Evaluation of LOC-DB

Criteria and preliminary results

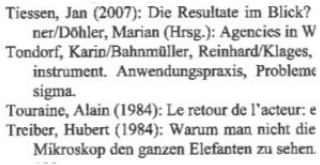
Linked Open Citation Database

Research Question

How much would it cost, with respect to resources,
if libraries catalogued everything and curated the citation graph?

Method

- Development of processes and tools based on linked data technologies to enable libraries to contribute to an open and interconnected citation graph
- Quantitative and qualitative evaluation, e.g., cost benefit analysis



Library Workflow

- Integrated into the standard library workflow
- Reuse of data from existing resources, e.g. from publishers, other projects, and standard library catalogs (high quality metadata)
- Automated as far as possible
  - Automatic reference extraction
  - Easy-to-use editorial system
- Distributed database and collaborative cataloging processes

Advantages of Semi-Automated Systems

- Human-level quality is guaranteed
- Higher data quality than fully-automated approaches, option to manually correct errors
- Libraries in control of curating citations, key measure for scientific impact

Quality + EFFICIENCY

Criteria

Efficiency (easy to measure)

- Citation Linking Time: Measure the time required to link a single citation
- First evaluation results, improvements, second evaluation results

Quality (harder to measure)

- Prevalence of identifiers
- Prevalence of important fields per resource type
- LOC-DB automatically generates statistics that support the quality assessment
- … (a topic for the afternoon sessions)

Criteria

Efficiency (easy to measure)

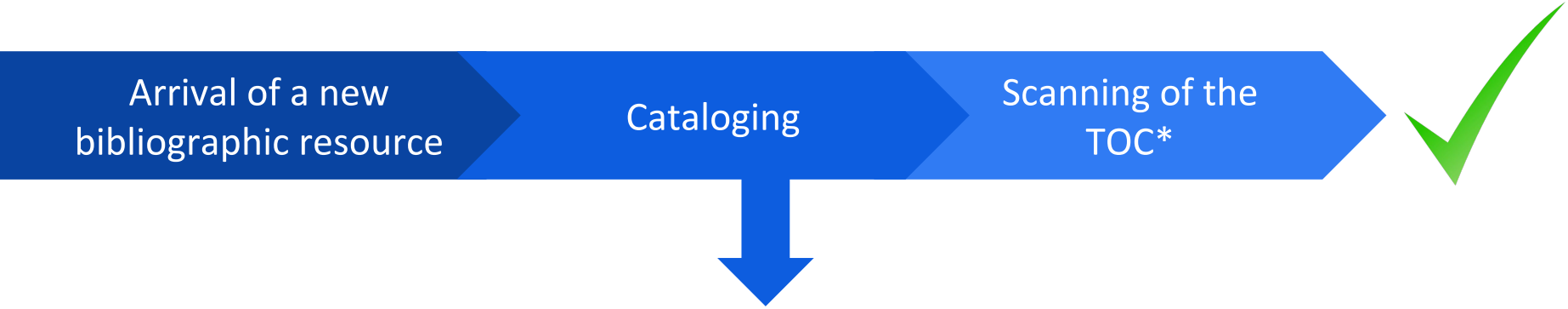
- Citation Linking Time: Measure the time required to link a single citation
- First evaluation results, improvements, second evaluation results

Quality (harder to measure)

- Prevalence of identifiers
- Prevalence of important fields per resource type
- LOC-DB automatically generates statistics that support the quality assessment
- … (a topic for the afternoon sessions)

Standard Library Workflow (UB Mannheim)

