gesis Leibniz-Institut für Sozialwissenschaften

hhu Heinrich Heine Universität Düsseldorf

HeiCAD
Heine Center for Artificial Intelligence and Data Science

Research Knowledge Graphs and scholarly information extraction @ BERD@NFDI & GESIS

Focused Tutorial on Capturing, Enriching, Disseminating Research Data Objects

Stefan Dietze, 25.11.2022

Leibniz Gemeinschaft

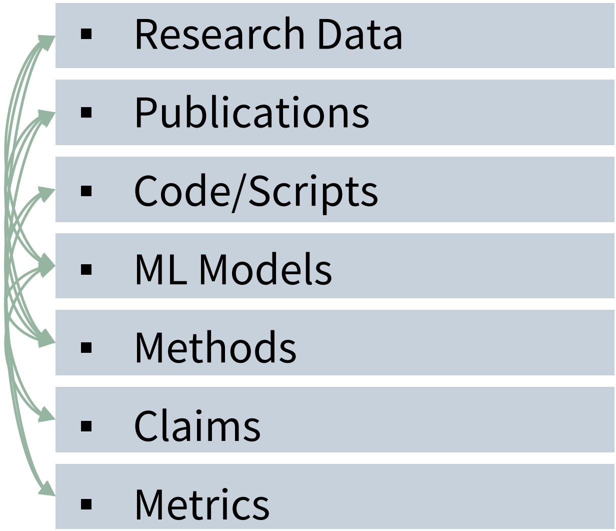
BERD@NFDI

GESIS @ National Research Data Infrastructure (NFDI)

Relevant consortia with GESIS in leading roles

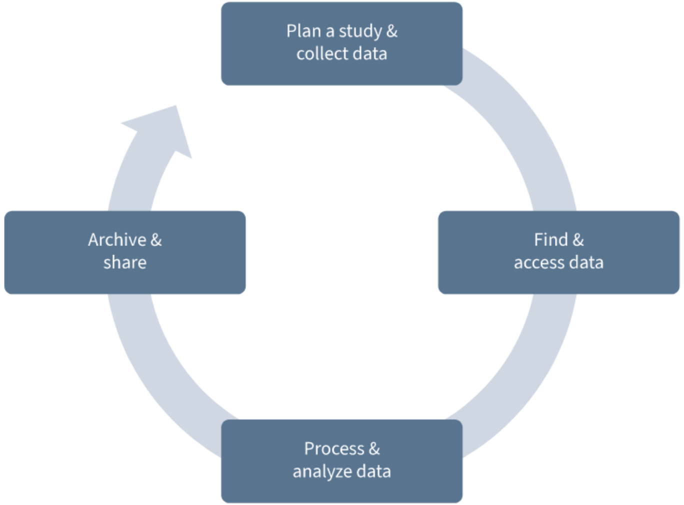
- BERD@NFDI
  https://www.berd-nfdi.de/
- NFD4DataScience – National Research Data Infrastructure for Data Science & AI
  https://www.nfdi4datascience.de/
- KonsortSWD
  https://www.konsortswd.de/en/
- Base4NFDI
  https://base4nfdi.de/

Provenance & Dependencies of Research Data, Resources, Knowledge



- Research Data
- Publications
- Code/Scripts
- ML Models
- Methods
- Claims
- Metrics

Relations between scientific resources, data, knowledge



Plan a study & collect data

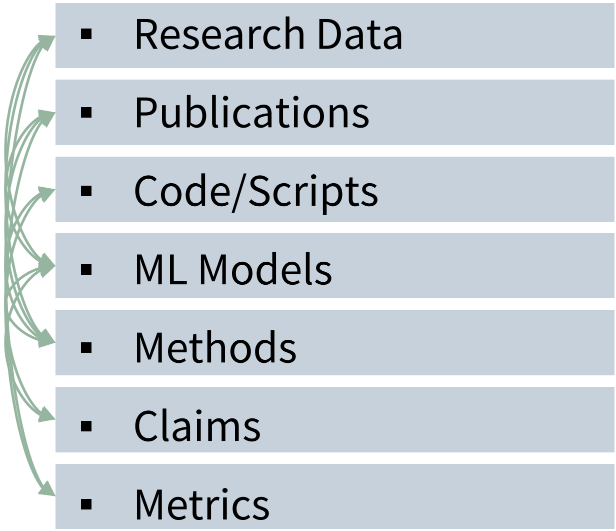
Find & access data

Process & analyze data

Archive & share

Research Data Cycle

Provenance & Dependencies of Research Data, Resources, Knowledge



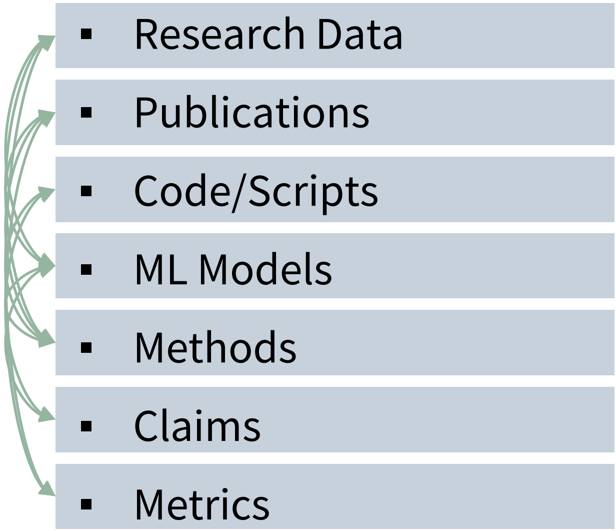
- Research Data
- Publications
- Code/Scripts
- ML Models
- Methods
- Claims
- Metrics

Relations between scientific resources, data, knowledge

Common questions for researchers

- Which top-tier publications cite which data/method? („dataset authority“)
- Which data was used to train/evaluate which method? Which method to produce what data?
- Which claims are supported/cited/rejected by what dataset or publication?

