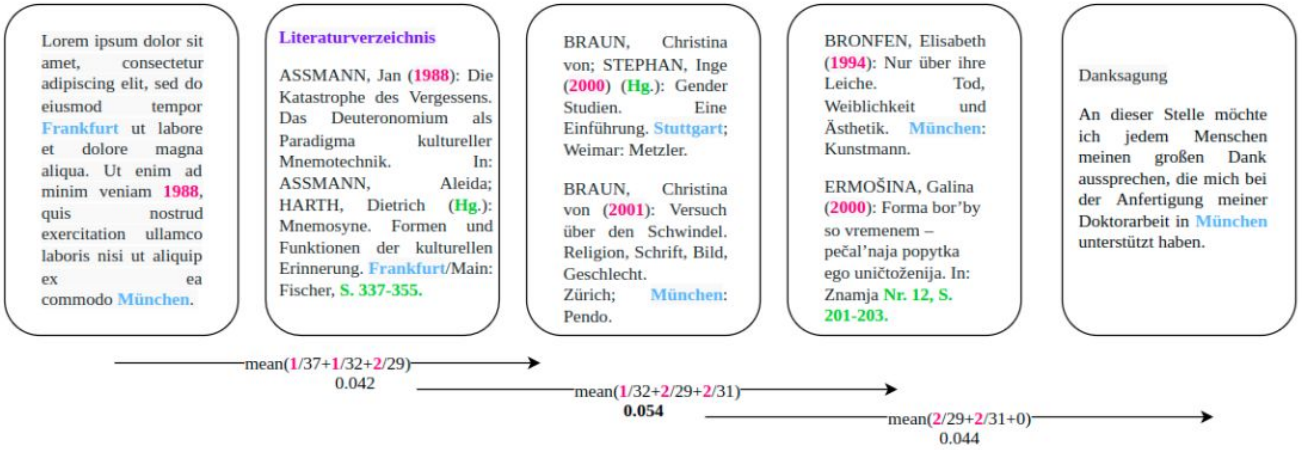


Figure 1: Example of sliding windows calculation.

Sliding Windows

01
02
03
04

04.

Results & next steps

What's next on this journey?

Features that help with prediction

feature	MDA
pubplaces_sw	32.3184181
dates_sw	24.87243224
dates	22.39062468
position	22.08909718
pubplaces	20.99889236

Table 2
Top 5 features by mean decrease accuracy (MDA)

01

02

03

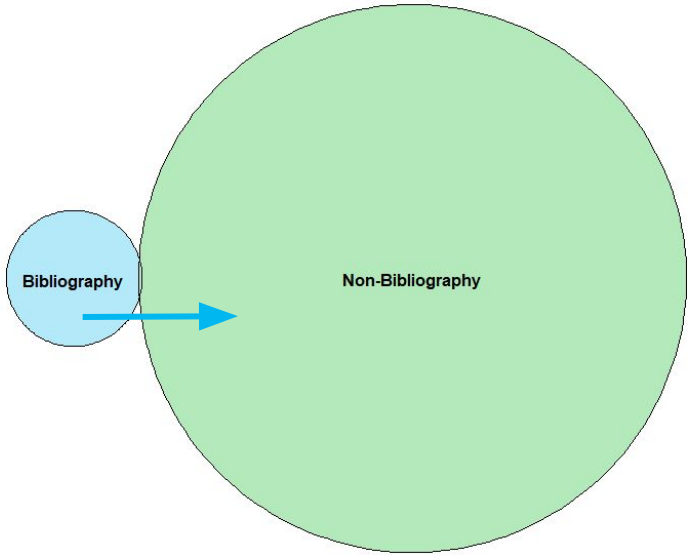
04

Confusion matrix

✔ 9 out of 1,145 reference section pages got erroneous predictions (false negative).

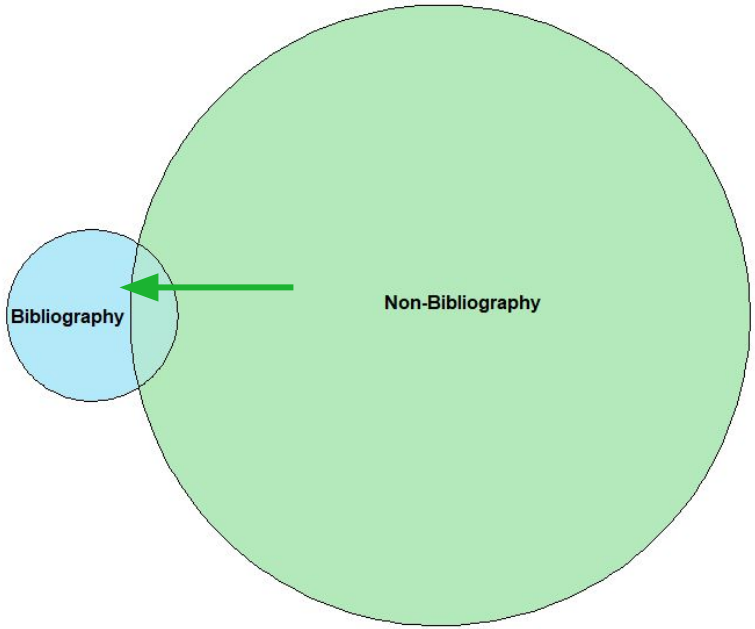
Confusion matrix

✔ 9 out of 1,145 reference section pages got erroneous predictions (false negative).



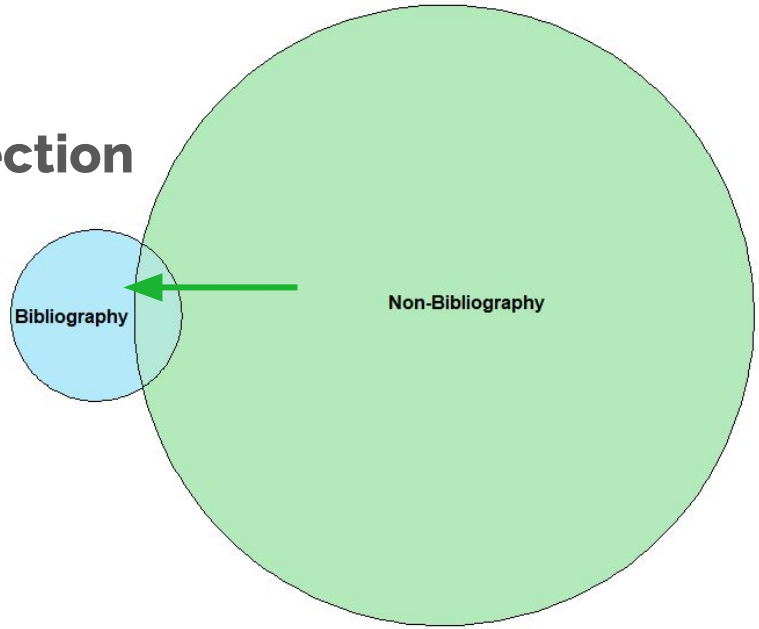
Confusion matrix

✗ 304 out of 18,544, false positives



Precision	0.79
Recall	0.99
F1	0.87

-Long footnotes
-SW: pages near ref. section
-Book lists



Future work

-New features (Author-name patterns)
-SHAP (SHapley Additive exPlanations)
-Regex (post-correction)

01

02

03

04

Danke!

Let's keep the conversation flowing
silviaegt@uni-leipzig.de