GESIS Leibniz Institute for the Social Sciences

Scholarly Document Processing in Action: Shared Tasks and Real-World Application

Leibniz Association

Philipp Mayr

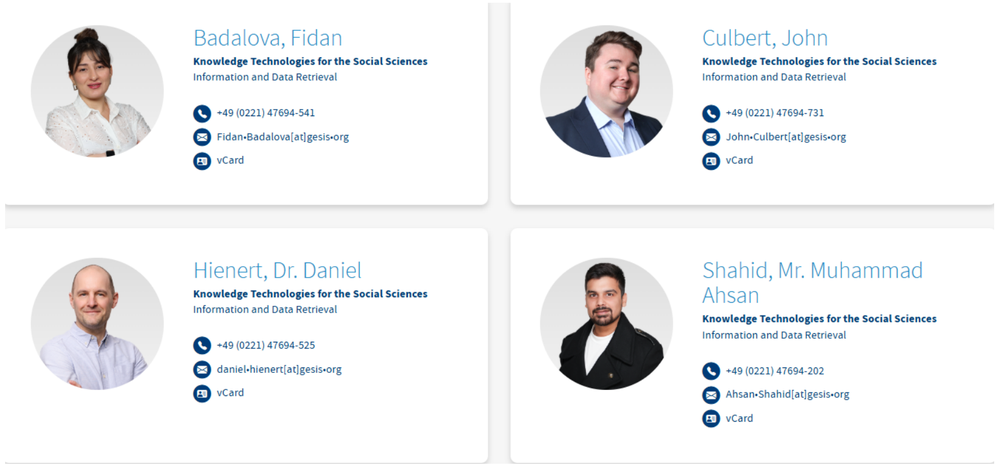
16th European Summer School on Information Retrieval

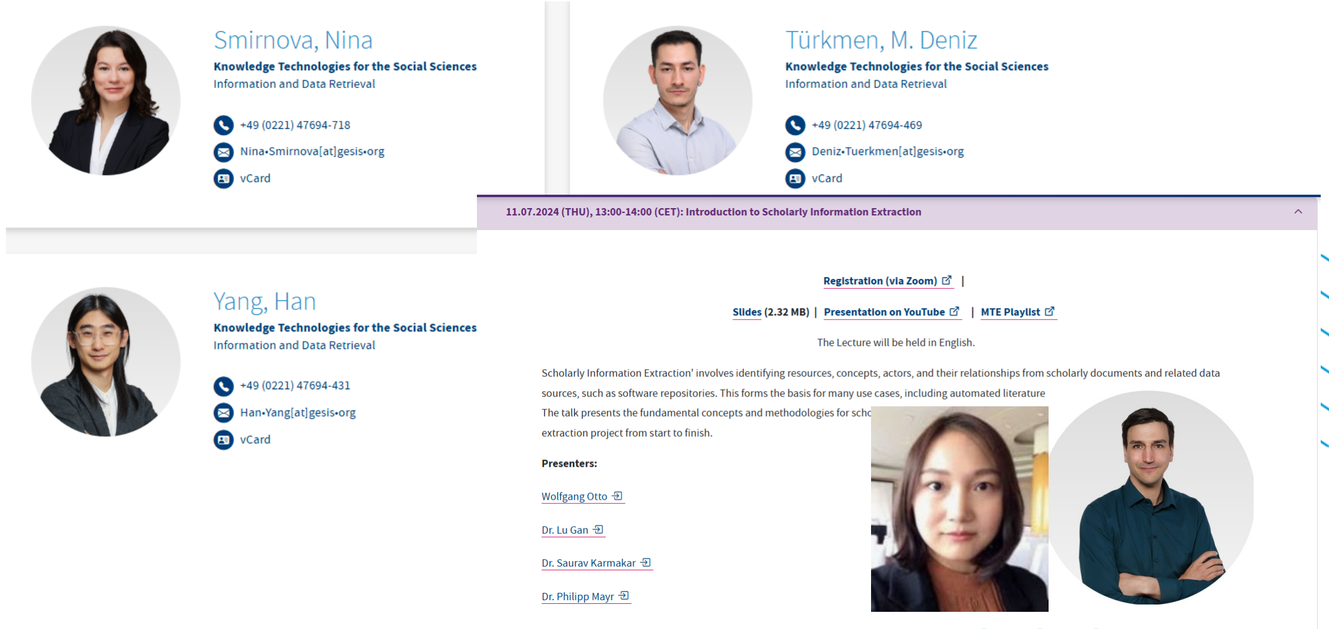
Wolverhampton, 10. July 2025

OMINO
Overcoming Multilevel INformation Overload

People to thank

- OMINO project (Ingo)
- my team **Information and Data Retrieval**
- **Lu Gan** and **Wolf Otto (**Meet the Experts lecture**)**





https://www.gesis.org/en/services/sharing-knowledge/meet-the-experts/meet-the-experts-season-6-knowledge-technologies-for-the-social-science-access-to-social-science-data-and-services

Overview

1. **Introduction to Scholarly Document Processing**
2. Selected Applications
    - Reference Extraction and Linking
    - Entity Extraction
3. A Guide to Scholarly Information Extraction
4. Shared Tasks
5. Summary & Outlook

GESIS Leibniz Institute for the Social Sciences

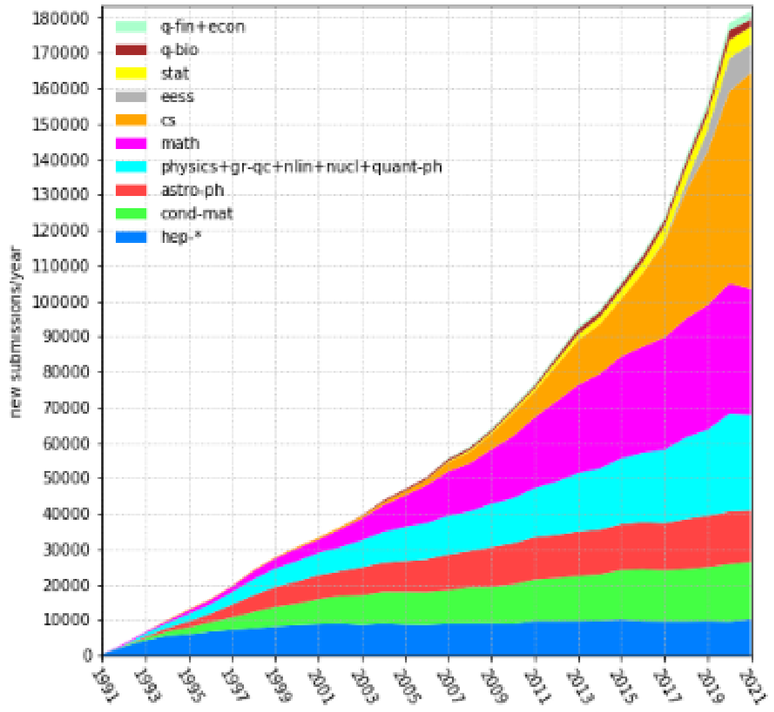
Motivation

Insights into the mechanics of science

- In the evolution of the sciences we find **growth** and increasing **differentiation**
- Leaving us with the problem to **review, evaluate and select**
- Precondition to any creation of (new) knowledge (in individual brains or for larger parts of academia) is to **find and understand**
- Machines (latest AI) have fostered growth, can they also support the knowledge production process? **This is a question of Information Retrieval**

Academic information overload

Data for 1991 through 2021, updated 3 January 2022.



https://info.arxiv.org/help/stats/2021_by_area/index.html

GESIS Leibniz Institute for the Social Sciences

Motivation

Insights into the mechanics of science

- Currently findable and accessible information is like the tip of the iceberg
- **Most of the valuable information is hidden in natural language inside publications**

Scholarly Information Extraction can …

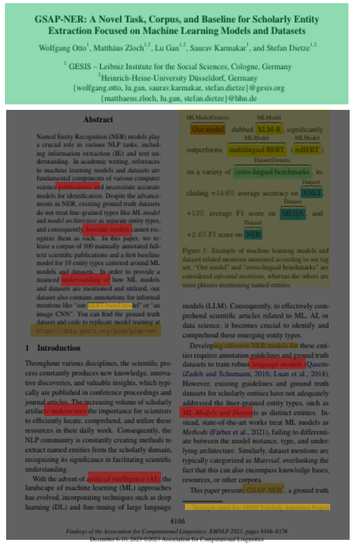
… enable a new quality of FAIR research (e.g. for datasets, software, models).

… help to foster research quality by support the verification of reproducibility.

… answer domain specific research questions.

Findable Accessible Interoperable Reusable

Findable and Accessible Research Information



GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets

Wolfgang Otto[1], Matthäus Zloch[1,2], Lu Gan[1,2], Saurav Karmakar[1], and Stefan Dietze[1,2]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[2] Heinrich-Heine-University Düsseldorf, Germany
{wolfgang.otto, lu.gan, saurav.karmakar, stefan.dietze}@gesis.org
{matthaeus.zloch, lu.gan, stefan.dietze}@hhu.de

Abstract

Named Entity Recognition (NER) models play a crucial role in various NLP tasks, including information extraction (IE) and text understanding. In academic writing, references to machine learning models and datasets are fundamental components of various computer science publications and necessitate accurate models for identification. Despite the advancements in NER, existing ground truth datasets do not treat fine-grained types like *ML model* and *model architecture* as separate entity types, and consequently, baseline models cannot recognize them as such. In this paper, we release a corpus of 100 manually annotated full-text scientific publications and a first baseline model for 10 entity types centered around ML models and datasets. In order to provide a nuanced understanding of how ML models and datasets are mentioned and utilized, our dataset also contains annotations for informal mentions like "our BERT-based model" or "an image CNN". You can find the ground truth dataset and code to replicate model training at https://data.gesis.org/gsap-ner.

1 Introduction

Throughout various disciplines, the scientific process constantly produces new knowledge, innovative discoveries, and valuable insights, which typically are published in conference proceedings and journal articles. The increasing volume of scholarly articles underscores the importance for scientists to efficiently locate, comprehend, and utilize these resources in their daily work. Consequently, the NLP community is constantly creating methods to extract named entities from the scholarly domain, recognizing its significance in facilitating scientific understanding.

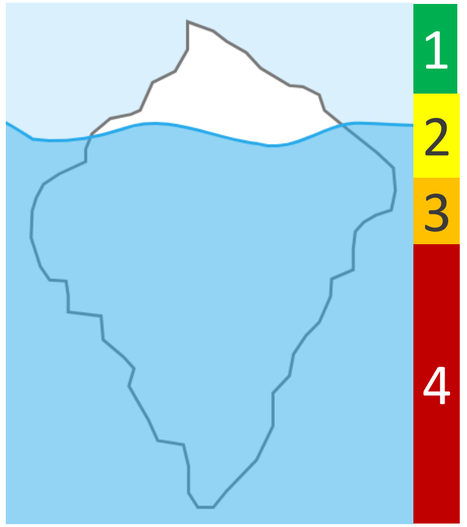
With the advent of artificial intelligence (AI), the landscape of machine learning (ML) approaches has evolved, incorporating techniques such as deep learning (DL) and fine-tuning of large language models (LLM). Consequently, to effectively comprehend scientific articles related to ML, AI, or data science, it becomes crucial to identify and comprehend these emerging entity types.

Developing effective NER models for these entities requires annotation guidelines and ground truth datasets to train robust supervised models (Qasemizadeh and Schumann, 2016; Luan et al., 2018). However, existing guidelines and ground truth datasets for scholarly entities have not adequately addressed the finer-grained entity types, such as *ML Models* and *Datasets* as distinct entities. Instead, state-of-the-art works treat ML models as *Methods* (Farber et al., 2021), failing to differentiate between the model instance, type, and underlying architecture. Similarly, dataset mentions are typically categorized as *Material*, overlooking the fact that this can also encompass knowledge bases, resources, or other corpora.

This paper presents GSAP-NER, a ground truth

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8166–8176
December 6-10, 2023 ©2023 Association for Computational Linguistics

8166



(1) Metadata
Title, Author, Institution, Keywords, ..

