david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib

A corpus and bibliography of metalexicography

David Lindemann
Fritz Kliche
Ulrich Heid

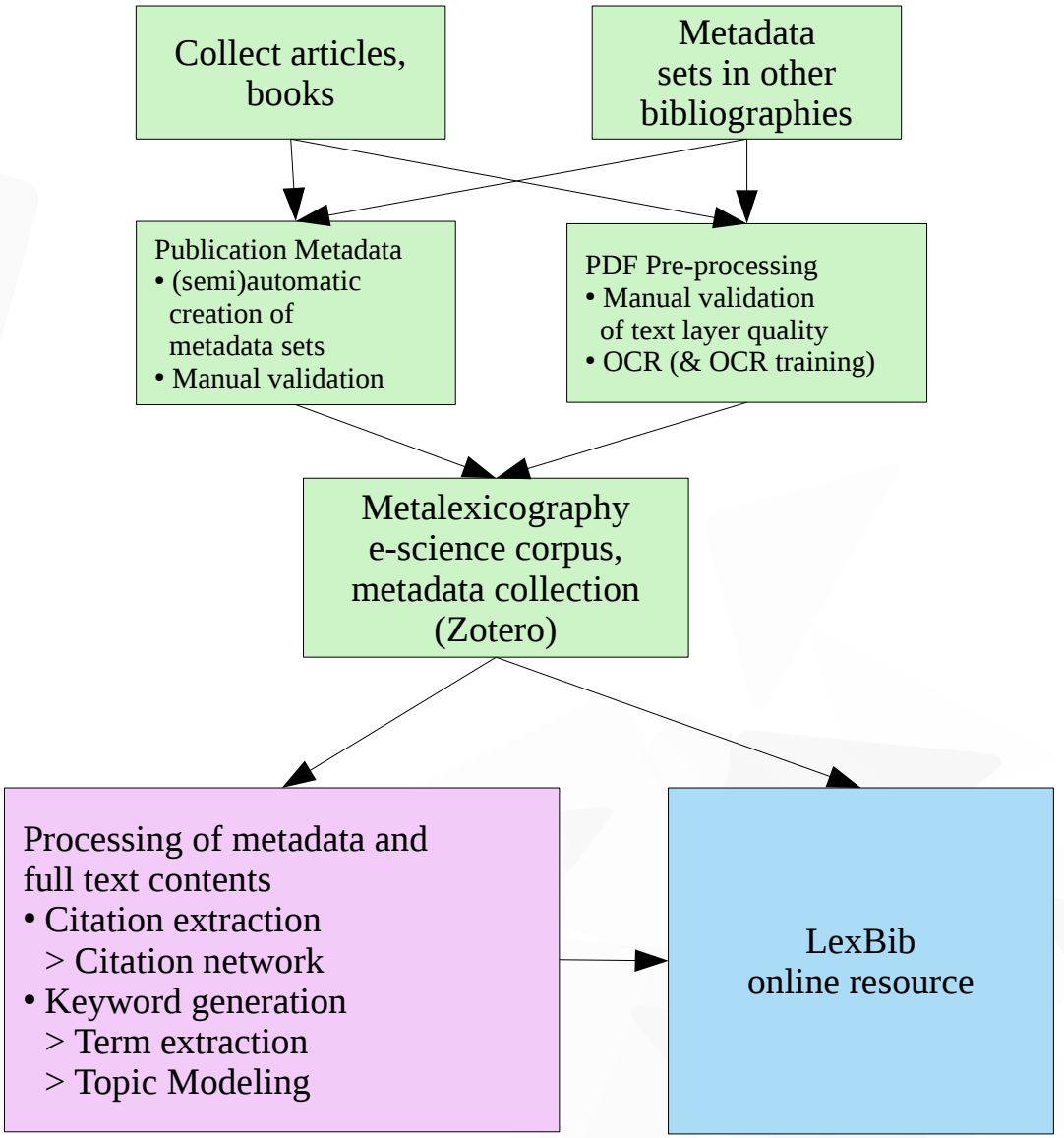
LexBib Goals

- Infrastructural aspect
  - Bibliography of Lexicography and Dictionary Research (Metalexicography)
  - complete, structured, validated
    - Full text collection (“e-Science Corpus”) > Computational text analytics
    - Publication metadata collection > Online-Bibliography
- Research Aspect
  - Computational methods, applied to full texts
    - Bibliometrics: Citation Extraction > Citation Network
    - Topic Modeling, Term Extraction > Keyword Indexation
    - Tests and Evaluation
  - Results: Additional publication metadata, to be included in online bibliography
- Motivation
  - Existing bibliographies far from state-of-the-art resources available for other disciplines
  - Hildesheim and IDS: Existing manually curated metadata collections and full text corpora as starting point
  - Relatively small discipline > Feasible for hand-validation and evaluation
    (predictions for the application of similar workflows in broader domains)

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib project: Main Modules

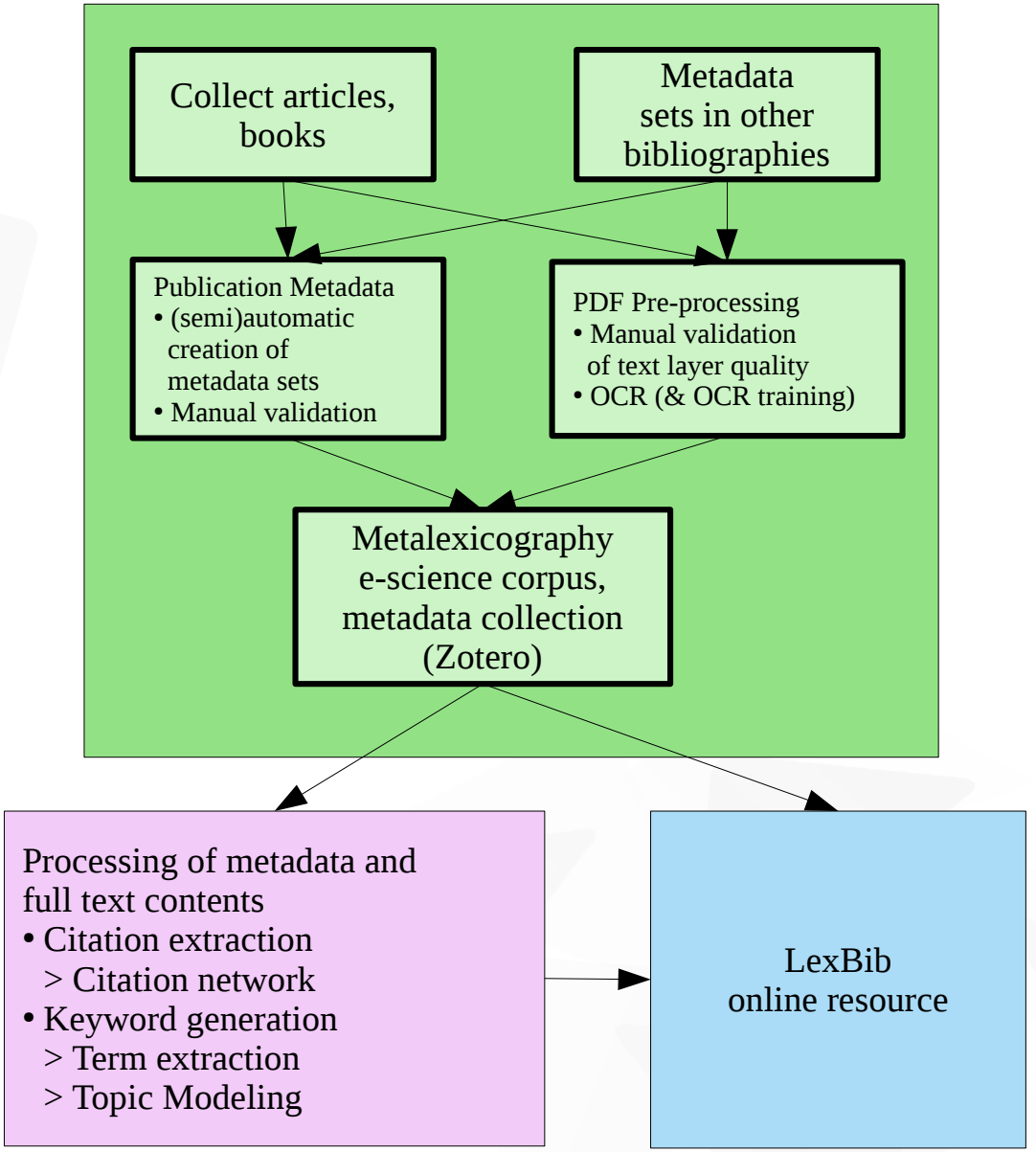
- Green: creation of e-science corpus, metadata and PDF pre-processing
  - Current LexBib collection: 2,500 items (July 2018)
- Red: Computational text analysis:
  - Creation of additional metadata
    - Citation Network
    - Keyword generation
- Blue: Creation of online bibliography
  - Models followed: DBLP, ACM