Provenance & Dependencies of Research Data, Resources, Knowledge



- Research Data
- Publications
- Code/Scripts
- ML Models
- Methods
- Claims
- Metrics

Relations between scientific resources, data, knowledge

Challenges

- Data & metadata about resources and concepts not represented in **structured, machine-interpretable, integrated manner** (hidden in publications, web pages etc)
- **Persistent identifiers** (e.g. DOIs) used inconsistently (e.g. on publications/datasets, to small degree on ML models)
- **Relations and semantics** not explicit
- **Reproducibility crisis** in CS/DS/AI

Knowledge Graphs for FAIR Research Data

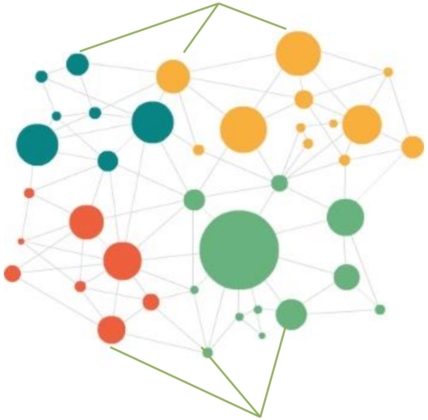
- **Data interoperability and reuse** through established W3C standards for data sharing (on the Web), e.g. RDF, JSON, shared vocabularies (e.g. schema.org, DCAT, DDI), APIs for data reuse and linking

- Making **links** between resources and concepts explicit & **machine-interpretable**
(e.g. which publications cite what dataset?)

- Consistent **use of persistent IDs** (e.g. URIs, DOIs) across all data, e.g. concepts, resources etc („*DOIs for all*“)

