Linked Open Citation Database

Enabling Libraries to Contribute to an Open

and Interconnected Citation Graph

Anne Lauscher, **Kai Eckert**, Lukas Galke, Ansgar Scherp, Szed Tahseen Raya Riyva,
Sheraz Ahmed, Andreas Dengel, Philipp Zumstein, Annette Klein
JCDL, 2018
Fort Worth, Texas

The origins: What were libraries cataloging?

- Everything you can put in a shelve, i.e. cataloging on the level of physical objects
- Dependent resources, e.g. journal articles, were not cataloged

Later: Resource Discovery Systems

- Combination of classical library content with large indexes of article data
- Indexes are dominantly run by commercial providers (Breeding, 2015)
- Some indexes are provided by library service centers (e.g., K10plus-Zentral)

Now: Citations?

- Commercial providers often include citation data in their catalogs (I4OC)
- Fully automated approaches are still erroneous
- .. libraries?

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

Tiessen, Jan (2007): Die Resultate im Blick?
ner/Döhler, Marian (Hrsg.): Agencies in W
Tondorf, Karin/Bahnmüller, Reinhard/Klages,
instrument. Anwendungspraxis, Probleme
sigma.
Touraine, Alain (1984): Le retour de l'acteur: e
Treiber, Hubert (1984): Warum man nicht die
Mikroskop den ganzen Elefanten zu sehen.

Agenda

1. Motivation
2. **Data**
3. Reference Linking Workflow
4. LOC-DB Infrastructure
5. Evaluation
6. Conclusion

Data

- Social sciences collection of Mannheim University Library
- 522 print books and collections acquired by in 2011: ~271,000 references
- Articles published in 2011 in 101 (mostly electronic) journals: ~298, 000 references
- New print acquisitions of the social sciences branch library from July 2017 on

Agenda

1. Motivation
2. Data
3. **Reference Linking Workflow**
4. LOC-DB Infrastructure
5. Evaluation
6. Conclusion

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

Standard Library Workflow (UB Mannheim)

For Print Resources
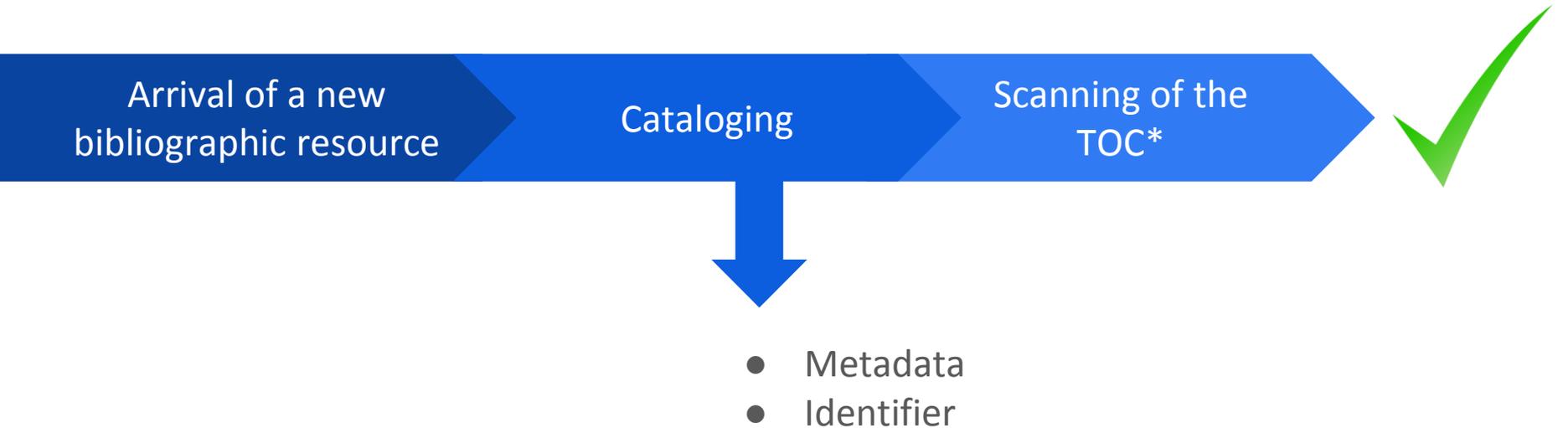
Arrival of a new bibliographic resource → Cataloging → Scanning of the TOC* ✓

- Metadata
- Identifier

*Table of Contents

Reference Linking

Extending the Standard Library Workflow

Scanning of the TOC* + List of references → Upload to LOC-DB → Reference Linking ✓

Scan of the list of references

- Metadata
- Identifier

*Table of Contents

Reference Linking
Extending the Standard Library Workflow

Scanning of the TOC*
+ List of references → Upload to LOC-DB → Reference Linking ✓

Scan of the list of references

- Metadata
- Identifier

Reference String → ?