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib project: Module 1

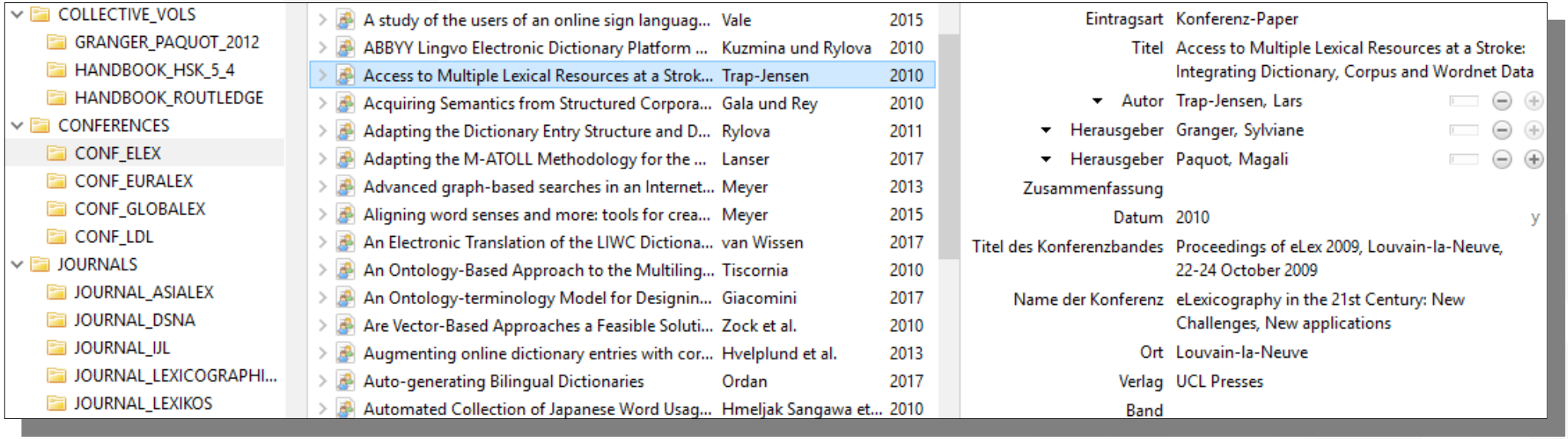
- Green: creation of e-science corpus, metadata and PDF pre-processing
  - 1st iteration: Spring 2018
    - 2,057 items
    - English texts 2000-2017
    - Core item types
      - Journal articles
      - Proceedings (eLex, Euralex)
      - Handbook articles



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

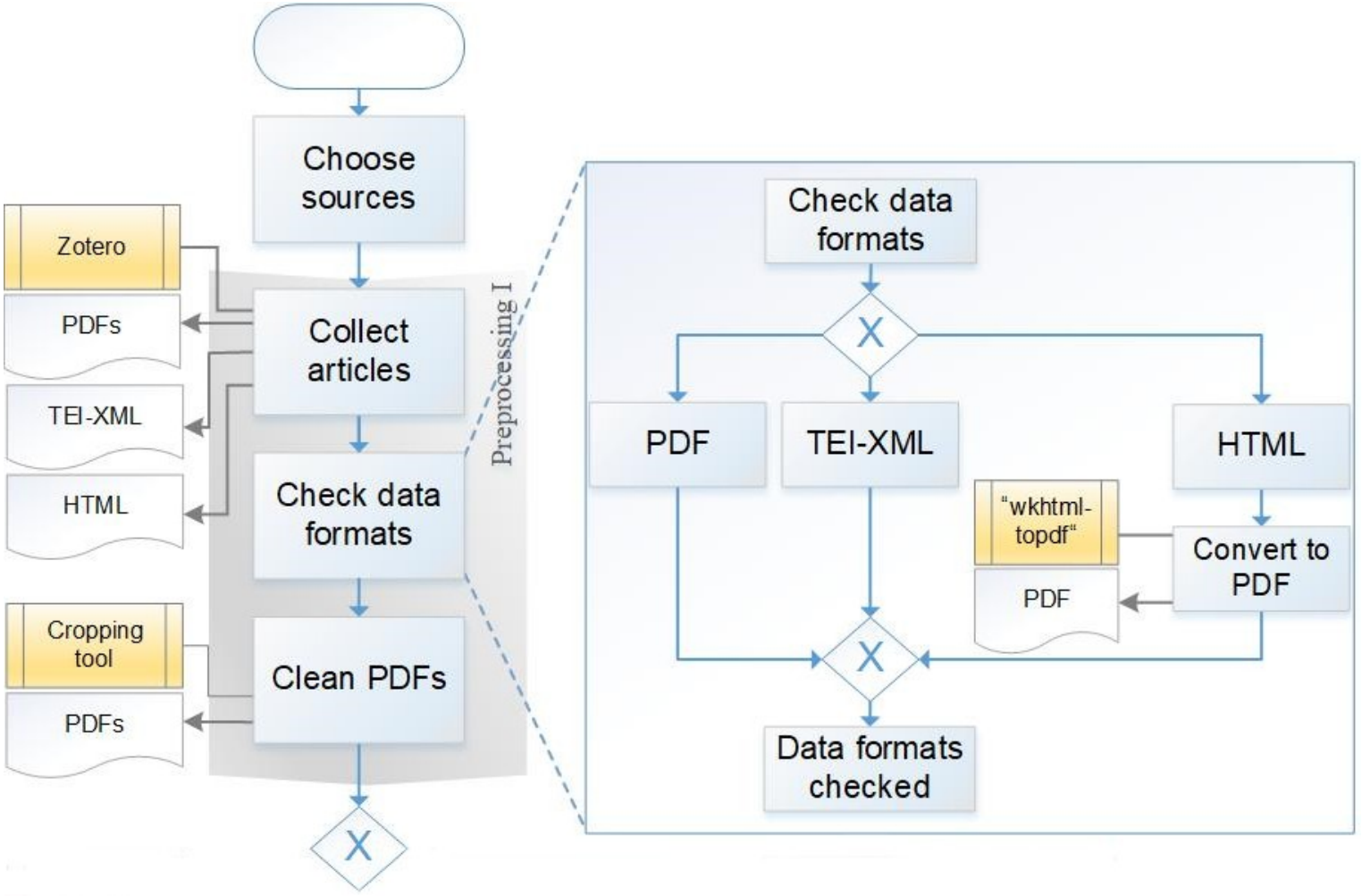
e-science corpus building: metadata & full text collection

- Collection of full texts and metadata using “Zotero”
  - From some online repositories (bulk) download of metadata and full texts possible
  - Zotero cloud: Enables collaborative editing of collections
  - Manual metadata validation/editing in Zotero GUI
  - Manual assessment of PDF quality
    - PDF manipulation: cropping of non-relevant content; running titles, page numbers, boilerplate
    - Re-digitization: OCR training



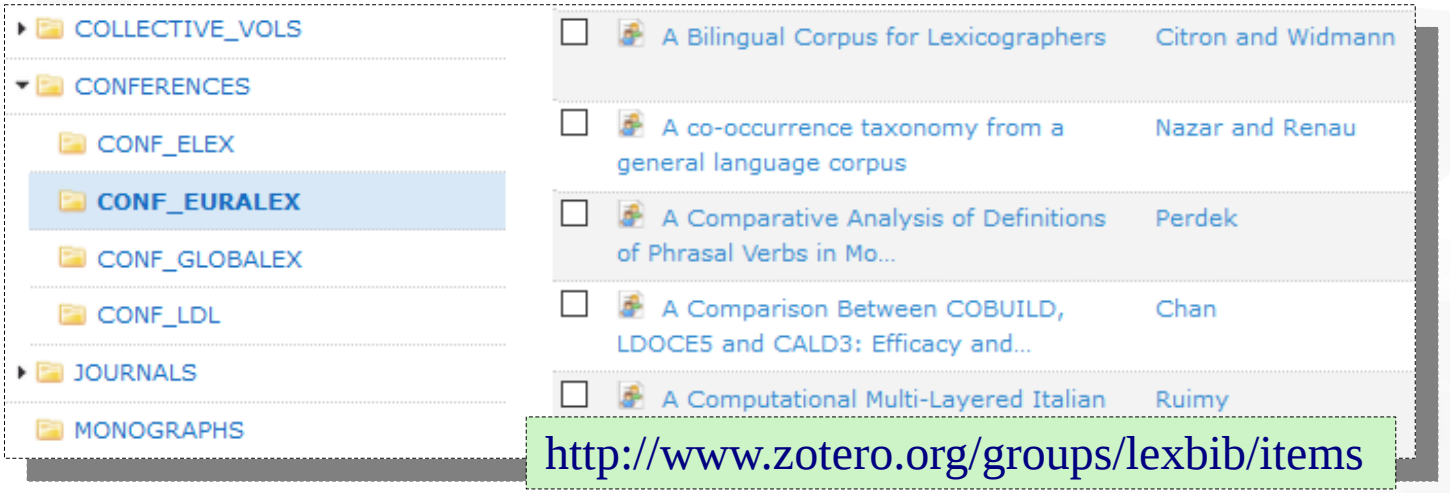
david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Processing Pipeline: Data collection



LexBib publication metadata

- Hand validated
  - Standard metadata (for citations), according to item type
  - Unique Identifiers: DOI, ISBN, ORCID
  - Item Relations “is_review_of“, “is_reviewed_by“
  - Item Relations “is_container_of“, “has_container“
  - Red module: Item Relations “citing“ and “is_cited_by“
- Compare and merge
  - LexBib_v1 (Lindemann et al. 2018)
  - IDS Obelex-Meta (Möhrs 2016)
  - Euralex-Bib (Dykstra & Hartmann)
  - WLWF-Bib (Kammerer, Gouws et al. in prep.)
  - Wiegand’s „Internationale Bibliographie“ (2006-2014)
  - Córdoba Rodríguez (2003)



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Some existing bibliographies

Title	*Scope (years)*	*Scope (domains)*	*Scope (languages)*	*# Items*	*Format*
LexBib Testset	2000-2017	Metalex	English	2,056	Structured database
EURALEX Bibliography	1600-2010	Lex/Metalex	Multiple	1,325	Unstructured (pub. as Wiki)
Obelex-meta	1982-2017	Metalex	Multiple	ca. 2,000	Structured database
WLWF	1420-2016	Lex/Metalex	Multiple	2,370	Unstructured (pub. as PDF)
Wiegand	1850-2014	Lex/Metalex	Multiple	33,339	Unstructured (pub. as PDF)
Hartmann	1930-2007	Metalex	Multiple	ca. 570	Unstructured (pub. as PDF)
Córdoba Rodríguez	1940-2003	Metalex	Multiple	10,192	Structured database
Ahumada	1535-2010	Metalex	Mainly Spanish	6,560	Structured database (in progress)

Wiegand: ‘Internationale Bibliographie der Lexikographie‘

- 33,000 items, 1850 to 2014
  - All standard item types + press articles; some with comments. Includes ref. to Zgusta 1988
- Lindemann & Khemakhem 2018 (in prep.): OCR, Parsing, repr. as XML

21923 **Voillat**, François: Le “Glossaire des patois de la Suisse romande” (GPSR). In: Actes du XVIII[e] Congrès International de Linguistique et de Philologie Romanes […] Tome VII, 1989↑, 338–345.

