Consortium:

UB Mannheim

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Hochschule der Medien

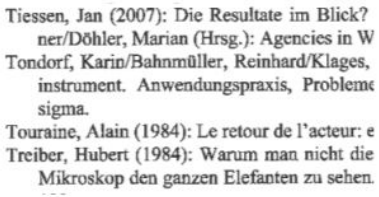
ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Funded by:

DFG

LOC-DB:
A **Linked Open Citation Database**
provided by Libraries.
Motivation and Challenges.

EXCITE Workshop 2017
March 30/31, 2017
GESIS, Cologne



Overview

1. **Motivation**
   Presenter: Kai Eckert, Stuttgart Media University

2. **Infrastructure**
   Presenter: Anne Lauscher, Stuttgart Media University

3. **Reference Extraction**
   Presenter: Akansha Bhardwaj, DFKI Kaiserslautern

Motivation

Scientists:

- ... search for **content**, not for books.
- ... need to check **citations**.
- ... need to know **who cites** the content.
- ... as a special case: need to know who cites **their** content.

Source: https://commons.wikimedia.org/wiki/File:Mad_scientist_transparent_background.svg

Libraries: What is in the catalog?

Basically: Things you can put on a shelf.

But there are databases!

Many commercial, incomplete databases.

Many smaller, focused databases.

Many library services, like the OnLine Contents Fachausschnitte (fka. OLC-SSG).

All are **incomplete** (most do not even try to be complete)

Most are **not freely available as Open Data.**

Most do **not contain citation references.**

What if...

... libraries would just do it?

- **Catalog** journals and collections (proceedings),
- with **all articles / chapters** as separate resources,
- with a structured form of all **citation references** in all articles,
- ideally with **links** to the cited resources.

Linked Open Citation Database

Goal: Answering the question:

What is needed (persons, resources, money)
for libraries to actually “do it”?

(i.e., not the creation of a complete and free citation database, at least for now...)

Method: Prototypical creation of a complete (in the sense of a subset) and free citation database with a focus on cataloging efficiency.

Consortium:

UB Mannheim

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Hochschule der Medien

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Funded by:

DFG

Focus on efficiency

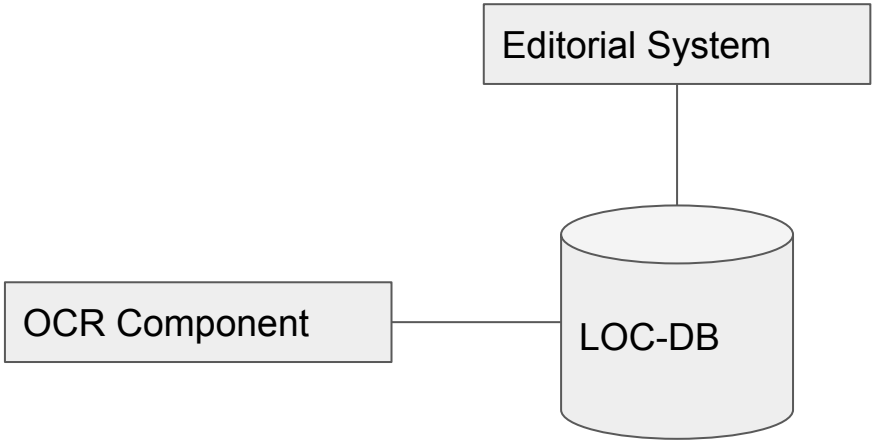
- Reuse existing data (e.g., from publishers or other projects).
- Use OCR.
- Semi-automatic data extraction and linking.
- Streamline and automate the process wherever possible.
- Distributed database and cataloging process.

Desired answer: If X libraries use Y persons to do the LOC-DB cataloging, we manage to get Z percent of the content.

Hopefully with low X and Y and high Z ;-)

Infrastructure

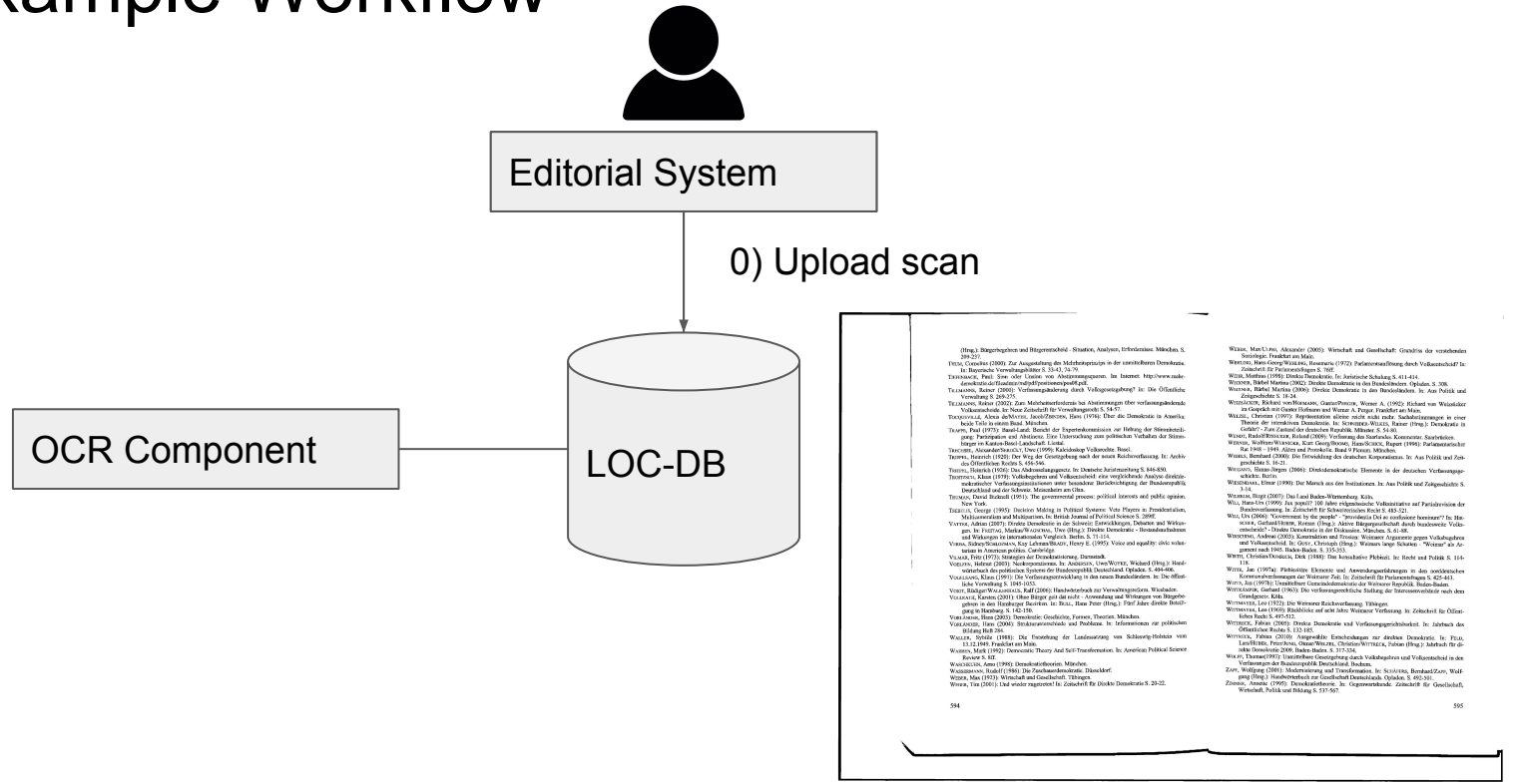
Overview



Main components:

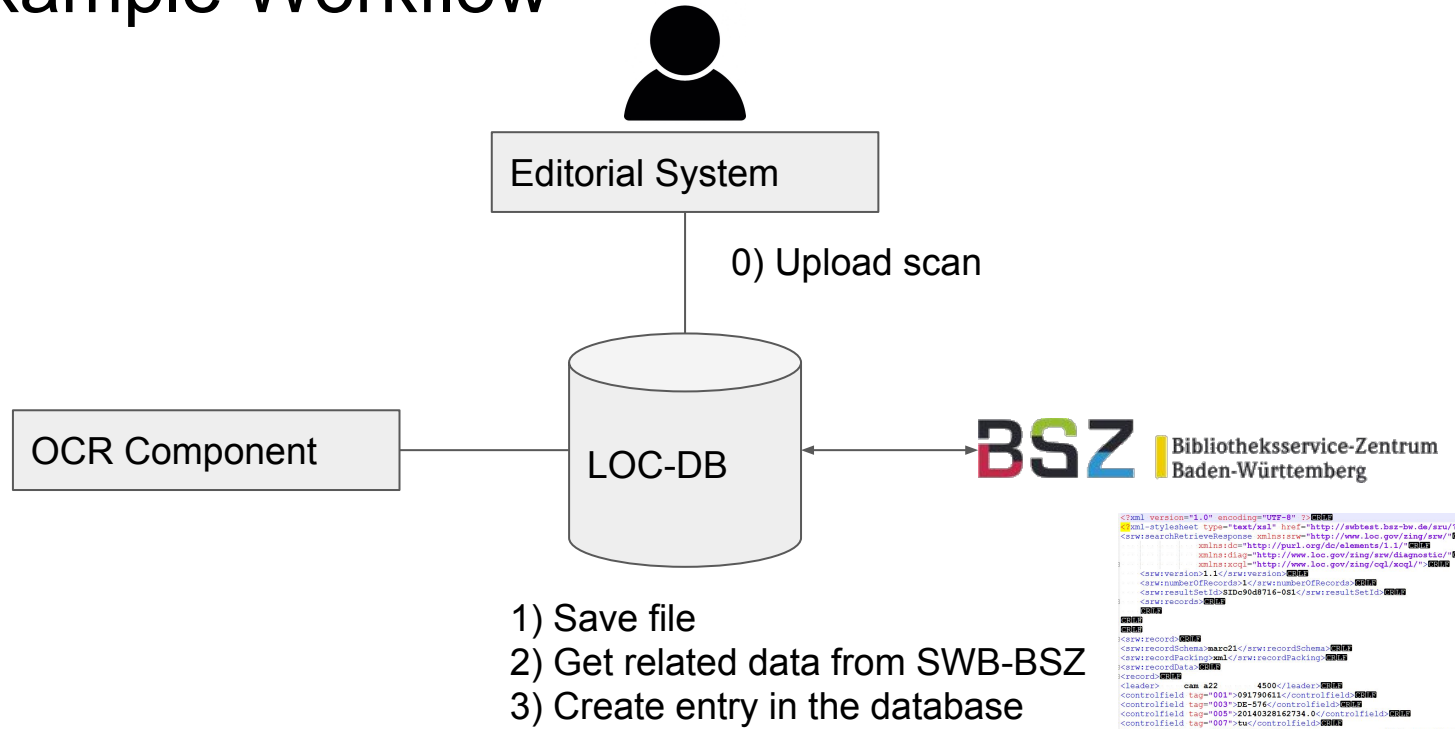
- Editorial System (GUI)
- Central Component (LOC-DB)
- OCR Component