*Table of Contents

Reference Linking
Extending the Standard Library Workflow

Scanning of the TOC* + List of references → Upload to LOC-DB → Reference Linking ✓

Supported by our infrastructure

Scan of the list of references

- Metadata
- Identifier

*Table of Contents

13
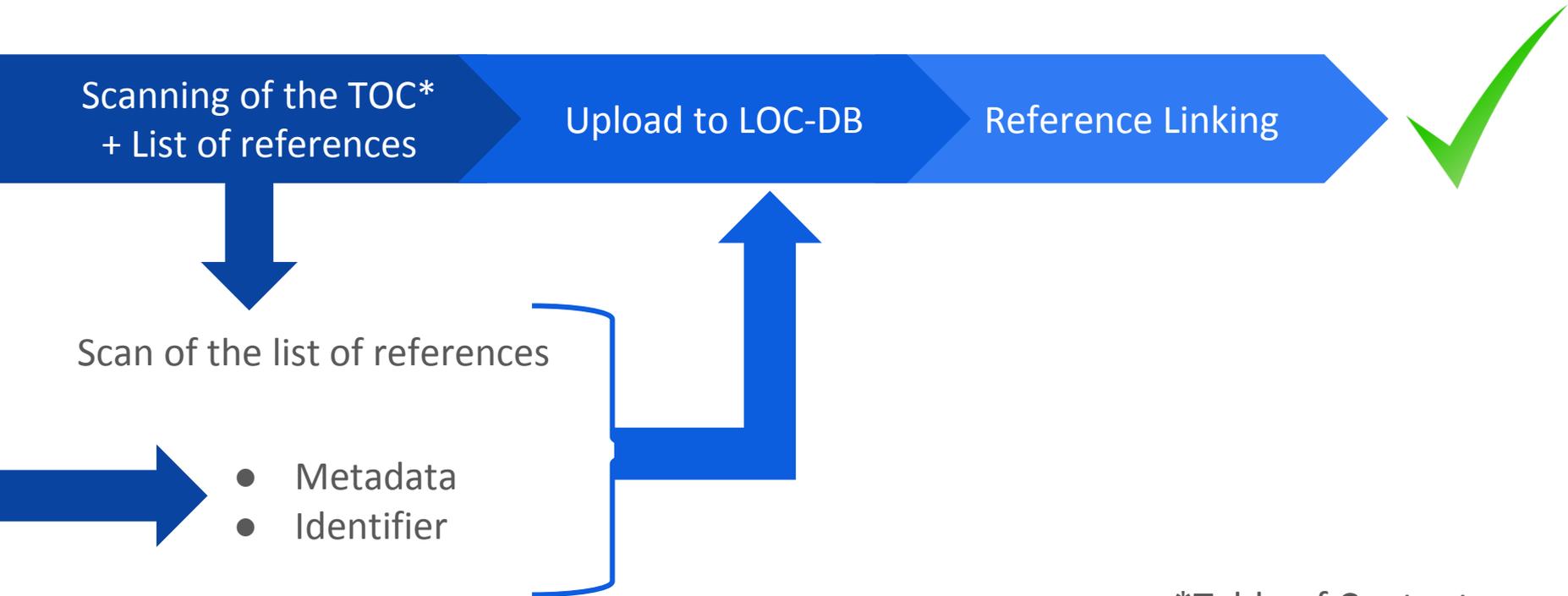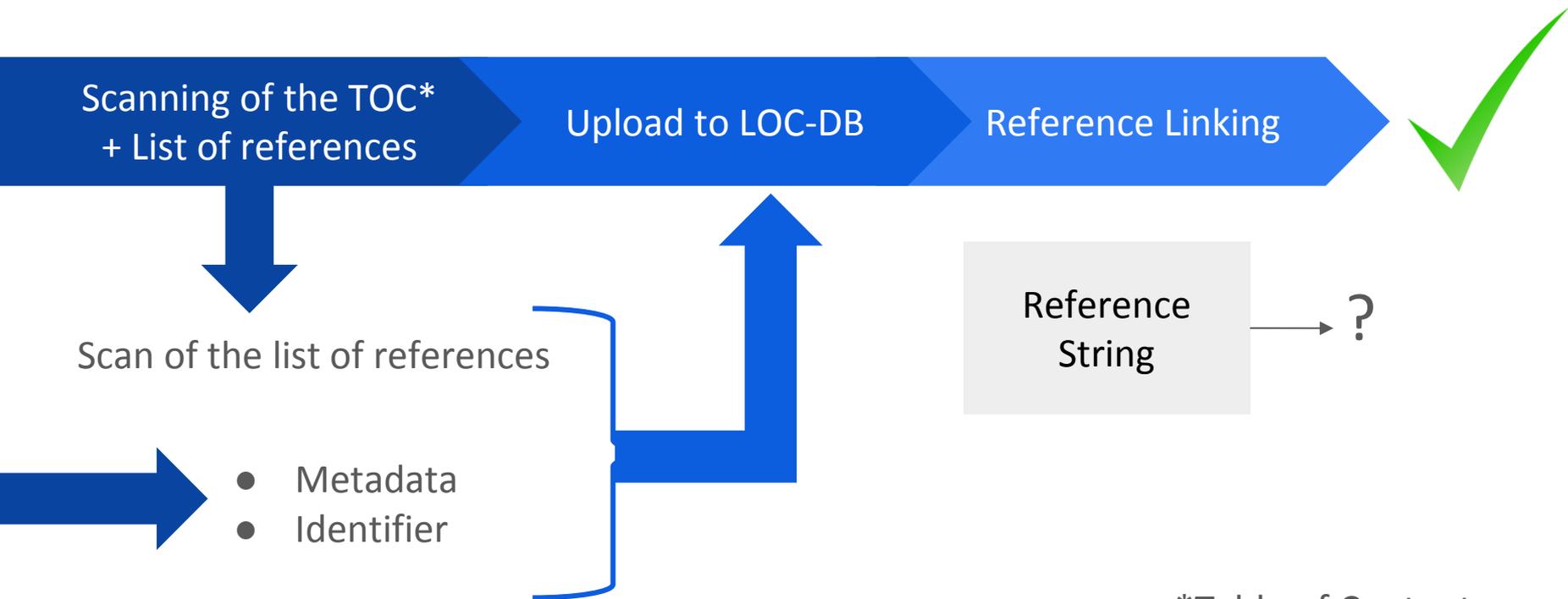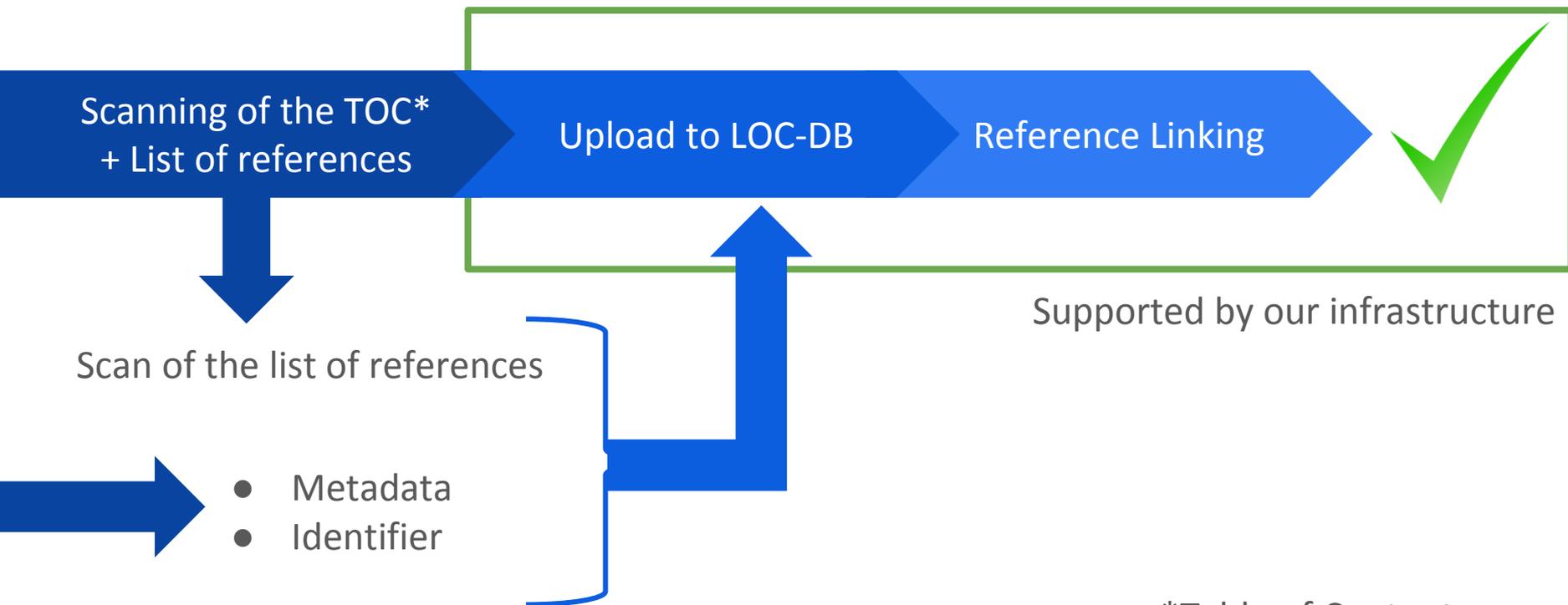
Agenda