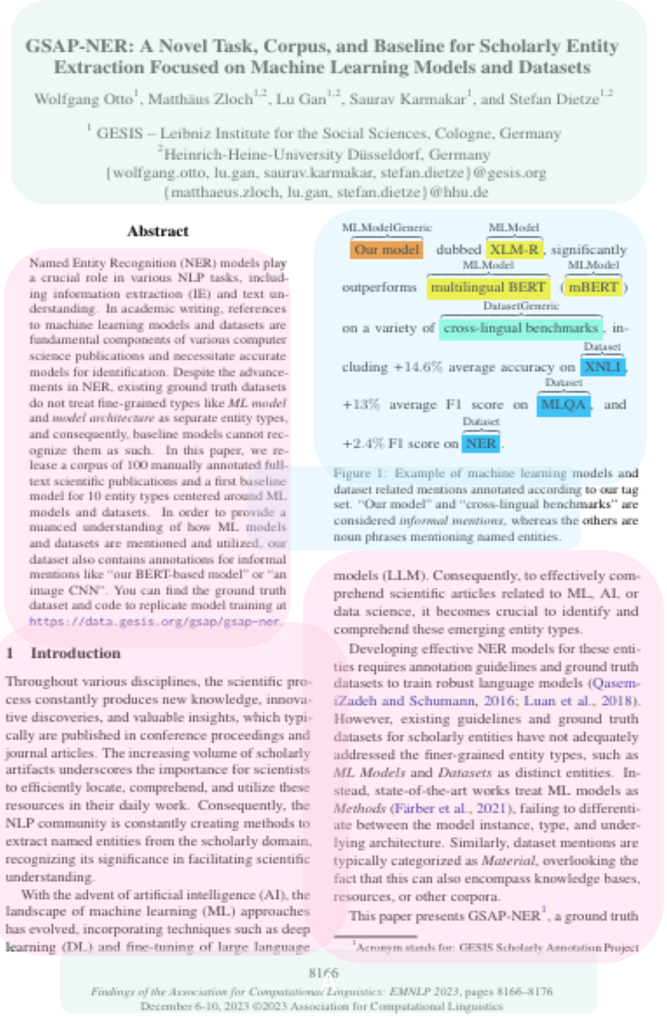
(2) Semi-structured
References, Images, Captions, …

(3) In-text (artifacts)
Datasets, Software, Methods, Variable

(4) In-text (in-domain)
Constitutional Institutions, Genes, Disease

GESIS Leibniz Institute for the Social Sciences

Scholarly Documents



GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets

Wolfgang Otto[1], Matthäus Zloch[1,2], Lu Gan[1,2], Saurav Karmakar[1], and Stefan Dietze[1,2]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[2]Heinrich-Heine-University Düsseldorf, Germany
{wolfgang.otto, lu.gan, saurav.karmakar, stefan.dietze}@gesis.org
{matthaeus.zloch, lu.gan, stefan.dietze}@hhu.de

Abstract

Named Entity Recognition (NER) models play a crucial role in various NLP tasks, including information extraction (IE) and text understanding. In academic writing, references to machine learning models and datasets are fundamental components of various computer science publications and necessitate accurate models for identification. Despite the advancements in NER, existing ground truth datasets do not treat fine-grained types like *ML model* and *model architecture* as separate entity types, and consequently, baseline models cannot recognize them as such. In this paper, we release a corpus of 100 manually annotated full-text scientific publications and a first baseline model for 10 entity types centered around ML models and datasets. In order to provide a nuanced understanding of how ML models and datasets are mentioned and utilized, our dataset also contains annotations for informal mentions like "our BERT-based model" or "an image CNN". You can find the ground truth dataset and code to replicate model training at https://data.gesis.org/gsap/gsap-ner.

MLModel(Generic) MLModel
Our model dubbed XLM-R, significantly
MLModel MLModel
outperforms multilingual BERT (mBERT)
DatasetGeneric
on a variety of cross-lingual benchmarks, in-
Dataset
cluding +14.6% average accuracy on XNLI,
Dataset
+13% average F1 score on MLQA, and
Dataset
+2.4% F1 score on NER.

Figure 1: Example of machine learning models and dataset related mentions annotated according to our tag set. "Our model" and "cross-lingual benchmarks" are considered *informal mentions*, whereas the others are noun phrases mentioning named entities.

1 Introduction

Throughout various disciplines, the scientific process constantly produces new knowledge, innovative discoveries, and valuable insights, which typically are published in conference proceedings and journal articles. The increasing volume of scholarly artifacts underscores the importance for scientists to efficiently locate, comprehend, and utilize these resources in their daily work. Consequently, the NLP community is constantly creating methods to extract named entities from the scholarly domain, recognizing its significance in facilitating scientific understanding.

With the advent of artificial intelligence (AI), the landscape of machine learning (ML) approaches has evolved, incorporating techniques such as deep learning (DL) and fine-tuning of large language models (LLM). Consequently, to effectively comprehend scientific articles related to ML, AI, or data science, it becomes crucial to identify and comprehend these emerging entity types.

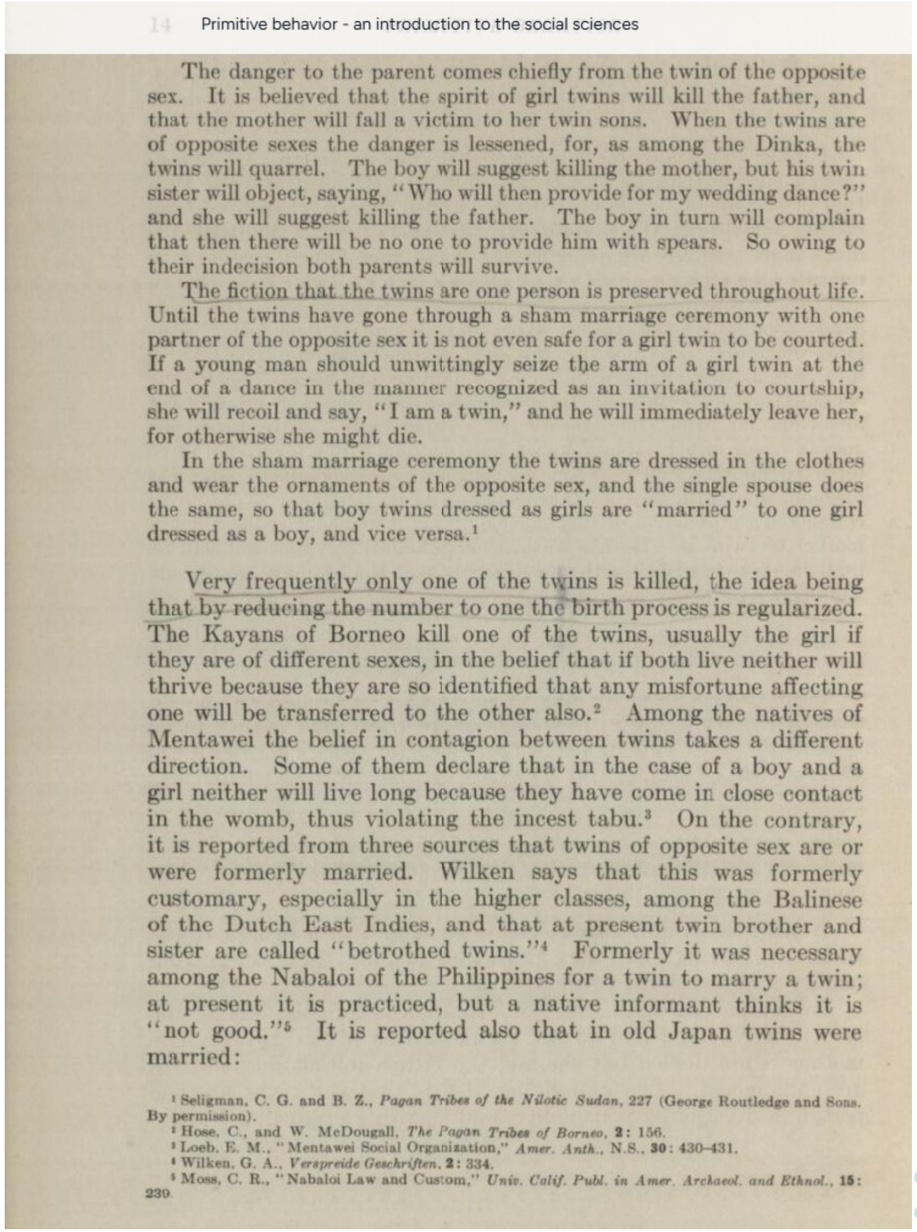
Developing effective NER models for these entities requires annotation guidelines and ground truth datasets to train robust language models (Qasemizadeh and Schumann, 2016; Luan et al., 2018). However, existing guidelines and ground truth datasets for scholarly entities have not adequately addressed the finer-grained entity types, such as *ML Models* and *Datasets* as distinct entities. Instead, state-of-the-art works treat ML models as *Methods* (Färber et al., 2021), failing to differentiate between the model instance, type, and underlying architecture. Similarly, dataset mentions are typically categorized as *Material*, overlooking the fact that this can also encompass knowledge bases, resources, or other corpora.

This paper presents GSAP-NER[1], a ground truth

[1]Acronym stands for: GESIS Scholarly Annotation Project

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8166–8176
December 6-10, 2023 ©2023 Association for Computational Linguistics

https://aclanthology.org/2023.findings-emnlp.548.pdf



The danger to the parent comes chiefly from the twin of the opposite sex. It is believed that the spirit of girl twins will kill the father, and that the mother will fall a victim to her twin sons. When the twins are of opposite sexes the danger is lessened, for, as among the Dinka, the twins will quarrel. The boy will suggest killing the mother, but his twin sister will object, saying, "Who will then provide for my wedding dance?" and she will suggest killing the father. The boy in turn will complain that then there will be no one to provide him with spears. So owing to their indecision both parents will survive.

The fiction that the twins are one person is preserved throughout life. Until the twins have gone through a sham marriage ceremony with one partner of the opposite sex it is not even safe for a girl twin to be courted. If a young man should unwittingly seize the arm of a girl twin at the end of a dance in the manner recognized as an invitation to courtship, she will recoil and say, "I am a twin," and he will immediately leave her, for otherwise she might die.

In the sham marriage ceremony the twins are dressed in the clothes and wear the ornaments of the opposite sex, and the single spouse does the same, so that boy twins dressed as girls are "married" to one girl dressed as a boy, and vice versa.[1]

Very frequently only one of the twins is killed, the idea being that by reducing the number to one the birth process is regularized. The Kayans of Borneo kill one of the twins, usually the girl if they are of different sexes, in the belief that if both live neither will thrive because they are so identified that any misfortune affecting one will be transferred to the other also.[2] Among the natives of Mentawei the belief in contagion between twins takes a different direction. Some of them declare that in the case of a boy and a girl neither will live long because they have come in close contact in the womb, thus violating the incest tabu.[3] On the contrary, it is reported from three sources that twins of opposite sex are or were formerly married. Wilken says that this was formerly customary, especially in the higher classes, among the Balinese of the Dutch East Indies, and that at present twin brother and sister are called "betrothed twins."[4] Formerly it was necessary among the Nabaloi of the Philippines for a twin to marry a twin; at present it is practiced, but a native informant thinks it is "not good."[5] It is reported also that in old Japan twins were married:

[1] Seligman, C. G. and B. Z., *Pagan Tribes of the Nilotic Sudan*, 227 (George Routledge and Sons. By permission).
[2] Hose, C., and W. McDougall, *The Pagan Tribes of Borneo*, **2**: 156.
[3] Loeb, E. M., "Mentawei Social Organization," *Amer. Anth.*, N.S., **30**: 430–431.
[4] Wilken, G. A., *Verspreide Geschriften*, **2**: 334.
[5] Moss, C. R., "Nabaloi Law and Custom," *Univ. Calif. Publ. in Amer. Archaeol. and Ethnol.*, **15**: 239.

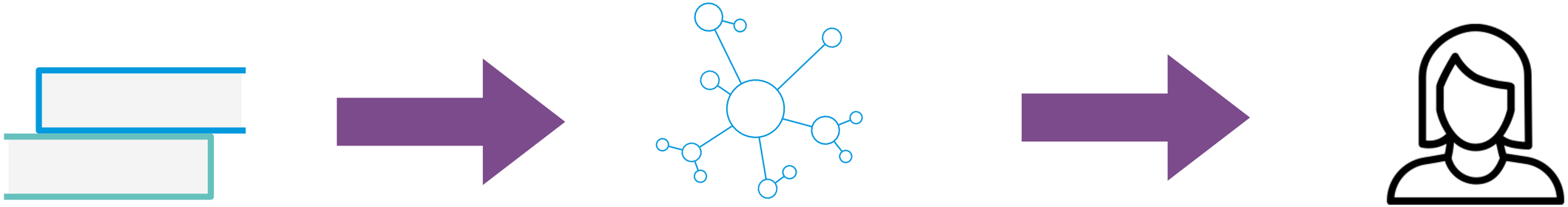
an introduction to the social sciences : Primitive behavior
Thomas, William Isaac ; aut
New York u.a : McGraw-Hill : 1937

https://digital.ub.uni-koeln.de/view/retro_9910338326 19706476_024978?p=23

GESIS Leibniz Institute for the Social Sciences

Motivation

Turning Scholarly Documents into Structured Knowledge



Enhanced Discoverability

- Quickly locate relevant research
- Efficiently navigate vast amounts of information

Comprehensive Insights

- Gain a broader perspective
- Identify trends and gaps in the literature

Enhance Reusability and Reproducability

- Direct access to research artifacts
- Support reproducibility of research

Improved Comprehension

- Accelerate literature reviews
- Easily understand key concepts and relationships

GESIS Leibniz Institute for the Social Sciences

Information Extraction and Subtasks

“NLP task of extracting relevant information from text documents”[1]

Named Entity Recognition (NER)

Identification and classification of named entities in texts.

Relation Extraction (RE)

Identification and classification of relations between named entities in text.