For Print Resources



- Metadata
- Identifier

*Table of Contents

Reference Linking

Extending the Standard Library Workflow

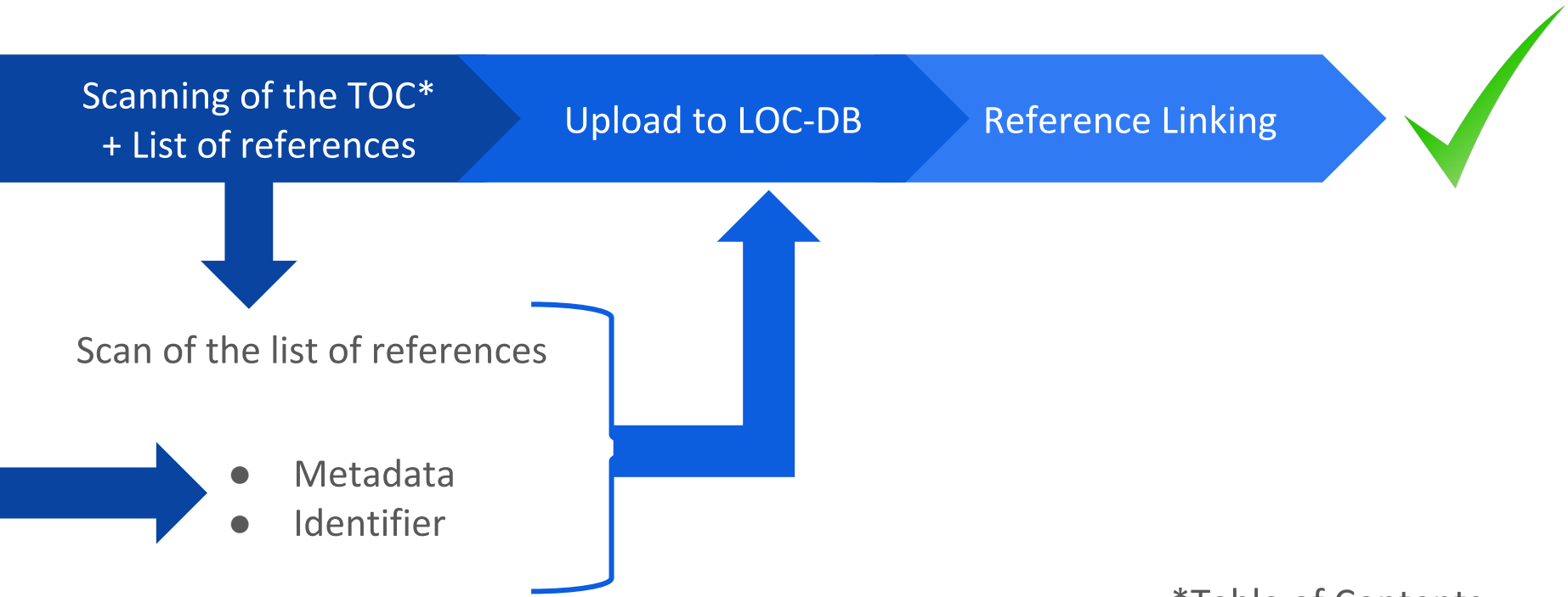


*Table of Contents

Reference Linking

Extending the Standard Library Workflow

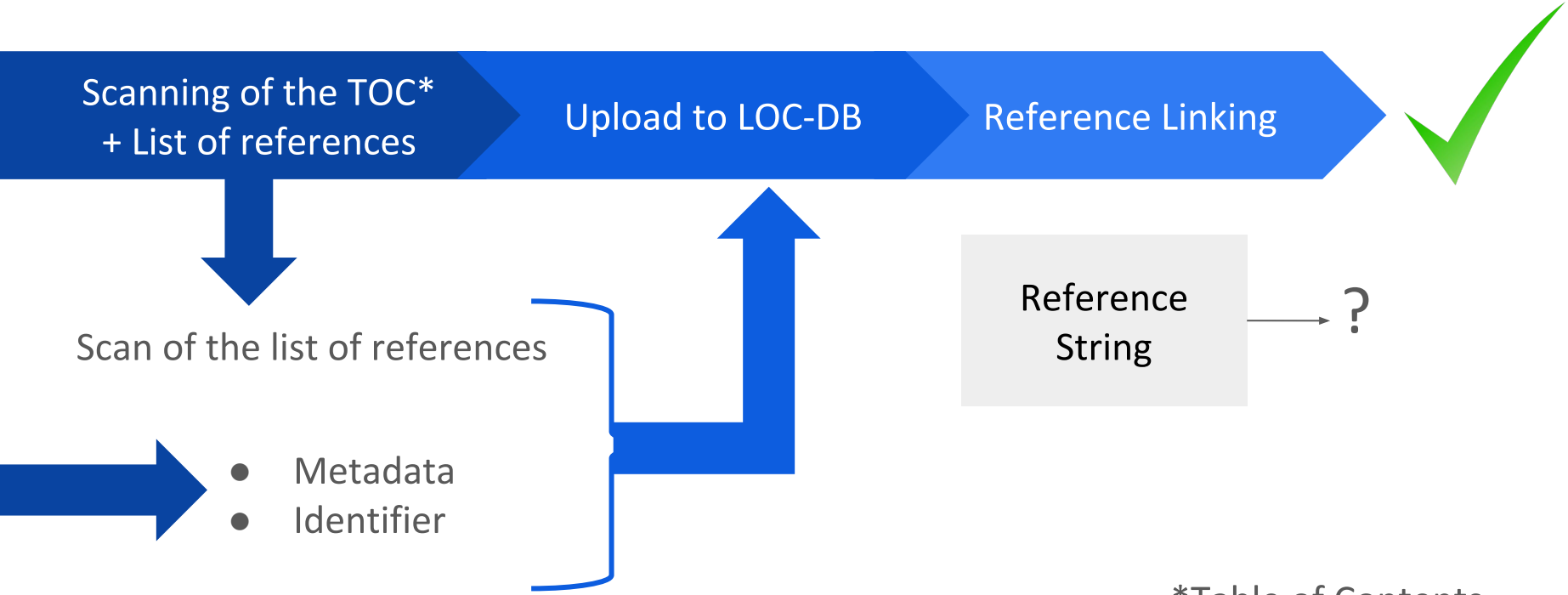


*Table of Contents

Reference Linking

Extending the Standard Library Workflow

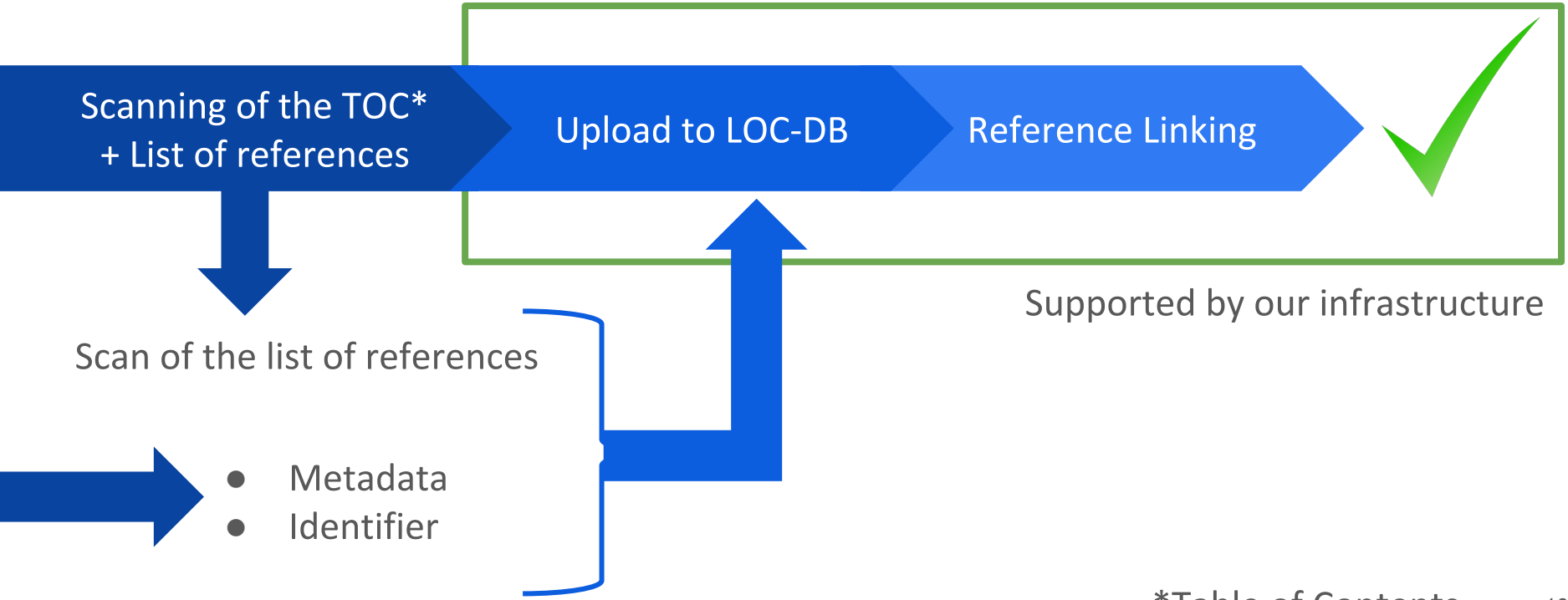


*Table of Contents

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

- > 100 pages per person per hour
- Upper bound ~ 15 minutes for scanning for an average book (26 pages of references)
- Prolongs the processing of a book on average by only 3 minutes

Lauscher et al.: **Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph,** JCDL 2018

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

- > 100 pages per person per hour
- Upper bound ~ 15 minutes for scanning for an average book (26 pages of references)
- Prolongs the processing of a book on average by only 3 minutes
- Additional scanning time does not significantly affect other processes in the library

Lauscher et al.: **Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph,** JCDL 2018

How much time does the whole process take?