1. Motivation
2. Data
3. Reference Linking Workflow
4. **LOC-DB Infrastructure**
5. Evaluation
6. Conclusion

LOC-DB Infrastructure

LOC-DB Infrastructure

PDF Web Print

Editorial System

Linked Open Data

BSZ | Bibliotheksservice-Zentrum Baden-Württemberg

ONLINE CONTENTS
Sozialwissenschaften OLC

Crossref

Automatic Reference Extraction

LOC-DB Instance 1

LOC-DB Instance 2

LOC-DB Instance N
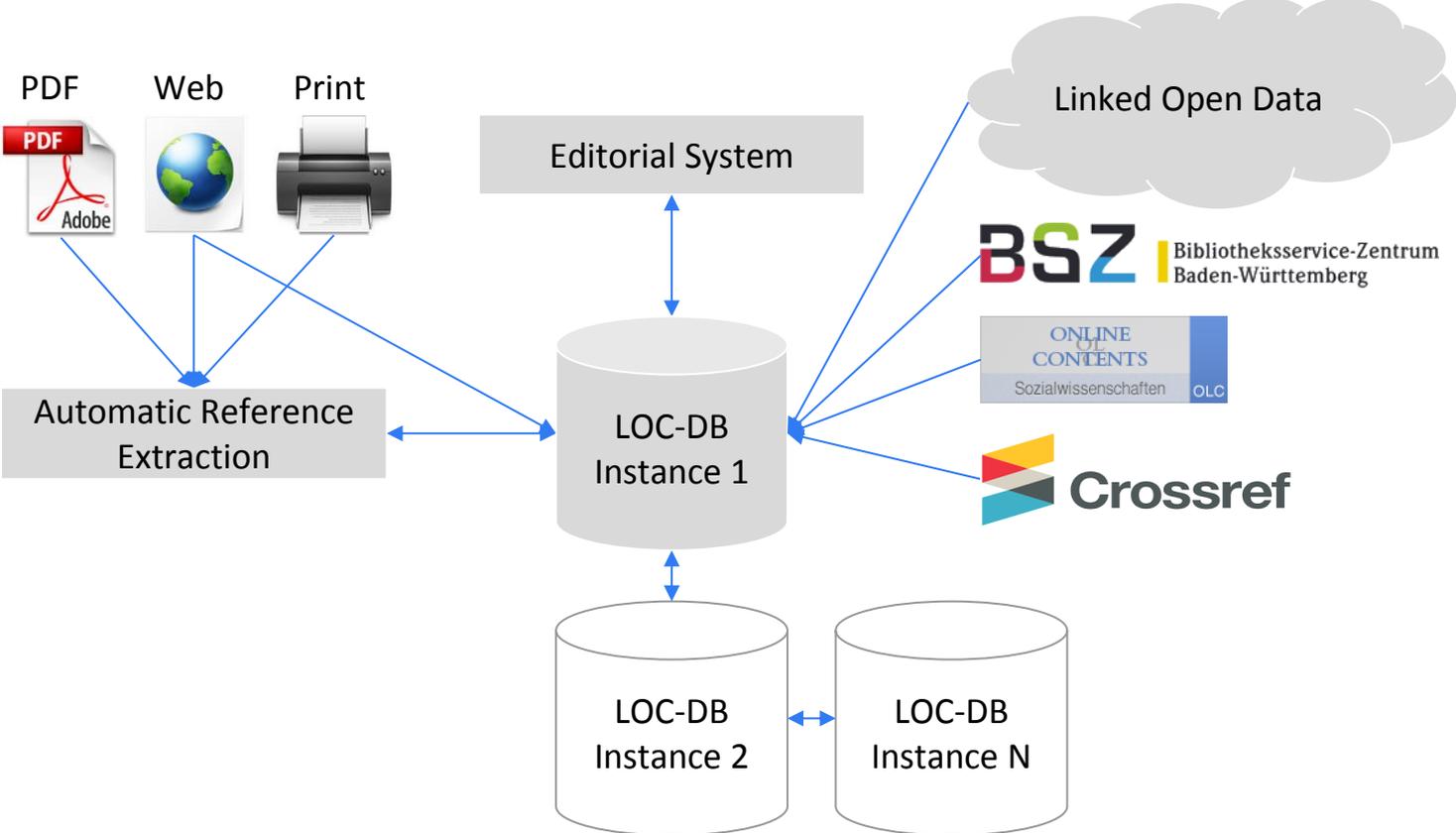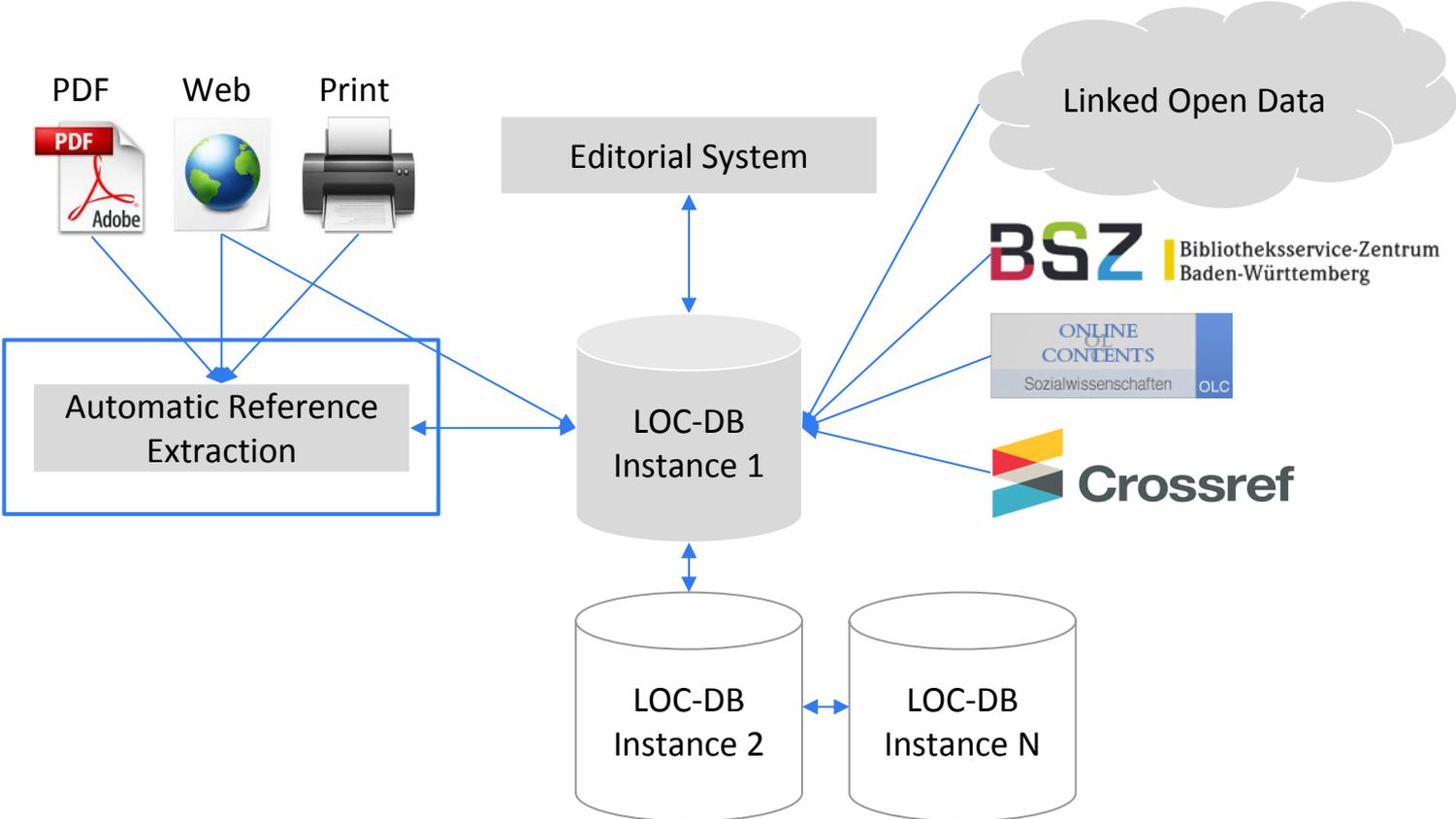
Automatic Reference Extraction

