

# Recent advances in the project EXCITE – Extraction of Citations from PDF Documents

Philipp Mayr

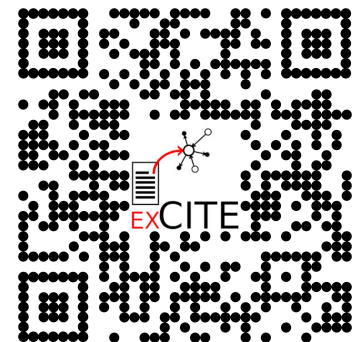
GESIS – Leibniz Institute for the Social Sciences

2018-09-03, Bologna



#opencitations

<http://excite.west.uni-koblenz.de/website/index.html>



## EXCITE team

- PI: Steffen Staab (WeST), Philipp Mayr (GESIS)
- Researchers: Behnam Ghavimi, Zeyd Boukhers
- Developer: Azam Hosseini
- Collaborators: Heinrich Hartmann, Martin Körner



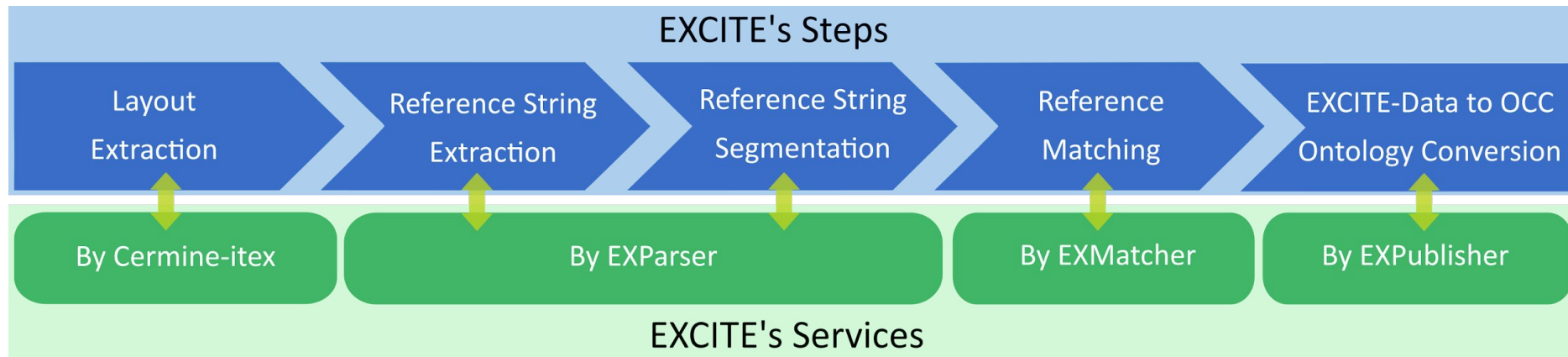
# EXCITE: Background

- We run productive search systems and research in information retrieval, recommendation systems and knowledge discovery
  - SSOAR <https://www.gesis.org/ssoar/> (48K full texts)
  - GESIS Search <https://search.gesis.org/> (242K data sets + further materials)
- **National literatures are not well represented** in major citation indices (like WoS, Scopus)
- **Shortage of citation data** for the international and German social sciences (Social Science Citation Index is not enough)
- **Open availability** of citation data **is improving** but still very limited

## EXCITE: Main objectives

- Develop **web services** to allow third-parties to extract citation data from arbitrary publications
- Develop a **toolchain** of reference extraction and matching software
- **Integrate and publish** the extracted citation data in reusable formats
- Narrow the supply gap of citation data in the social sciences

# EXCITE: toolchain



- (1) Extraction of text from source documents (PDFs),
- (2) Identification of reference sections and other forms of embedded reference information within the text,
- (3) Segmentation of individual references into its constituent fields such as author, title, etc.,
- (4) Matching of reference strings against bibliographic databases,
- (5) Export and publication of matched references to reusable formats (convert to OCC)

# EXCITE: recent advances

- All components are available as reusable components, see <https://github.com/exciteproject>
  - EXparser – tool to extracting and segment references (see talk by Zeyd Boukhers: “**A Generic Approach for Reference Extraction from PDF Documents**” tomorrow)
  - **Annotators** and Gold standards – tools to annotate references and different gold standards to train and test the tools
  - **EXmatcher** – tool to match references to bibliographic databases which base on solr, elasticsearch
  - **EXpublisher** – tool to convert EXCITE data to JSON-LD
- **Public demo** <http://excite.west.uni-koblenz.de:8081/excite>
- Extracted and matched data in productive systems, e.g. <https://search.gesis.org/publication/gesis-ssoar-10004>