Resources

- Datasets
- Publications
- Code
- Software

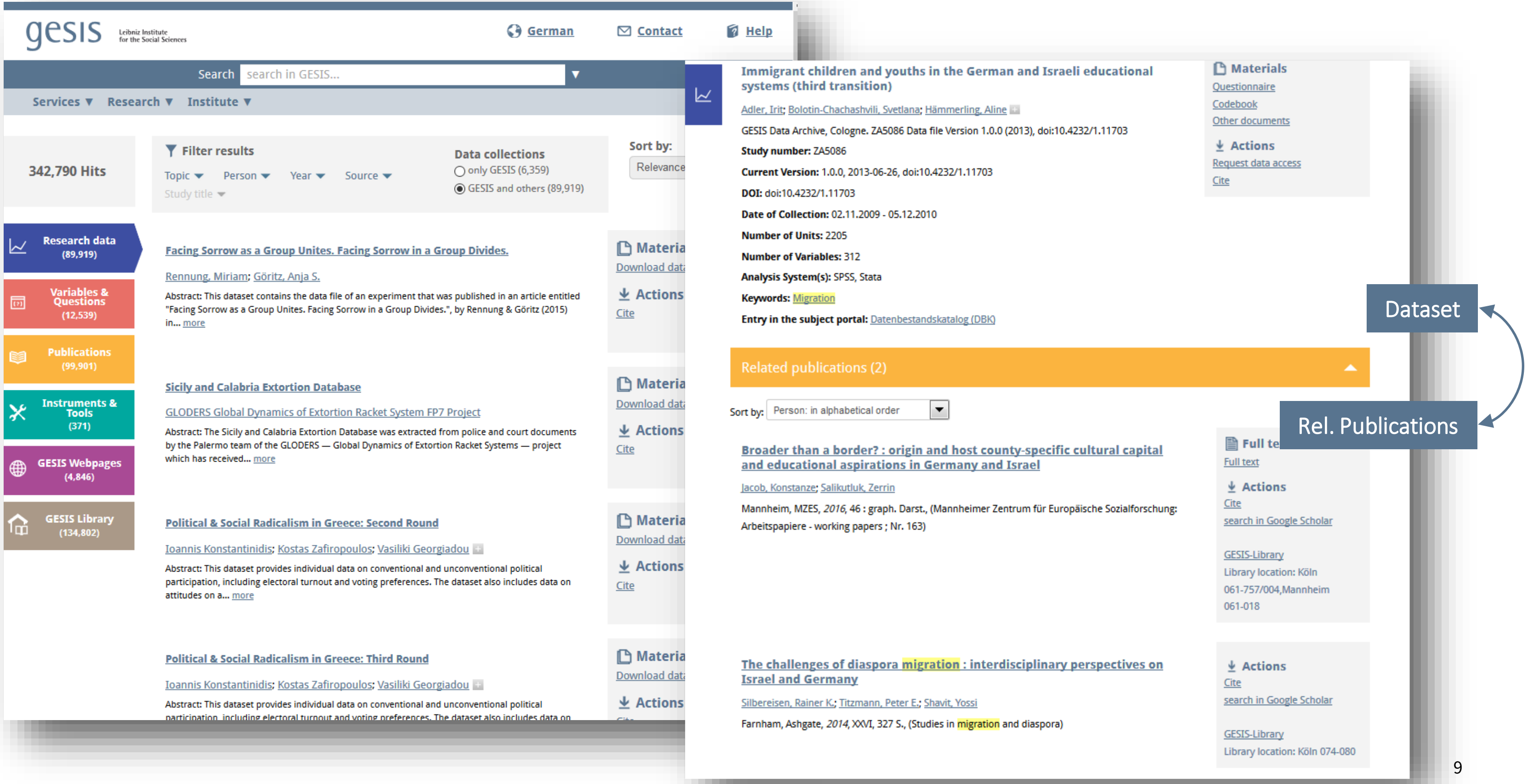


Concepts

- Terms & Definitions
- Claims
- Methods
- Topics
- Entities

Research KGs in Practice: integrated search @ GESIS

https://search.gesis.org/



gesis Leibniz Institute for the Social Sciences

German Contact Help

Search search in GESIS...

Services ▼ Research ▼ Institute ▼

342,790 Hits

Filter results

Topic ▼ Person ▼ Year ▼ Source ▼

Study title ▼

Data collections

○ only GESIS (6,359)

● GESIS and others (89,919)

Sort by: Relevance

Research data (89,919)

Variables & Questions (12,539)

Publications (99,901)

Instruments & Tools (371)

GESIS Webpages (4,846)

GESIS Library (134,802)

Facing Sorrow as a Group Unites. Facing Sorrow in a Group Divides.

Rennung, Miriam; Göritz, Anja S.

Abstract: This dataset contains the data file of an experiment that was published in an article entitled "Facing Sorrow as a Group Unites. Facing Sorrow in a Group Divides.", by Rennung & Göritz (2015) in... more

Materia Download dat Actions Cite

Sicily and Calabria Extortion Database

GLODERS Global Dynamics of Extortion Racket System FP7 Project

Abstract: The Sicily and Calabria Extortion Database was extracted from police and court documents by the Palermo team of the GLODERS — Global Dynamics of Extortion Racket Systems — project which has received... more

Materia Download dat Actions Cite

Political & Social Radicalism in Greece: Second Round

Ioannis Konstantinidis; Kostas Zafiropoulos; Vasiliki Georgiadou

Abstract: This dataset provides individual data on conventional and unconventional political participation, including electoral turnout and voting preferences. The dataset also includes data on attitudes on a... more

Materia Download dat Actions Cite

Political & Social Radicalism in Greece: Third Round

Ioannis Konstantinidis; Kostas Zafiropoulos; Vasiliki Georgiadou

Abstract: This dataset provides individual data on conventional and unconventional political participation, including electoral turnout and voting preferences. The dataset also includes data on

Materia Download dat Actions

Immigrant children and youths in the German and Israeli educational systems (third transition)

Adler, Irit; Bolotin-Chachashvili, Svetlana; Hämmerling, Aline

GESIS Data Archive, Cologne. ZA5086 Data file Version 1.0.0 (2013), doi:10.4232/1.11703

Study number: ZA5086

Current Version: 1.0.0, 2013-06-26, doi:10.4232/1.11703

DOI: doi:10.4232/1.11703

Date of Collection: 02.11.2009 - 05.12.2010

Number of Units: 2205

Number of Variables: 312

Analysis System(s): SPSS, Stata

Keywords: Migration

Entry in the subject portal: Datenbestandskatalog (DBK)

Materials
Questionnaire
Codebook
Other documents

Actions
Request data access
Cite

Related publications (2)

Sort by: Person: in alphabetical order

Broader than a border? : origin and host county-specific cultural capital and educational aspirations in Germany and Israel

Jacob, Konstanze; Salikutluk, Zerrin

Mannheim, MZES, *2016*, 46 : graph. Darst., (Mannheimer Zentrum für Europäische Sozialforschung: Arbeitspapiere - working papers ; Nr. 163)

Full te
Full text

Actions
Cite
search in Google Scholar

GESIS-Library
Library location: Köln
061-757/004,Mannheim
061-018

The challenges of diaspora migration : interdisciplinary perspectives on Israel and Germany

Silbereisen, Rainer K.; Titzmann, Peter F.; Shavit, Yossi

Farnham, Ashgate, *2014*, XXVI, 327 S., (Studies in migration and diaspora)