Combination of text-driven
and layout-driven extraction
using deep learning techniques
("DeepBIBX"; Bhardwaj et al., 2017)

LOC-DB Infrastructure

PDF Web Print

Editorial System

Linked Open Data

BSZ | Bibliotheksservice-Zentrum Baden-Württemberg

ONLINE CONTENTS — Sozialwissenschaften OLC

Crossref

Automatic Reference Extraction

LOC-DB Instance 1

LOC-DB Instance 2

LOC-DB Instance N

Agenda

☐ **NOT_OCR_PROCESSED**   ☑ **OCR_PROCESSING**   ☐ **OCR_PROCESSED**   ☐ **EXTERNAL**   **Refresh**

EDITED_BOOK **2017** **Kritische politische Bildung: Standpunkte und Perspektiven /**
, Wochenschau, *In:* Görtler, Michael; Lotz, Mathias; Partetzke, Marc; Poma Poma, Sara; Winckler, Marie (ed.), Kritische politische Bildung: Standpunkte und Perspektiven /, Wochenschau, .

✎ Edit   ☰ References

BOOK_CHAPTER *In:* Görtler, Michael; Lotz, Mathias; Partetzke, Marc; Poma Poma, Sara; Winckler, Marie (ed.), Kritische politische Bildung: Standpunkte und Perspektiven /, Wochenschau, .

✎ Edit   ☰ References

14 - 16 PRINT

- 5b13ffbd48b24e6c788da1f3.pdf

9

Edit Resource

`JOURNAL_ARTICLE` Platt, Jennifer `1974` **The Phenomenon of Sociology: A Reader in the Sociology of Sociology** , SAGE Publications .

Type

JOURNAL_ARTICLE ▼

Title

The Phenomenon of Sociology: A Reader in the Sociology of Sociology

Subtitle

Contributors

✕   AUTHOR ▼   Platt, Jennifer

✕   PUBLISHER ▼   SAGE Publications

Add Contributor

Username: name   **Password:** ••••   Sign in

Die Zukunft Europas :

Back   Reference list   Scan view

Literaturverzeichnis

Albers, Detlev / Haseler, Stephen / Meyer, Henning: Social Europe - A Continent's Answer to Market Fundamentalism; European Research Forum at London Metropolitan University, London 2006.

Baum-Ceising, Alexandra / Faber, Anne (Hrsg.): Soziales Europa? Perspektiven des Wohlfahrtsstaates im Kontext von Europäisierung und Globalisierung: Festschrift für Klaus Busch; Vs Verlag, Wiesbaden 2005.

Beck, Ulrich / Grande, Edgar: Das kosmopolitische Europa; Suhrkamp Verlag, Frankfurt am Main 2004.

Bundeszentrale für politische Bildung: Vertrag von Lissabon, Bonn 2008.

Collignon, Stefan: Bundesrepublik Europa? Die demokratische Herausforderung und Europas Krise; Vorwärts Buch, Berlin 2007.

Delors, Jacques: Erinnerungen eines Europäers; Parthas Verlag, Berlin 2004.

Dullien, Sebastian / Herr, Hansjörg / Kellermann, Christian: Der gute Kapitalismus; Transcript Verlag, Bielefeld 2009.

Eichenhofer, Eberhard: Geschichte des Sozialstaats in Europa: Von der „sozialen Frage" bis zur Globalisierung; C.H. Beck Verlag, München 2007.

Europäische Kommission: Der Vertrag von Lissabon: Einführung, Vertrag über die Europäische Union, Vertrag über die Arbeitsweise der Europäischen Union, Charta der Grundrechte der Europäischen Union, Protokolle, Erklärungen, Deutsches Recht; Autoren der Einführung: Brok, Elmar/Leinen, Jo; Bundeszentrale für politische Bildung, Bonn 2008.