## EXAnnotators: Reference Identification

The screenshot shows the EXCITE web application interface for reference identification. The interface includes a top navigation bar with buttons like "Choose Files (.txt/.pdf)", "Insert Tag (whole line)", "Insert Tag (part)", "Insert Tag (other)", "Remove Tag", "Export To XML", "Reload page", and "Help". The main area is divided into two panes: "Text File: 1181.xml (81.96 KB)" and "PDF File: 1181.pdf (776.92 KB)". The "Text File" pane displays a list of references with XML tags, including "Aron, Raymond/Dominique Schnapper (1988): Power, modernity, and sociology : selected sociological writings. Aldershot, Hants, England Brookfield, Vt., USA: E. Elgar ; Collins, Harry (2004): Gravity's shadow : the search for gravitational waves. Chicago: University of Chicago Press; Collins, Harry M. (1981): Stages in the Empirical Programme of Relativism. In: Social Studies of Science, 11 S. 3-10; Collins, Harry M. (1983): An Empirical Relativist Programme in the Sociology of Scientific Knowledge. In: K.D. Knorr-Cetina/M. Hulsey (Hrsg.): Science observed. Perspectives on the social study of science. London: Sage: S. 85-113; Collins, Harry M. (1985): Changing order : replication and induction in scientific practice, London u.a.: Sage; Collins, Harry/Trevor Pinch (1999): Geleit der Forschung (Der). Wie unsere Wissenschaft die Natur erfindet. Berlin: Berlin Verlag; Dosi, Giovanni (1982): Technological Paradigms and Technological Trajectories. A Suggested Interpretation of the Determinants and Directions of Technical Change. In: Research Policy, 11 S. 147-162; Esser, Hartmut (1993): Soziologie. Allgemeine Grundlagen, Frankfurt/SLASHINTEXTMain u.a.: Campus; Gower Pub. Co.; Hacking, Ian (1999): Was heißt "soziale Konstruktion"? Zur Konjunktur einer Kampfbibel in der Wissenschaft, Frankfurt/SLASHINTEXTMain: Fischer; Kuhn, Thomas S. (1976): Die Struktur wissenschaftlicher Revolutionen. Frankfurt/SLASHINTEXTMain: Suhrkamp; Luhmann, Niklas (1990): Die Wissenschaft der Gesellschaft, Frankfurt/SLASHINTEXTMain: Suhrkamp; MacKenzie, Donald (1989): From Kuvajalein to Amagaddon? Testing and the Social Construction of Missile Accuracy. In: D.G.v. al. (Hrsg.): S. 409-435; Meyer, Uli (2004): Die Kontroverse um Neuronale Netze. Zur sozialen Aushandlung der wissenschaftlichen Methoden eines Forschungsgegenstandes. Wiesbaden: Deutscher Universitäts-Verlag; Minsky, Marvin (1985): SLASHINTEXTSeymour Papert (1989): Perceptrons: an introduction to computational geometry. Cambridge, Mass.: MIT Press;". The "PDF File" pane shows the same references in a PDF format. A "Preview section (Text + XML Tags)" at the bottom displays the XML tags for the selected text.



# EXAnnotators: Reference Segmentation


Choose File (txt/xml/csv)

Load Last Session

Export to XML

Reload page

Help



Tag Reference String: [ 1181.xml ] - [ File Size : 5.55 KB ] - [ References Number: 18 ]

**Annotated Reference String**

Esser, Hartmut (1993): Soziologie. Allgemeine Grundlagen, Frankfurt/Main u.a.: Campus, Gower Pub. Co.

Given Name	Title	Year	Identifier	First Page	Others
Surname	Editor	Volume	Source	Last Page	Remove Tag
	Publisher	Issue	URL		

**Preview Reference String with XML tags:**

```
<author><surname>Esser</surname>, <given-names>Hartmut</given-names></author> (<year>1993</year>): <title>Soziologie. Allgemeine Grundlagen</title>,<other>
Frankfurt/Main u.a.</other>: <publisher>Campus, Gower Pub. Co.</publisher>
```

First

Prev

8/18

Next

Last



## EXCITE: Demo



**EXCITE** (Extraction of Citations from PDF Documents) is a toolchain of citation extraction software and particular focus on the German-language social sciences and this is a public service for the project. In the background of this page we are using [Refex](#) for pdfs processing and [Exparsen](#) for reference string segmentation.

### How to start the process:

- **First:** Choose a pdf file by click on "Choose File" button. (size of the file should be less than 5 MB)
- **Second:** Click on "Upload file" button to start the process.
- **tip:** After uploading a file a code will be displayed on the screen. This code is necessary for follow up the result of reference extraction. If you would like to load last follow up code related to last uploaded file in your browser.

### How to check the result:

- **First:** Enter the follow-up code in the appropriate box in the right-hand of page then click on "Display References" button.
- **Second:** The result will be displayed on screen if it available.
- **tip:** Extracting References process will take a little time, at least 30 seconds, and it completely depends on the size of the file.
- **tip:** Click on "Load last follow up code" If you would like load last follow up code related to last uploaded file in your browser.