Actions
Cite
search in Google Scholar

GESIS-Library
Library location: Köln 074-080

Dataset

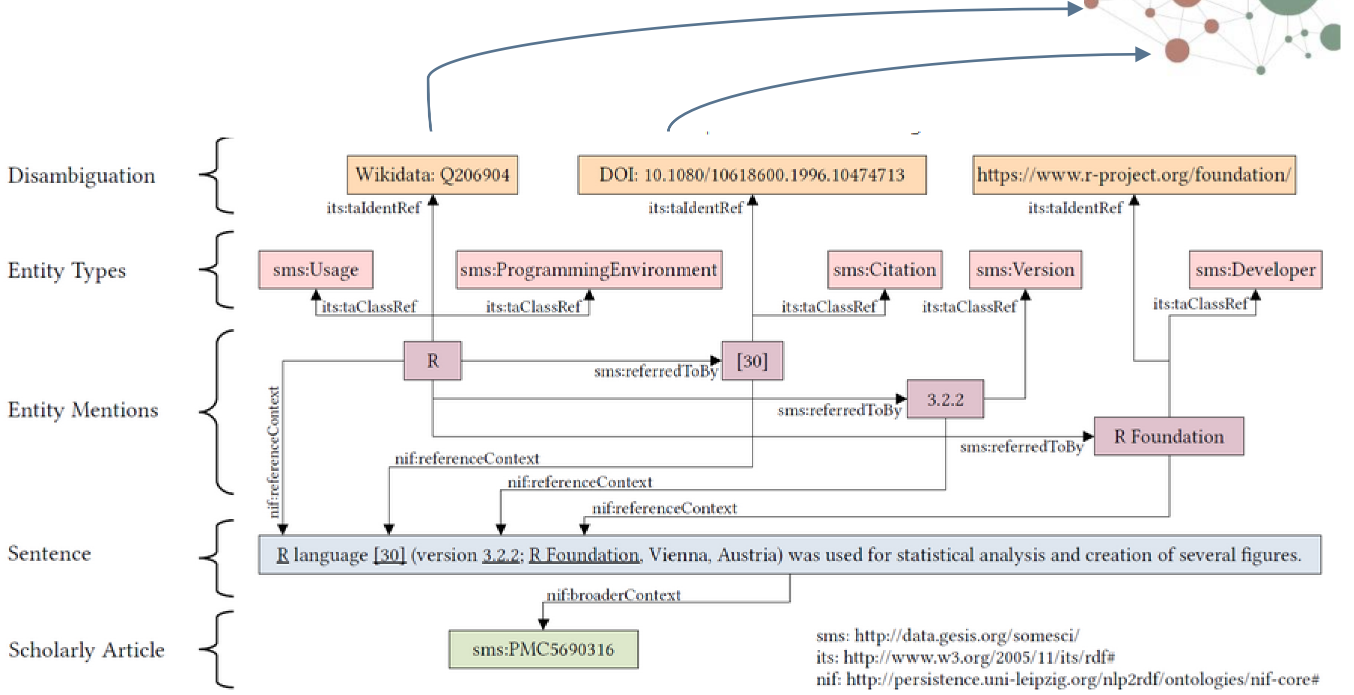
Rel. Publications

From publications to machine-interpretable metadata KGs

Disambiguation of dataset & software/script citations

https://data.gesis.org/softwarekg
https://data.gesis.org/somesci

- Manual annotation (“SomeSci”)
- Training deep learning-based model for extraction software & data references in large-scale data (3.5 M publications)
- Data lifting into KG (“SoftwareKG”)
- 300+ M triples / statements
- Search across data/software/publications (GESIS Search)



Schindler, D., Bensmann, F., Dietze, S., Krüger, F., SoMeSci—A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles, (CIKM2021), ACM 2021

From publications to machine-interpretable metadata KGs

Understanding scientific software/data usage

https://data.gesis.org/softwarekg

(Schindler et al., CIKM2021)

- Understanding SW usage, citation habits and their evolution across disciplines
- Rise of data science = rise of software usage

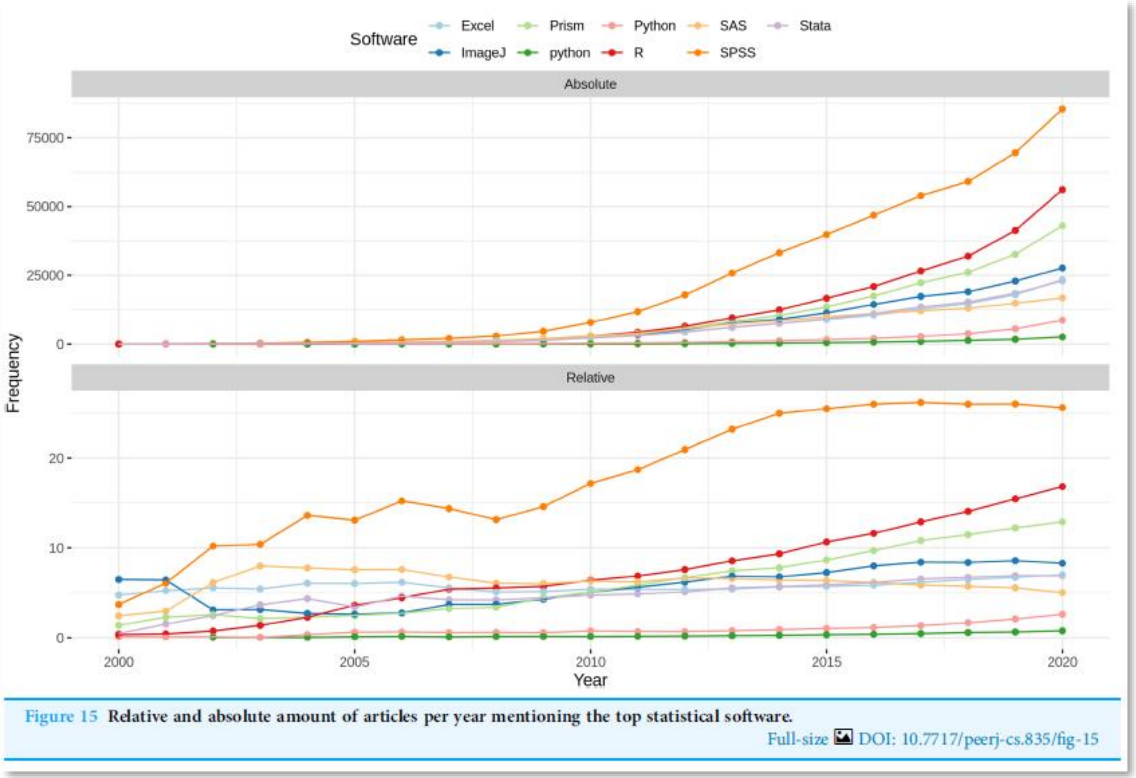


Figure 15 **Relative and absolute amount of articles per year mentioning the top statistical software.**