Europäische Union: Amtsblatt der Europäischen Union C 306, 17.12.2007: Vertrag von Lissabon;

Bundesrepublik Europa? Die demokratische Herausforderung und Europas Krisc; Vorwärts Buch, Stefaan CCollignon

Confidence threshold for internal ... and for external Suggestions

1   0,5

Search   Ask Others

Suggestions

External suggestions

Stefan Collignon: Bundesrepublik Europa? Die demokratische Herausforderung und Europas Krise. **CrossRef**

New resource from reference metadata

Bundesrepublik Europa? Die demokratische Herausforderung und Europas Krisc; Vorwärts Buch, **Local**

Edit

21

Ingest Resolve Browse

Scan view   **Reference list**

BOOK_CHAPTER .

Theodor W Adorno. 1967. Resume über Kulturindustrie. In: Adorno,Theodor W. Ohne Leitbild. Parva Aesthetica. FrankfurtIM.: Suhrkarmp,. In: S.   Edit

Theodor W Adorno. 1971. Nas bedeutet: Aufarbeitung der Vergangenheit. In: Adorno,Theodor W.: Erziehung zur Mündigkeit. Vortrige und Gespriche mit Hellmut Becker 1959-1969. FrankfurtIM.: Suhrkamp,. In: S.   Edit

Ien Ang. 1999. Culture and Communications,'Towards an EthnographicCritique of MediaConsumption in theTransnational Media System. In: Edit

Ien Ang. 2005. Who needs cultural esearch? In: Leistyna, Pepi (Hrsg.): Cultural Studies, From 'Theory to Action. Osford: Edit

Carsten Bünger. 2010. Poliische Bildung nach dem „Tod desSubjekts'.. In: In: Losch, Bettinal'Thimmel, Andreas(Hrsg.) Kritischepolitische Bildung. Ein Handbuch. Schwalbachl'Te.: Wochenschau, S.   Edit

Halil Can. 2011. Demokratiearbeit und Empowerment gegen Diskriminierung und Rassismus in selbstbestimmten People of Color-Räumen. In: Castro Varela, Maria do MarlDhawan, Nikita (Hrsg.): Soziale (Un)Gerechtigkeit. Kritische Perspektiven auf Diversity, Intersektionalitit und Antidiskriminierung.   Edit

Resume über Kulturindustrie. In: Adorno,Theodor W. Ohne Leitbild. Parva Aesthetica. FrankfurtIM.: Suhrkarmp, 1967   🔍

Confidence threshold for internal   ... and for external Suggestions

0,8   0,5   Ask Others

Suggestions

External Suggestions

BOOK Adorno, Theodor W. **1970** OHNE LEITBILD. PARVA AESTHETICA. VON THEODOR W. ADORNO. 4. AUFL. FRANKFURT A.M.: SUHRKAMP 1970. 193 S. , **GVI** .   Edit

New resource from reference metadata

Create Resource

22

LOC-DB Infrastructure

PDF   Web   Print

Linked Open Data

Editorial System

BSZ | Bibliotheksservice-Zentrum Baden-Württemberg

ONLINE CONTENTS
Sozialwissenschaften   OLC

Crossref

Automatic Reference Extraction

LOC-DB Instance 1

LOC-DB Instance 2

LOC-DB Instance N
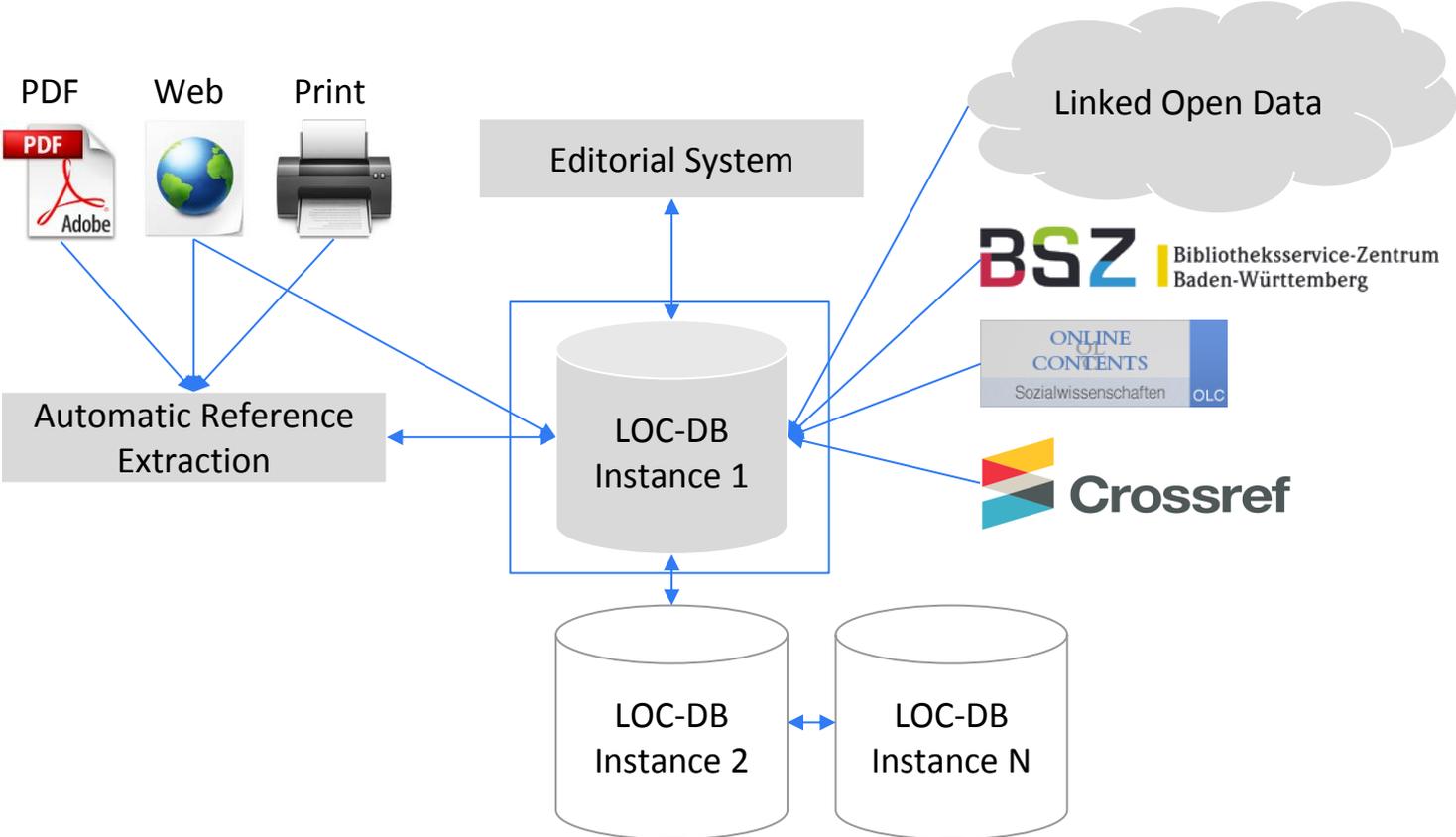
Reference Target Suggestions