Example Workflow



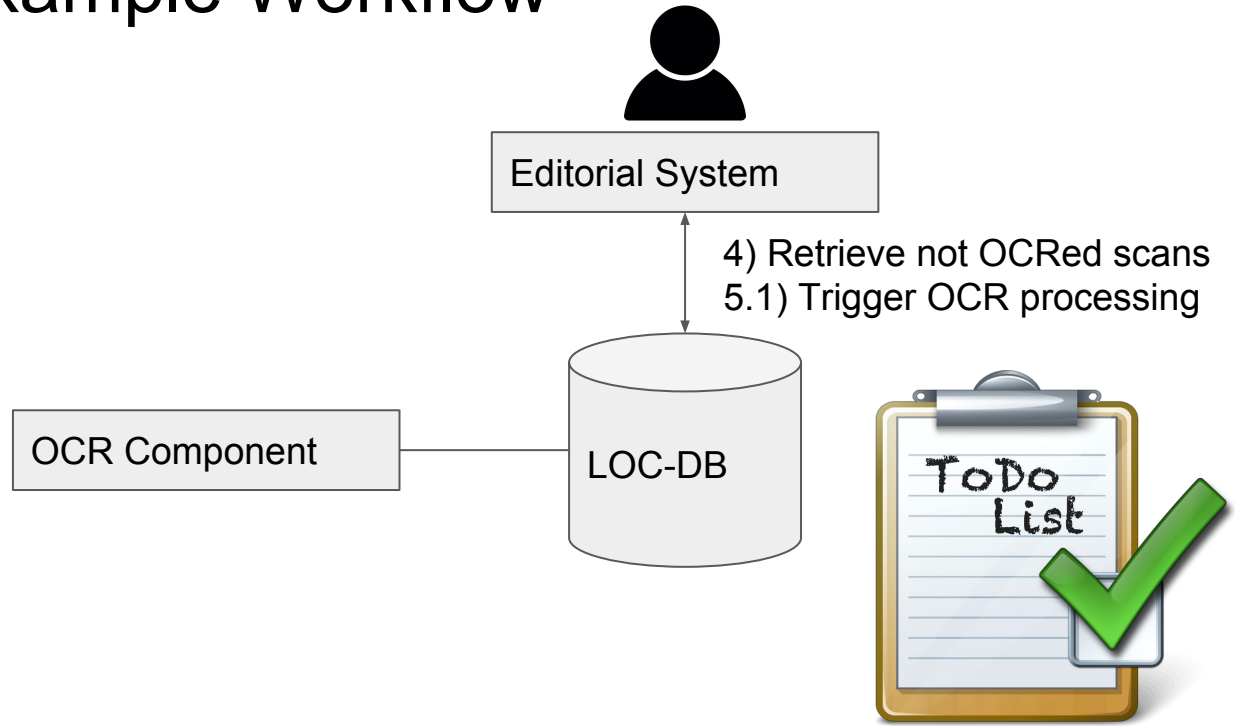
Sources of the images used: [1]

Example Workflow



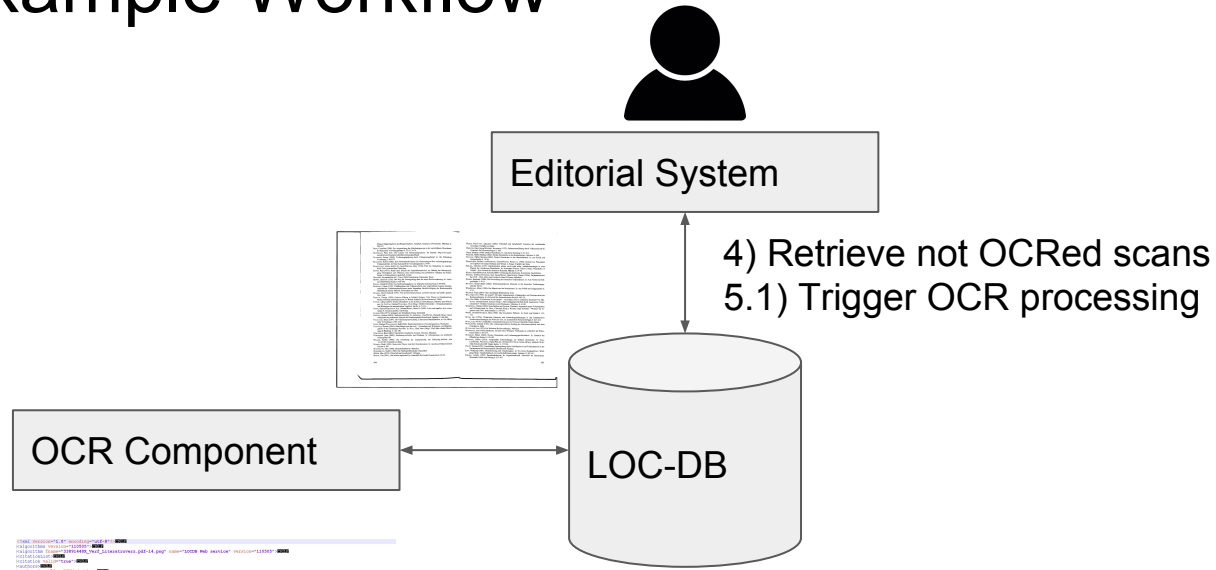
Sources of the images used: [1], [2]

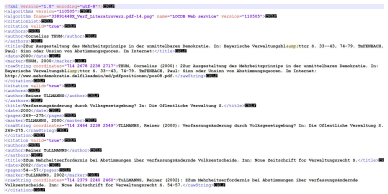
Example Workflow



Sources of the images used: [1], [3]