- Batch upload
- Background processing for meta data retrieval and reference extraction
  → Does not affect the process in the library

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)			
Internal Suggestion Retrieval (s)			
External Suggestion Retrieval (s)			
# Searches per Reference			

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	9.93	557.20	89.45
Internal Suggestion Retrieval (s)	0.02	0.5	0.06
External Suggestion Retrieval (s)	0.50	95.65	0.89
# Searches per Reference	1	36	2

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	9.93	557.20	89.45
Internal Suggestion Retrieval (s)	0.02	0.5	0.06
External Suggestion Retrieval (s)	0.50	95.65	0.89
# Searches per Reference	1	36	2

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	9.93	557.20	89.45
Internal Suggestion Retrieval (s)	0.02	0.5	0.06
External Suggestion Retrieval (s)	0.50	95.65	0.89
# Searches per Reference	1	36	2

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	9.93	557.20	89.45
Internal Suggestion Retrieval (s)	0.02	0.5	0.06
External Suggestion Retrieval (s)	0.50	95.65	0.89
# Searches per Reference	1	36	2

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	9.93	557.20	89.45
Internal Suggestion Retrieval (s)	0.02	0.5	0.06
External Suggestion Retrieval (s)	0.50	95.65	0.89
# Searches per Reference	1	36	2

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	9.93	557.20	89.45
Internal Suggestion Retrieval (s)	0.02	0.5	0.06
External Suggestion Retrieval (s)	0.50	95.65	0.89
# Searches per Reference	1	36	2

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

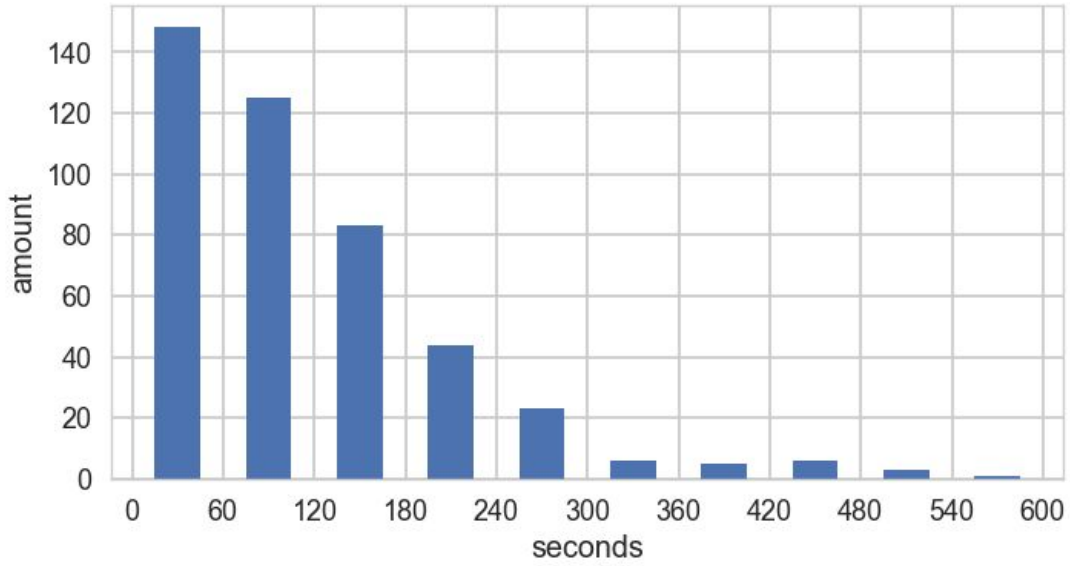
Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	9.93	557.20	89.45
Internal Suggestion Retrieval (s)	0.02	0.5	0.06
External Suggestion Retrieval (s)	0.50	95.65	0.89
# Searches per Reference	1	36	2

Minimum, maximum, and median time in seconds for the reference linking step

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking



Histogram of reference linking times, N = 444

How much would it cost if libraries catalogued everything and curated the citation graph?

Estimation about the number of full-time employees needed to process all literature of social sciences bought in 2011 by Mannheim University Library, depending on the time t in seconds to resolve a reference.

t	1	5	10	20	30	60	120
# employees	0.1	0.5	1	2	3	5.9	11.9

Evaluation in January 2018

How much would it cost if libraries catalogued everything and curated the citation graph?