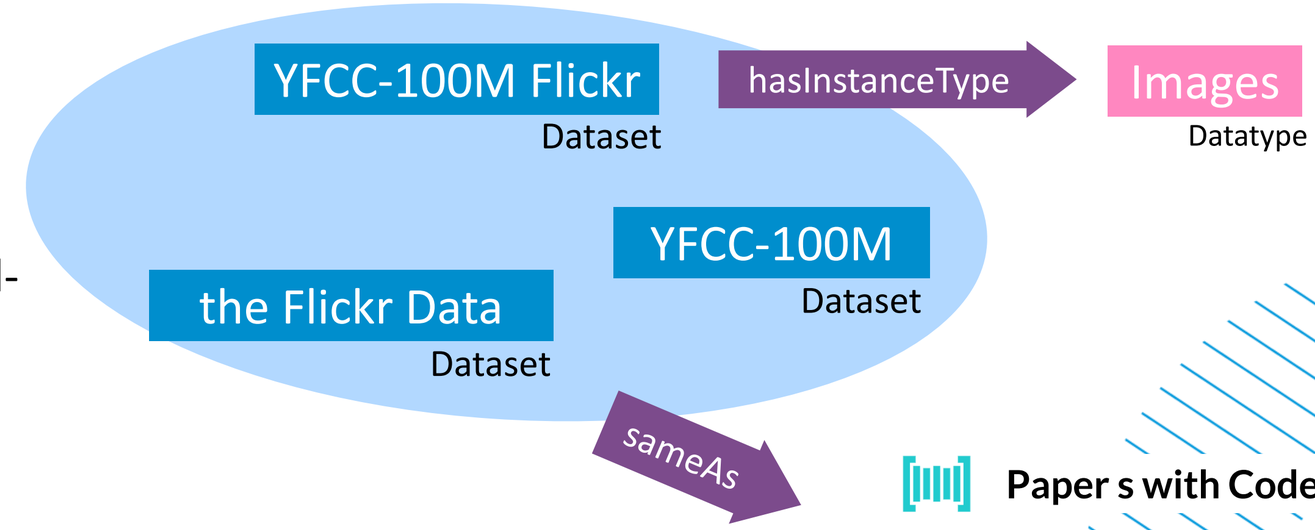
Entity Resolution

The process where all mentions of the same real-world entity are unified into a single, unique instance in a dataset.

Entity Linking

Link with existing resources (e.g. databases or knowledge graphs).

Images were collected from the YFCC-100M Flickr dataset and labeled with gender, and age groups.



1: Practical Natural Language Processing (2020) Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. O‘Reiley Media.

GESIS Leibniz Institute for the Social Sciences

Related Tasks on Scholarly Publications

Summarization

- Structured (IMRaD)
- Extreme (Example Semantic Scholar)

Text Classification

- Research Field Classification
- Citation Intention

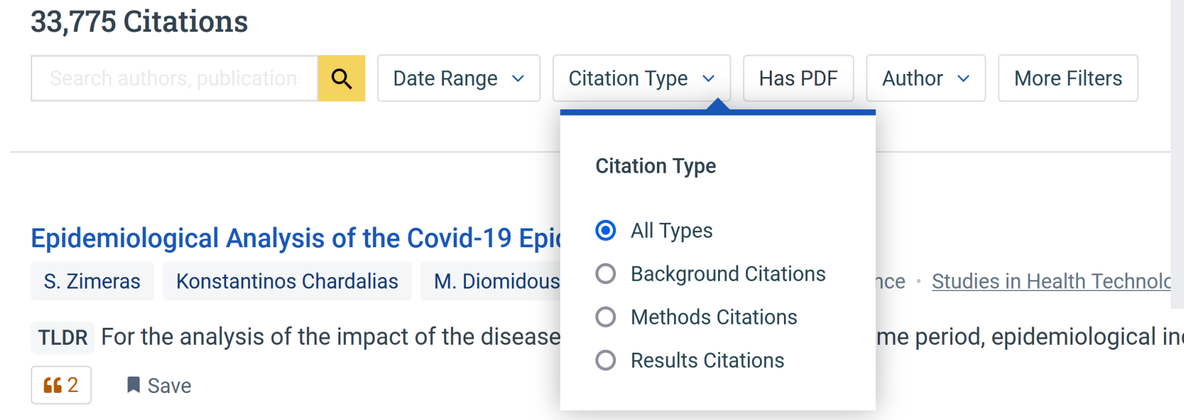
Question Answering

Layout Extraction

- Image + Table Extraction

Example: Semantic Scholar

- Variety of supporting information
- TLDR
- Citation Context Classification



GESIS Leibniz Institute for the Social Sciences

References and Links

- Beltagy, I., Cohan, A., Feigenblat, G., Freitag, D., Ghosal, T., Hall, K., Herrmannova, D., Knoth, P., Lo, K., Mayr, P., Patton, R., Shmueli-Scheuer, M., de Waard, A., Wang, K., & Wang, L. (2021). Overview of the Second Workshop on Scholarly Document Processing. Proceedings of the Second Workshop on Scholarly Document Processing, 159–165. https://aclanthology.org/2021.sdp-1.22
- Otto, W., Zloch, M., Gan, L., Karmakar, S., & Dietze, S. (2023). GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8166–8176. https://doi.org/10.18653/v1/2023.findings-emnlp.548
- Hołyst, J. A., Mayr, P., Thelwall, M., Frommholz, I., Havlin, S., Sela, A., Kenett, Y. N., Helic, D., Rehar, A., Maček, S. R., Kazienko, P., Kajdanowicz, T., Biecek, P., Szymanski, B. K., & Sienkiewicz, J. (2024). Protect our environment from information overload. *Nature Human Behaviour*, *8*, 402–403. https://doi.org/10.1038/s41562-024-01833-8
- https://ominoproject.eu/

Overview

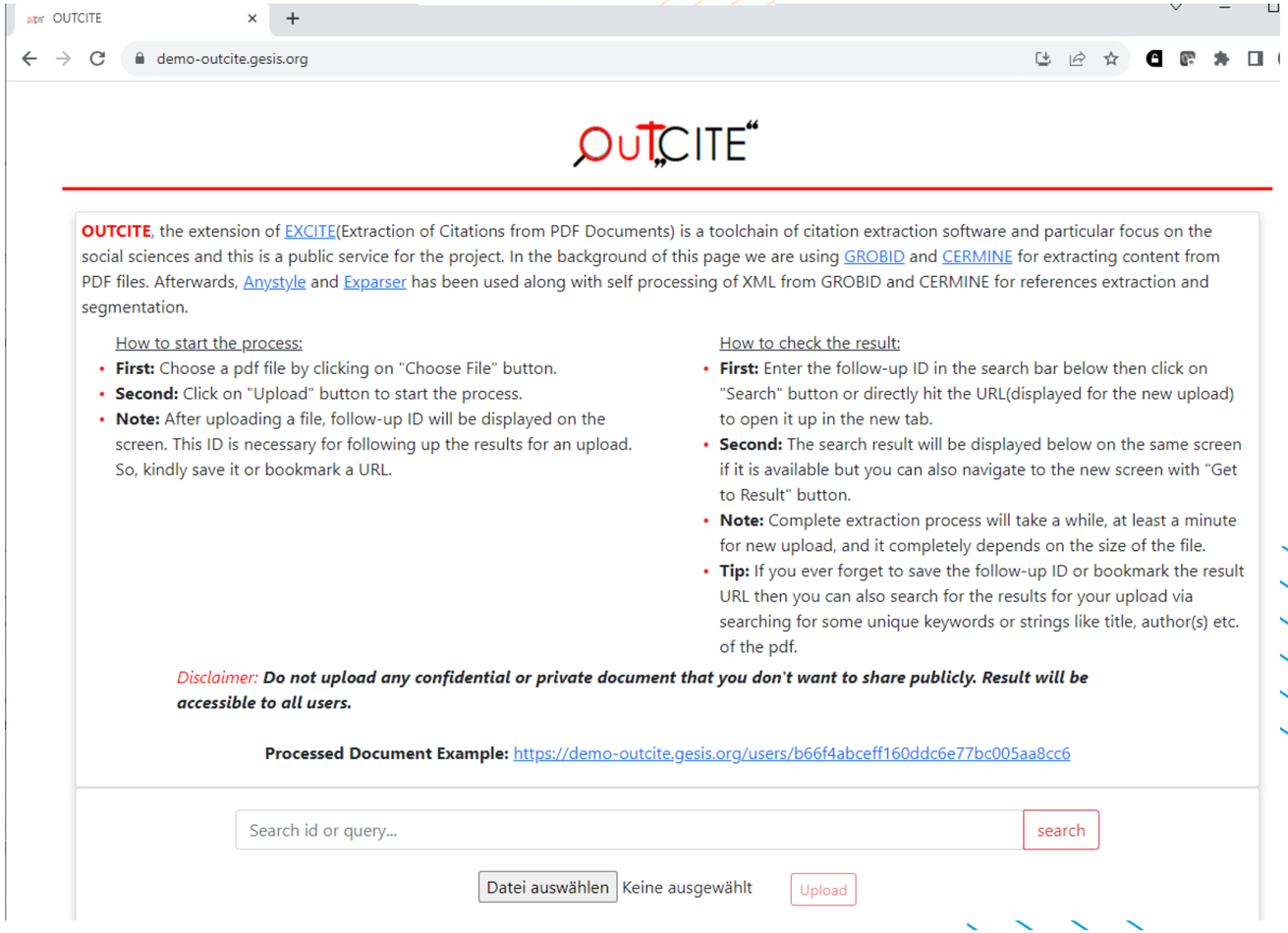
1. Introduction to Scholarly Document Processing
2. **Selected Applications**
    - **Reference Extraction and Linking**
    - Entity Extraction
3. A Guide to Scholarly Information Extraction
4. Shared Tasks
5. Summary & Outlook

GESIS Leibniz Institute for the Social Sciences

Reference Extraction and Linking

SSOAR

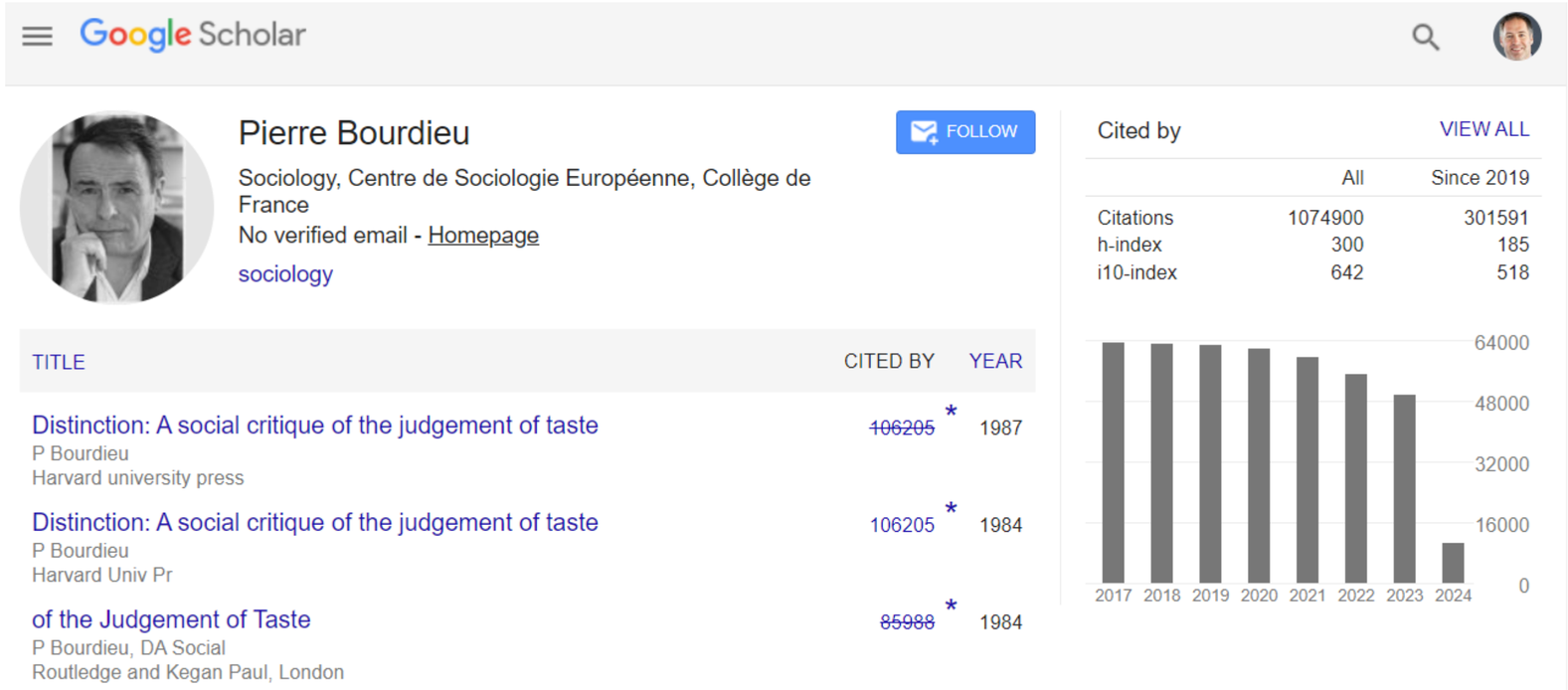
- *Project:* Reference Understanding in the Social Sciences (OUTCITE)
- *Funding*: DFG
- *Background*: Footnote chasing as a widely used discovery approach
- *Task*: Reference identification, extraction, segmentation and linking
- *Tools*: Anystyle, Grobid



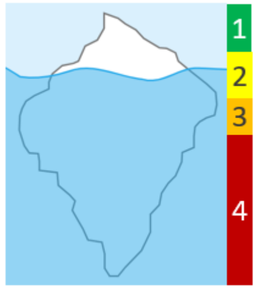
GESIS Leibniz Institute for the Social Sciences

Reference Extraction and Linking: Motivation

- Citation data for the social sciences are largely missing
- Many literature references are not matched with existing databases/catalogs
- Providing open tools, data pipelines and open data



Reference Extraction and Linking



GESIS Leibniz Institute for the Social Sciences

Semi-structured

Named Entity Recognition (NER)

Entity Linking

OUTCITE

SSOAR

References

Abels, Gabriele, 2011: 90 Jahre Frauenwahlrecht: Zum Wandel von Geschlechterverhältnissen in der deutschen Politik. In: Abels, Gabriele (Ed.), 2011: Deutschland im Jubiläumsjahr 2009: Blick zurück nach vorn. Baden-Baden, 197-219.