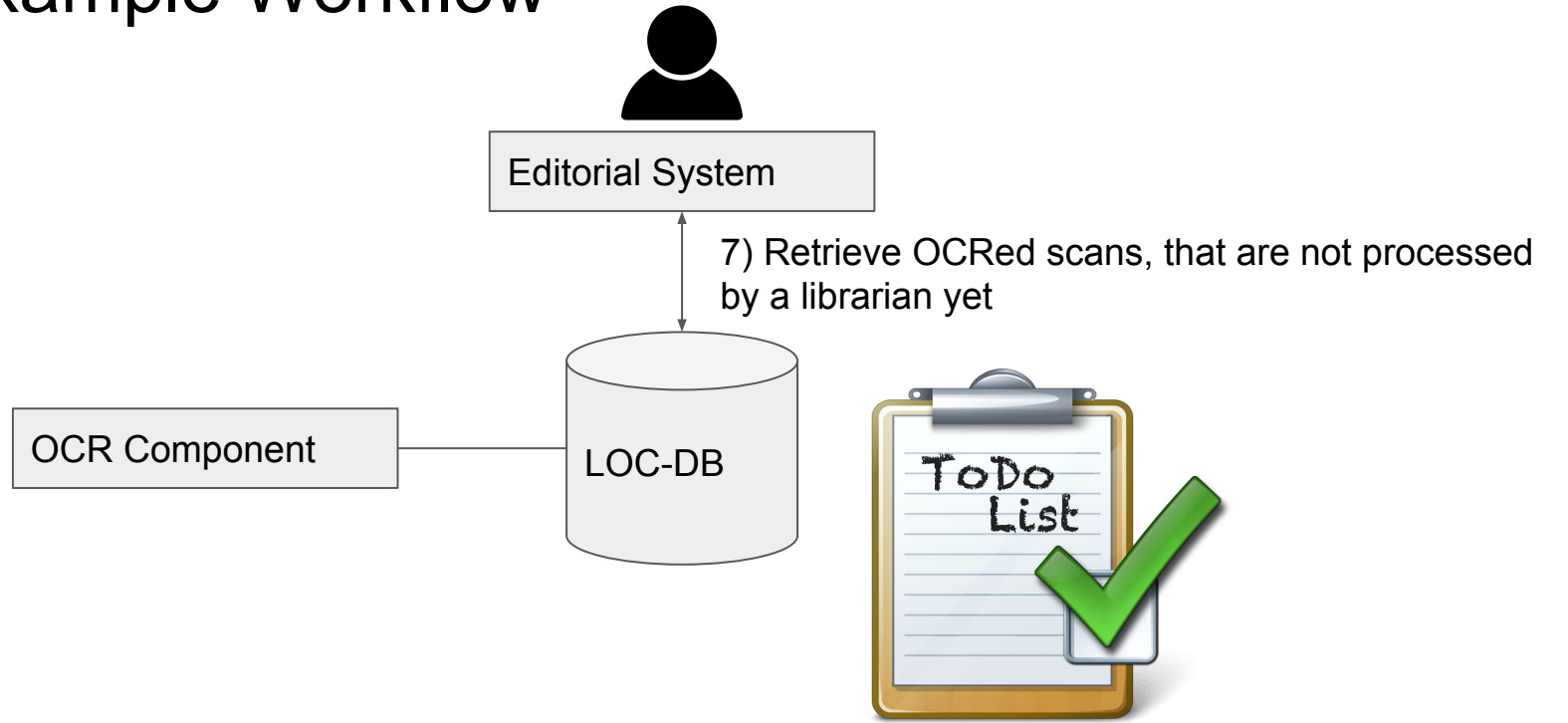
Example Workflow





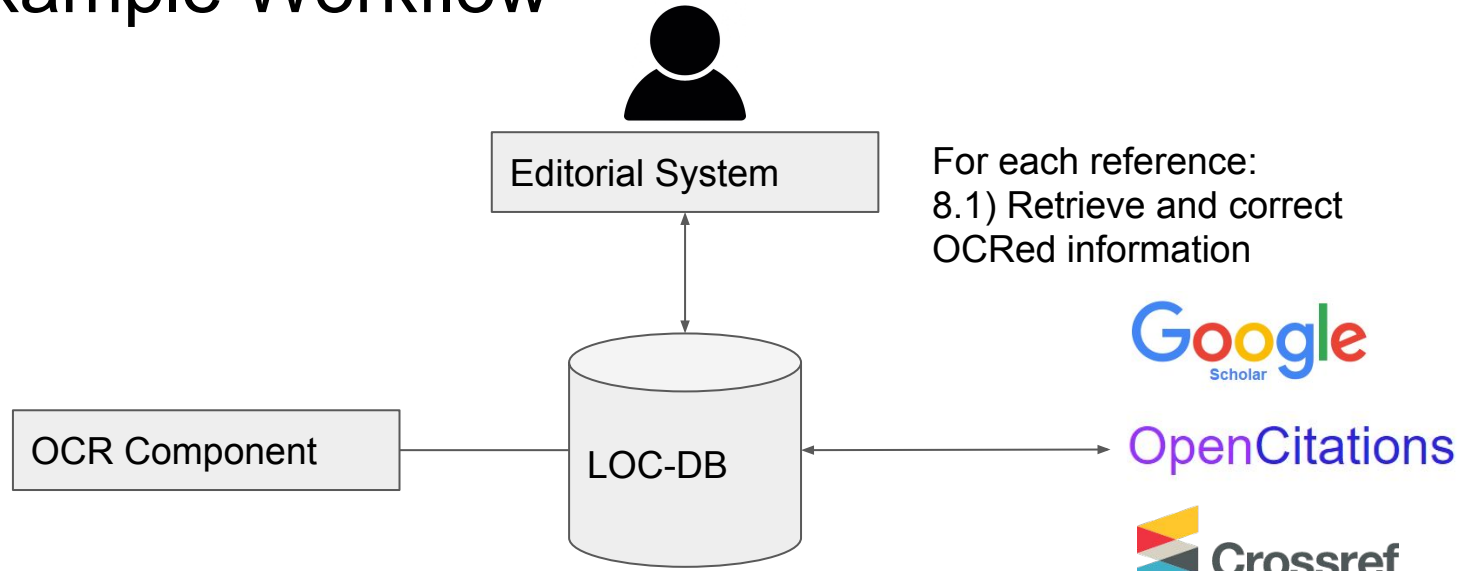
5.2) Trigger OCR processing
6) Retrieve OCRed data (e.g. coordinates of the references on the scanned page) and save it

Example Workflow



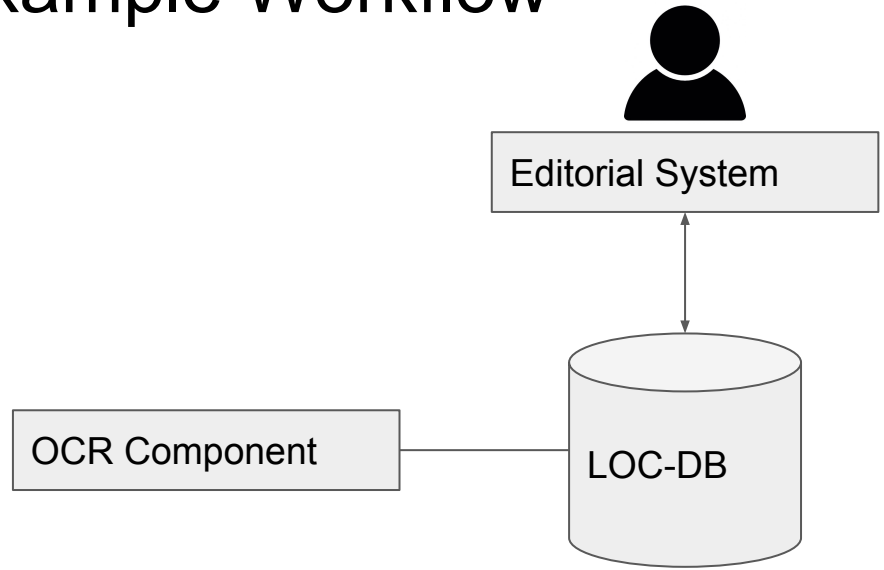
Sources of the images used: [1], [3]

Example Workflow



Sources of the images used: [1], [4], [5], [6]

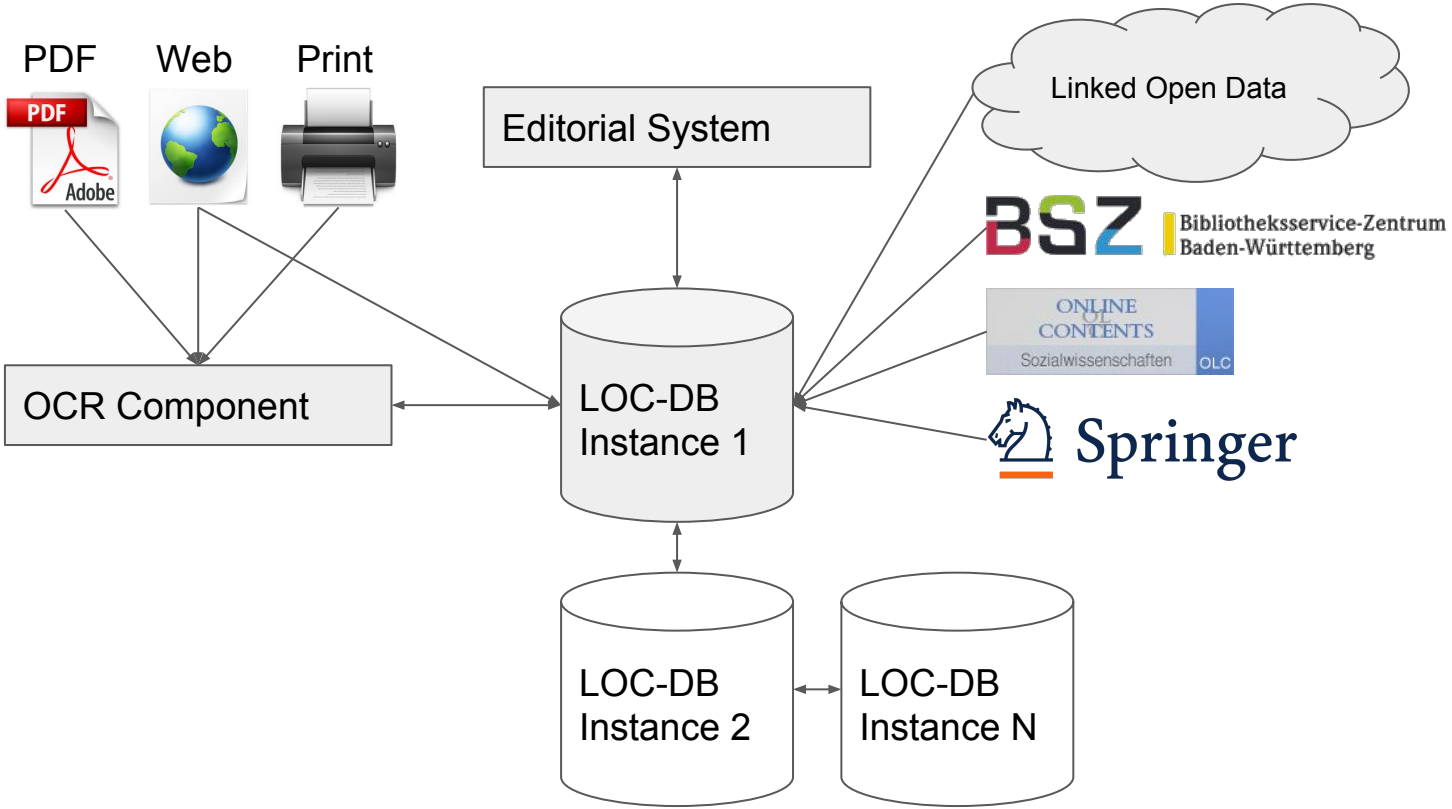
Example Workflow



Result:
A linked open citation database extracted from reference lists created by librarians