15475 **Nyhlén**, Lars-Olaf: Wie sagt man in Österreich? In: Moderna Språk 68. 1974, 275–281. [Bes. zu: Jakob Ebner: Duden. Wie sagt man in Österreich? Wörterbuch der österreichischen Besonderheiten. Mannheim 1969 (Duden Taschenbücher 8)].

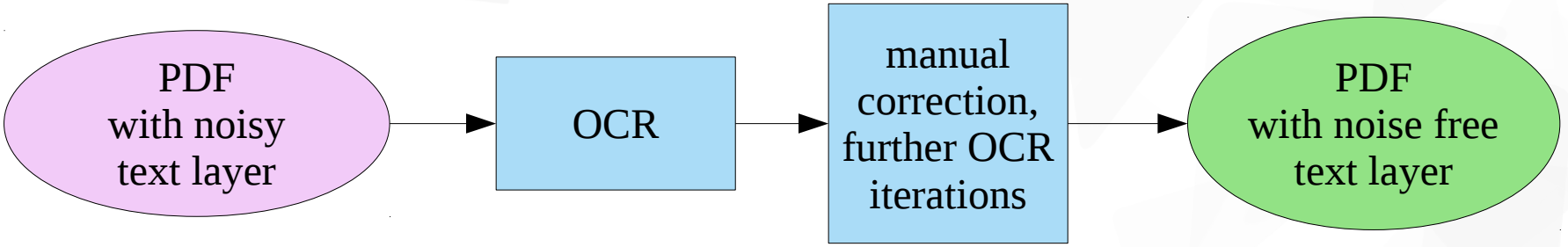
15397 Norden 11.–14. maj 1993. Red. af Anna Garde op Pia Jarvad. Oslo 1993 (Skrifter udgivet av Nordisk Forening for Leksikografi. Skrift nr. 2). [Daraus: 363, 1003, 2755, 3101, 3870, 5222, 5251, 5390, 6058, 6758, 9206, 9743, 10158, 11521, 11553, 11843, 12600, 12940, 12974, 13146, 13399, 14192, 15316, 15345, 15402, 15474, 15651, 15944, 17322, 20818, 21371].

15427 **Novikov**, L. A.: K probleme omonimii. [Zum Problem der Homonymie]. In: Leksikografičeskij sbornik 1960, 93–102.

Wiegand: ‘Internationale Bibliographie der Lexikographie‘

21923 **Voillat**, François: Le “Glossaire des patois de la Suisse romande” (GPSR). In: Actes du XVIII[e] Congrès International de Linguistique et de Philologie Romanes [...] Tome VII, 1989↑, 338–345.

Original Text Layer	21923·····Voillat,···F·r·a·n·c·i·s·:···Le···"Glossaire···des···patois···de···la···Suisse···r·o·m·a·n·d·e·"···(GPSR).···In:· Actes···du···X·v·i̊·i̊·Γ···Congres···International···de···Linguistique···et···de···Philologie···R·o·m·a·n·e·s····[...]· T·o·m·e··VII,···1·9·8·9·|·,···3·3·8·–·3·4·5·.
Transkribus OCR	21923·Voillat,·François:·Le·“Glossaire·des·patois·de·la·Suisse·romande”·(GPSR).·In: Actes·du·XVIIIe·Congrès·International·de·Linguistique·et·de·Philologie·Romanes·[...] Tome·VII,·1989Î,·338-345.
Manual correction	21923·Voillat,·François:·Le·“Glossaire·des·patois·de·la·Suisse·romande”·(GPSR).·In: Actes·du·XVIII[e]·Congrès·International·de·Linguistique·et·de·Philologie·Romanes·[...] Tome·VII,·1989↑,·338-345.



Wiegand: Internationale Bibliographie der Lexikographie

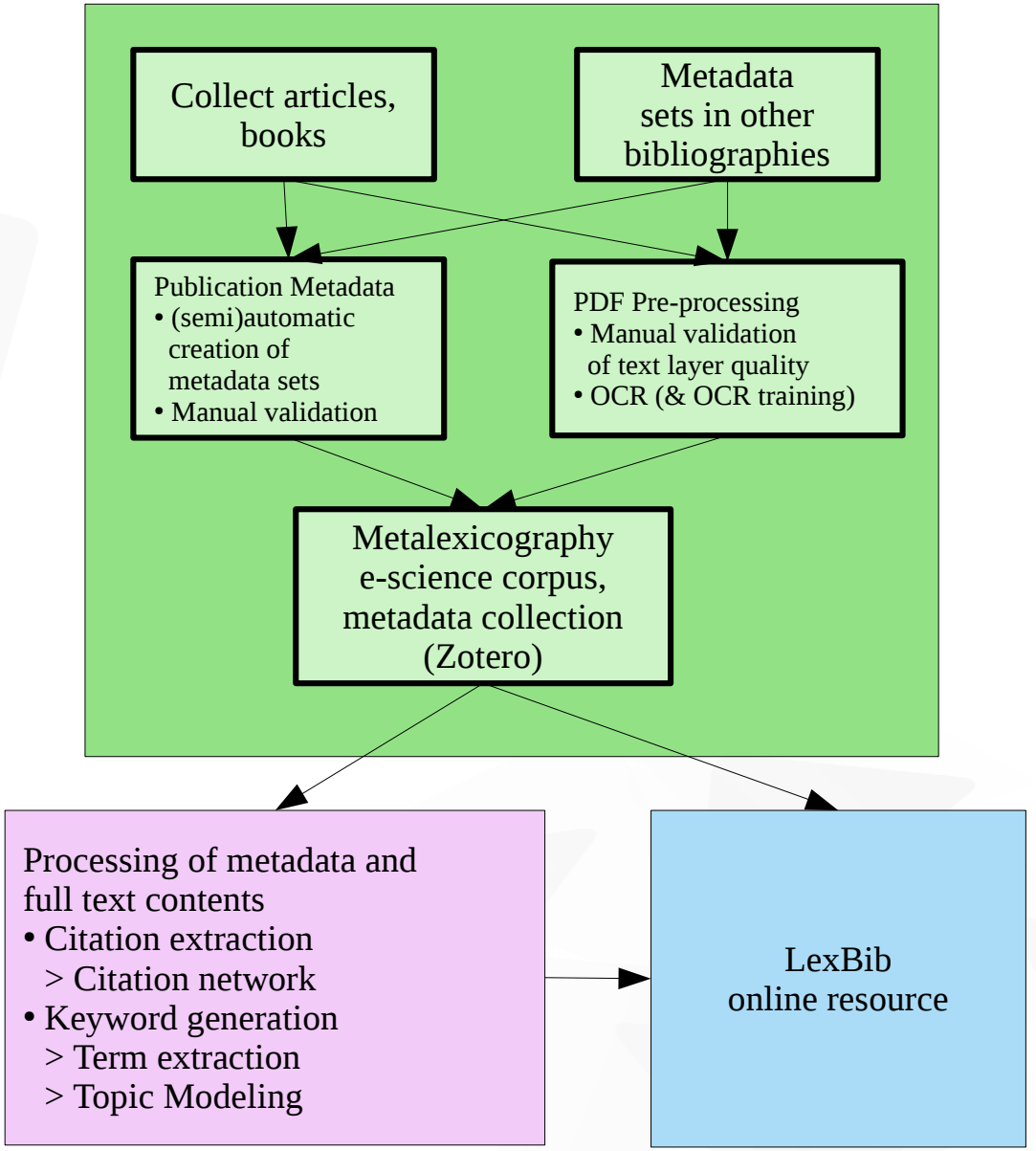
21923 **Voillat**, François: Le "Glossaire des patois de la Suisse romande" (GPSR). In: Actes du XVIII[e] Congrès International de Linguistique et de Philologie Romanes [...] Tome VII, 1989↑, 338–345.

```
<biblStruct xml:id="Wiegand21923">
    <analytic>
        <author>
            <persName>
                <surname>Voillat</surname><forename>François</forename>
            </persName>
        </author>
        <title level="a">Le "Glossaire des patois de la Suisse romande" (GPSR)</title>
    </analytic>
    <monogr>
        <title level="m">Actes du XVIIIe Congrès International de Linguistique et de
        Philologie Romanes [...] Tome VII</title>
        <imprint><date when="1989"/></imprint>
        <biblScope unit="pp" from="338" to="345">338-345</biblScope>
    </monogr>
</biblStruct>
```