Annesley, Claire/**Gains**, Francesca, 2010: The Core Executive: Gender, Power, and Change. In: Political Studies. 58, 909-929.

Beckett, Clare, 2006: Thatcher (British Prime Ministers of the 20th Century). London.

Berghahn, Sabine/**Fritzsche**, Andrea, 1991: Frauenrecht in Ost und West Deutschland. Berlin.

Carless, Sally A., 1998: Gender Differences in Transformational Leadership: An Examination of Superior, Leader and Subordinate Perspectives. In: Sex Roles. 39 (11/12), 887-902.

Clemens, Clay, 2006: From the Outside In: Angela Merkel as Opposition Leader, 2000-2005. In: German Politics & Society. 24 (3), 1-19.

Dahlerup, Drude, 1988: From a Small to a Large Minority: Women in Scandinavian Politics. In: Scandinavian Political Studies. 11 (4), 275-298.

Identification

Annesley, Claire/**Gains**, Francesca, 2010: The Core Executive: Gender, Power, and Change. In: Political Studies. 58, 909-929.

Extraction

Title:
The Core Executive: Gender, Power, and Change

Authors:
Claire Annesley, Francesca Gains

Year: **2010**

Journal: **Political Studies**

Volume: **58**

Pages: **909-929**

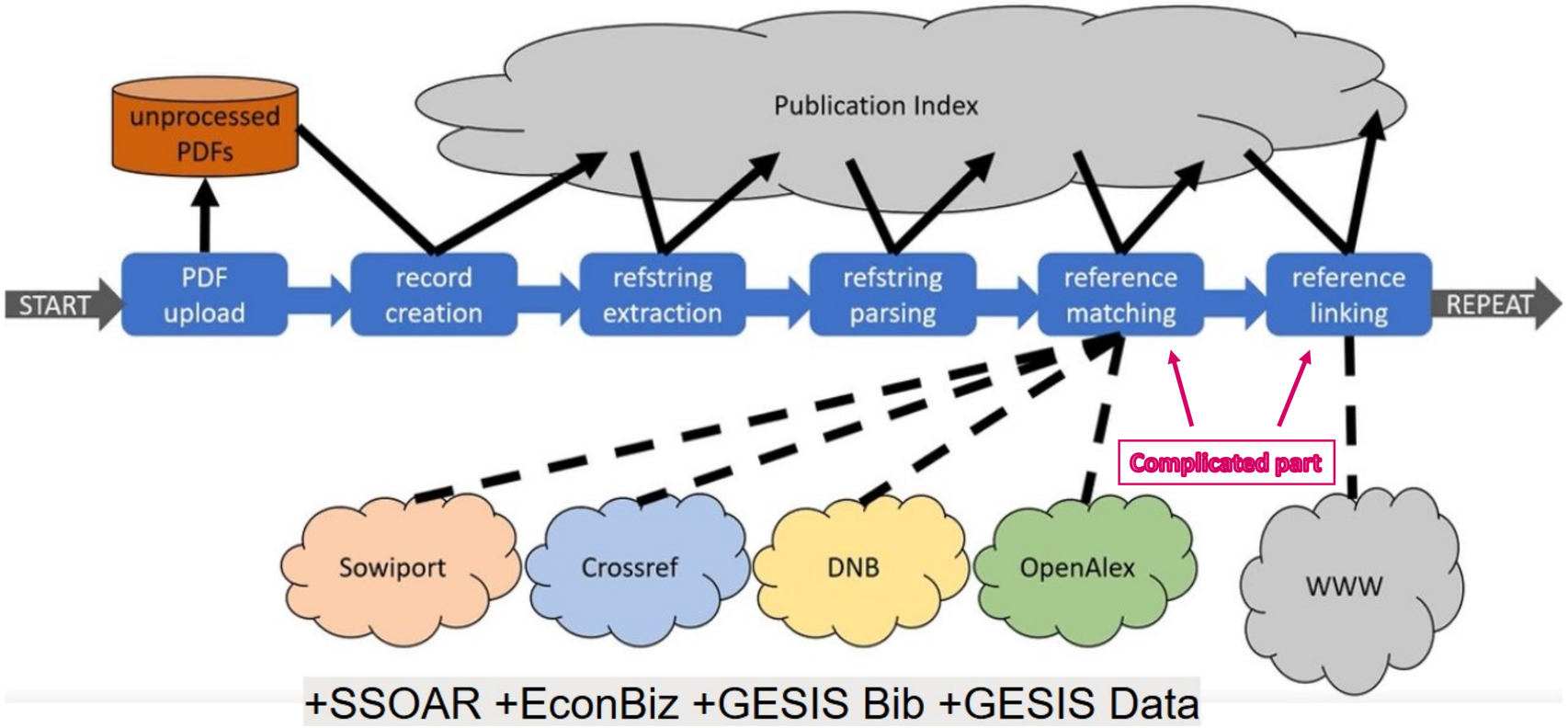
Metadata Catalog

e.g. OpenAlex Crossref

Linking

https://demo-outcite.gesis.org/

Reference Extraction and Linking

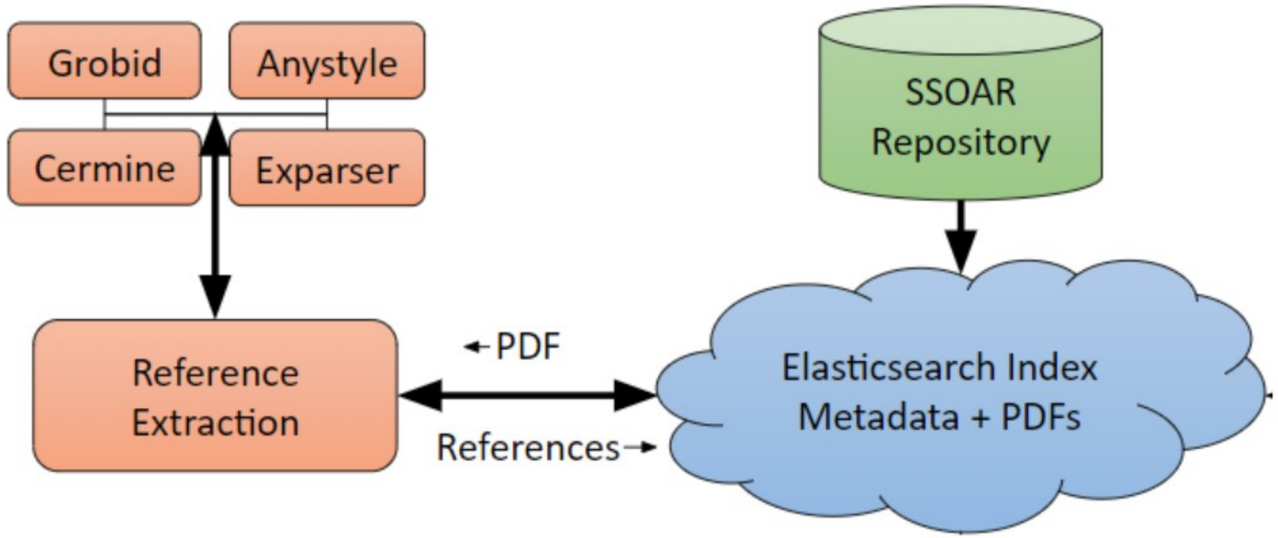


Tasks*:*

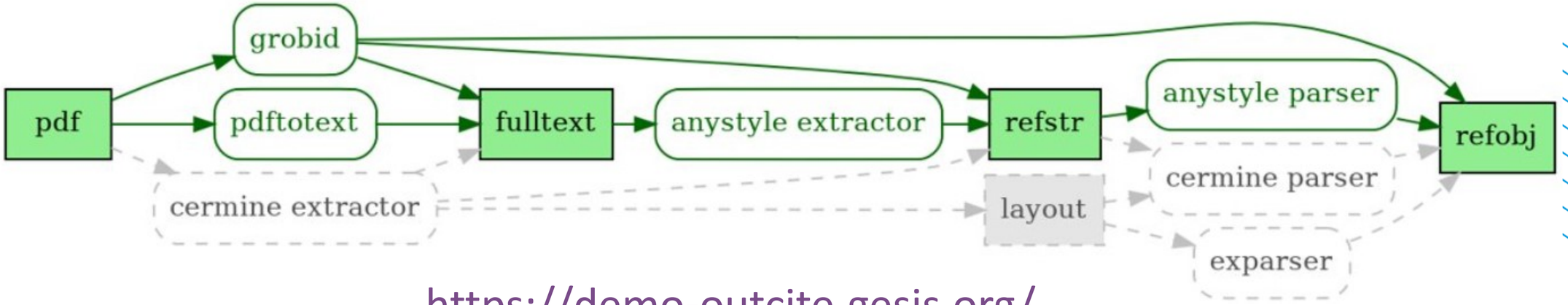
- Identification
- Segmentation Parsing
- Linking Matching

GESIS Leibniz Institute for the Social Sciences

Reference Extraction Pipeline



Tool combinations for the extraction part



https://demo-outcite.gesis.org/

GESIS Leibniz Institute for the Social Sciences

Goldstandards

- Generation of goldstandards to evaluate the tools from different domains
- Manual curation is costly
- Often done by students

	# Docs	# Refs	Mean	Lowest	Highest	0-19	20-39	40-59	60-79	80-99	100-119	120-139	140-159
GEOCITE	178	4729	27	0	109	86	48	29	7	5	3	0	0
EXCITE	348	8679	25	1	151	162	111	63	8	2	0	1	1
CIOFFI	56	2536	45	10	113	5	23	14	9	4	1	0	0

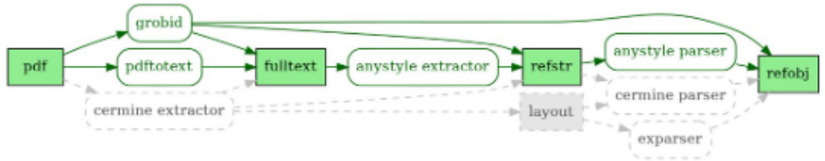
```
<author><surname>Bluhm</surname>
</author>,
<author><given-names>K.</given-names>
</author>
(<year>2001</year>):
<title>Exporting or Abandoning the
“German Model”? Labour Policies of
German Manufacturing Firms in Central
Europe</title>.
<source>European Journal of Industrial
Relations</source>,
<volume>7</volume> (<issue>2</issue>):
<fpage>153</fpage> - <lpage>74</lpage>.
```

https://github.com/exciteproject/EXgoldstandard/tree/master/Goldstandard_EXparser