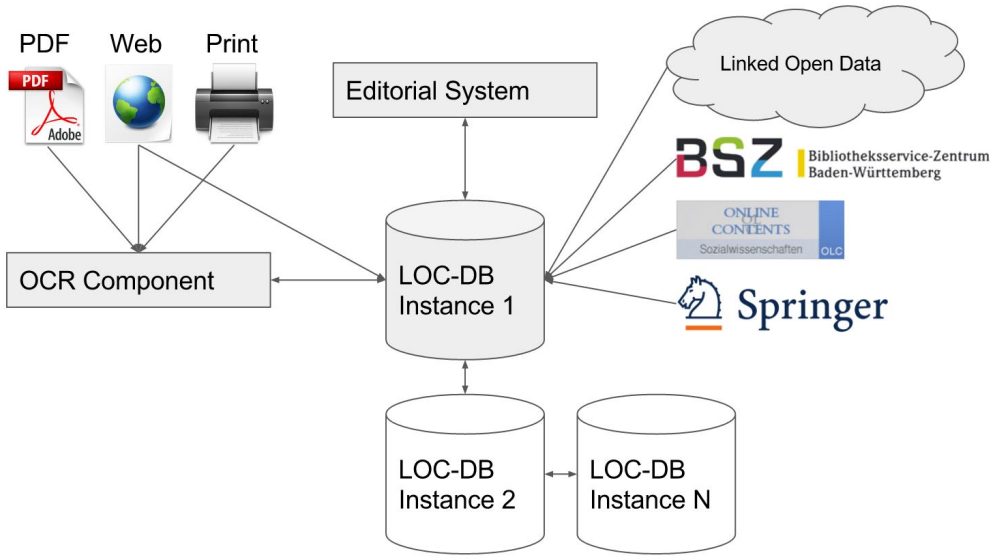
Sources of the images used: [1]

A bit more than just three components..



Sources of the images used: [1], [2], [7], [8], [9], [10],[11]

Technologies



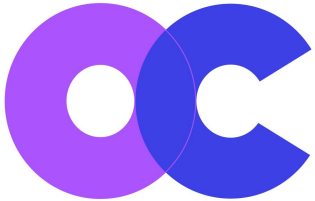
Main components:

- Editorial System (GUI)
  - Angular.io
  - Typescript
  - Bootstrap
- Central Component (LOC-DB)
  - Swagger
  - Node.js
  - MongoDB
- OCR Component

Sources of the images used: [1], [2], [7], [8], [9], [10],[11]

Data Model

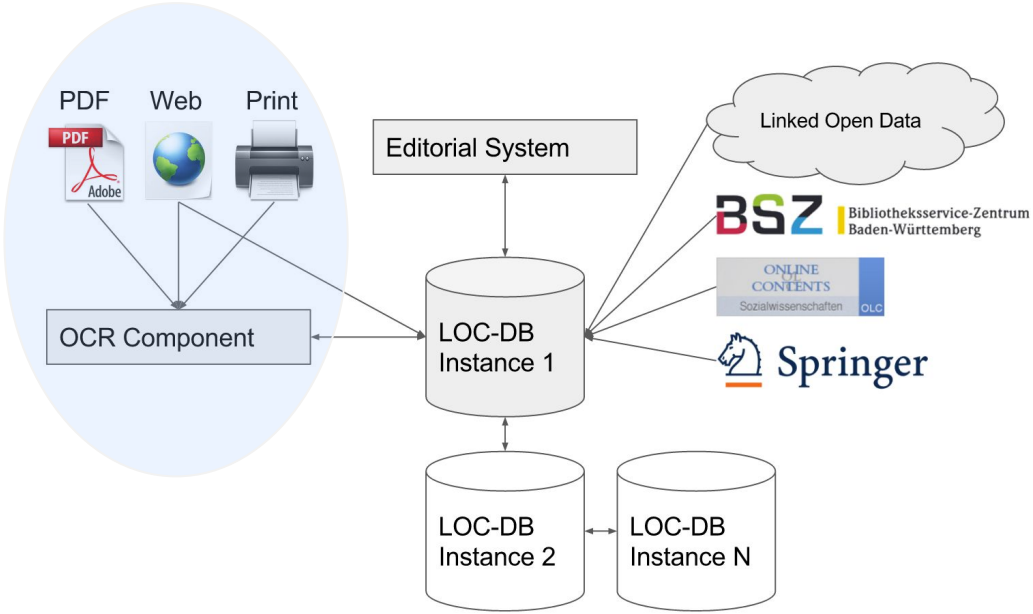
- Inspired by the OpenCitations Data Model
- Extensions to support the management of scans and detected references in different status
- Example: BibliographicEntry (“a single reference”)
  - Identifier of the corresponding scan
  - Status
  - OCRed information, e.g. coordinates of the reference on the scan, title etc.



Sources of the images used: [12]

Reference Extraction

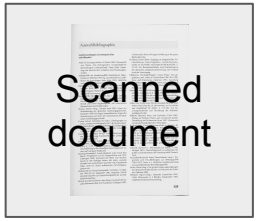
Overview

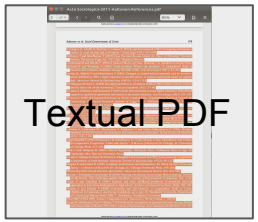


Main components:

- Editorial System (GUI)
- Central Component (LOC-DB)
- OCR Component

Sources of the images used: [1], [2], [7], [8], [9], [10],[11]