Estimation about the number of full-time employees needed to process all literature of social sciences bought in 2011 by Mannheim University Library, depending on the time t in seconds to resolve a reference.

t	1	5	10	20	30	60	120
# employees	0.1	0.5	1	2	3	5.9	11.9

Evaluation in January 2018

Improvements

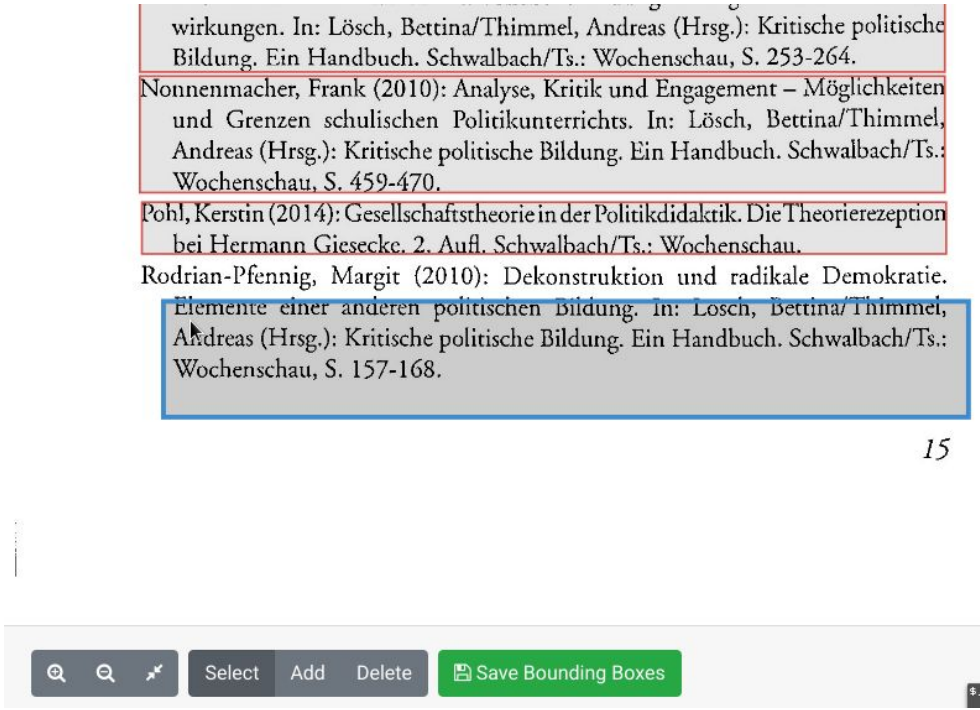
- **Improved automatic reference extractor:** improved readability, less need for manually creating bibliographic entries, better suggestions
- **Support for multi-page scans:** Less overhead while ingesting data, increased readability
- **Resizable bounding boxes:** to improve layout detection step
- **Precalculated suggestions:** faster suggestions
- **Add filtering capabilities:** find the correct suggestions faster
- **Support for container resources:** shared metadata, less overhead
- **Identifier handling:** migration between resources, author identifiers
- **More external sources:** more suggestions (more in the afternoon sessions)
- **Various UI improvements** (more in the afternoon sessions)

Support for Multi-Page Scans

- Multi-page scans are automatically split by the reference extractor
- Users can conveniently advance through the pages in the front-end