GESIS Leibniz Institute for the Social Sciences

Evaluation of reference extraction

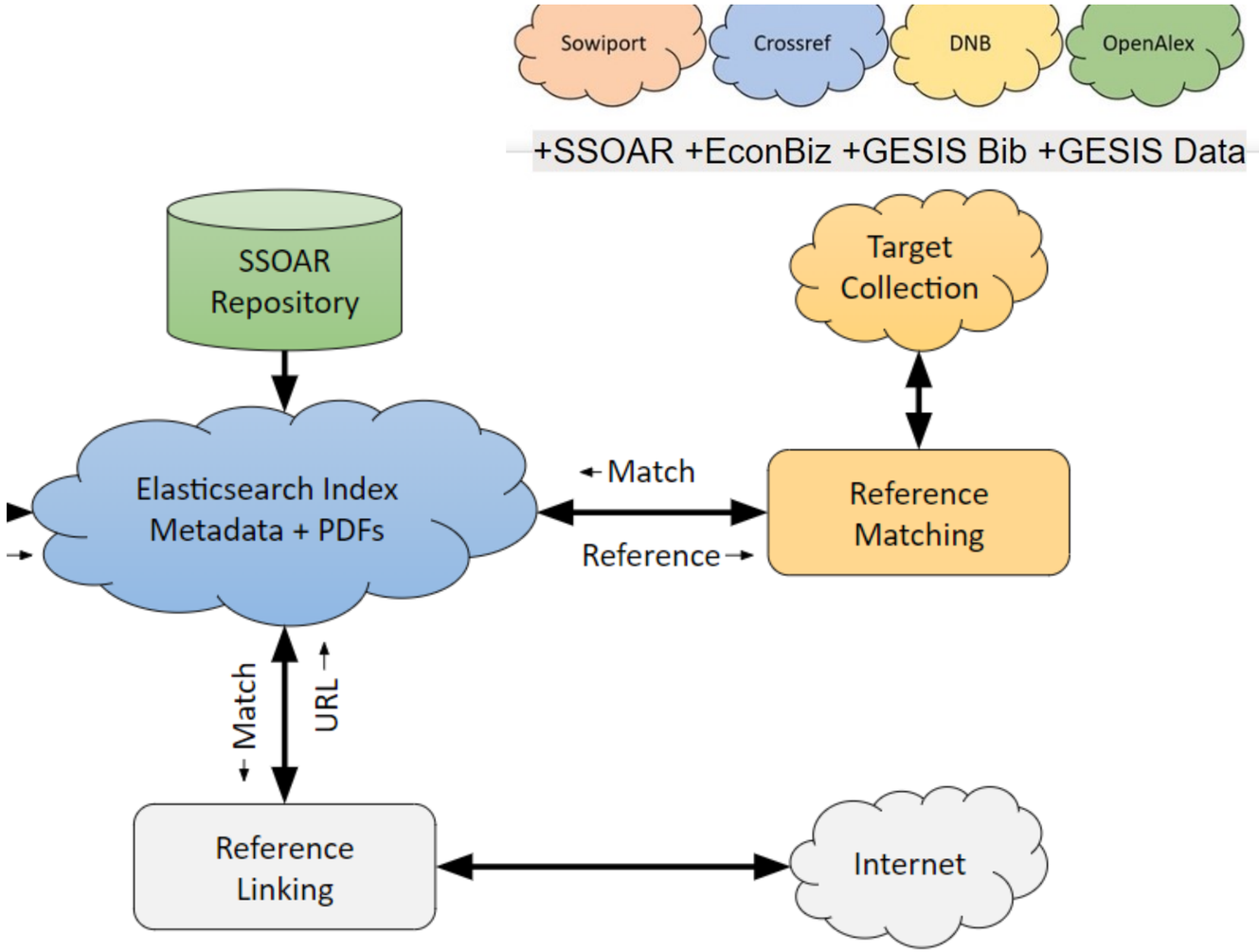
Evaluated against **EXCITE, GEOCITE**, (CIOFFI) gold standards



performance	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R
GEOCITE	**refstr**			**title**			**year**			**author**			**editor**			**publ**			**source**			**vol**			**issue**			**startp**			**endp**		
cerm.txt–anyst	95	**77**	65	86	**69**	58	93	**73**	62	65	**62**	60	47	**32**	26	22	**19**	18	69	**59**	52	70	**57**	48	41	**36**	33	72	**63**	58	76	**65**	58
cerm.ref–anyst	90	**83**	77	85	**75**	68	91	**80**	73	64	**67**	70	48	**38**	32	24	**21**	21	67	**61**	56	71	**61**	54	41	**37**	34	74	**70**	67	81	**72**	66
grob.txt–anyst	94	**86**	80	89	**81**	75	94	**82**	73	66	**71**	77	58	**45**	38	26	**24**	25	71	**68**	66	74	**63**	56	43	**41**	39	80	**77**	74	84	**78**	73
grob.ref–anyst	90	**87**	83	89	**83**	78	92	**84**	78	67	**73**	80	57	**46**	39	26	**25**	27	72	**69**	67	78	**66**	58	47	**44**	43	80	**78**	76	84	**79**	75
pdftotxt–anyst	95	**81**	72	87	**74**	65	91	**77**	67	64	**65**	67	49	**39**	33	24	**21**	22	70	**63**	58	72	**59**	51	47	**42**	38	74	**68**	64	77	**69**	62
cermine	90	**83**	77	80	**68**	60	90	**78**	69	56	**57**	59	0	–	0	0	–	0	47	**37**	32	57	**50**	46	29	**22**	19	57	**52**	49	73	**59**	49
grob.ref–cerm	90	**87**	83	89	**79**	72	91	**82**	75	69	**60**	54	0	–	0	0	–	0	52	**45**	41	58	**53**	50	31	**27**	26	61	**57**	56	76	**64**	56
cerm.layo–exp	88	**74**	65	84	**70**	61	85	**69**	58	59	**57**	55	45	**34**	28	16	**11**	11	58	**52**	48	54	**50**	46	40	**37**	36	58	**52**	48	60	**51**	45
grobid	90	**87**	83	86	**78**	71	89	**82**	78	71	**75**	81	47	**26**	18	25	**18**	17	67	**56**	49	73	**65**	58	37	**29**	24	82	**78**	74	83	**78**	74
EXCITE	**refstr**			**title**			**year**			**author**			**editor**			**publ**			**source**			**vol**			**issue**			**startp**			**endp**		
cerm.txt–anyst	93	**85**	78	90	**81**	74	93	**84**	76	82	**79**	76	67	**56**	50	52	**50**	49	72	**67**	63	73	**62**	55	46	**44**	44	77	**71**	65	80	**70**	64
cerm.ref–anyst	89	**82**	76	85	**78**	73	90	**81**	75	78	**76**	75	60	**52**	46	47	**45**	44	71	**67**	64	69	**61**	55	42	**38**	37	75	**71**	67	79	**72**	68
grob.txt–anyst	87	**84**	82	83	**79**	77	89	**83**	79	75	**77**	80	65	**58**	53	49	**49**	50	65	**66**	68	72	**63**	56	44	**43**	43	76	**74**	73	80	**77**	74
grob.ref–anyst	83	**85**	87	81	**81**	82	88	**85**	83	75	**79**	84	63	**57**	53	49	**51**	53	66	**67**	69	71	**64**	58	45	**44**	45	75	**75**	74	79	**77**	75
pdftotxt–anyst	90	**87**	85	88	**84**	81	90	**86**	82	79	**81**	83	67	**59**	53	50	**51**	53	71	**69**	69	70	**63**	58	43	**43**	44	77	**75**	73	80	**75**	72
cermine	89	**82**	76	81	**71**	63	89	**80**	73	66	**61**	57	0	–	0	0	–	0	53	**46**	41	61	**55**	50	36	**31**	28	63	**60**	57	77	**63**	55
grob.ref–cerm	83	**85**	87	80	**77**	74	88	**84**	81	69	**58**	50	0	–	0	0	–	0	50	**46**	43	58	**56**	55	39	**36**	34	65	**63**	61	75	**67**	61
cerm.layo–exp	87	**79**	72	85	**77**	71	87	**75**	66	74	**67**	62	59	**55**	53	50	**44**	40	65	**62**	59	57	**56**	56	43	**42**	42	59	**56**	54	61	**57**	53
grobid	83	**85**	87	77	**77**	76	84	**83**	82	74	**78**	84	51	**33**	25	47	**41**	37	57	**51**	47	68	**63**	60	39	**32**	28	74	**74**	75	75	**73**	73

GESIS Leibniz Institute for the Social Sciences

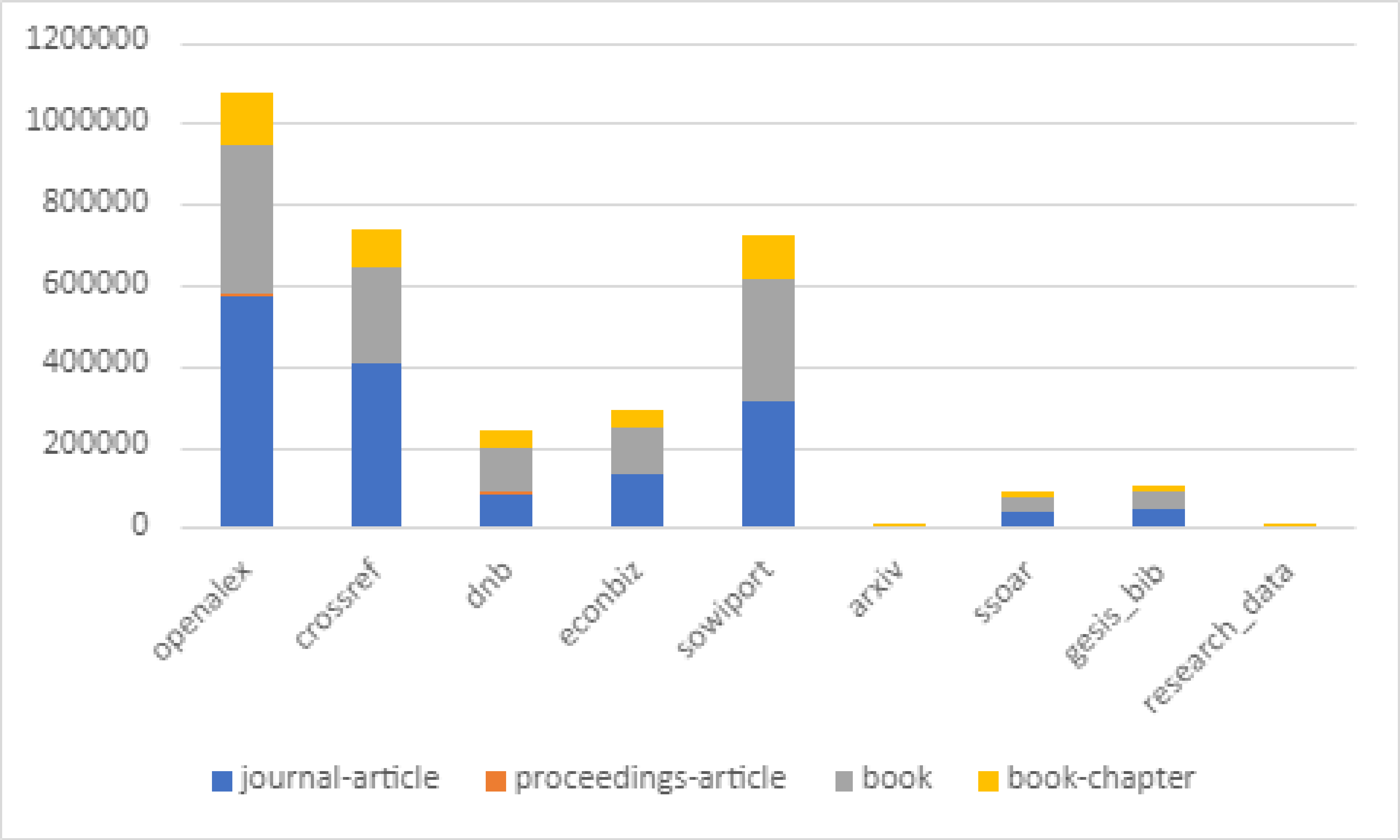
Reference Matching and Linking

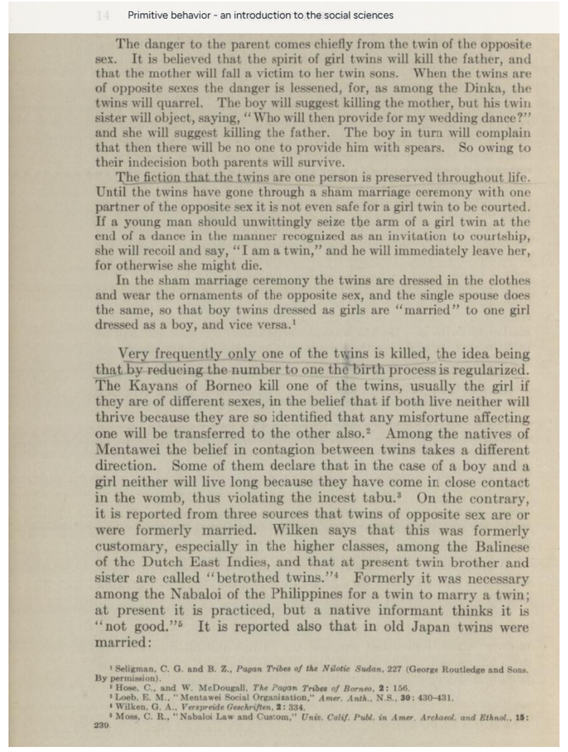


Reference Matching and Linking

Linking results

- 2.6 million references extracted
- **1.5 million references matched and linked**
- multiple matches for a reference possible (demo)
- data ingested to Open Citations





GESIS Leibniz Institute for the Social Sciences

Reference Extraction and Linking: lessons learned

Reference Extraction & Parsing

- Looked up open-source tools for reference extraction and **created tool combinations (pipeline)**
- Three datasets to **evaluate** these reference extraction and parsing tools
- Can recommend **best tool chains** → overall good performance
- Older and less standardized documents (SSOAR, GEOCITE) are harder to work on

Reference Matching and Linking to target collections

- Focus of the project was to match more references to existing metadata
- Done using Elasticsearch queries to **locally replicated target collections** like Openalex, etc.
- Title or reference string queries followed by a rule-based check for actual matching

Reference Matching to Web (the remaining 1.1 mio unmatched refs)

- Another goal was to match references to targets on the internet
- **BING API was feasible and cost-effective. Search functionality is fully implemented and integrated**
- First, we were reminded of the updated terms of use and we noticed that we were not allowed to store the mapping between the search query and the result, which would be necessary
- Next, we were reminded of the API's new pricing which increased the cost manifold
- Had to resort to only **providing query links to the user in the demonstrator for unresolved references**

GESIS Leibniz Institute for the Social Sciences

OUTCITE

OUTCITE, the extension of EXCITE(Extraction of Citations from PDF Documents) is a toolchain of citation extraction software and particular focus on the social sciences and this is a public service for the project. In the background of this page we are using GROBID and CERMINE for extracting content from PDF files. Afterwards, Anystyle and Exparser has been used along with self processing of XML from GROBID and CERMINE for references extraction and segmentation.