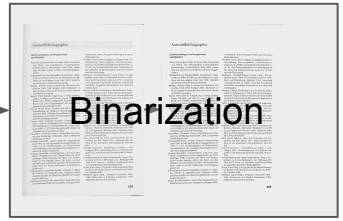




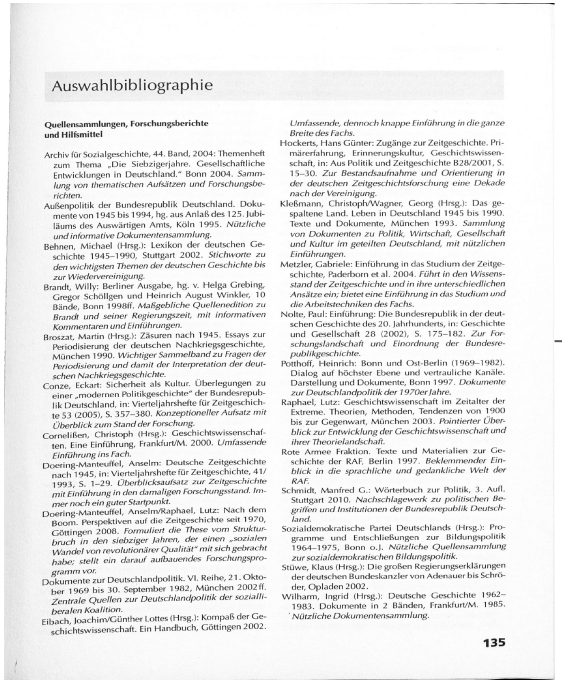
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
−<Publisher xml:lang="en">
  +<PublisherInfo></PublisherInfo>
  −<Journal OutputMedium="Online">
    −<JournalInfo JournalProductType="ArchiveJournal"
     NumberingStyle="Unnumbered">
       <JournalID>41109</JournalID>
       <JournalElectronicISSN>2364-8228</JournalElectronicISSN>
       <JournalTitle>Applied Network Science</JournalTitle>
       <JournalAbbreviatedTitle>Appl Netw
       Sci</JournalAbbreviatedTitle>
      −<JournalSubjectGroup>
         <JournalSubject Code="SCP"
         Type="Primary">Physics</JournalSubject>
         <JournalSubject Code="SCP33010" Priority="1"
         Type="Secondary">Complex Networks</JournalSubject>
         <JournalSubject Code="SCT11022" Priority="2"
         Type="Secondary">Complexity</JournalSubject>
         <JournalSubject Code="SCI21025" Priority="3"
         Type="Secondary">Simulation and Modeling</JournalSubject>
         <SubjectCollection Code="Physics and
         Astronomy">SC12</SubjectCollection>
       </JournalSubjectGroup>
    </JournalInfo>
    −<Volume OutputMedium="Online">
      −<VolumeInfo TocLevels="0" VolumeType="Regular">
         <VolumeIDStart>1</VolumeIDStart>
         <VolumeIDEnd>1</VolumeIDEnd>
         <VolumeIssueCount>1</VolumeIssueCount>
       </VolumeInfo>
      −<Issue IssueType="Regular" OutputMedium="Online">
        +<IssueInfo IssueType="Regular" TocLevels="0"></IssueInfo>
        +<Article ID="s41109-016-0011-2" OutputMedium="Online">
         </Article>
       </Issue>
    </Volume>
  </Journal>
  +<ns0:Info></ns0:Info>
</Publisher>
```

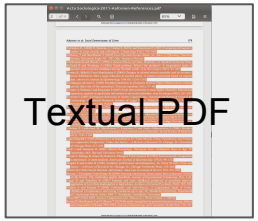
Scanned document



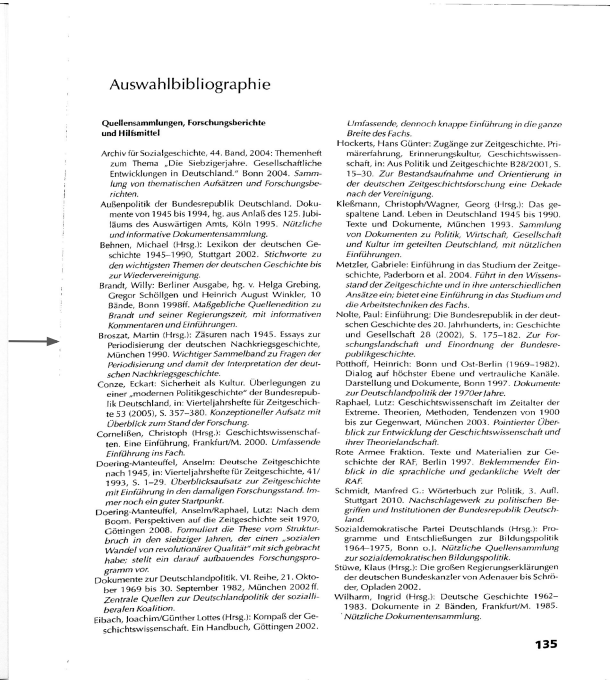
Binarization



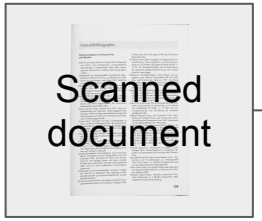
Textual PDF



Structured XML



Scanned document



Binarization

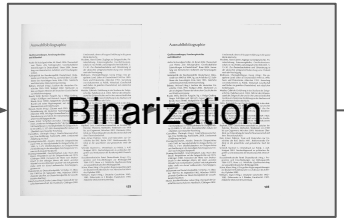
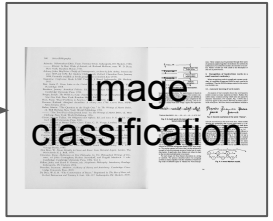
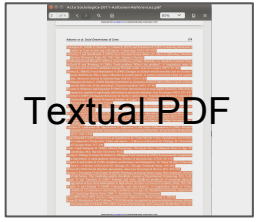


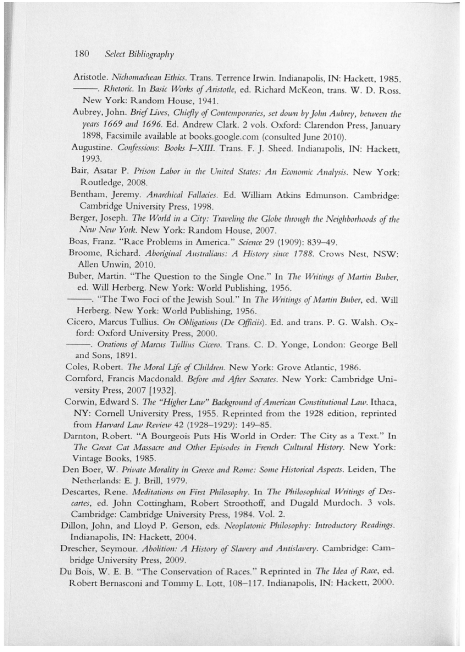
Image classification



Textual PDF



Structured XML



180 *Select Bibliography*

Aristotle. *Nichomachean Ethics*. Trans. Terrence Irwin. Indianapolis, IN: Hackett, 1985.

———. *Rhetoric*. In *Basic Works of Aristotle*, ed. Richard McKeon, trans. W. D. Ross. New York: Random House, 1941.

Aubrey, John. *Brief Lives, Chiefly of Contemporaries, set down by John Aubrey, between the years 1669 and 1696*. Ed. Andrew Clark. 2 vols. Oxford: Clarendon Press, January 1898, Facsimile available at books.google.com (consulted June 2010).

Augustine. *Confessions: Books I–XIII*. Trans. F. J. Sheed. Indianapolis, IN: Hackett, 1993.

Bair, Asatar P. *Prison Labor in the United States: An Economic Analysis*. New York: Routledge, 2008.

Bentham, Jeremy. *Anarchical Fallacies*. Ed. William Atkins Edmunson. Cambridge: Cambridge University Press, 1998.

Berger, Joseph. *The World in a City: Traveling the Globe through the Neighborhoods of the New New York*. New York: Random House, 2007.

Boas, Franz. "Race Problems in America." *Science* 29 (1909): 839–49.

Broome, Richard. *Aboriginal Australians: A History since 1788*. Crows Nest, NSW: Allen Unwin, 2010.

Buber, Martin. "The Question to the Single One." In *The Writings of Martin Buber*, ed. Will Herberg. New York: World Publishing, 1956.

———. "The Two Foci of the Jewish Soul." In *The Writings of Martin Buber*, ed. Will Herberg. New York: World Publishing, 1956.

Cicero, Marcus Tullius. *On Obligations (De Officiis)*. Ed. and trans. P. G. Walsh. Oxford: Oxford University Press, 2000.

———. *Orations of Marcus Tullius Cicero*. Trans. C. D. Yonge, London: George Bell and Sons, 1891.

Coles, Robert. *The Moral Life of Children*. New York: Grove Atlantic, 1986.

Cornford, Francis Macdonald. *Before and After Socrates*. New York: Cambridge University Press, 2007 [1932].