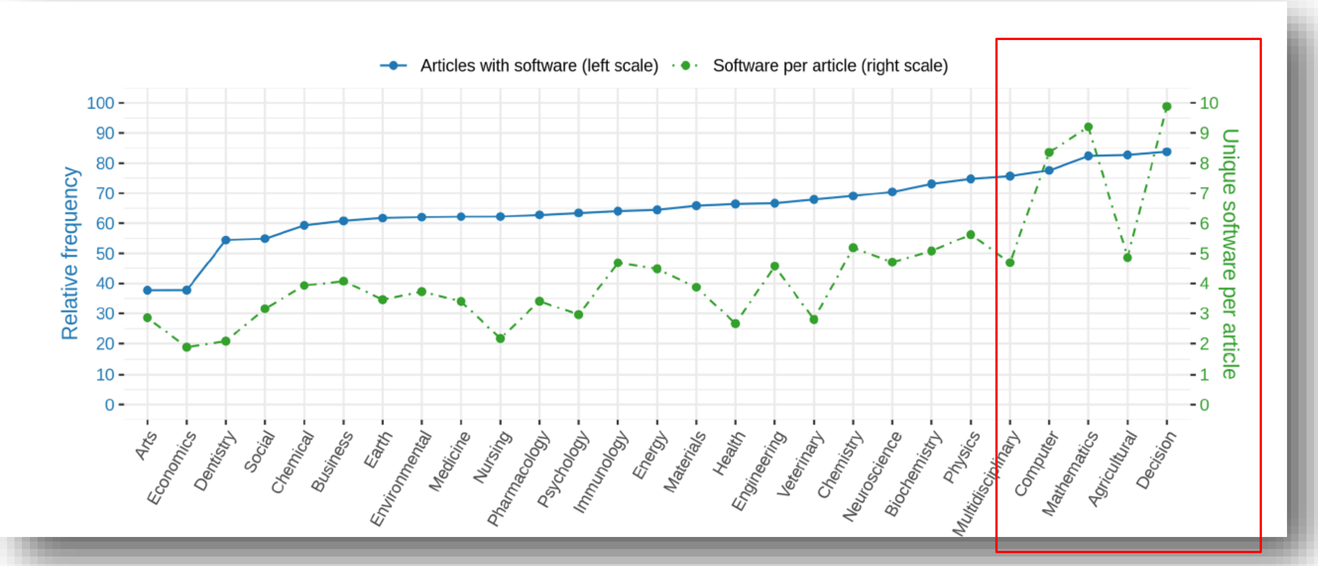
Full-size DOI: 10.7717/peerj-cs.835/fig-15

Schindler D, Bensmann F, Dietze S, Krüger F., The role of software in science: a knowledge graph-based analysis of software mentions in PubMed Central. **PeerJ Computer Science 8:e835**

From publications to machine-interpretable metadata KGs

Understanding scientific software/data usage

https://data.gesis.org/softwarekg



- Top adopters of data science/AI/software…

Schindler D, Bensmann F, Dietze S, Krüger F., The role of software in science: a knowledge graph-based analysis of software mentions in PubMed Central. **PeerJ Computer Science 8:e835**