Resizable Bounding Boxes

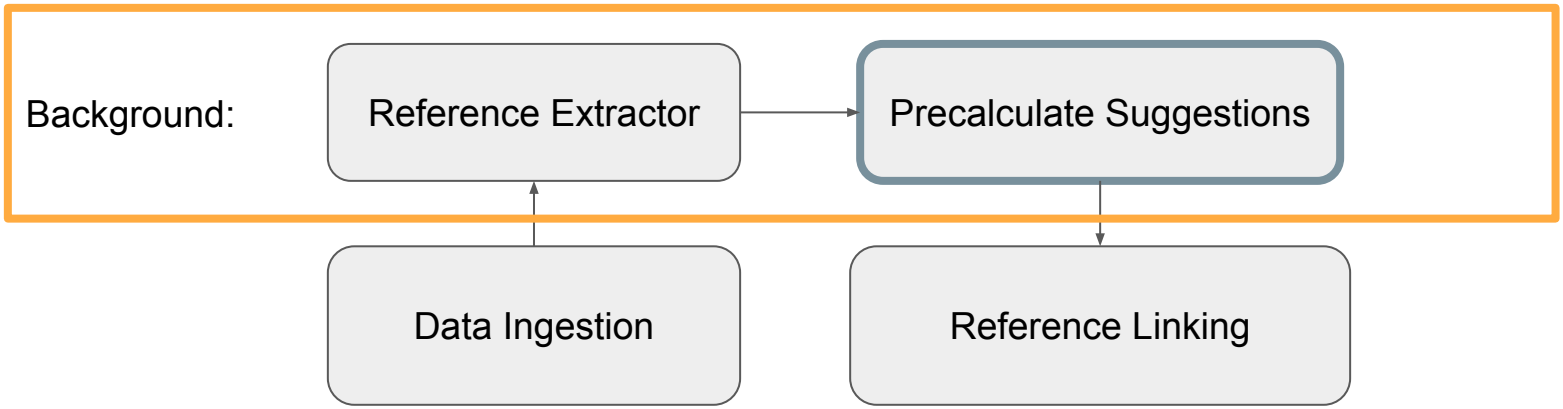
- Bounding boxes can be inserted, deleted, or adjusted
- New boxes are saved and the suggestions will be updated.
- Manually adjusted bounding box data can be used as training data for layout detection algorithms



Precalculated Suggestions

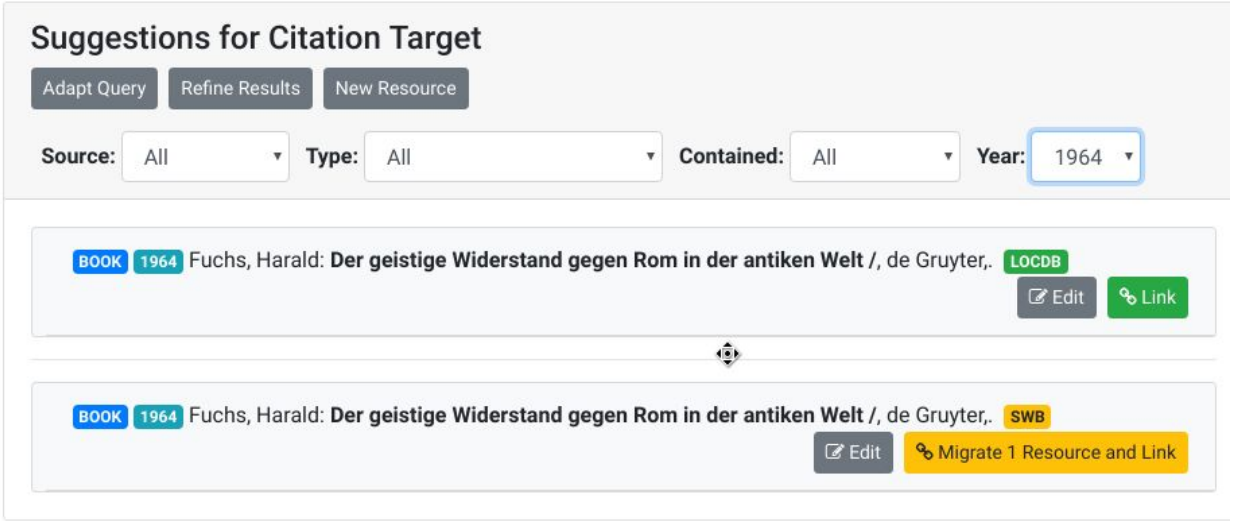
As soon as data is ingested into the system:

1. Reference extractor yields structured meta-data
2. Structured metadata is used to query external sources
3. Pre-calculated are stored and supplied to the user when needed



Filtering Suggestions

- Generate many precalculated suggestions (30)
- Filter them dynamically after retrieval
- Reduce the need for imposing new queries (expensive)