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib project: Module 1

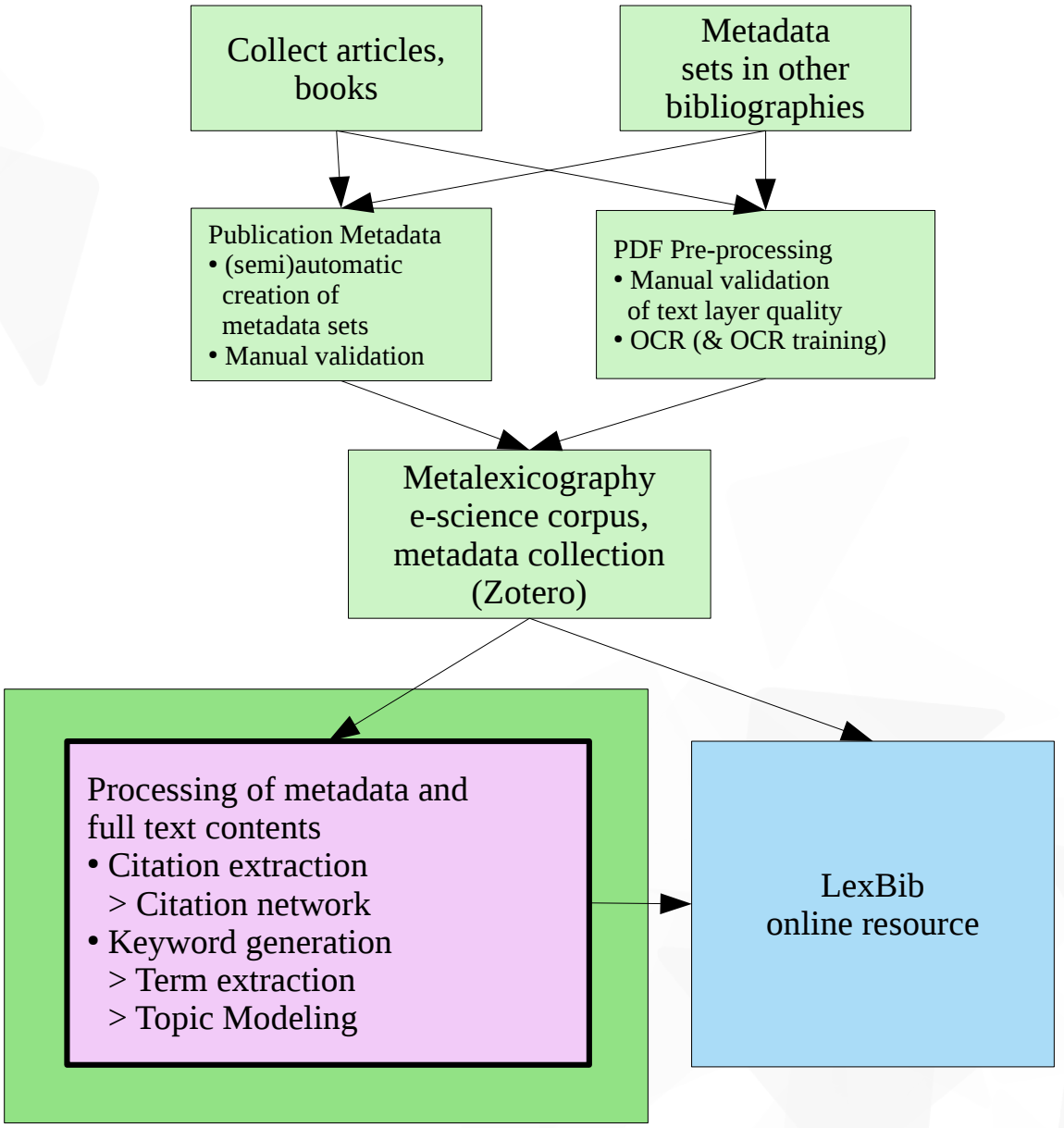
- Green: e-science corpus: 2nd iteration
  - Papers from other journals
  - Papers from other conferences
    - Retrieval also by citation network created in 1st iteration
  - Other languages: DE, ES, FR
    - Metadata, Full Texts
  - Items from before 2000
    - PDF manipulation
    - OCR training
  - Other item types
    - Books
    - Dissertations
    - Press articles
    - Blog posts, websites



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib project: Module 2

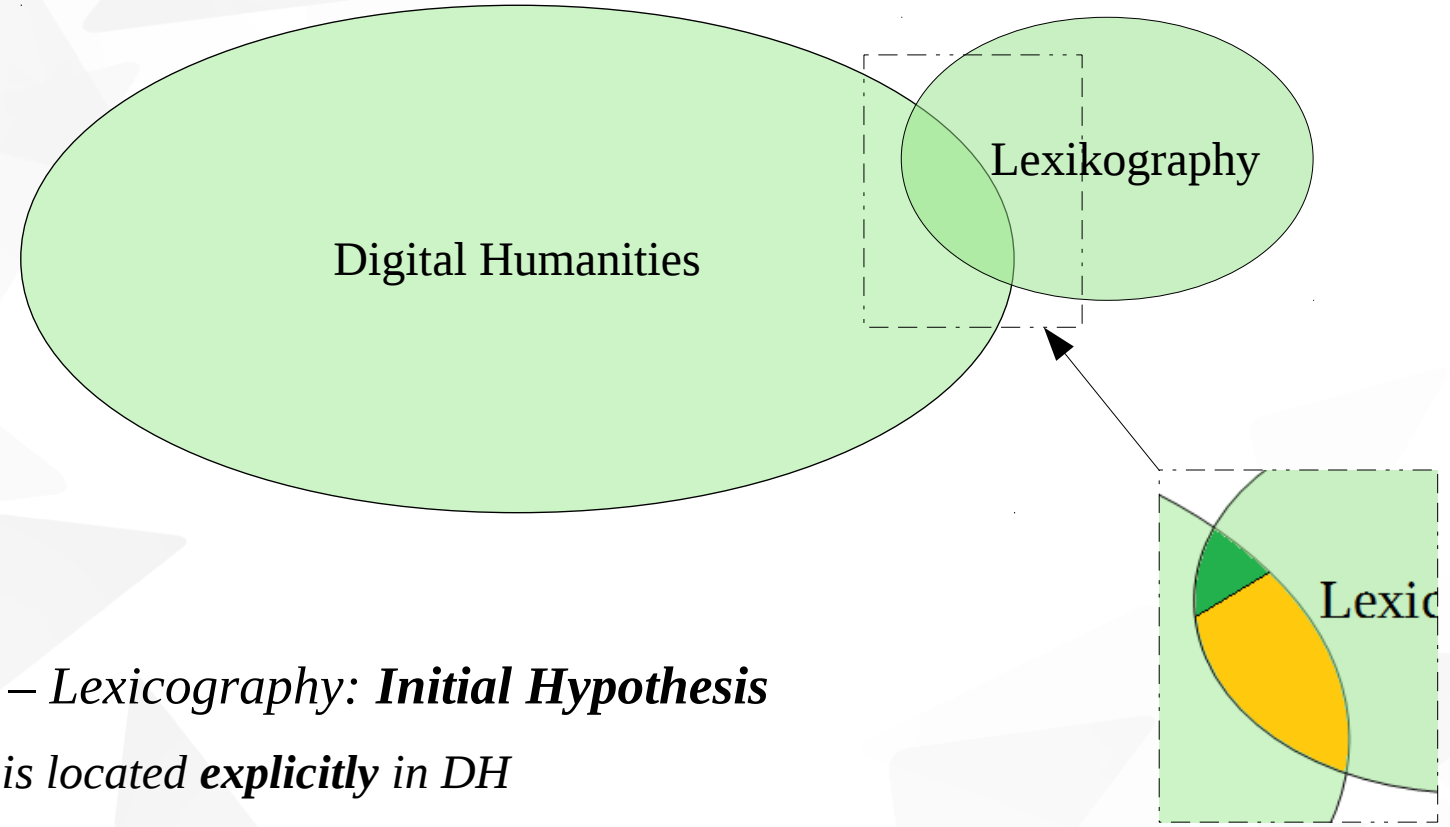
- Red: Computational text analysis
- Citation extraction
  - PDF parsing: “GROBID“
    - Manual annotation of training material
  - Citation network
    - Item-Relations “citing“, “is_cited_by“
- Topic Modeling
  - LDA (“Mallet”)
- Term Extraction
  - “Trex“ (Schäfer et al.)



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Preliminary Study: Lexicography and DH (DHd 2018)

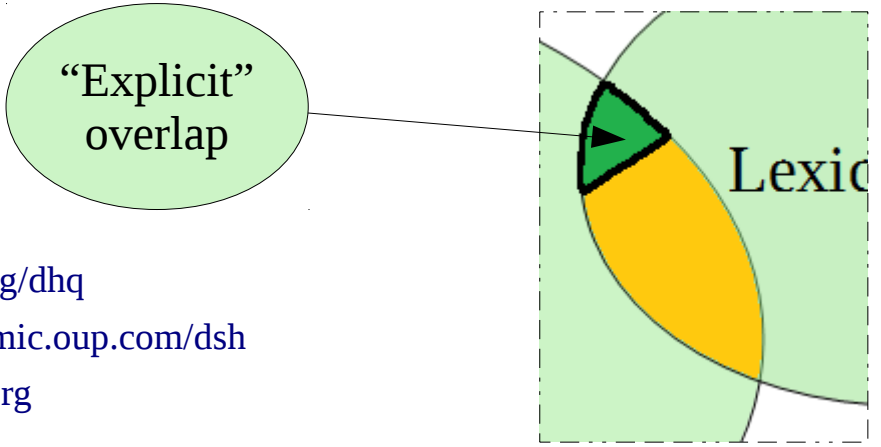
Lindemann, Kliche & Kutzner 2018



- *Overlaps DH – Lexicography:* ***Initial Hypothesis***
  - ***Only a part*** *is located* ***explicitly*** *in DH*
    - *...and publishes in a DH-journal or at a DH conference*
    - *...and describes own research as DH.*
  - ***The bigger part*** *may be located* ***implicitly*** *in DH.*

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Manual search: Lexicography @ DH



- Resources:
  - Digital Humanities Quarterly: http://www.digitalhumanities.org/dhq
  - Digital Scholarship in the Humanities (ex LLC): https://academic.oup.com/dsh
  - TEI Journal of the Text Encoding Initiative: http://jtei.revues.org
  - DHCommons: http://dhcommons.org
  - DH Conferences: http://digitalhumanities.org/dh-abstracts/search | https://dh2017.adho.org/program/abstracts/

- Keywords of 31 relevant contributions:

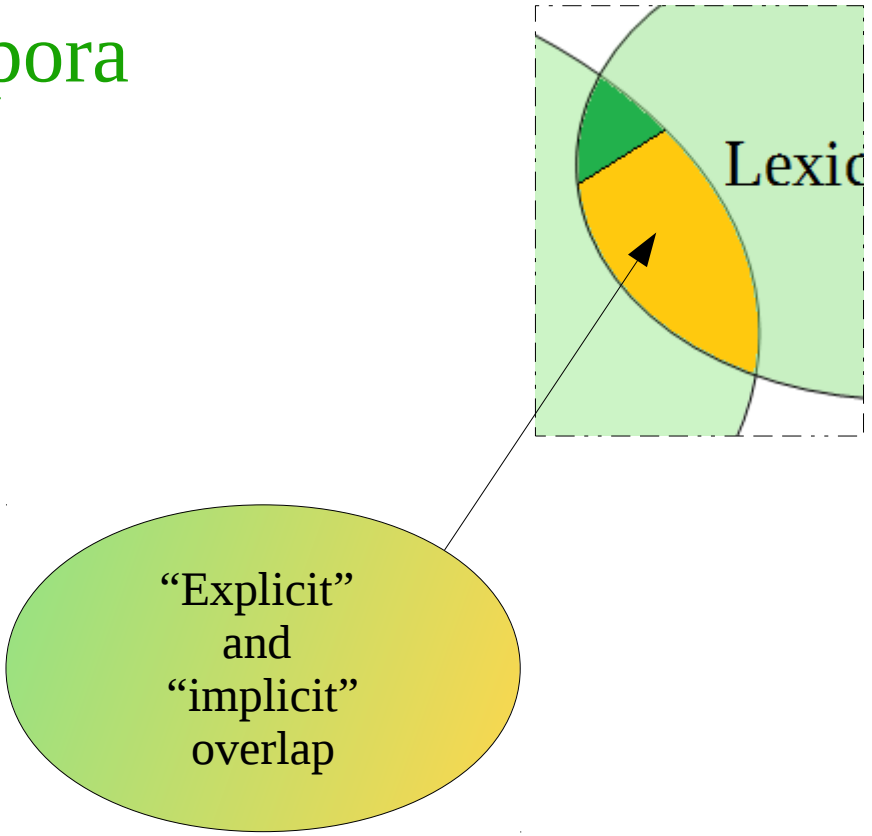
Text Encoding / Markup / Encoding Formats / XML	31
E-dictionaries / Visualization of lexical data	14
Historical Lexicography	10
Corpus Linguistics	9
Linking Lexical Resources	4
NLP-Lexicon	2
Bilingual Dictionary Drafting	2
Author dictionaries	2
Dialectology	1

Manual keyword indexation

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

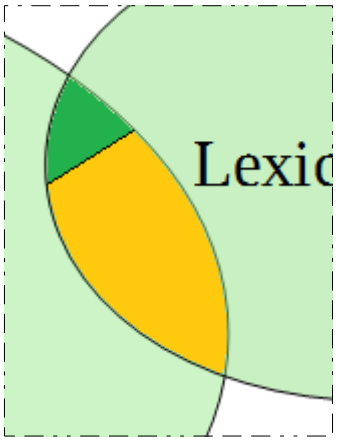
Lexikography, DH e-science corpora

- Definition of overlaps using DH methods
  - two e-science corpora
    - (1) DH corpus
      - DH journal articles
      - DH conference papers
      - DH handbook articles
    - (2) Lexicography corpus
      - Lexicography journal articles
      - Lexicography conference papers
      - Lexicography handbook articles



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

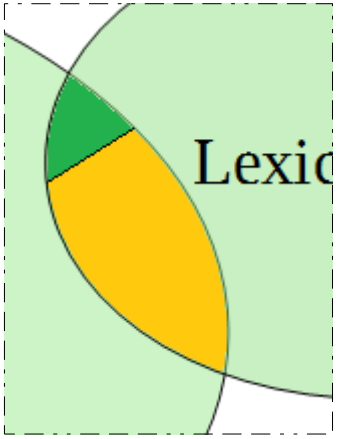
Lexicography corpus v1 (Spring 2018)



Lexicography: e-science corpus

Proceedings of Euralex	782
Lexikos Journal	376
International Journal of Lexicography	282
Journal of the Dictionary Society of North America (DSNA)	257
Proceedings of eLex	202
HSK 5/4: Dictionaries: An International Encyclopedia of Lexicography (Gouws et al., 2013)	110
Routledge Handbook of Lexicography (Fuertes-Olivera, 2018)	47

DH corpus v1 (2018)

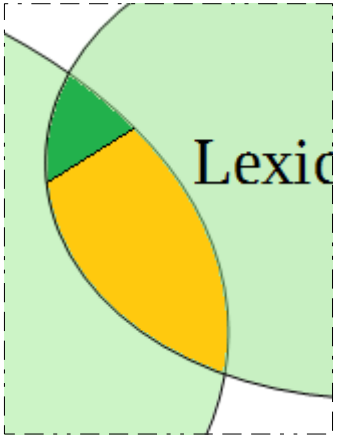


Digital Humanities e-science corpus

Digital Scholarship in the Humanities (DSH/LLC)	886
Digital Humanities Quarterly (DHQ)	284
Digital Studies/Le champ numérique	152
Journal of the Text Encoding Initiative	63
Blackwell Companion to Digital Humanities (Siemens et al. 2004)	37

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Lex_DH corpus v1 (2018)



Full text corpora

Digital Humanities Corpus	1,380 Full texts
Lexicography Corpus	1,919 Full texts (out of 2,057)
Total	3,299 Full texts

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Text Analytics: 3+ Methods

- Topic Modeling:
  - LDA: Mallet
  - Lemmatized Text
  - Suppression of stop words

McCallum 2002

- Citation Analysis:
  - GROBID:
    - "GeneRation Of BIbliographic Data"
    - Input: PDF full texts
    - Output: TEI-XML

Romary & Lopez 2015

- Term Extraction:
  - „TrEx": Term extraction for CWB corpora
    - NN, NN-NN and NN-NN-NN patterns
    - Weirdness ratio: $\frac{\textit{Relative Frequency in Lex\_DH Corpus}}{\textit{Relative Frequency in Ref.-Corpus}}$
    - Reference corpus: BNC

Schäfer et al. 2015

- Term Extraction:
  - GROBID

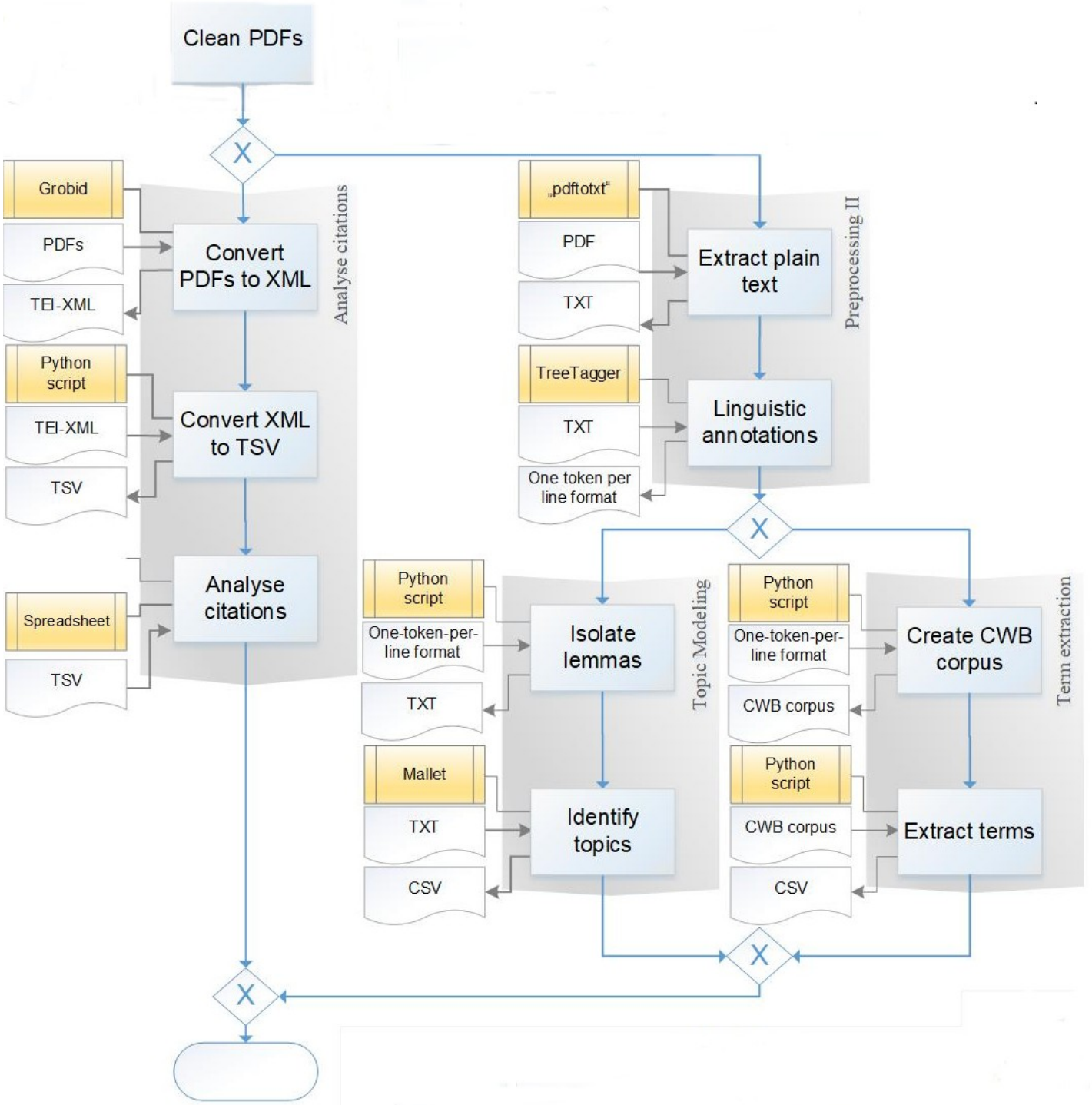
Lopez & Romary 2010

- Argumentative Zoning
- Citation Function

Teufel et al. 2002, 2006

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Lex_DH e-science corpus text analytics workflow model



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Citations

Normalization
Author = Last name of first author, lowercase, [a-z]
Title = lowercase, [a-z]

Validation category 1
Triple Author, Year, Title
found in Lex_DH corpus as item

Validation category 2
Triple Author, Year, Title
found in Lex_DH corpus as item, and:

- $Lev(Author_i, Author_j) \leq 2$
- $\Delta(Year_i, Year_j) \leq 1$
- $Lev(Titel_i, Titel_j) \leq 8$

```
wiegand 2013 textualstructuresinprinteddictionariesan
wiegand 2013 textualstructuresinprinteddictionaries

wiegand 2013 macrostructuresinprinteddictionariesgouw
wiegand 2013 microstructuresinprinteddictionaries

kypridemou 2013 narrativesimilarityascommonsummarypaperp
kypridemou 2014 narrativesimilarityascommonsummary

mccarty 2005 treeturfcentrearchipelagoorwildacremetap
mccarty 2006 treeturfcentrearchipelagoorwildacre

siepmann 2006 collocationcolligationandencodingdiction
siepmann 2005 collocationcolligationandencodingdictionaries

orli 2006 theoxforddzscomprehensiveenglishslovenia
orli 2006 theoxforddzscomprehensiveenglishslovenian

bergenholtz 2003 userorientedunderstandingofdescriptivepr
bergenholtz 2003 userorientedunderstandingofdescriptive

fuertesolivera 2016 acambrianexplosioninlexicographysomerefl
fuertesolivera 2016 acambrianexplosioninlexicography
```

Result counts:

val_cat 1: 760
val_cat 2: 1639

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Lex_DH Citation Network

1,108	citing articles	i.e. articles in Lex_DH corpus that cite one or more articles in Lex_DH corpus
	709 Lex 399 DH	
2,432	citation pairs	i.e. citations in Lex_DH corpus that point to another Lex_DH article
	1,674 Lex 758 DH	
1,057	cited articles	i.e. articles in Lex-DH corpus that are cited in one or more Lex_DH articles
	715 Lex 342 DH	

742	**Pairs**	**DH → DH**
16	**Pairs**	**DH → Lex**
1,659	**Pairs**	**Lex → Lex**
15	**Pairs**	**Lex → DH**

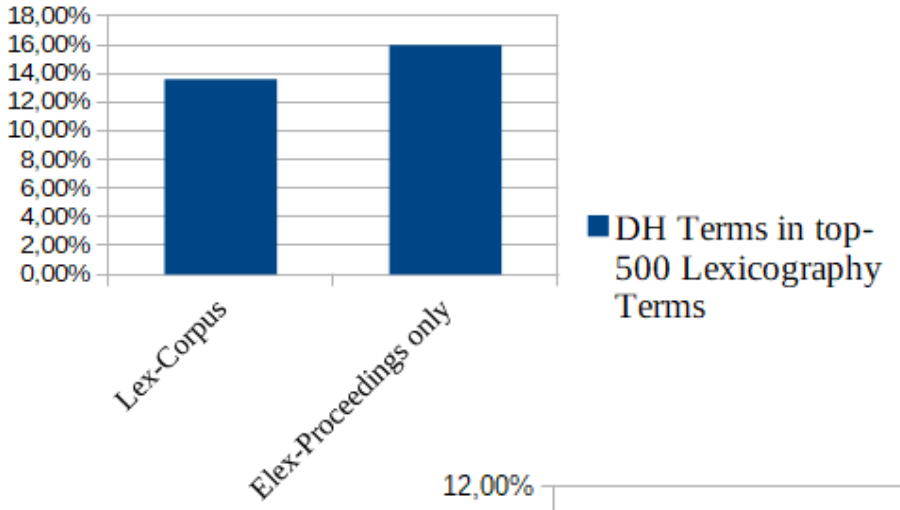
david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

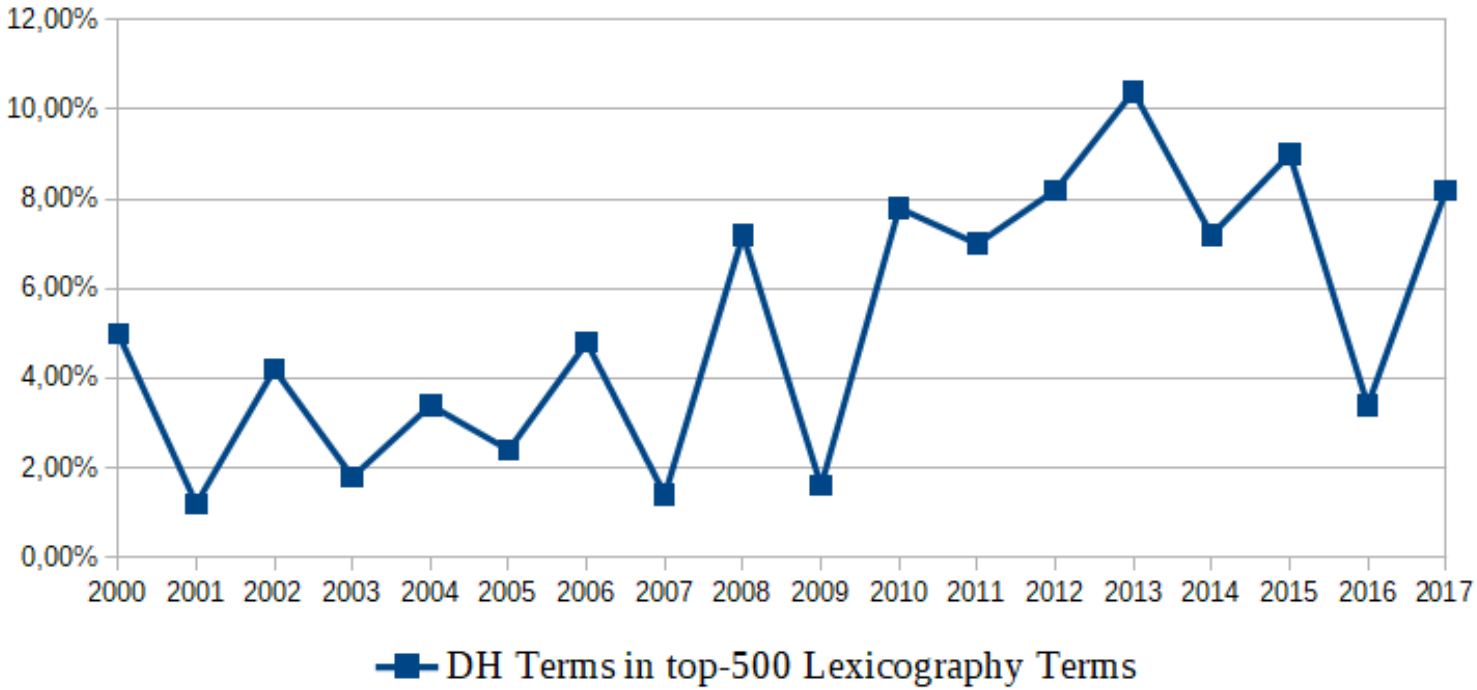
Results: Term Extraction Lex_DH

DHTerms sortHits	DHTerms sortWeird	LexTerms sortHits	LexTerms sortWeird	Top LexTerms found in DH
markup	website	lexicography	dictionary article	website
internet	pdf	lemma	access structure	lemmatization
website	xml	internet	dictionary user	wordnet
authorship attribution	stemma	dictionary use	lemma sign	reference corpus
metadata	text mining	macrostructure	text reception	corpus query
digitization	authorship attribution	polysemy	multiword	search engine
pdf	blog	multiword	website	internet
modeling	text classification	dictionary article	dictionary consultation	web site
text analysis	cyberinfrastructure	multi-word	word formation	web page
xml	search engine	access structure	lexicography	text box
stemma	feature selection	source language	corpus evidence	print version
blog	url	text production	text production	subcorpus
stylometry	classification accuracy	website	lemmatization	frequency list
cyberinfrastructure	web page	word formation	dictionary research	crowdsourcing
url	web site	dictionary user	function theory	web interface
dh	php	lemmatisation	article stretch	corpus research
search engine	crowdsourcing	lemma sign	word sketch	language documentation
avatar	open-source	lemmatization	wordnet	word alignment
curation	base text	text reception	reference corpus	hyperlink
part-of-speech	internet	wordnet	translation equivalent	wikipedia
intertextuality	text categorization	translation equivalent	dictionary information	search interface
text mining	test text	pdf	pdf	blog
word frequency	text reuse	definition	lemma list	source word
open-source	book history	target user	dictionary making	text genre
php	metadata	dictionary consultation	definition	word sense disambiguation

Universität Hildesheim
david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

