gesis
Leibniz Institute for the Social Sciences

"Challenges in Extracting and Managing References"

EXCITE Workshop 2017

Philipp Mayr

GESIS – Leibniz Institute for the Social Sciences

2017-03-30

#excitews2017

http://west.uni-koblenz.de/en/research/excite/workshop-2017

Who we are

- PI: Steffen Staab (WeST), Philipp Mayr (GESIS)
- Researchers: Behnam Ghavimi, Martin Körner
- Collaborator: Heinrich Hartmann (Independent)





EXCITE: Background

- We run productive search systems and research in information retrieval, recommendation systems and knowledge discovery
  - Sowiport http://sowiport.gesis.org/
  - Related-work http://dev.related-work.net/
- **Shortage of citation data** for the international and German social sciences
- **Open availability of citation data is still very limited**
  - Open Citation Corpus http://opencitations.net/
  - CitEc: Citations in Economics http://citec.repec.org/

EXCITE: Background

Service	Scope of the database	Coverage of the Social Sciences	Open Source/Data
Web of Science (WoS) webofknowledge.com	interdisciplinary	marginal	no/no
Scopus scopus.com	interdisciplinary	marginal	no/no
Google Scholar (GS) scholar.google.com	interdisciplinary	marginal	no/no
Microsoft Academic Search (MAS) Deprecated in 2013	interdisciplinary	marginal	no/no
CiteSeerX citeseerx.ist.psu.edu	Computer Science, Mathematics, Physics	no coverage	yes/yes
ArnetMiner arnetminer.org	Computer Science	no coverage	no/yes
CitEc citec.repec.org	Economics	marginal	no/yes
Mendeley mendeley.com	interdisciplinary	no documentation	no/no
ResearchGate researchgate.net	interdisciplinary	no documentation	no/no
SSRN http://www.ssrn.com/	Social Sciences	no documentation	no/no