Corwin, Edward S. *The "Higher Law" Background of American Constitutional Law*. Ithaca, NY: Cornell University Press, 1955. Reprinted from the 1928 edition, reprinted from *Harvard Law Review* 42 (1928–1929): 149–85.

Darnton, Robert. "A Bourgeois Puts His World in Order: The City as a Text." In *The Great Cat Massacre and Other Episodes in French Cultural History*. New York: Vintage Books, 1985.

Den Boer, W. *Private Morality in Greece and Rome: Some Historical Aspects*. Leiden, The Netherlands: E. J. Brill, 1979.

Descartes, Rene. *Meditations on First Philosophy*. In *The Philosophical Writings of Descartes*, ed. John Cottingham, Robert Stroothoff, and Dugald Murdoch. 3 vols. Cambridge: Cambridge University Press, 1984. Vol. 2.

Dillon, John, and Lloyd P. Gerson, eds. *Neoplatonic Philosophy: Introductory Readings*. Indianapolis, IN: Hackett, 2004.

Drescher, Seymour. *Abolition: A History of Slavery and Antislavery*. Cambridge: Cambridge University Press, 2009.

Du Bois, W. E. B. "The Conservation of Races." Reprinted in *The Idea of Race*, ed. Robert Bernasconi and Tommy L. Lott, 108–117. Indianapolis, IN: Hackett, 2000.

Auswahlbibliographie

Quellensammlungen, Forschungsberichte und Hilfsmittel

Archiv für Sozialgeschichte, 44. Band, 2004: Themenheft zum Thema „Die Siebzigerjahre. Gesellschaftliche Entwicklungen in Deutschland." Bonn 2004. *Sammlung von thematischen Aufsätzen und Forschungsberichten.*

Außenpolitik der Bundesrepublik Deutschland. Dokumente von 1945 bis 1994, hg. aus Anlaß des 125. Jubiläums des Auswärtigen Amts, Köln 1995. *Nützliche und informative Dokumentensammlung.*

Behnen, Michael (Hrsg.): Lexikon der deutschen Geschichte 1945–1990, Stuttgart 2002. *Stichworte zu den wichtigsten Themen der deutschen Geschichte bis zur Wiedervereinigung.*

Brandt, Willy: Berliner Ausgabe, hg. v. Helga Grebing, Gregor Schöllgen und Heinrich August Winkler, 10 Bände, Bonn 1998ff. *Maßgebliche Quellenedition zu Brandt und seiner Regierungszeit, mit informativen Kommentaren und Einführungen.*

Broszat, Martin (Hrsg.): Zäsuren nach 1945. Essays zur Periodisierung der deutschen Nachkriegsgeschichte, München 1990. *Wichtiger Sammelband zu Fragen der Periodisierung und damit der Interpretation der deutschen Nachkriegsgeschichte.*

Conze, Eckart: Sicherheit als Kultur. Überlegungen zu einer „modernen Politikgeschichte" der Bundesrepublik Deutschland, in: Vierteljahrshefte für Zeitgeschichte 53 (2005), S. 357–380. *Konzeptioneller Aufsatz mit Überblick zum Stand der Forschung.*

Cornelißen, Christoph (Hrsg.): Geschichtswissenschaften. Eine Einführung, Frankfurt/M. 2000. *Umfassende Einführung ins Fach.*

Doering-Manteuffel, Anselm: Deutsche Zeitgeschichte nach 1945, in: Vierteljahrshefte für Zeitgeschichte, 43/1993, S. 1–29. *Überblicksaufsatz zur Zeitgeschichte mit Einführung in den damaligen Forschungsstand. Immer noch ein guter Startpunkt.*

Doering-Manteuffel, Anselm/Raphael, Lutz: Nach dem Boom. Perspektiven auf die Zeitgeschichte seit 1970, Göttingen 2008. *Formuliert die These vom Strukturbruch in den siebziger Jahren, der einen „sozialen Wandel von revolutionärer Qualität" mit sich gebracht habe; stellt ein darauf aufbauendes Forschungsprogramm vor.*

Dokumente zur Deutschlandpolitik. VI. Reihe, 21. Oktober 1969 bis 30. September 1982, München 2002ff. *Zentrale Quellen zur Deutschlandpolitik der sozialliberalen Koalition.*

Eibach, Joachim/Günther Lottes (Hrsg.): Kompaß der Geschichtswissenschaft. Ein Handbuch, Göttingen 2002. *Umfassende, dennoch knappe Einführung in die ganze Breite des Fachs.*

Hockerts, Hans Günter: Zugänge zur Zeitgeschichte. Primärerfahrung, Erinnerungskultur, Geschichtswissenschaft, in: Aus Politik und Zeitgeschichte B28/2001, S. 15–30. *Zur Bestandsaufnahme und Orientierung in der deutschen Zeitgeschichtsforschung eine Dekade nach der Vereinigung.*