Speeding up the linking process
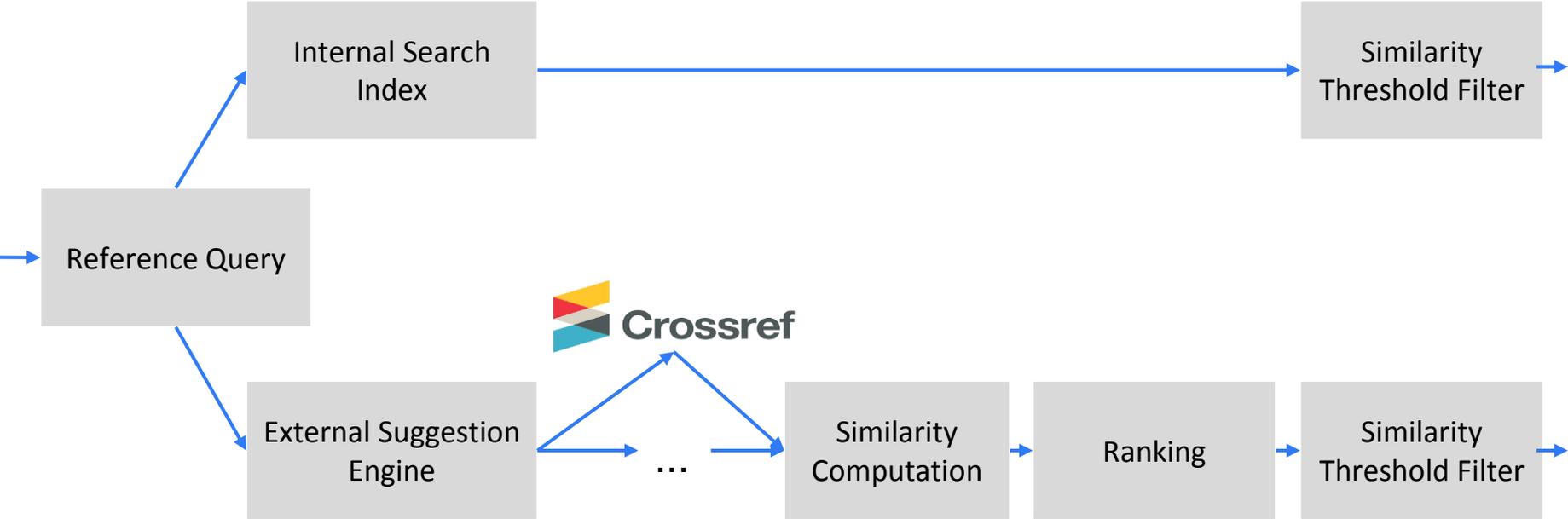
Data Model and Publishing

Ensuring optimal reusability
and interoperability of the produced data

- Adaption of the OpenCitations metadata model
  (Peroni and Shotton, 2016)
- Publishing of the Data in RDF format
  by using the Semantic Publishing
  and Referencing (SPAR) Ontologies (Peroni, 2014)
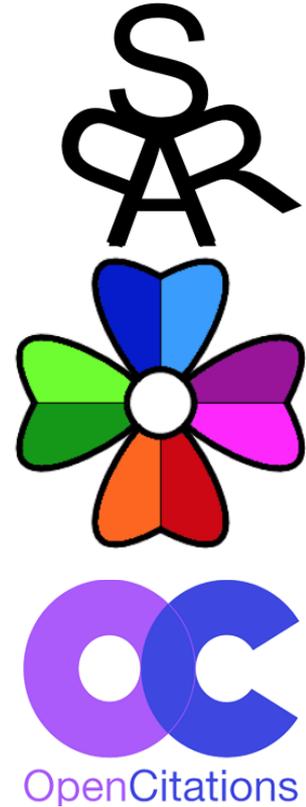
Agenda