### Uploading File

Choose File No file chosen

Enter your Email address.(optional)

Upload File

### Display References

If you already have a code, Enter code to load data.

180827134030-EWJ0MP9X1T

Load Last follow up code

Display References

Result: (52) references are extracted

### Result

<http://excite.west.uni-koblenz.de:8081/excite>

Reference String

Reference Segment (XML)

Reference BibTeX

Anderson, J. A., Wright, E. R., Kooreman, H. E., Mohr, W. K., & Russell, L. A. (2003). The dawn project: A model for responding to the needs of children with emotional and behavioral challenges and their families. *Community Mental Health Journal*, 39, 63-74.

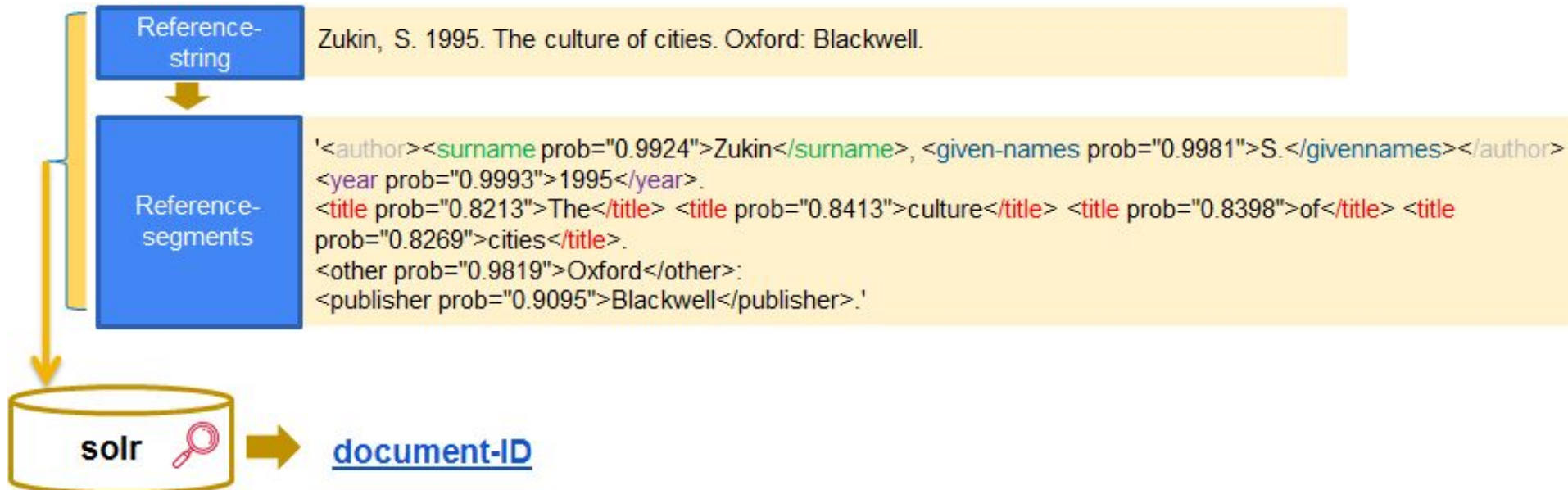
Ashford, J. B., & LeCroy, C. W. (1990). Juvenile recidivism: A comparison of three prediction instruments. *Adolescence*, 25, 441-450.

Bazemore, G., & Walgrave, L. (Eds.). (1999). *Restorative juvenile justice: Repairing the harm of youth crime*. Monsey, NY: Criminal Justice.

Boesky, L. M. (2002). *Juvenile offenders with mental health disorders: Who are they and what do we do with them?* Lanham, MD: American Correctional Association.