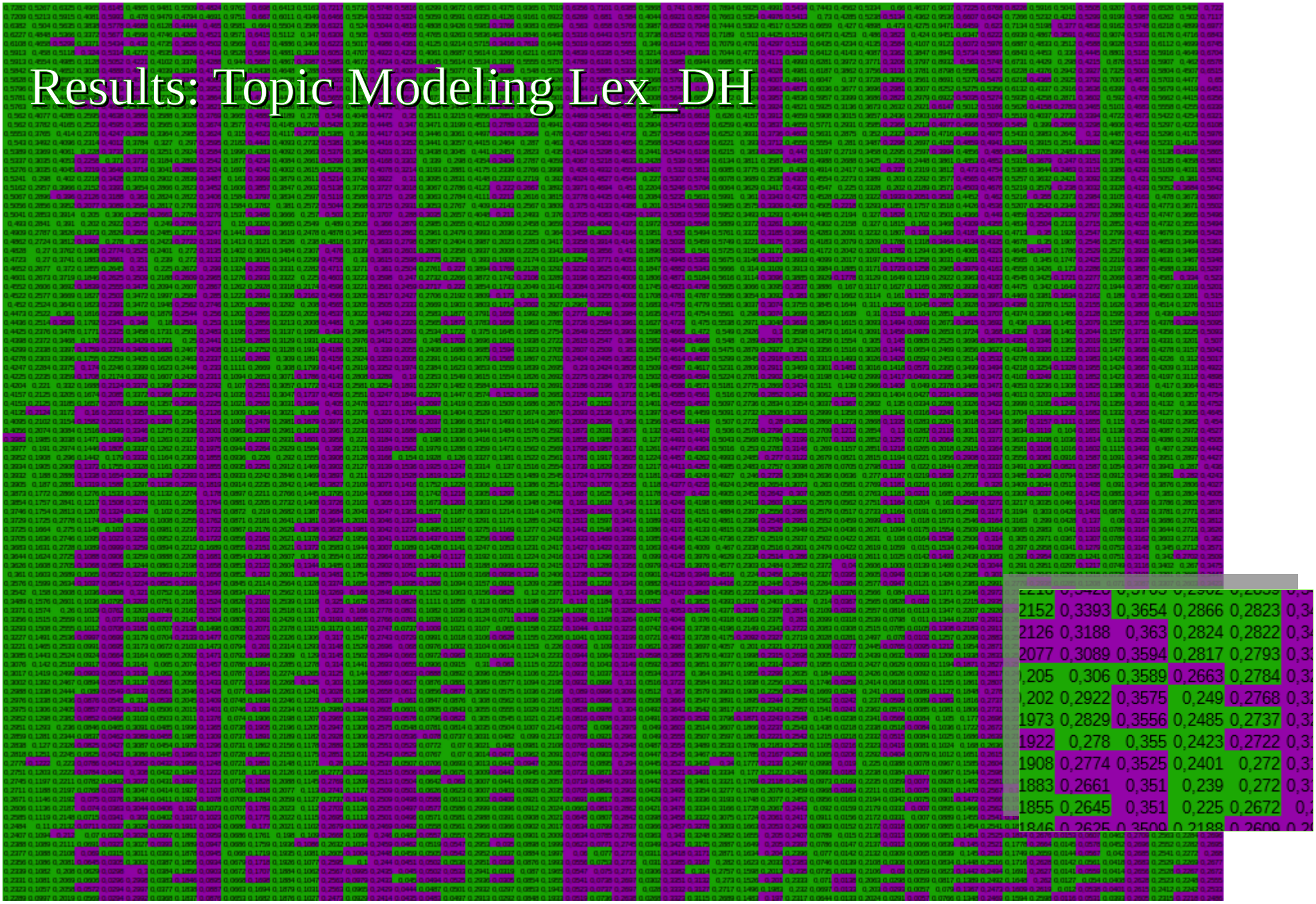
Results: Term Extraction Lex_DH





david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Results: Topic Modeling Lex_DH



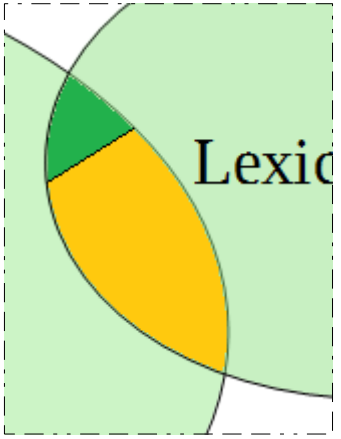
david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Results: Topic Modeling

#	DH	LEX	
25	100	0	digital library collection research information datum http project access resource material www metadata web record provide document object user archive
27	100	0	digital medium text reading work literary literature technology press form space au
46	100	0	digital library collection research information
15	99	0	digital medium text reading work literary litera
49	99	0	
8	95	0	digital humanity humanities computing resea
23	95	1	text edition manuscript textual electronic digit
16	90	1	
41	89	1	text word author feature analysis authorship r
40	87	5	image visual film visualization figure sound m
13	84	5	
44	84	5	shakespeare word poem line play text poetry
4	81	10	
24	64	36	dialect map datum linguistic distance variation area analysis spatial variant
10	60	40	card table number page figure total show high compare study sample result journal list average percentage type distribution author score
31	56	44	corpus word text language frequency english linguistics datum study linguistic analysis speech list discourse include speak research table genre million
5	54	46	card case make give point present approach information study work form type part fact analysis question find provide result structure
32	48	52	database system tool text software datum project file computer user application design work interface create code figure process develop xml
28	44	56	card word language corpus method result system text proceeding sentence computational evaluation approach error automatic algorithm number extract tag datum
36	44	56	legal law copyright business term court political property company acronym register public act trade government state land market financial fair
3	33	67	woman slang gender female male man term sex fem
42	33	67	card book work history page author write year publis
20	31	69	chinese arabic irish english character china canadia
39	27	73	word form rule morphological noun lexicon language
7	26	74	word make people geen thing time find good sense
17	24	76	english johnson london language john johnsons ed
34	24	76	estonian adv language hindi arabic indian proper in
43	23	77	proverb japanese boyer french page royal expressio
18	20	80	spanish card del diccionario los una madrid por cata
37	20	80	gabonese gabon french stein part polis des work ma
38	18	82	des french les card dan dictionnaire une qui langue
6	15	85	italian della dizionario del anglicism lingua italiana ch
12	14	86	translation equivalent language bilingual translate e
14	12	88	dutch van language dictionary slovene frisian datab
21	11	89	der german die und card das von des den mit ein zur
48	10	90	user search online card access electronic information http figure www web internet datum resource result link tool content query provide
26	5	95	idiom polish czech expression phraseological idiomatic usage syn maiv idioms variant słownik mienh light unit variation maaih praha figurative nyei
2	4	96	semantic lexical sense card meaning word relation definition language concept wordnet synonym metaphor category lexicon semantics conceptual thesaurus cognitive structure
30	4	96	english oed quotation dictionary word citation oxford editor murray evidence source historical middle etymology supplement entry language press med american
45	3	97	article card item structure dictionary text access wiegand give address type cross-reference outer microstructure form lemma reference follow partial relation
1	2	98	swedish language danish dictionary lemma word norwegian project romanian list form card editor version croatian lexicon publish compound nordic german
29	2	98	card dictionary learner word student study english language test vocabulary information learners participant
0	1	99	language card dictionary african south shona africa speaker english ndebele community lexicography culture
47	1	99	dictionary user card text datum lexicography information lexicographic function situation tarp bergenholtz the
9	0	100	die van word wat nie vir afrikaans woordeboek meet hierdie dit woordeboeke card dat heat aan kan gebruik
11	0	100	term terminology concept domain terminological field card information language specialized definition knowle
19	0	100	dictionary word entry card english definition include meaning language sense headword information label list
22	0	100	card dictionary lexicography language international lexicographic linguistics corpus research dictionaries edit
33	0	100	dictionary word sotho class noun stem lemma form northern african schryver language zulu user prinsloo ver
35	0	100	collocation card corpus collocate lexical verb word noun ofthe pattern combination collocational expression sketch adjective unit phrase meaning base semantic

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Preliminary Study Lex/DH: Results



- Contribution to mutual consideration of research communities
  - Citation overlap (“explicit overlap“) smaller than “implicitly“ overlapping Topics
- Overlapping Topics
  - Common Methods
    - Examples: Corpus Linguistics; Language Documentation
  - Common “third“ research interests
    - Examples: Gender; juridical issues (copyright etc.)
- Outlook
  - Close reading in overlapping / disjunctive sets
    - If terms / topics overlap: Which texts are we talking about?
    - If citations overlap, but extracted terms / topics do not: Are we talking about the same, using different words?

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Term extraction tests

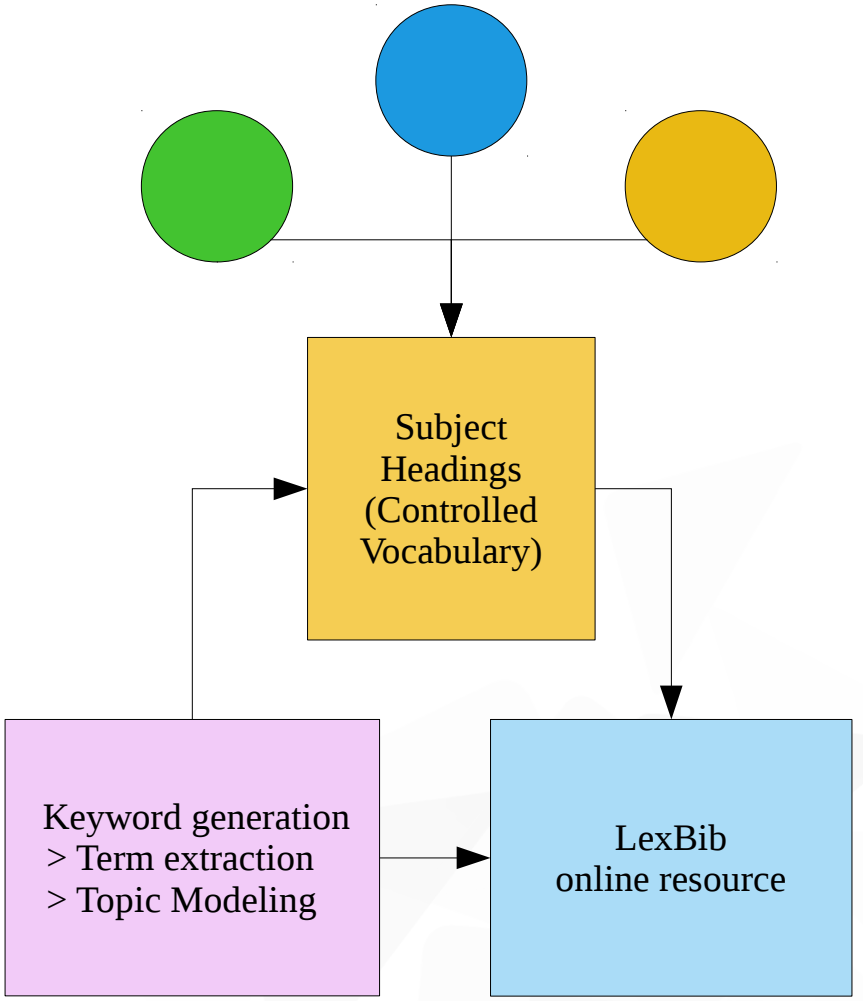
- Example: Top 20 terms, extracted from a single full text (here, a 2014 Euralex keynote speech)
  - Reference corpus BNC
  - Reference corpus LexBib
- Mapping to existing keyword lists
  - Obelex-meta
    - as silver standard for evaluation
    - as first basic grid for keyword ontology
  - [etc.]