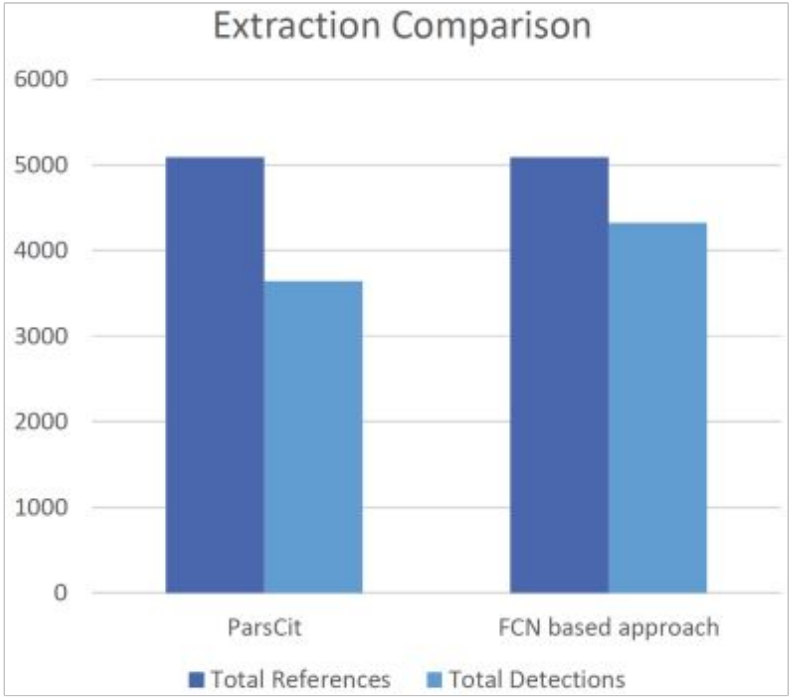


Improved Automatic Reference Extractor

First Workshop November 2017

- Dataset consisted of 514 scans containing 16629 references:

	Trainset	Val.set	Testset
Number of Images	572	50	286
Number of References	10631	904	5094



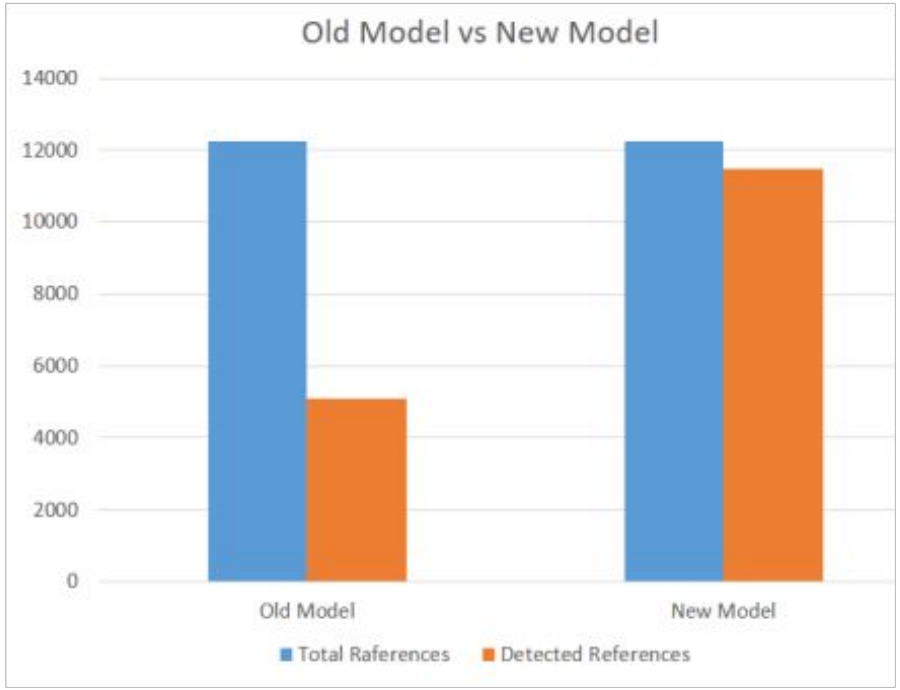
Second Workshop November 2018

- Dataset consisted of 2401 scans containing 38863 references:

	Trainset	Val.set	Testset
Number of Images	1513	132	756
Number of References	24606	2013	12244

Old vs Current Model on Current Dataset

	Old Model	New Model
Number of Detections	5108	11501
Accuracy	41.71%	93.8%



Evolution of Automatic Reference Extractor

ParsCit

Old Model (2017)

Current Model

How much time does the whole process take?