# EXmatcher

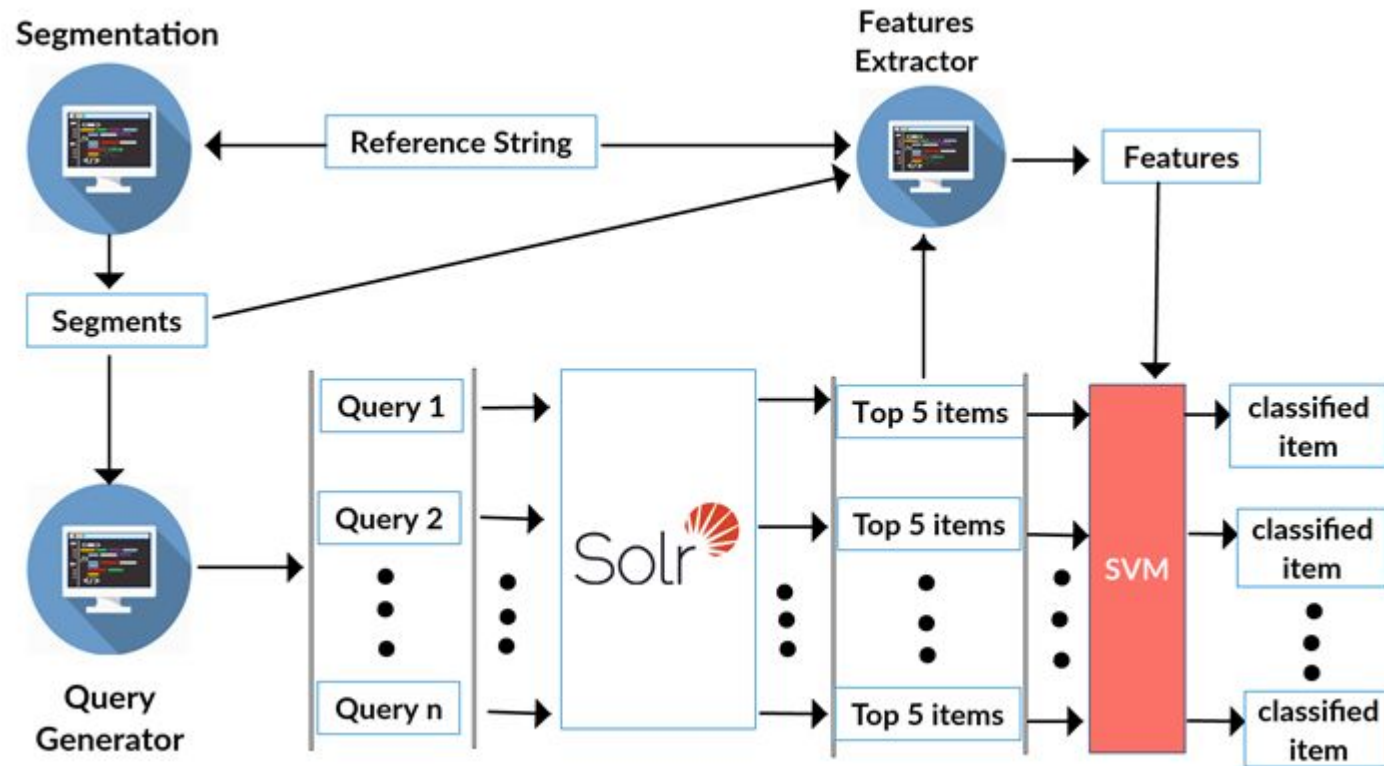
- Input are segmented reference strings with probabilities for each segment
- Output are matched document ids



## EXmatcher

**hybrid approach** -  
combination of  
blocking  
techniques  
and a classifier  
algorithm

Input: strings,  
segments,  
probabilities



# EXPublisher

- Converting extracted and matched data to the OCC ontology
- Enrichment of the reference information by external metadata

## Details about data:

Besides EXCITE data availability in OCC portal, the bulk download is accessible via EXCITE server:

- [Version 1 \(17/07/2018\)](#)

In this version of data we have a part of extracted references from [SSOAR](#) PDF corpus (about 24 k of SSOAR PDFS):

1. The total number of brs: 1,045,189
2. The total number of bes: 1,146,213

**EXMatcher and ExPublisher will be included in the demo soon!**

<https://github.com/exciteproject/EXpublisher>

## Next steps in EXCITE

- EXCITE data published in OpenCitationCorpus
- Public EXCITE API for testing (to be public soon)
- Reference Matching to Crossref to be added in the demo/API
- Gold Standards (German/English/Reference Section/Footnotes) to be completed
- Extractions models for German and English texts
- More Social Science data to be processed and released
- Processing of complete Arxiv??

# Thank you

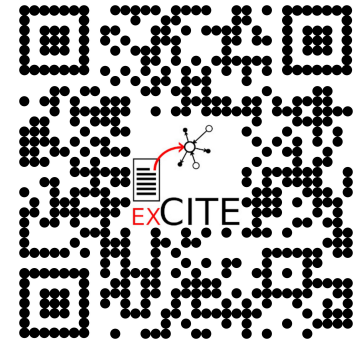
## Contact:

Dr Philipp Mayr

GESIS - Leibniz Institute for the Social Sciences, Germany

Email: [philipp.mayr@gesis.org](mailto:philipp.mayr@gesis.org)

Twitter: @philipp\_mayr



- Project website  
<http://excite.west.uni-koblenz.de/website/index.html>
- EXCITE mailing list: [Subscribe to our Newsletter](#).
- Demo <http://excite.west.uni-koblenz.de:8081/excite>
- GIT <https://github.com/exciteproject/>

# EXCITE: Toolchain