Kleßmann, Christoph/Wagner, Georg (Hrsg.): Das gespaltene Land. Leben in Deutschland 1945 bis 1990. Texte und Dokumente, München 1993. *Sammlung von Dokumenten zu Politik, Wirtschaft, Gesellschaft und Kultur im geteilten Deutschland, mit nützlichen Einführungen.*

Metzler, Gabriele: Einführung in das Studium der Zeitgeschichte, Paderborn et al. 2004. *Führt in den Wissensstand der Zeitgeschichte und in ihre unterschiedlichen Ansätze ein; bietet eine Einführung in das Studium und die Arbeitstechniken des Fachs.*

Nolte, Paul: Einführung: Die Bundesrepublik in der deutschen Geschichte des 20. Jahrhunderts, in: Geschichte und Gesellschaft 28 (2002), S. 175–182. *Zur Forschungslandschaft und Einordnung der Bundesrepublikgeschichte.*

Potthoff, Heinrich: Bonn und Ost-Berlin (1969–1982). Dialog auf höchster Ebene und vertrauliche Kanäle. Darstellung und Dokumente, Bonn 1997. *Dokumente zur Deutschlandpolitik der 1970er Jahre.*

Raphael, Lutz: Geschichtswissenschaft im Zeitalter der Extreme. Theorien, Methoden, Tendenzen von 1900 bis zur Gegenwart, München 2003. *Pointierter Überblick zur Entwicklung der Geschichtswissenschaft und ihrer Theorielandschaft.*

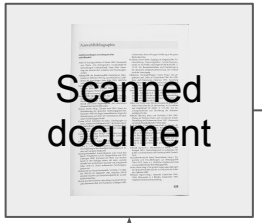
Rote Armee Fraktion. Texte und Materialien zur Geschichte der RAF. Berlin 1997. *Bokkemmender Einblick in die sprachliche und gedankliche Welt der RAF.*

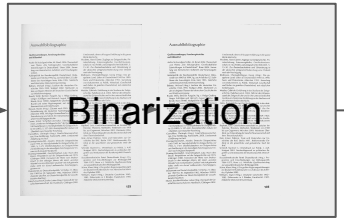
Schmidt, Manfred G.: Wörterbuch zur Politik, 3. Aufl. Stuttgart 2010. *Nachschlagewerk zu politischen Begriffen und Institutionen der Bundesrepublik Deutschland.*

Sozialdemokratische Partei Deutschlands (Hrsg.): Programme und Entschließungen zur Bildungspolitik 1964–1975, Bonn o.J. *Nützliche Quellensammlung zur sozialdemokratischen Bildungspolitik.*

Stüwe, Klaus (Hrsg.): Die großen Regierungserklärungen der deutschen Bundeskanzler von Adenauer bis Schröder, Opladen 2002.

Wilharm, Ingrid (Hrsg.): Deutsche Geschichte 1962–1983. Dokumente in 2 Bänden, Frankfurt/M. 1985. *Nützliche Dokumentensammlung.*


Scanned document


Binarization

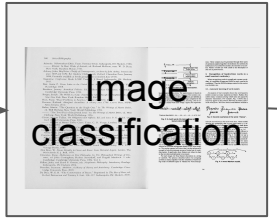
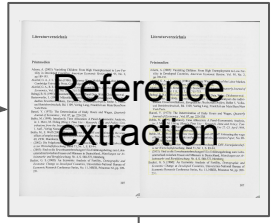
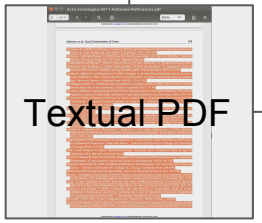
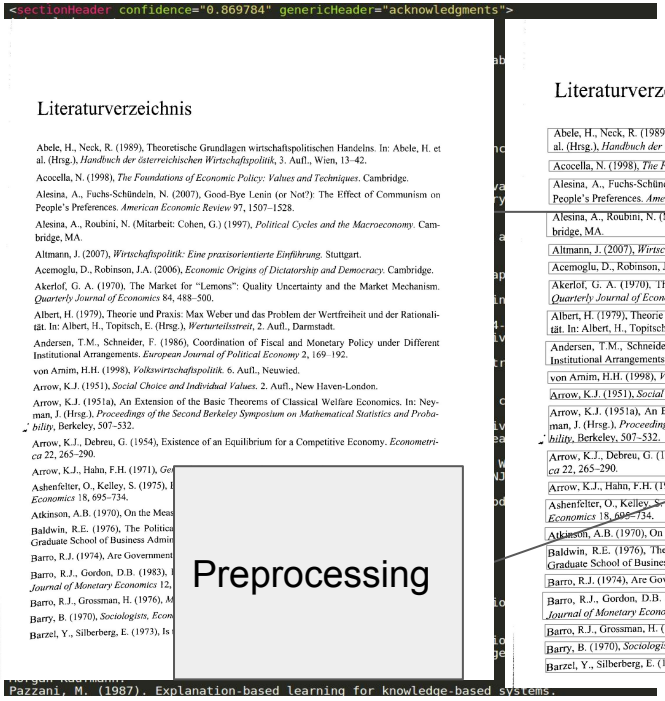


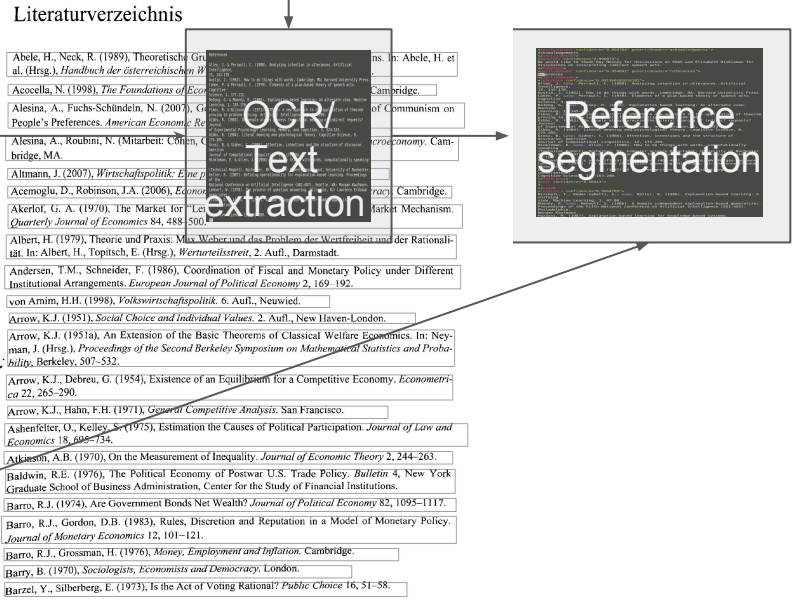
Image classification


Reference extraction

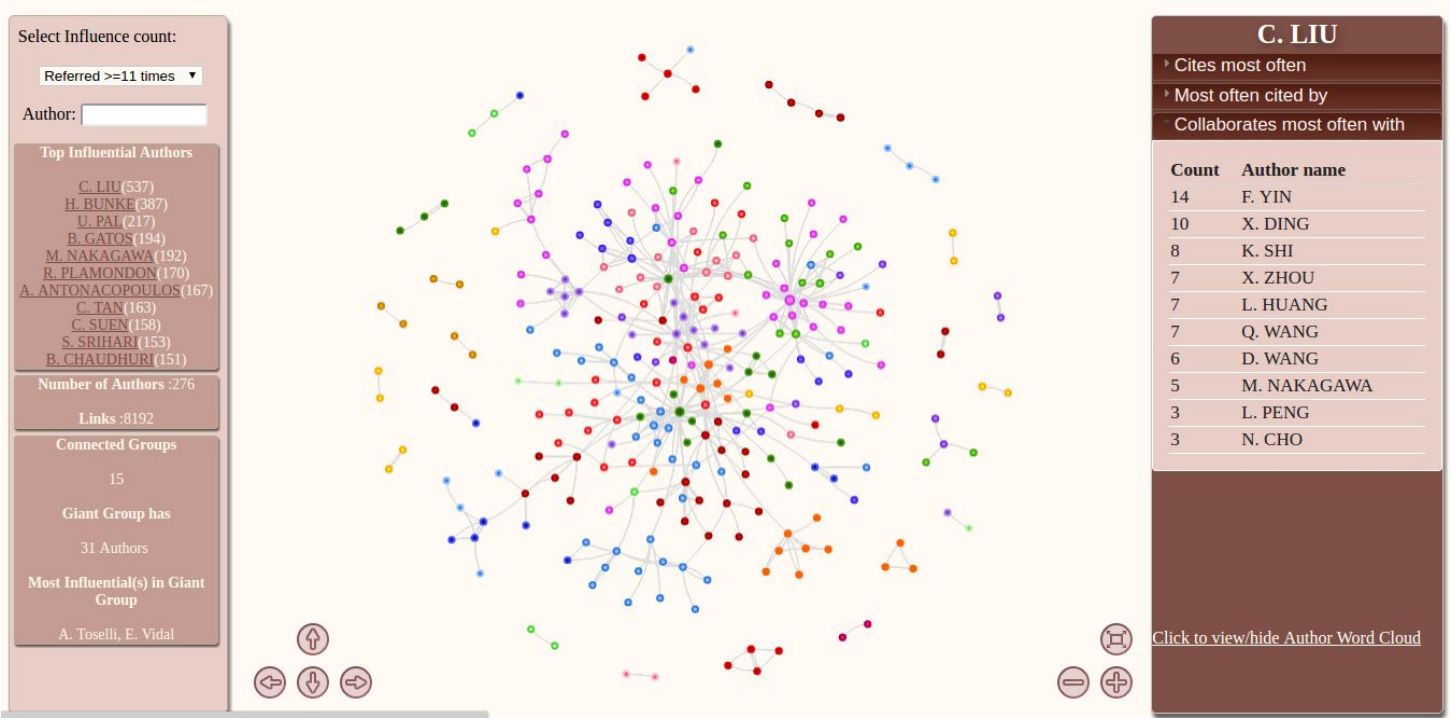

Textual PDF


Structured XML

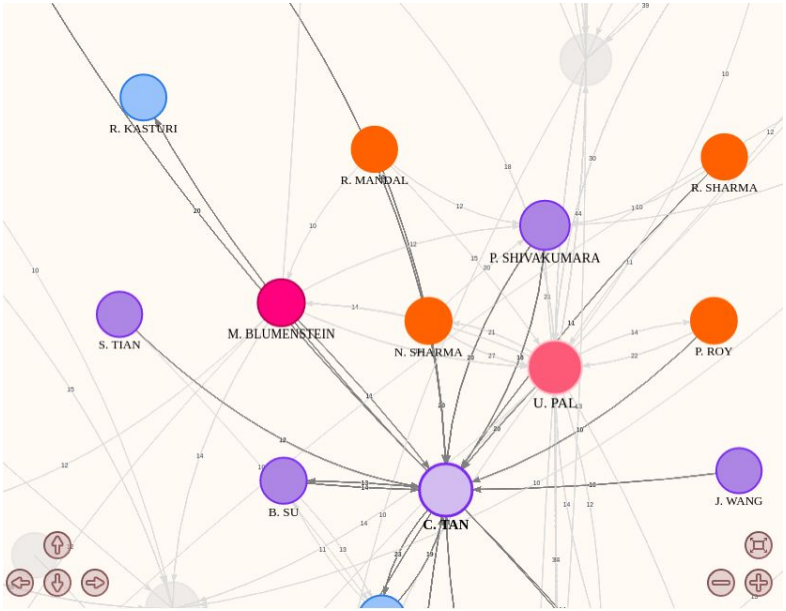

<sectionHeader confidence="0.869784" genericHeader="acknowledgments">
Literaturverzeichnis
Preprocessing

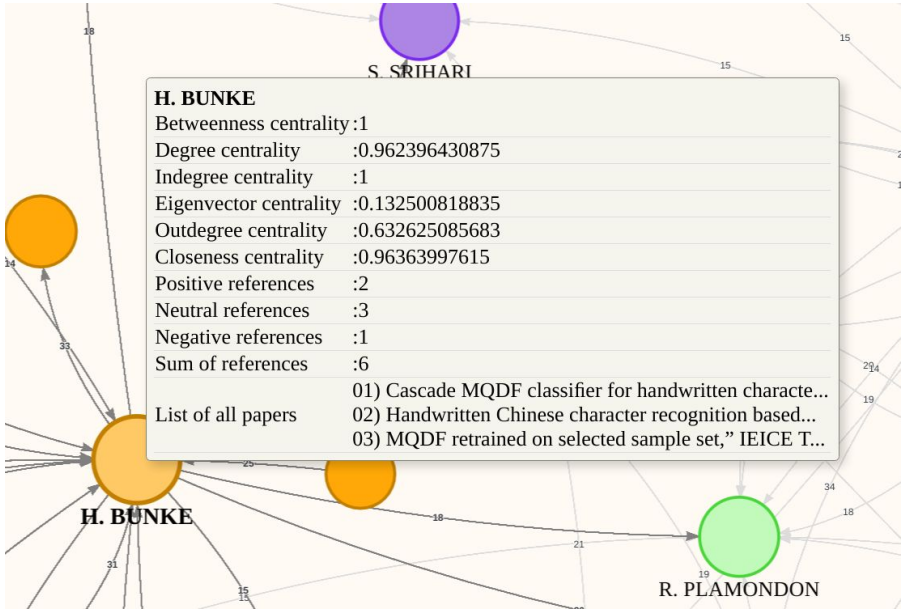

Literaturverzeichnis
OCR / Text extraction
Reference segmentation

Appendix

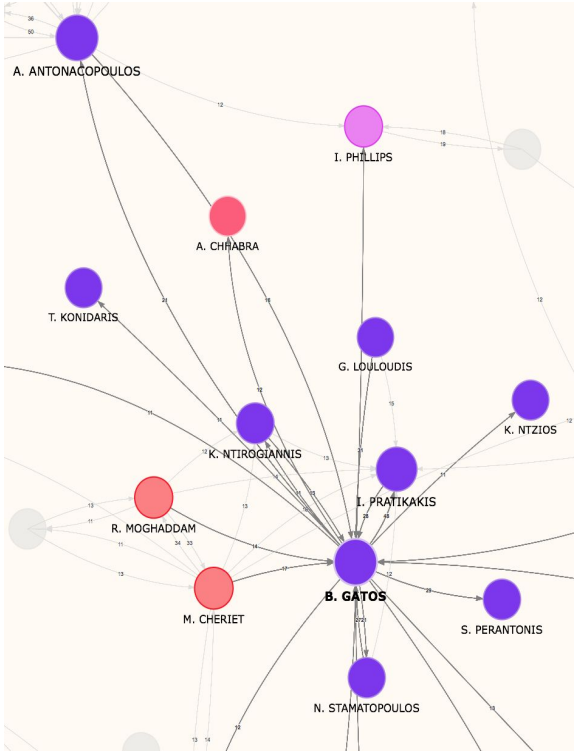

Select Influence count:
Referred >=11 times
Author:
Top Influential Authors
C. LIU(537)
H. BUNKE(387)
U. PAL(217)
B. GATOS(194)
M. NAKAGAWA(192)
R. PLAMONDON(170)
A. ANTONACOPOULOS(167)
C. TAN(163)
C. SUEN(158)
S. SRIHARI(153)
B. CHAUDHURI(151)
Number of Authors :276
Links :8192
Connected Groups
15
Giant Group has
31 Authors
Most Influential(s) in Giant Group
A. Toselli, E. Vidal
C. LIU
Cites most often
Most often cited by
Collaborates most often with
Count
Author name
14 F. YIN
10 X. DING
8 K. SHI
7 X. ZHOU
7 L. HUANG
7 Q. WANG
6 D. WANG
5 M. NAKAGAWA
3 L. PENG
3 N. CHO
Click to view/hide Author Word Cloud

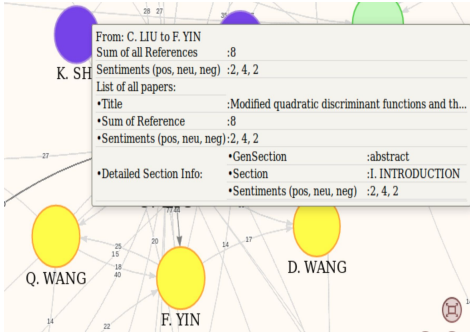
Appendix

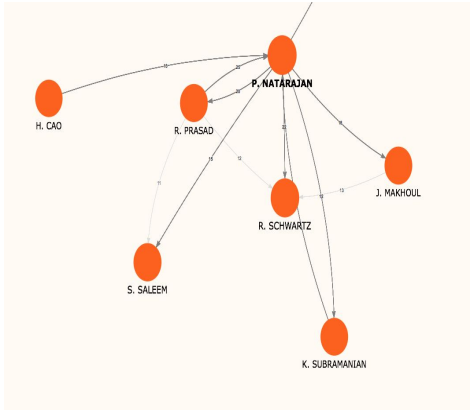

R. KASTURI
R. MANDAL
R. SHARMA
P. SHIVAKUMARA
S. TIAN
M. BLUMENSTEIN
N. SHARMA
P. ROY
U. PAL
B. SU
C. TAN
J. WANG

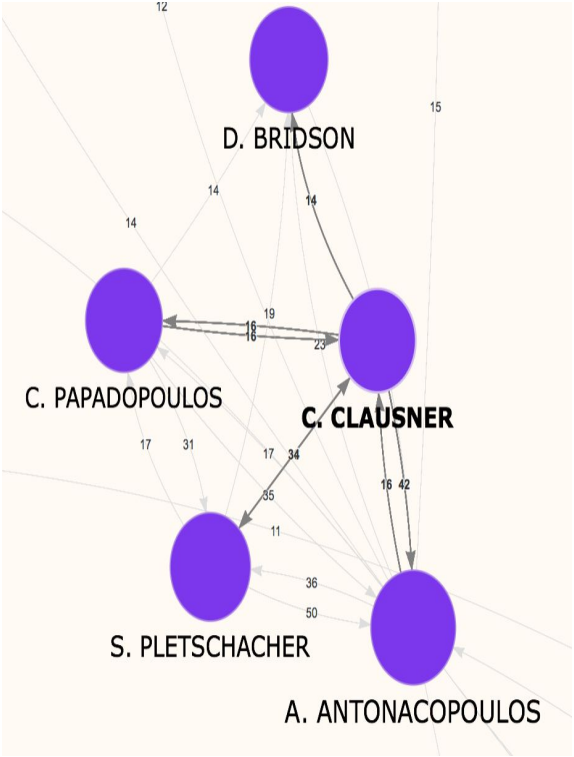

S. SRIHARI
H. BUNKE
Betweenness centrality :1
Degree centrality :0.962396430875
Indegree centrality :1
Eigenvector centrality :0.132500818835
Outdegree centrality :0.632625085683
Closeness centrality :0.96363997615
Positive references :2
Neutral references :3
Negative references :1
Sum of references :6