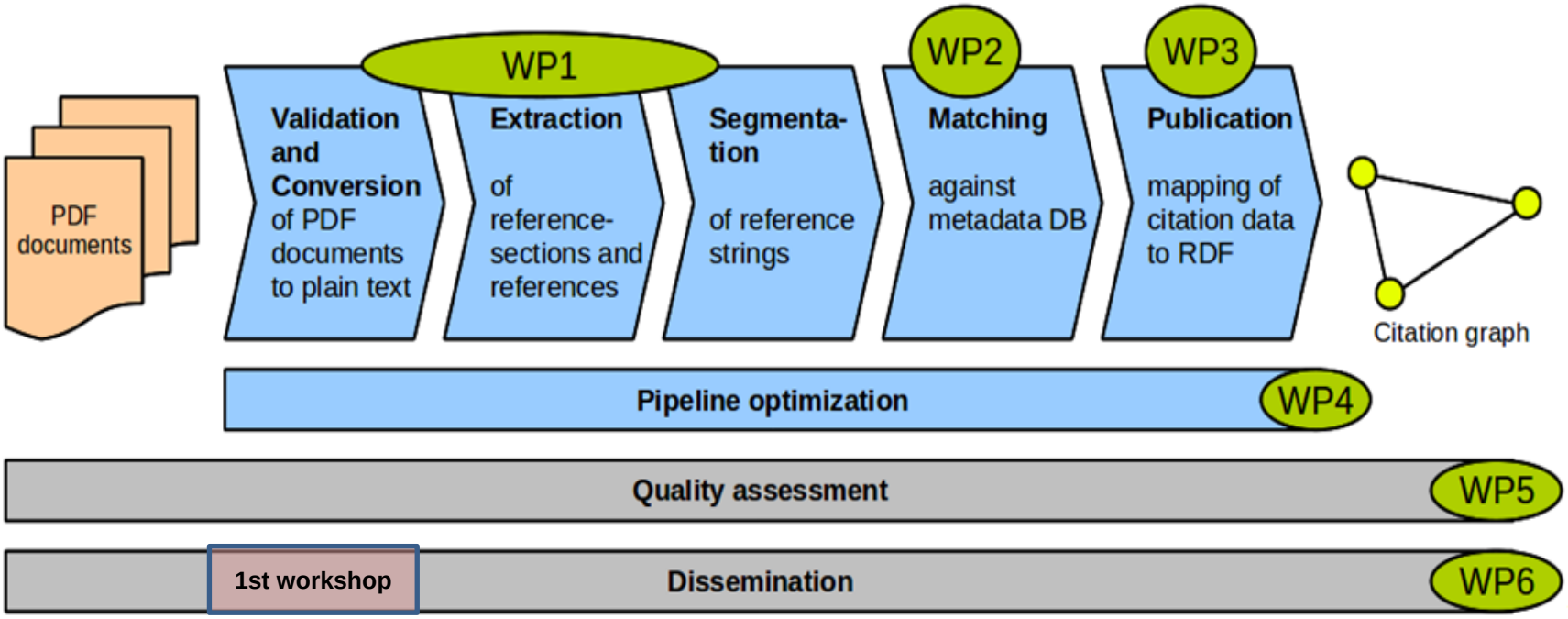
EXCITE: Main objectives

- Narrow the supply gap of citation data in the social sciences
- Improve the breadth and accuracy of current reference extraction systems
- Develop web service API to allow third-parties to extract citation data from arbitrary publications
- Integrate and publish the extracted citation data

General objectives

- Developing a toolchain of citation extraction and matching software
- Tools and data will be made available to researchers

EXCITE: five steps



(1) extraction of text from source documents,
(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 databases of bibliographic information,
(5) the export of matched references to reusable formats

EXCITE: Outcomes

- Open reference extraction tools for PDF documents
- Open datasets of segmented references
- Web service API for citation extraction
- Assessment of the overall quality of the extraction and matching pipeline
- Open gold standard testbed
- Improved infrastructures and services

gesis
Leibniz Institute for the Social Sciences

Workshop Agenda: Day 1

12:20	Information Extraction out of Born-Digital Scientific Articles	Roman Kern, TU Graz
12:40	Advanced citation matching and large-scale full-text analysis	Nees Jan van Eck, Leiden U
13:00	**Lunch Break (Cafeteria)**	
14:20	APIs for third parties to extract and deposit output executions of automated extraction pipelines (via **videoconferencing**)	Min-Yen Kan, NU Singapore
14:40	Extracting references from scientific articles in CERMINE system	Dominika Tkaczyk, U Warsaw
15:00	**Coffee Break (Cafeteria)**	
15:30	CitEc to CitEcCyr. A stab at distributed citation systems. (via **videoconferencing**)	Thomas Krichel, Open Library Society, NYC
15:50	EXCITE project: Status report	Behnam Ghavimi, GESIS Martin Körner, WeST Heinrich Hartmann, Circonus
16:10	Processing of in-text References: Towards a Semantic Analysis	Marc Bertin, U Toulouse
16:30	Citations in Utopia Documents	David Thorne, U Manchester
16:50	**Coffee Break (Cafeteria)**	
17:20	Research around the Tagging System BibSonomy	Andreas Hotho, U Würzburg
17:50	LOC-DB: A Linked Open Citation Database provided by Libraries. Motivation and Challenges.	Kai Eckert, HDM Stuttgart Anne Lauscher, HDM Stuttgart Akansha Bhardwaj, DFKI
18:20	tbd (via **videoconferencing**)	Lee Giles, Penn State U

Workshop Agenda: Day 2

- Hand-on sessions

9:00	**Second Day Kickoff (Room: West II)**		
9:15	Extraction Result Discussion Group	Gold Standard Discussion Group	Collaboration Discussion Group
11:15	**Coffee Break (Cafeteria)**		
11:30	Extraction Result Discussion Group	Gold Standard Discussion Group	Collaboration Discussion Group
12:30	**Closing Talks (Room: West II)**		
13:00	**End**		

Thank you

Contact:

Dr Philipp Mayr
GESIS - Leibniz Institute for the Social Sciences, Germany
Email: philipp.mayr@gesis.org
Twitter: @philipp_mayr

- Workshop website http://west.uni-koblenz.de/en/research/excite/workshop-2017
- GIT https://github.com/exciteproject/