From publications to machine-interpretable metadata KGs

Understanding scientific software/data usage

https://data.gesis.org/softwarekg

- Top adopters of data science/AI/software…
- …follow the worst citation habits

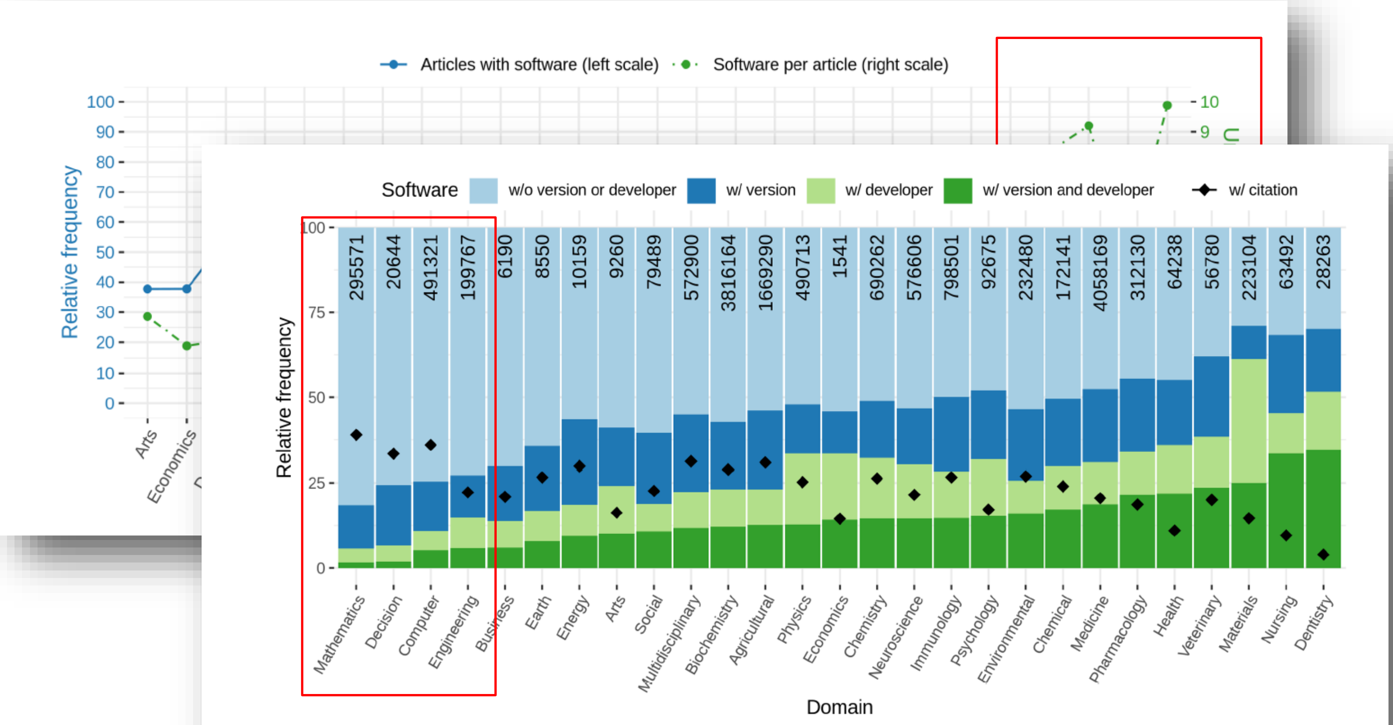


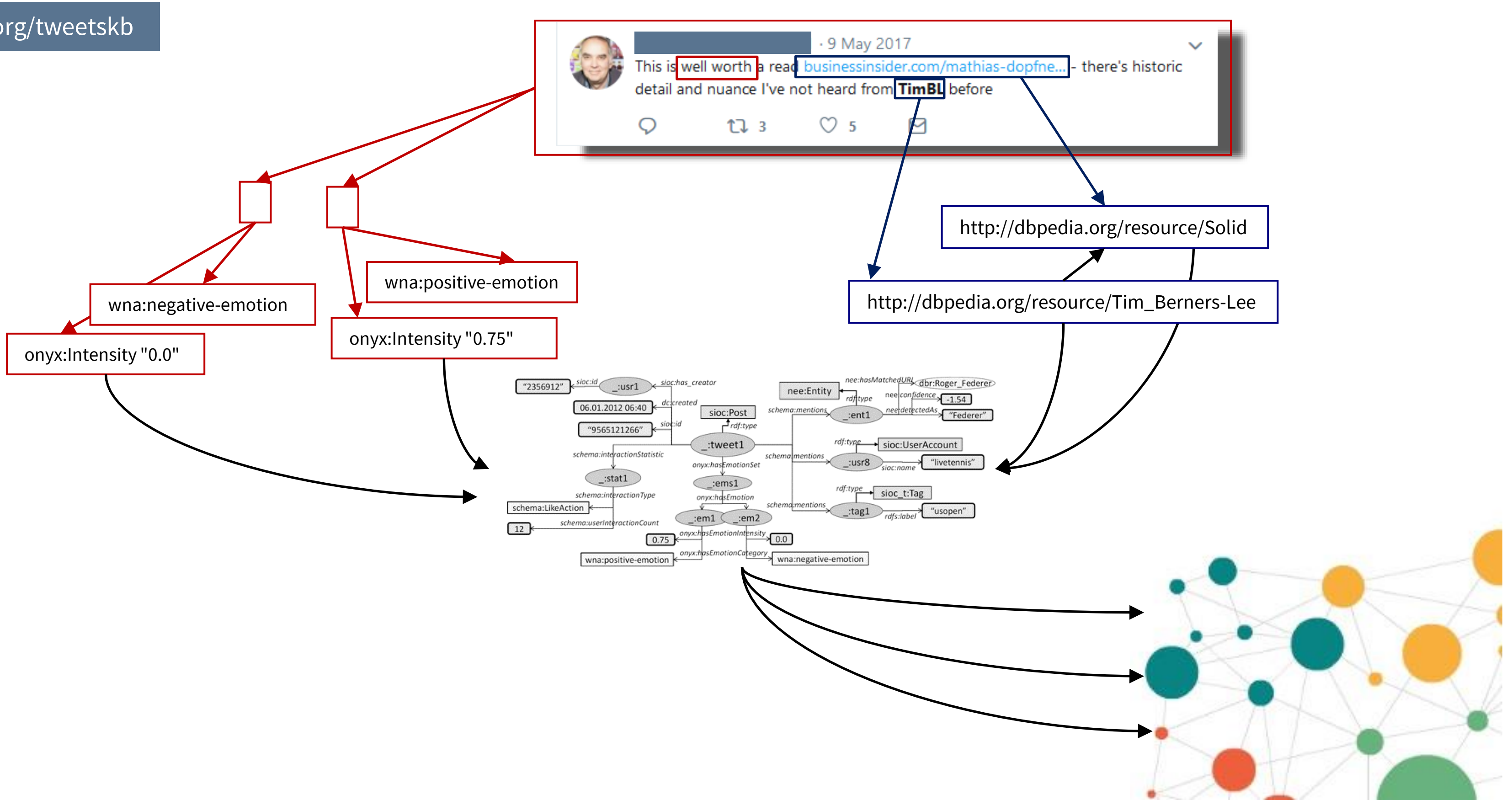
Figure 12 Distribution of software completeness per research domain. The numbers at the top of the bars represent the absolute numbers of software considered per domain. Please note that articles may belong to multiple categories. Full-size DOI: 10.7717/peerj-cs.835/fig-12