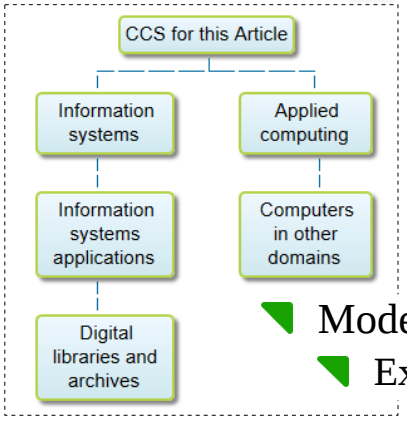
Top 20 Terms (Ref. BNC)	*Top 20 Terms (Ref. LexBib)*
text reception	information-on-demand
text production	on-demand
dictionary function	data repository
multiword	user friendliness
word formation	user orientation
production dictionary	text reception
user orientation	production dictionary
information-on-demand	text production
data repository	dictionary function
dictionary entry	repository
user friendliness	valency
internet	guidance
markup	orientation
on-demand	concord
valency	word formation
concord	scenario
dictionary	markup
corpus	production
collocation	classification
language processing	advance

Universität Hildesheim
david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

Additional Module: Keyword / Subject heading ontology

- Development and application of an ontological controlled vocabulary for subject headings
- Sources:
  - e-science corpus: extracted terms
  - e-science corpus: author keywords
  - Obelex-meta keywords
  - Wiegand‘s concept taxonomies
  - Library of Congress Subject Headings
  - GND (Gemeinsame Normdatei)



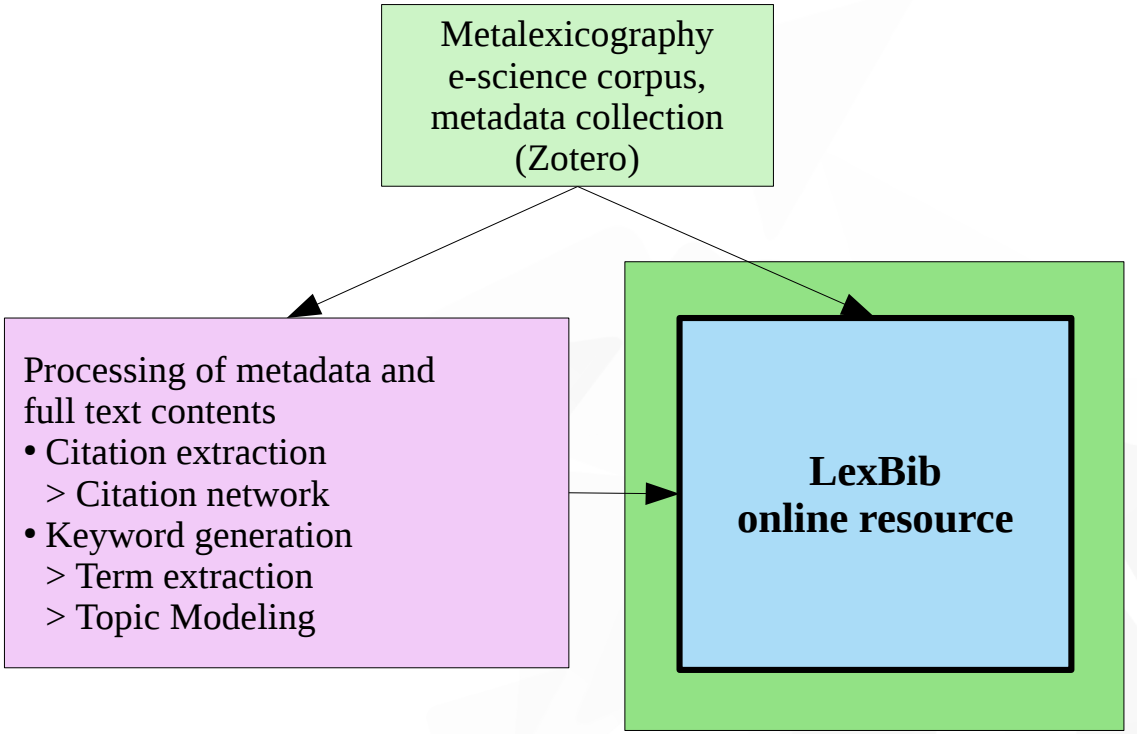


- Model: ACM CCS
  - Example Bamman & Crane 2008 @ ACM

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib project: Module 3

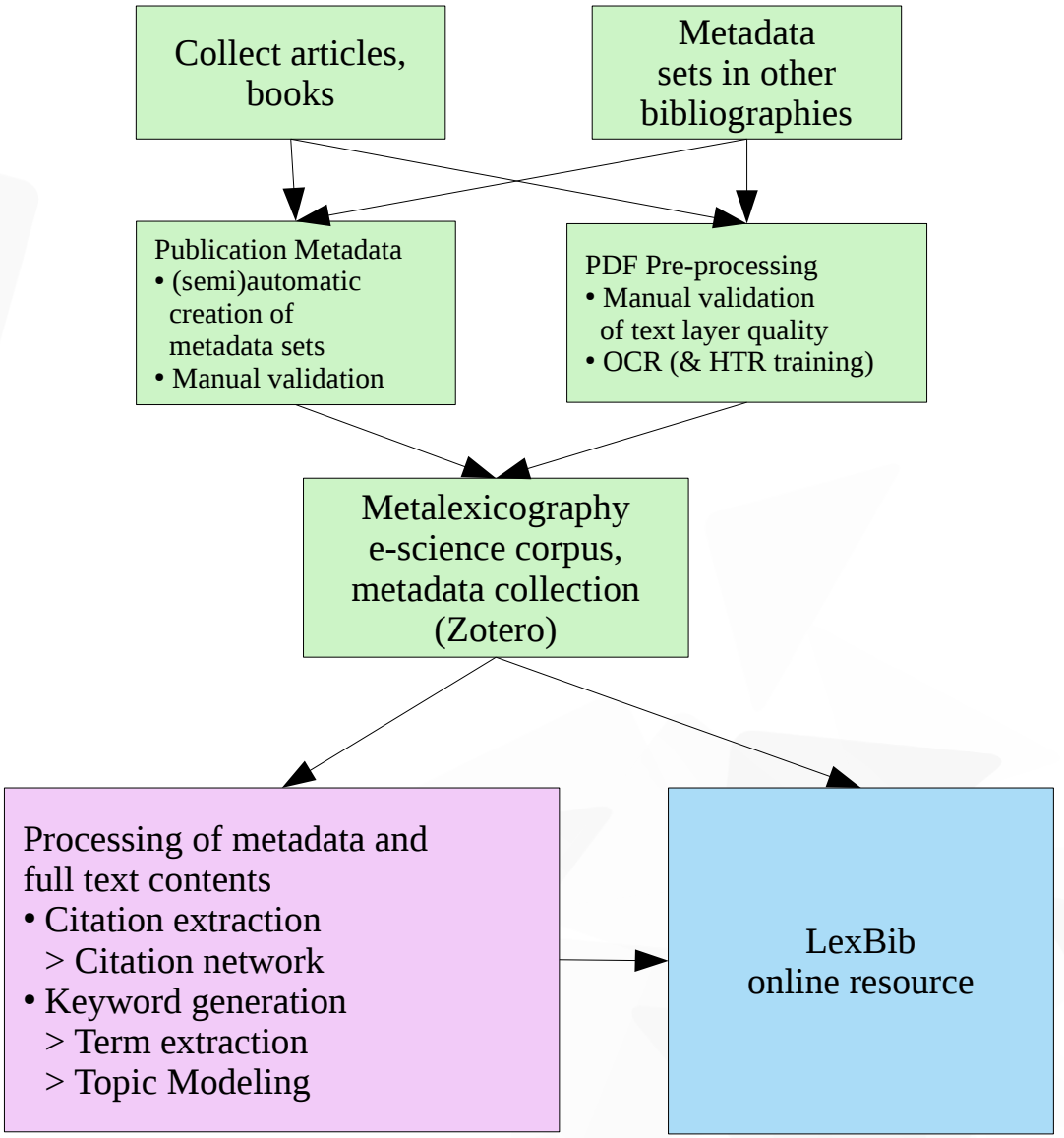
- Blue: Creation of online bibliography
  - Features:
    - Enhanced search functions
    - Links to full text download pages
    - Persons / events: Profile pages
    - API for metadata, BibTeX, RDF
    - is-container-of / is-contained-in
    - Browsable citation network
    - is-review-of / is-reviewed-by
    - Keywords: Results of red module
    - Browsable keyword ontology



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib project: Partners

- Green
  - Elexis
  - Euralex
  - IDS
  - GLex UDC
  - EMLex
  - de Gruyter
- Red
  - Romary, Khemakhem (Inria, CMB)
- Blue
  - Clarin-D (IDS)
  - Euralex
  - Elexis



david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

LexBib project: Long-term perspective

- Complete collections (before 2020)
  - various languages: EN, DE, ES, FR
    - publication metadata
    - full texts
- LexBib Online Resource
  - Bibliography items with unique IDs
    - publications
    - persons
    - events, places
  - Keyword indexation
    - automatically extracted
    - mapped to controlled vocabulary
- After 2020
  - Blue: Long-term hosting options
    - CLARIN server @ IDS
    - ELEXIS
  - Green: Curation team and/or community?
    - Author pages
      - editing of own page?
    - Keyword indexation of new items
      - editing of own items (as in ACM)?
  - Red: Tools as services
    - Application to e-science corpora of other domains

david.lindemann@uni-hildesheim.de
fritz.kliche@uni-hildesheim.de
heid@uni-hildesheim.de

http://www.zotero.org/groups/lexbib/items

David Lindemann
Fritz Kliche
Ulrich Heid

Thank you for your attention

See detailed bibliography in the proceedings