How to start the process:

- **First:** Choose a pdf file by clicking on "Choose File" button.
- **Second:** Click on "Upload" button to start the process.
- **Note:** After uploading a file, follow-up ID will be displayed on the screen. This ID is necessary for following up the results for an upload. So, kindly save it or bookmark a URL.

How to check the result:

- **First:** Enter the follow-up ID in the search bar below then click on "Search" button or directly hit the URL(displayed for the new upload) to open it up in the new tab.
- **Second:** The search result will be displayed below on the same screen if it is available but you can also navigate to the new screen with "Get to Result" button.
- **Note:** Complete extraction process will take a while, at least a minute for new upload, and it completely depends on the size of the file.
- **Tip:** If you ever forget to save the follow-up ID or bookmark the result URL then you can also search for the results for your upload via searching for some unique keywords or strings like title, author(s) etc. of the pdf.

Disclaimer: ***Do not upload any confidential or private document that you don't want to share publicly. Result will be accessible to all users.***

Processed Document Example: https://demo-outcite.gesis.org/users/b66f4abceff160ddc6e77bc005aa8cc6

Search id or query... search

Datei auswählen Keine ausgewählt Upload

https://demo-outcite.gesis.org/

References and Links

• Backes, T., Iurshina, A., Shahid, M. A., & Mayr, P. (2024). Comparing Free Reference Extraction Pipelines. International Journal on Digital Libraries. https://link.springer.com/article/10.1007/s00799-024-00404-6

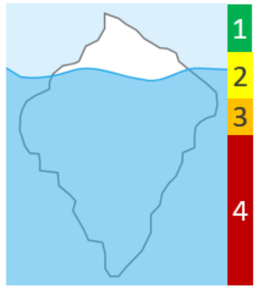
• Hosseini, A., Ghavimi, B., Boukhers, Z., & Mayr, P. (2019). EXCITE - A toolchain to extract, match and publish open literature references. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2019, 432–433. https://doi.org/10.1109/JCDL.2019.00105

• OUTCITE demo: https://demo-outcite.gesis.org/

• Published components: EXCITE: https://github.com/exciteproject
OUTCITE: https://github.com/OUTCITE/outcite-refextract

Overview

1. Introduction to Scholarly Document Processing
2. **Selected Applications**
    - Reference Extraction and Linking
    - **Entity Extraction**
3. A Guide to Scholarly Information Extraction
4. Shared Tasks
5. Summary & Outlook



GESIS Leibniz Institute for the Social Sciences

In-text (Artifacts)

In-text (in-domain)

GSAP-NER

Entity Extraction Focused on Machine Learning Models and Datasets

Idea of GSAP:
Improving the reproducibility and reusability of ML-research

Tracking Models and Data:

- Which model is trained on which dataset?
- What is the data source of each dataset?

Documenting Model Architectures:

- Which model architectures are applied to specific tasks?

GSAP-NER is a key component to address these questions by detecting mentions of relevant entities.

BERD @NFDI

NFDI4 DataScience

Unknown Data

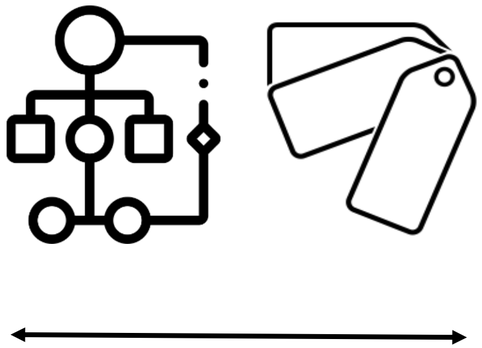
DFG funded

https://data.gesis.org/gsap/gsap-ner

GESIS Leibniz Institute for the Social Sciences

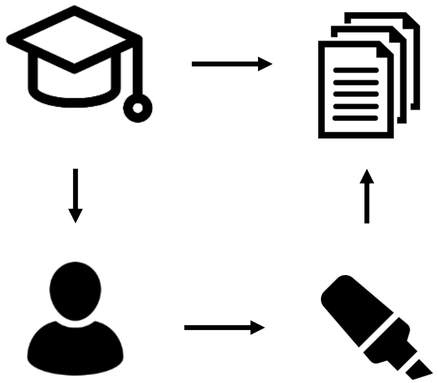
GSAP-NER Contribution

Data Model and Tag Set



Scholarly entity tag set
and guideline
for data driven research

Corpus



Annotated full text corpus
computer science domain

Model

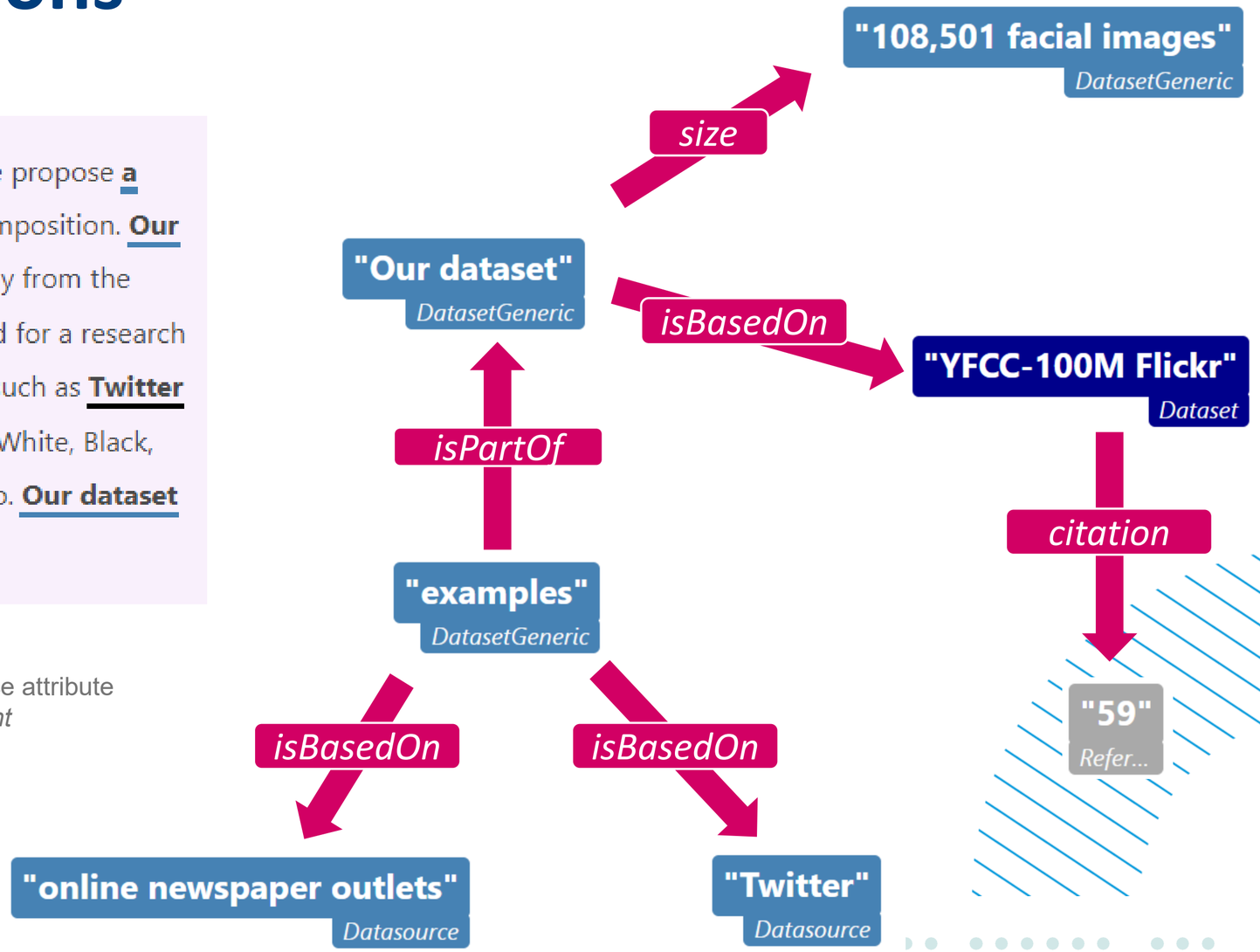
Fine-tuned NER
baseline models
for automatic extraction

GESIS Leibniz Institute for the Social Sciences

Example: Dataset Mentions

To mitigate the race bias in **the existing face datasets**, we propose **a novel face dataset** with an emphasis of balanced race composition. **Our dataset** contains **108,501 facial images** collected primarily from the **YFCC-100M Flickr** dataset [**59**], which can be freely shared for a research purpose, and also includes **examples** from other sources such as **Twitter** and **online newspaper outlets**. We define 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino. **Our dataset** is well-balanced on these 7 groups (See Figure 3 and 2)

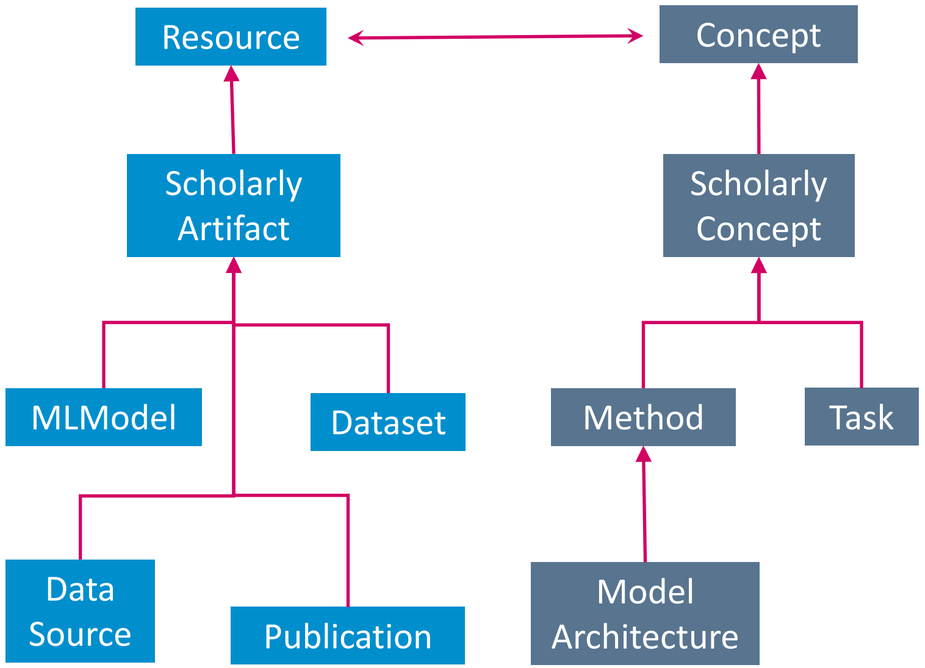
Paragraph from:
Kärkkäinen, Kimmo, and Jungseock Joo (2019). FairFace: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*



GESIS Leibniz Institute for the Social Sciences

From Data Model to Tag Set

Conceptional data model



Tag set applicable for annotation

Model Related

- MLModel
- MLModelGeneric
- ModelArchitecture
- Method
- Task

Data Related

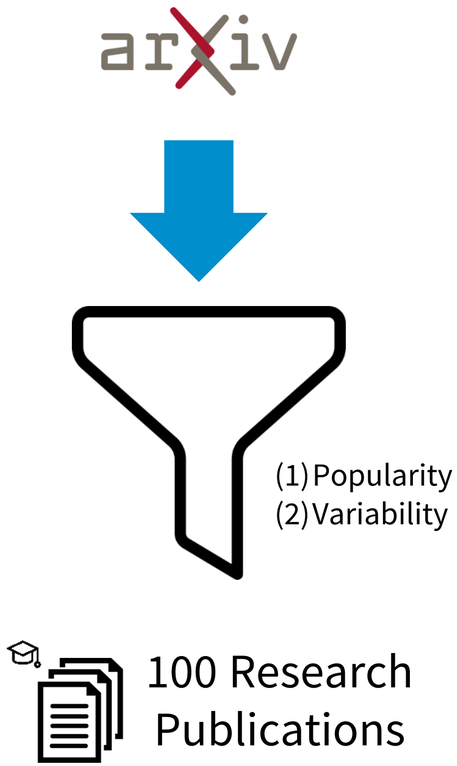
- Dataset
- DatasetGeneric
- DataSource

Anchors

- ReferenceLink
- URL

GESIS Leibniz Institute for the Social Sciences

Corpus Creation



Choose Data Source:

- Relevance for research domain
- Open Access Repositories

Choose sampling method:

Popularity:

- Based on mentions in Readme files of the most popular HuggingFace models

Variability:

- Random selection with defined filters (ML Category)
- Timespan (last 5 year)
- model and data mention (string match)

Preprocess the documents:

- Use Grobid for PDF to Text
- Choose Inception as Annotation Tool

https://grobid.readthedocs.io/
https://inception-project.github.io/

GESIS Leibniz Institute for the Social Sciences

Annotation Process

100 Research Publications

Manual Annotation

Annotators

Interrater Agreement

	mutual F1 exact-match	mutual F1 partial-match
MLModel	72.1	74.6
MLModelGeneric	60.7	67.6
ModelArchitecture	23.7	34.4
Method	47.0	60.7
Task	51.4	55.2
Dataset	84.1	86.7
DatasetGeneric	56.2	65.8
DataSource	55.3	62.7
ReferenceLink	90.5	94.8
URL	86.1	94.1
all	61.4	69.3

Table 3: Interrator agreement as measured by the average mutual F1 of three annotators on the 14% co-annotated publications.