Schindler D, Bensmann F, Dietze S, Krüger F., The role of
PeerJ Computer Science 8:e835

From social media to machine-interpretable research data KGs

Building a public research knowledge graph from Twitter data

https://data.gesis.org/tweetskb



[redacted] · 9 May 2017

This is well worth a read businessinsider.com/mathias-dopfne... - there's historic detail and nuance I've not heard from TimBL before

3 5

wna:negative-emotion

onyx:Intensity "0.0"

wna:positive-emotion

onyx:Intensity "0.75"

http://dbpedia.org/resource/Solid

http://dbpedia.org/resource/Tim_Berners-Lee

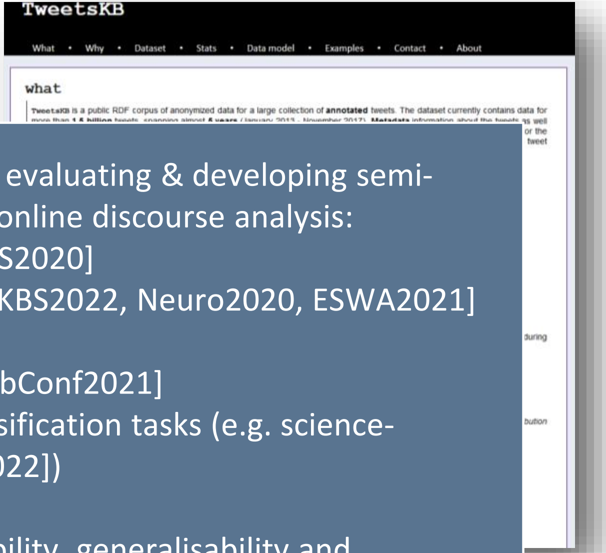
From social media to machine-interpretable research data KGs

TweetsKB – a large-scale research KG of societal opinions

https://data.gesis.org/tweetskb

- **Harvesting & archiving of** 10 Billion tweets (permanent collection from Twitter 1% sample since 2013)

- **Information extraction pipeline** to build a KG of entities, interactions & sentiments (distributed Map/Reduce batch processing)
  - Entity linking with knowledge graph/DBpedia (*“president”/“potus”/”trump”* => *dbp:DonaldTrump)*
  - Sentiment analysis/annotation
  - Geotagging
  - Lifting into knowledge graph schema



TweetsKB

What • Why • Dataset • Stats • Data model • Examples • Contact • About

what

TweetsKB is a public RDF corpus of anonymized data for a large collection of **annotated** tweets. The dataset currently contains data for more than **1.5 billion** tweets, spanning almost **5 years** (January 2013 - November 2017). **Metadata** information about the tweets as well … or the … tweet … during … bution

KTS research focused on evaluating & developing semi-supervised methods for online discourse analysis:

- Stance detection [IJIS2020]
- Sentiment analysis [KBS2022, Neuro2020, ESWA2021]
- Entity linking
- Georeferencing [WebConf2021]
- More fine-grain classification tasks (e.g. science-relatedness [CIKM2022])

But: focus here on scalability, generalisability and robustness towards evolving data/vocabulary => unsupervised approaches

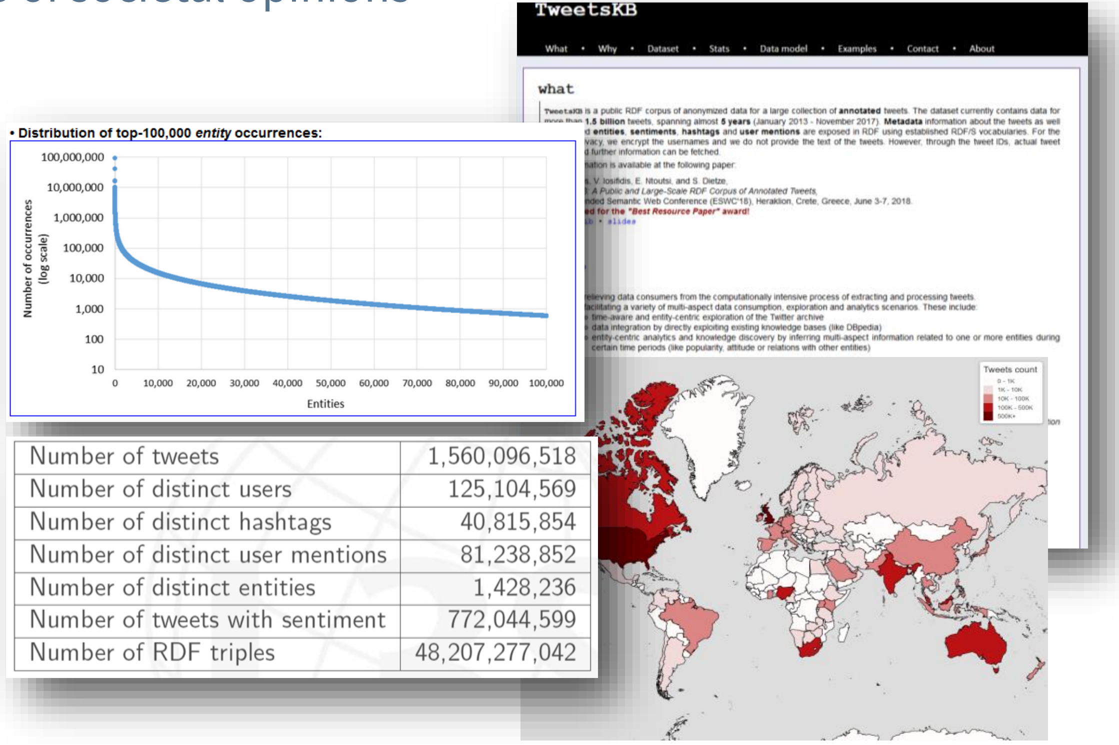
Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., Dietze, S., TweetsCOV19 – A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic, **CIKM2020**