1. Motivation
2. Data
3. Reference Linking Workflow
4. LOC-DB Infrastructure
5. **Evaluation**
6. Conclusion

Results of our Public Workshop (Wilke and Retter, 2017)

- Cataloging citation data is
  an important task for libraries
- Data from union catalogs should be used
- Suggestions help to minimize
  the time required to resolve a link
- Acquired citation data should be integrated
  in existing systems

Save the date: The LOC-DB Workshop 2018 will take place on the **6th November, 2018** in Mannheim!

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)
- Additional scanning time does not significantly affect other processes in the library
- Prolongs the processing of a book on average by only 3 minutes

How much time does the whole process take?

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

- 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)	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
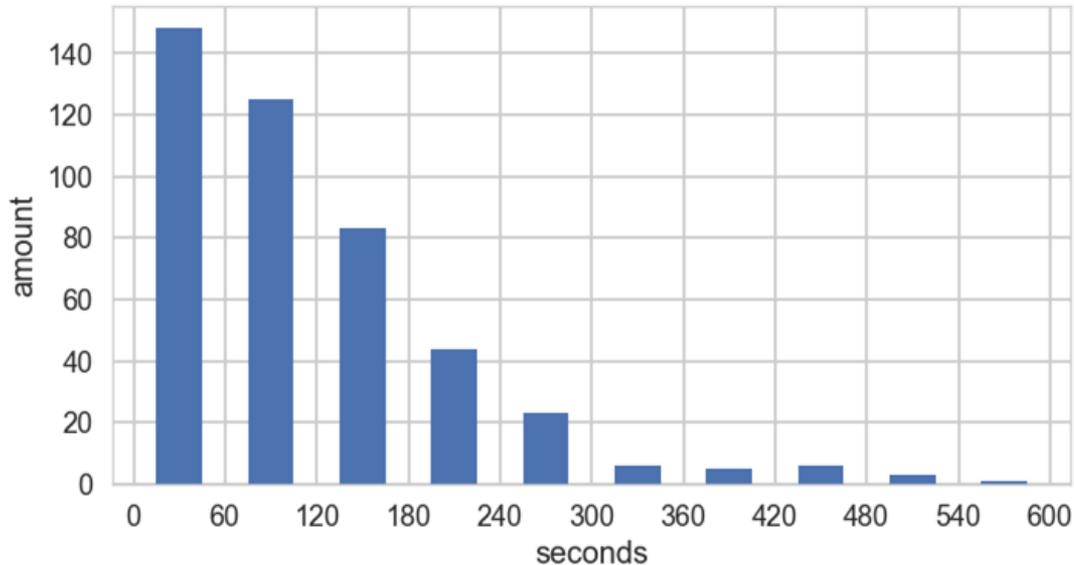
Histogram of reference linking times

Agenda

1. Motivation
2. Data
3. Reference Linking Workflow
4. LOC-DB Infrastructure
5. Evaluation
6. **Conclusion**

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 suggest general feasibility of the approach
Continuous improvement on the infrastructure and processes ongoing

Semi-automated approach ensures human-level quality
of the generated data

Thank you for your interest

For more information please visit
https://locdb.bib.uni-mannheim.de
https://github.com/locdb/
http://wisslab.org
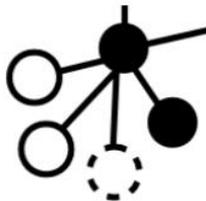
Anne Lauscher & Kai Eckert

Tiessen, Jan (2007): Die Resultate im Blick?
ner/Döhler, Marian (Hrsg.): Agencies in W
Tondorf, Karin/Bahnmüller, Reinhard/Klages,
instrument. Anwendungspraxis, Probleme
sigma.
Touraine, Alain (1984): Le retour de l'acteur: e
Treiber, Hubert (1984): Warum man nicht die
Mikroskop den ganzen Elefanten zu sehen.

Bibliography

- Marshall Breeding. 2015. Future of Library Discovery Systems. Information Standards Quarterly 27, 1 (2015), 24. https://doi.org/10.3789/isqv27no1.2015.04
- Christian Wilke and Regina Retter. 2017. Zitationsdaten extrahieren: halbautomatisch, offen, vernetzt. Ein Workshopbericht. Informationspraxis 3, 2 (Dec. 2017). https://doi.org/10.11588/ip.2017.2.43235
- Silvio Peroni and David Shotton. 2016. Metadata for the OpenCitations Corpus. Technical Report. https://dx.doi.org/10.6084/m9.figshare.3443876
- Akansha Bhardwaj, Dominik Mercier, Andreas Dengel, and Sheraz Ahmed. 2017. DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction. Springer International Publishing, Cham, 286–293. https://doi.org/10.1007/978-3-319-70096-0_30

Bibliography

- Silvio Peroni. 2014. The Semantic Publishing and Referencing Ontologies. In Semantic Web Technologies and Legal Scholarly Publishing. Springer, Cham, 121–193. https://doi.org/10.1007/978-3-319-04777-5_5