Quality Assurance

GSAP-NER Corpus

- > 54,000 Annotations
- 546 Annotations/Document
- > 25,000 Sentences

Corpus Statistics

	# spans	# unique spans
Method	12,826	6,547
DatasetGeneric	9,838	5,781
MLModelGeneric	8,521	4,238
ReferenceLink	7,172	2,257
MLModel	5,012	944
Task	4,143	1,478
Dataset	3,898	883
ModelArchitecture	2,612	985
DataSource	508	185
URL	68	61
Total	54,598	23,359

Table 4: Text span statistics in our GSAP-NER dataset ordered by the number of spans per entity type.

GESIS Leibniz Institute for the Social Sciences

Model Selection

Supervised Approach

- state of the art domain specific NER
- Scalable approach (vs. LLMs)
- Transformer language models
- Classify each word in the publication
  - Goal: Reproduce annotation

Domain Specific Model Pre-training

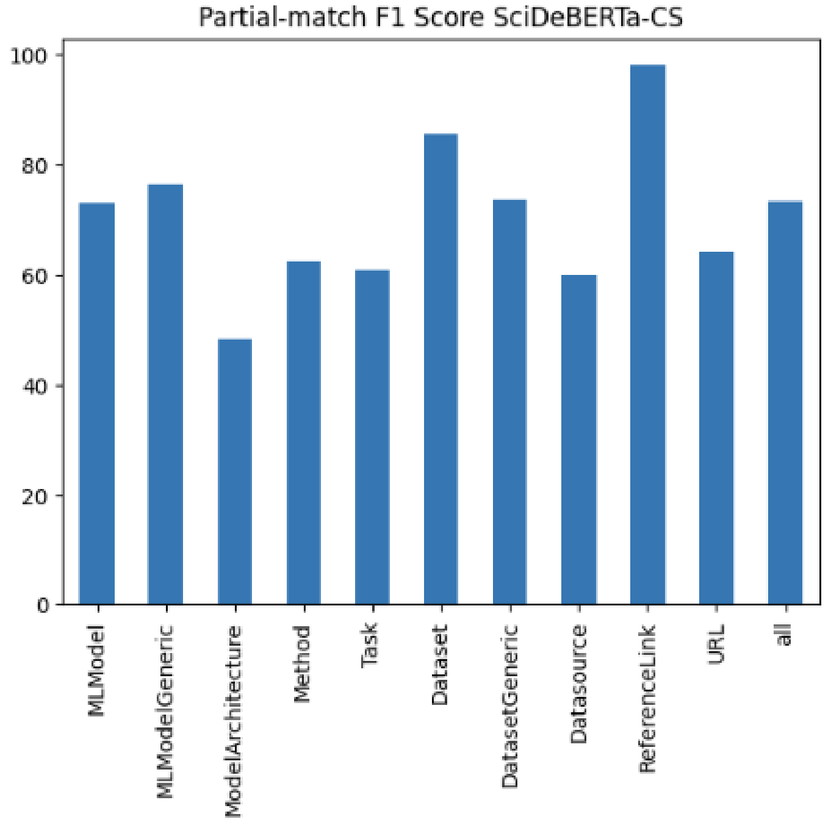
- Select and compare models trained inside the domain

Model Architecture	MLModel	In-domain pre-training
BERT	SciBERT	Science
RoBERTa	$\text{RoBERTa}_{\text{base}}$	-
RoBERTa	$\text{RoBERTa}_{\text{large}}$	-
DeBERTa	**SciDeBERTa-CS**	**Computer Science**

GESIS Leibniz Institute for the Social Sciences

Experiments

MLModel	NER Fine-Tuning	10-Fold F1-Score
SciBERT	GSAP-NER Corpus	70.6
$RoBERTa_{base}$		72.0
$RoBERTa_{large}$		72.7
SciDeBERTa-CS		**73.4**



Evaluation target

- How correct are the extractions (incl. label)?
- How many of the annotated text spans could be reproduced

GESIS Leibniz Institute for the Social Sciences

Next steps

Further development of the GSAP Pipeline

- Relate the detected entities
- Disambiguate the entity mentions
- Link the entities to external sources

Apply our pipeline on larger corpora

- Linguistics (ACL), ML-learning (arXiv ML)
- Test domain adaptability on publications of different research domains (e.g. social sciences)

Building a GSAP Knowledge Graph

- Integrate the results of the GSAP in machine-readable form
- Ready to query
- Ready to support generative LLMs with traceable information

Integrate the results into GESIS Services

- Integrate results in the Gesis Knowledge Graph infrastructure.
- E.g., enriching method descriptions and tutorials in *GESIS Method Hub* *https://methodshub.gesis.org/*

GSAP demo will be available soon!

Overview

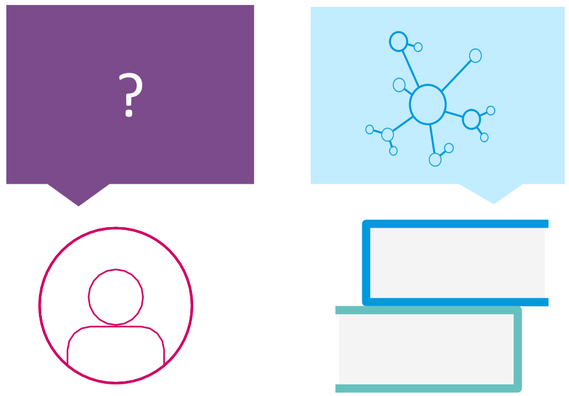
1. Introduction to Scholarly Document Processing
2. Selected Applications
    - Reference Extraction and Linking
    - Entity Extraction
3. **A Guide to Scholarly Information Extraction**
4. Shared Tasks
5. Summary & Outlook

GESIS Leibniz Institute for the Social Sciences

How to Extract Scholarly Information

Start with specific research questions or use cases
(often domain dependent)

End with structured information that can serve to answer the questions



Steps:

- Data modeling
- Data collection
- Information extraction
- Annotation
- Model Evaluation
- Deployment

GESIS Leibniz Institute for the Social Sciences

Data Modeling

Assumption:

You have defined research questions or use cases

Goal:

Define the related entity types

→ specificity vs. general applicability

Define the expected interaction of the entities (relations)

Helpful hint:

Related Work
Search for existing approaches on similar tasks
Search for existing extraction models

Make use of
→ existing data models
→ existing models and tools

How do I extract the information?
→ Is the method extracting the information I want?

GESIS Leibniz Institute for the Social Sciences

Data collection

Aspects of accessing scholarly text

- Fulltext vs. Abstract
- Open Access is an opportunity
  - Publisher Version
  - Preprint Version (Repositories)
- Legal Issues
  - License of original data, and think about redistribution
  - OA: Not every free accessible publication can be used for text and data mining

Helpful Hint: Ask your local librarian!

Example resources to get open access full-texts

Service/Corpus	Size	domain	description
SSOAR	83.711 publications	Social science	Links to PDFs
OpenAlex	6M works in open access status (only links)	Cross discipline	Contains metadata incl. links to open access versions of publications
S2ORC	81.1M English-language scholarly articles	Cross discipline (bias to natural science)	Contains metadata, incl. Abstracts, resolved references and fulltexts
Pubmed Central	10M articles	Biomedical	Provides XML, TXT, PDF possibilities for text mining
ArXiv	2.4M articles	Cross discipline	provides PDF and latex versions

https://github.com/allenai/s2orc
https://openalex.org/
https://gesis.org/ssoar
https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/
https://arxiv.org/

Information Extraction Methods

Transform the information need into an extraction task

- Task difficulty level varies, e.g. sentence classification vs. token classification vs. question answering

Choose models or tools for corresponding task

- In-context learning with pretrained models
- Fine-tune pretrained models with task-specific corpus
- The choice is made based on task performance on your targeted data

Task	Goal	Input	Output
Named Entity Recognition (NER)	Identify named entities in the text	Text	A list of target entities with their types
Relation Extraction (RE)	Identify (binary) relations of entities in the text	Text + optional identified entities in the given text	A list of entity pairs and the relation type between them.
Entity Resolution	Identify mentions referring to the same entity	Text + optional entity knowledge base	Clusters of entity mentions
Question Answering (QA)	Extract information using defined extraction Question	Text + predefined questions	A list of related answers In text or generated based on context

GESIS Leibniz Institute for the Social Sciences

Annotation Phase

Select Annotators

- Are experts needed?

Annotation is an iterative process

- Time for re-definition/adjustment of data model is needed

Annotation is needed, even if you use out of the box models

Alternatives to full manual annotation:

- Distant supervision
- Active learning approaches
- Automatic Annotation

Model Evaluation

How good is my model performance?

- Metrics
  - Which metrics can reflect the performance?
    - Accuracy metric: precision, recall, F1 etc.
    - Alignment metric (to compare multi-version annotations): interrater agreement
  - To which benchmark or ground-truth can we compare the model?
- Human evaluation

GESIS Leibniz Institute for the Social Sciences

Deployment

Apply your model on a larger corpus

- Data collection of the larger corpus (see data collection)
- Model performance analysis
- When an in-production system is needed, please reach out for help

- At GESIS we have a process to bring methods into production system

GESIS Leibniz Institute for the Social Sciences

References and Links

- Otto, W., Zloch, M., Gan, L., Karmakar, S., & Dietze, S. (2023). GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8166–8176. https://doi.org/10.18653/v1/2023.findings-emnlp.548
- https://data.gesis.org/gsap/gsap-ner

Overview

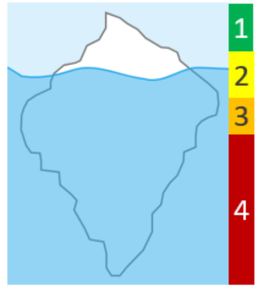
1. Introduction to Scholarly Document Processing
2. Selected Applications
    - Reference Extraction and Linking
    - Entity Extraction
3. A Guide to Scholarly Information Extraction
4. **Shared Tasks**
5. Summary & Outlook

GESIS Leibniz Institute for the Social Sciences

SOMD2025: Shared Task for Software Related Information Extraction

Motivation

- Scientific research increasingly relies on software for analysis and data interpretation.
- Tracing software usage is crucial for reproducibility and collaboration.
- Software mentions in papers are heterogeneous and informal, requiring robust IE methods.



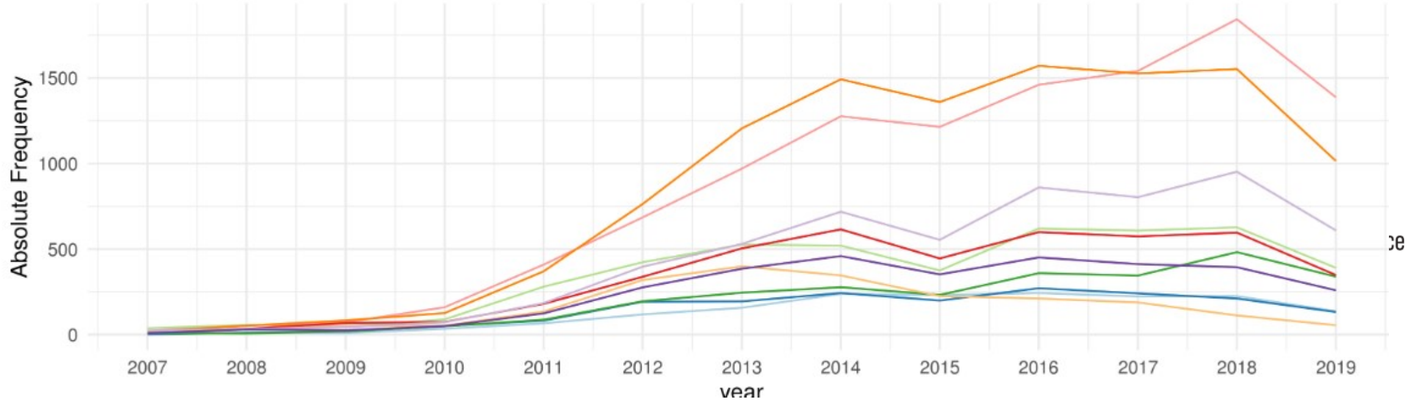
In-text (Artifacts)

Named Entity Recognition (NER)

Relation Extraction (RE)

Entity Linking

Software Usage over time



BERD @NFDI

https://sdproc.org/2025/somd25.html

Software Mention Detection (SOMD) 2025

GESIS Leibniz Institute for the Social Sciences

SOMD2025: Shared Task for Software Related Information Extraction

Objective: The goal is to jointly detect software mentions and their attributes in scholarly texts and classify their relationships.