List of all papers
01) Cascade MQDF classifier for handwritten characte...
02) Handwritten Chinese character recognition based...
03) MQDF retrained on selected sample set," IEICE T...
H. BUNKE
R. PLAMONDON

Appendix

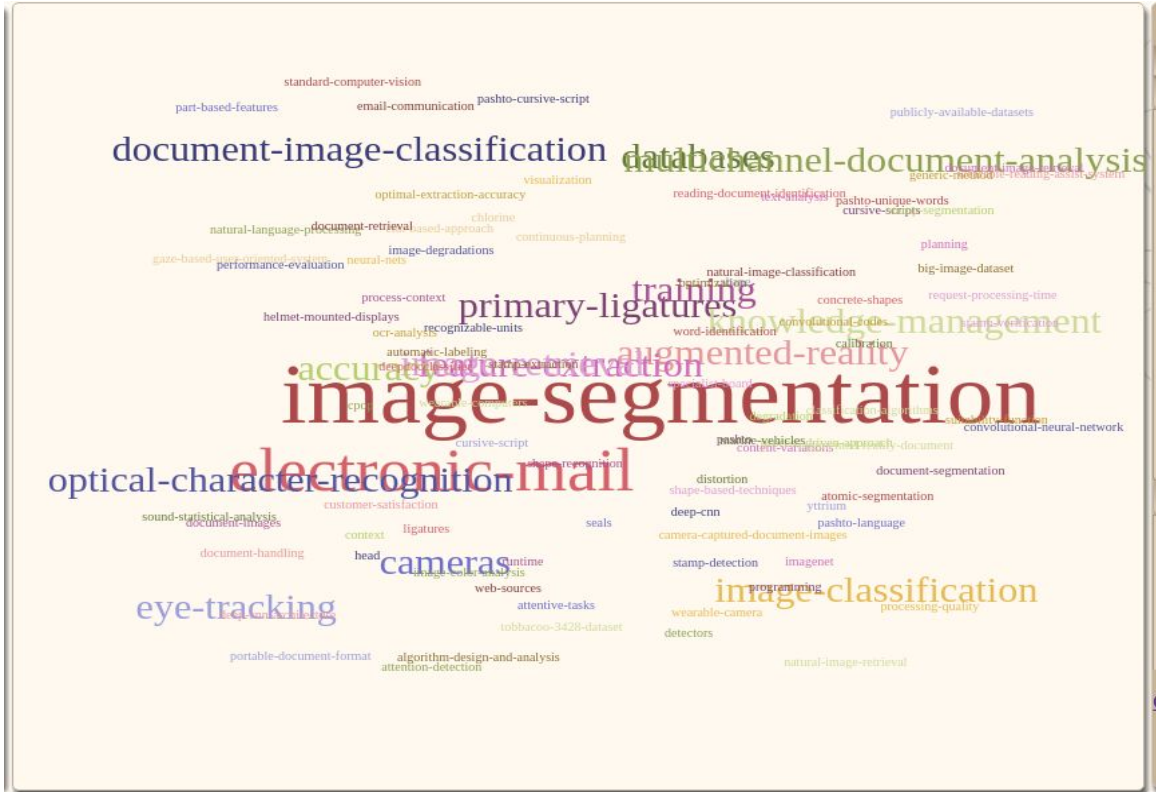

A. ANTONACOPOULOS
I. PHILLIPS
A. CHHABRA
T. KONIDARIS
G. LOULOUDIS
K. NTZIOS
K. NTIROGIANNIS
I. PRATIKAKIS
R. MOGHADDAM
B. GATOS
M. CHERIET
S. PERANTONIS
N. STAMATOPOULOS


From: C. LIU to F. YIN
Sum of all References :8
Sentiments (pos, neu, neg) :2, 4, 2
List of all papers:
•Title :Modified quadratic discriminant functions and th...
•Sum of Reference :8
•Sentiments (pos, neu, neg) :2, 4, 2
•Detailed Section Info: •GenSection :abstract
•Section :I. INTRODUCTION
•Sentiments (pos, neu, neg) :2, 4, 2
Q. WANG
F. YIN
D. WANG


P. NATARAJAN
H. CAO
R. PRASAD
J. MAKHOUL
R. SCHWARTZ
S. SALEEM
K. SUBRAMANIAN


D. BRIDSON
C. PAPADOPOULOS
C. CLAUSNER
S. PLETSCHACHER
A. ANTONACOPOULOS

Appendix



A. Dengel

Cites most often

Most often cited by

Count	Author name
12	M. Liwicki
8	K. Kise
7	S. Ahmed
6	M. Malik
6	M. Gillmann
5	J. Llados
5	S. Uchida
5	L. las
5	R. Huang
4	M. Weber

Collaborates most often with

Count	Author name
9	M. Liwicki
5	S. Ahmed
3	F. Shafait
3	M. Malik

Click to view/hide Author Word Cloud

Appendix

C. TAN

Cites most often

Count	Author name
67	C. TAN
23	S. LU
16	P. SHIVAKUMARA
14	A. JAIN
13	R. KASTURI
13	B. SU
11	I. PRATIKAKIS
11	B. GATOS
10	Y. ZHOU
9	H. BUNKE

Most often cited by

Count	Author name
67	C. TAN
20	N. SHARMA
20	P. SHIVAKUMARA
20	U. PAL
19	S. LU
14	B. SU
13	W. HUANG
12	M. CHERIET
12	S. TIAN
11	M. BLUMENSTEIN

Collaborates most often with

Count	Author name
11	S. LU
6	B. SU
6	P. SHIVAKUMARA
5	L. LI
4	S. TIAN
4	W. HUANG
4	Y. LU
3	B. YUAN
3	J. WANG
3	L. FAN

U. PAL

Cites most often

Count	Author name
109	U. PAL
44	B. CHAUDHURI
34	F. KIMURA
21	N. SHARMA
21	P. SHIVAKUMARA
20	C. TAN
20	M. BLUMENSTEIN
15	S. SRIHARI
14	R. PLAMONDON
14	P. ROY

Most often cited by

Count	Author name
109	U. PAL
31	B. CHAUDHURI
30	B. ROAD
30	F. KIMURA
27	N. SHARMA
22	P. ROY
19	M. BLUMENSTEIN
18	S. CHANDA
15	R. MANDAL
14	T. WAKABAYASHI

Collaborates most often with

Count	Author name
10	F. KIMURA
8	B. ROAD
7	B. CHAUDHURI
7	P. ROY
5	M. BLUMENSTEIN
5	N. SHARMA
4	A. ALAEI
4	J. LLADOS
4	P. NAGABHUSHAN
4	T. WAKABAYASHI

V. GOVINDARAJU

Cites most often

Count	Author name
64	V. GOVINDARAJU
26	S. SRIHARI
16	Z. SHI
11	H. BUNKE
11	R. PLAMONDON
10	L. SCHOMAKER
9	G. KIM
9	S. SETLUR
9	C. RAMAIAH
8	A. JAIN

Most often cited by

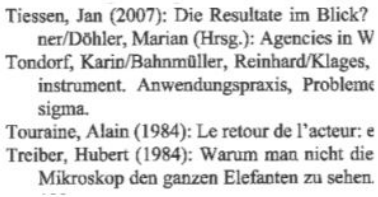
Count	Author name
64	V. GOVINDARAJU
20	C. RAMAIAH
20	S. SETLUR
20	Z. SHI
18	H. BUNKE
14	A. SHIVRAM
13	B. GATOS
12	P. NATARAJAN
11	S. SRIHARI
10	C. SUEN

Collaborates most often with

Count	Author name
10	S. SETLUR
8	Z. SHI
4	A. SHIVRAM
4	S. SRIHARI
3	C. RAMAIAH
3	S. MADHVANATH
3	S. TULYAKOV
2	B. ZHU
2	I. NWOGU
2	M. NAKAGAWA

Thank you.

https://locdb.bib.uni-mannheim.de/blog/de/



Images

[1] https://image.freepik.com/free-icon/male-user-shadow_318-34042.jpg

[2] https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTuDltGNoXCvMPu3jwTMa-lio-DeK2TZljFZRK9THBduoIliarvKA

[3] https://img.clipartfest.com/c123ecda18b92c3b0a147a994e57e0e8_to-do-list-clipart-things-to-do-list-clip-art_1024-1024.png

[4] https://scholar.google.de/

[5] http://opencitations.net/

[6] https://us.intacct.com/sites/default/files/press-images/crossref_logo.png

[7] https://articles-images.sftcdn.net/wp-content/uploads/sites/8/2008/03/pdf.png

[8] http://www.iconarchive.com/download/i22892/kyo-tux/phuzion/File-Web.ico

[9] https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcREEssBPV3h0H5JbJ1yrU_t5NRKhjyH5Hgnxd9iCkX3iExBnCXzAA

[10] http://gsowww.gbv.de/images/logos/2_152.gif

[11] https://upload.wikimedia.org/wikipedia/en/thumb/e/eb/Springer.svg/1280px-Springer.svg.png

[12] http://opencitations.net/static/img/logo.png