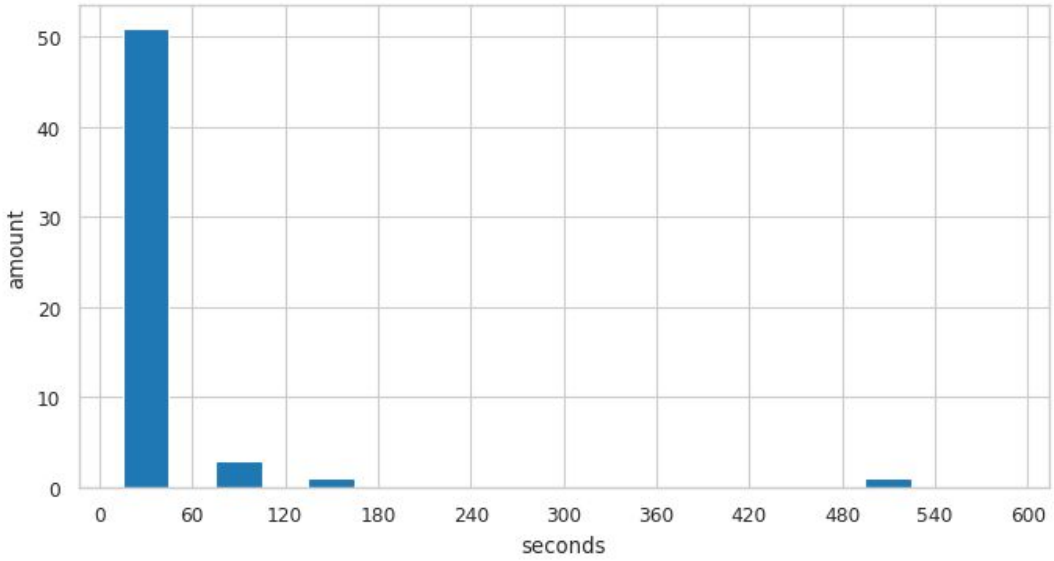
Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking

Criterion	Minimum	Maximum	Median
Citation Linking (s)	3.777	535.33	12.55
Internal Suggestion Retrieval (s)	0.14	34.77	0.34
External Suggestion Retrieval (s)	0.08	0.59	0.33
# Extra Searches per Reference	0	4	0

Minimum, maximum, and median reference linking time, N=56 (preliminary results)

How much time does the whole process take?

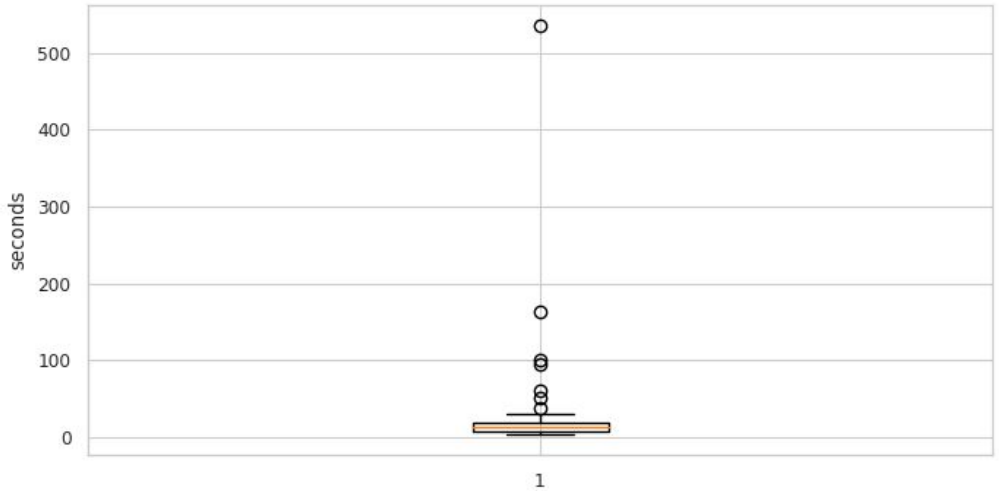
Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking



Histogram of reference linking times, N = 56

How much time does the whole process take?

Scanning of the list of references (only print) → Upload to LOC-DB → Reference Linking



Box plot of reference linking times, N = 56

How much would it cost if libraries catalogued everything and curated the citation graph?

Estimation about the number of full-time employees needed to process all literature of social sciences bought in 2011 by Mannheim University Library, depending on the time t in seconds to resolve a reference.

t	1	5	10	20	30	60	120
# employees	0.1	0.5	1	2	3	5.9	11.9

How much would it cost if libraries catalogued everything and curated the citation graph?

Preliminary results from January suggested general feasibility of the approach
Continuous improvement on the infrastructure and processes led to further efficiency improvements (and this can be even further improved)!

How much would it cost if libraries catalogued everything and curated the citation graph?

Thank you!

Preliminary results from January suggested general feasibility of the approach
Continuous improvement on the infrastructure and processes led to further efficiency improvements (and this can be even further improved)!