Data Source: The dataset consists of annotated sentences from full-text scholarly documents.

Annotations: Entities (e.g., software names, URLs, versions) and their interrelations are manually labeled in the data.

Focus/Task: Participants must build models for joint Named Entity Recognition (NER) and Relation Extraction (RE) evaluated in two phases.

NFDI4 DataScience

BERD @NFDI

https://sdproc.org/2025/somd25.html

GESIS Leibniz Institute for the Social Sciences

SciVQA

SciVQA: Scientific Visual Question Answering

Objective: Develop multimodal question answering (QA) systems capable of interpreting scientific figures (e.g., charts, diagrams) from scholarly articles.

Data Source: Figures extracted from ACL Anthology and arXiv papers.

Annotations: Each figure is annotated with seven QA pairs and includes metadata such as captions, figure IDs, figure types (e.g., compound, line graph, bar chart, scatter plot), and QA pair types.

Focus/Task : The task emphasizes closed-ended questions, both visual (addressing attributes like color, shape, size, height) and non-visual (not addressing figure visual attributes).

https://sdproc.org/2025/scivqa.html

GESIS Leibniz Institute for the Social Sciences

SciHal2025: Hallucination Detection for Scientific Content

Objective: Develop systems to detect hallucinated (unsupported) claims in AI-generated answers to scientific questions.

Data Source: Research-oriented questions from experts, with answers generated by retrieval-augmented generation (RAG) systems indexing millions of academic abstracts.

Annotations: Each answer is labeled with a three-class scheme (entailment, neutral, contradiction) and a fine-grained scheme encompassing over 10 categories.

Focus/Task: Participants classify claims into appropriate categories, with evaluation metrics focusing on precision, recall, and F1 score for detecting unsupported claims.

https://sdproc.org/2025/scihal.html

GESIS Leibniz Institute for the Social Sciences

Overview of SDP shared tasks

- Outlook “NFDI4DS Shared Tasks for Scholarly Document Processing” preprint available soon
- More shared tasks at past SDP workshops -> https://sdproc.org/2025/previousworkshops.html

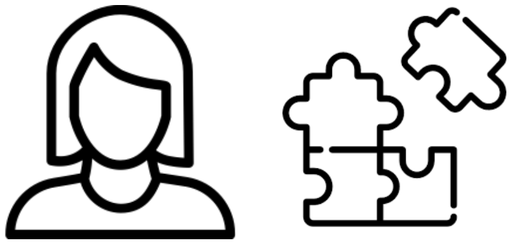
Overview

1. Introduction to Scholarly Document Processing
2. Selected Applications
    - Reference Extraction and Linking
    - Entity Extraction
3. A Guide to Scholarly Information Extraction
4. Shared Tasks
5. **Summary & Outlook**

GESIS Leibniz Institute for the Social Sciences

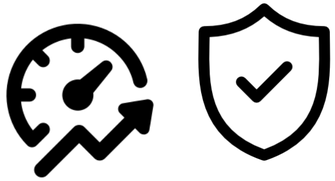
Make use of Scholarly Information

Scholarly Entity Extraction for open infrastructures



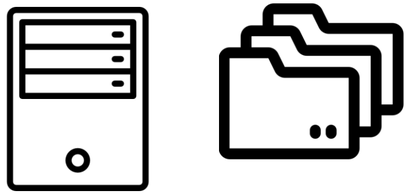
1. Usability and Extensibility

Usability and Interface: User-friendly interface or API
Extensibility: Modular architecture



2. Scalability and Reliability

Scalability: Applicability on larger corpora of thousands of publications
Error Handling and Recovery: Robust mechanisms for error detection, handling, and recovery



3. Data and System Integration

Compatibility with existing Systems: Smooth integration with existing enterprise systems and workflows.
Robust Data Integration: Clean, deduplicate and unify different sources and formats

GESIS Leibniz Institute for the Social Sciences

Example: GESIS Search

< Back

Publications

The Reluctant Feminist: Angela Merkel and the Modernization of Gender Politics in Germany

Mushaben, Joyce Marie

In: Femina Politica - Zeitschrift für feministische Politikwissenschaft, 27, 2018, 2, 83-95

Abstract: Academic studies regarding the impact of various forms of gender representation focus largely on quantitative evidence that women in power can make a difference, downplaying qualitative case studies that can establish causal links between women's participation in government and better policies for women. Analyzing policy changes initiated by Germany's first female Chancellor since 2005, the paper argues that despite her CDU-affiliation, Angela Merkel has contributed more to gender equality in Germany than all previous chancellors, even though she refuses to label herself a feminist. The author ... more

Institution(s): Verlag Barbara Budrich

Topics: Frau | Politikerin | politische Entscheidung | Entscheidungsfindung | Repräsentation | Intersektionalität | Gleichstellung | Merkel, A. | Geschlechterpolitik | Feminismus | Bundesrepublik Deutschland

Document type: Zeitschriftenartikel

Database: SSOAR - Social Science Open Access Repository

100 JAHRE FRAUENWAHLRECHT – UND WO BLEIBT DIE GLEICHHEIT?

The Reluctant Feminist: Angela Merkel and the Modernization of Gender Politics in Germany

JOYCE MARIE MUSHABEN

The adoption of female suffrage across multiple western nations in the early 1900s was accompanied by the expectation that women's ability to vote would eventually lead to their direct involvement in governance. It was further assumed that by boosting more of their own kind into positions of power, female suffrage could and would make a significant difference in the laws and policies being adopted, thus allowing women to shape their own lives (cf. Cress in this volume). The last 100 years have unfortunately supplied much evidence to the contrary, leading countless scholars to investigate women's irregular paths to power, the institutional barriers they face, the stereotypical role expectations that hinder their progress, and new mechanisms seeking to equalize their participation in politics (in Germany: Davidson-Schmich 2016; Kolinsky 1991; Roll 2005; Scholz 2007; Clemens 2006).[1] These studies, in turn, have led us to theorize about different types of representation in an effort to explain when, where and how more women *in* politics might generate better, far-reaching policies *for* women.

Female suffrage may be a necessary condition, but it is clearly not a sufficient one in fostering gender equality. The 100th anniversary of women's right to vote in Germany (cf. Abels 2011) provides a unique opportunity to re-assess the metrics scholars use to determine whether the politicians female voters help to elect do, in fact, adopt policies enhancing the balanced participation of women and men in public life (Carless 1998; Geissel 2000). While many comparativists focus on the quantitative dimensions, known as descriptive representation (the number of women in

https://search.gesis.org/publication/gesis-ssoar-60425

Research data (verified links) (1)

The links to the following research data have been reviewed and confirmed by our staff. The research data are either mentioned in the publication or the publication was reported to us as primary literature for the research data.

International Social Survey Programme: Role of Government I - ISSP 1985

Department of Sociology, Research School of Social Sciences, The Australian National University, Canberra; Eurisko, Ricerca Sociale e di Marketing, Milan; Institut für Soziologie, Universität Graz

GESIS Data Archive, Cologne. ZA1490 Data file Version 1.0.0, https://doi.org/10.4232/1.1490, Date(s) of Data Collection: 02.1985 - 06.1986

Abstract: The International Social Survey Programme (ISSP) is a continuous programme of cross-national collaboration running annual surveys on topics important for the social scien ... more

Downloads
Datasets
Questionnaires
Codebook
Other documents

Actions
Bookmark
Cite

References (49)

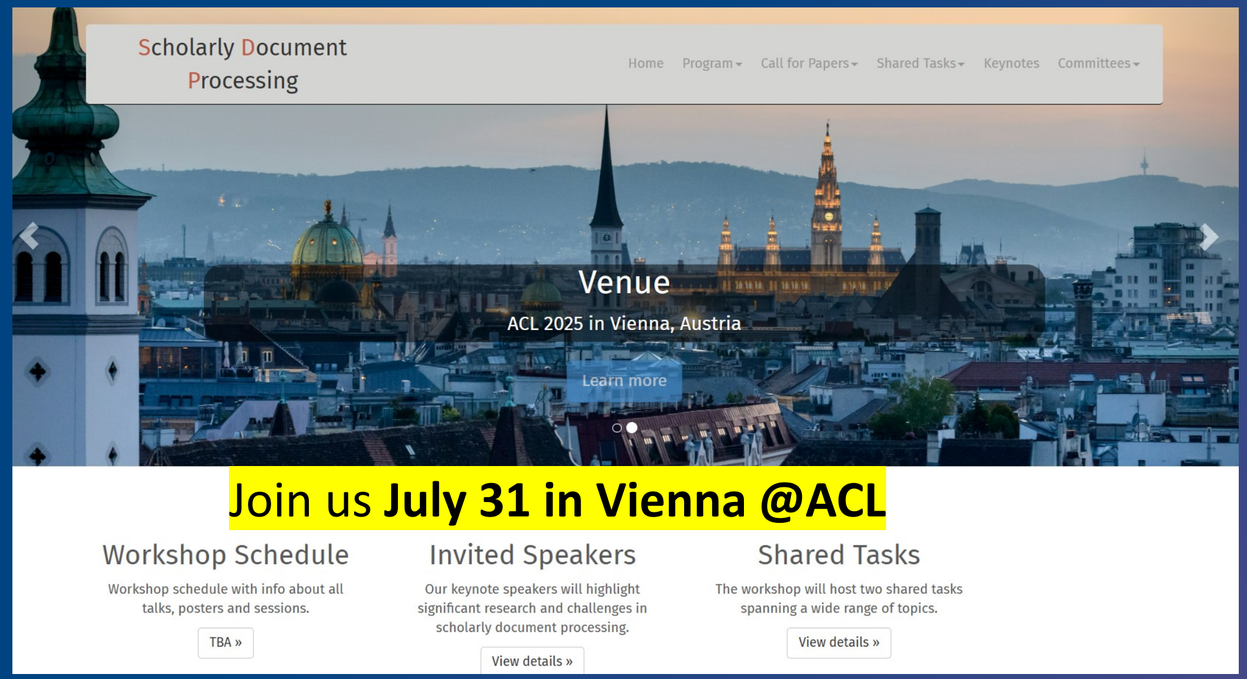
Die Referenzen zu dieser Publikation wurden automatisch aus dem Volltext extrahiert und wo möglich zu Quelle verlinkt.

- Abels, G. (2011). 90 Jahre Frauenwahlrecht in Deutschland. Zum Wan- del von Geschlechterverhältnissen in der Politik. Baden-Baden: Nomos. Deutschland im Jubiläumsjahr 2009. Blick zurück nach vorn. Theodor Eschenburg-Vorlesung, pp. 197-219.
- Annesley, C. G., F. (2010). The Core Executive: Gender, Power and Change. Political Studies, 58(5), pp. 909-929. PDF
- Beckett, C. (2006). Thatcher (British Prime Ministers of the 20th Century. London:.
- Wilpert, B., A. (1991). Frauenrecht in Ost-und Westdeutschland, Bi¬ lanz, Ausblick. Berlin: Bilanz-Ausblick. io Management Zeitschrift.
- Carless, S. A. (1998). Gender Differences in Transformational Leadership: An Examination of Superior, Leader and Subordinate Perspectives. Sex Roles, 39(11), pp. 887-902.
- Clemens, C. (2006). From the outside in: Angela Merkel as opposition leader, 2000-2005. German Politics & Society, 24(3), pp. 41-81.
- Dahlerup, D. (1988). From a Small to a Large Minority: Women in Scandinavian Po- litics" en Scandinavian Political Studies. Asmara, Eritrea: Africa World Press. Estocolmo, Suecia, Institute of Political Science, 11(4), pp. 275-298.
- Davidson-Schmich, L. K. (2016). Gender Quo- tas and Democratic Participation. Recruiting Candi- dates for Elective Offices in Germany. Ann Arbor, MI: University of Michigan Press.

Summary: Key take aways

- Scholarly information extraction (IE) is able to extract metadata and content information from scientific documents
- Tools and frameworks like OUTCITE and GSAP-NER demonstrate scalable, accurate extraction and linking for real-world scholarly use cases
- Scholarly IE can help to implement the FAIR principles
- Results can be used for new services to support researchers
- If you are interested in conducting scholarly information extraction: Requests Welcome!

Leibniz Association



https://sdproc.org/2025/

Contact:

Philipp Mayr
philipp.mayr@gesis.org
https://philippmayr.github.io/

Leibniz Association