From social media to machine-interpretable research data KGs

TweetsKB – a large-scale research KG of societal opinions

https://data.gesis.org/tweetskb

- **Harvesting & archiving of** 10 Billion tweets (permanent collection from Twitter 1% sample since 2013)
- **Information extraction pipeline** to build a KG of entities, interactions & sentiments (distributed Map/Reduce batch processing)
  - Entity linking with knowledge graph/DBpedia (*“president”/“potus”/”trump” => dbp:DonaldTrump)*
  - Sentiment analysis/annotation
  - Geotagging
  - Lifting into knowledge graph schema
- **Public, privacy-aware, large-scale research corpus of public opinions and their evolution => interdisciplinary research**



• Distribution of top-100,000 *entity* occurrences:

Number of tweets	1,560,096,518
Number of distinct users	125,104,569
Number of distinct hashtags	40,815,854
Number of distinct user mentions	81,238,852
Number of distinct entities	1,428,236
Number of tweets with sentiment	772,044,599
Number of RDF triples	48,207,277,042

TweetsKB

What • Why • Dataset • Stats • Data model • Examples • Contact • About

what

TweetsKB is a public RDF corpus of anonymized data for a large collection of **annotated** tweets. The dataset currently contains data for more than **1.5 billion** tweets, spanning almost **5 years** (January 2013 - November 2017). **Metadata** information about the tweets as well d **entities**, **sentiments**, **hashtags** and **user mentions** are exposed in RDF using established RDF/S vocabularies. For the vacy, we encrypt the usernames and we do not provide the text of the tweets. However, through the tweet IDs, actual tweet d further information can be fetched.

nation is available at the following paper:

s, V. Iosifidis, E. Ntoutsi, and S. Dietze,
: *A Public and Large-Scale RDF Corpus of Annotated Tweets*,
nded Semantic Web Conference (ESWC'18), Heraklion, Crete, Greece, June 3-7, 2018.
ed for the *"Best Resource Paper"* award!
ib • slides

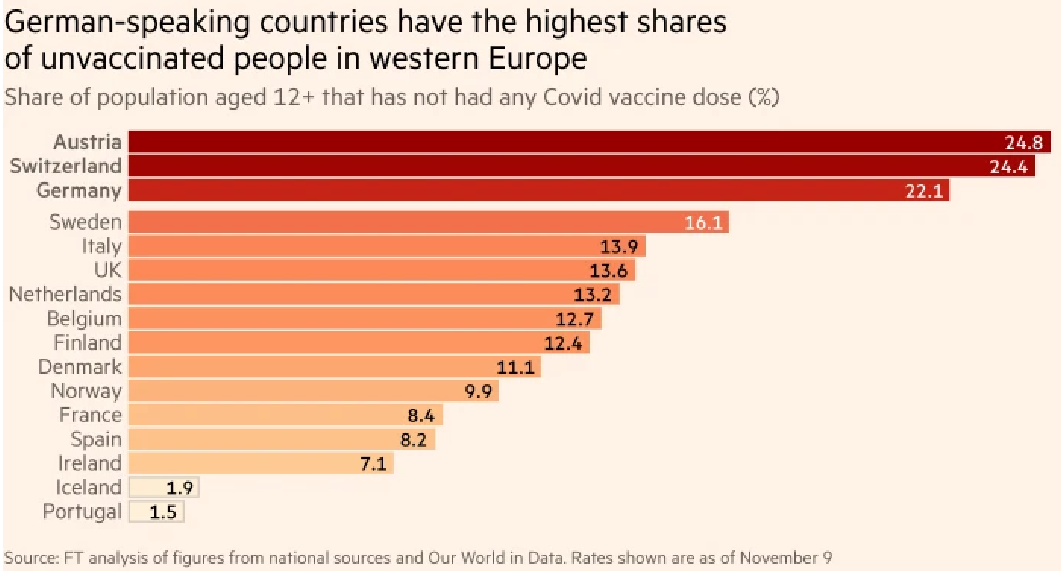
relieving data consumers from the computationally intensive process of extracting and processing tweets.
facilitating a variety of multi-aspect data consumption, exploration and analytics scenarios. These include:
- time-aware and entity-centric exploration of the Twitter archive
- data integration by directly exploiting existing knowledge bases (like DBpedia)
- entity-centric analytics and knowledge discovery by inferring multi-aspect information related to one or more entities during certain time periods (like popularity, attitude or relations with other entities)

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., Dietze, S., TweetsCOV19 – A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic, **CIKM2020**

RKG-based social science research using TweetsKB

Investigating Vaccine Hesitancy in DACH countries

https://dd4p.gesis.org



German-speaking countries have the highest shares of unvaccinated people in western Europe

Share of population aged 12+ that has not had any Covid vaccine dose (%)

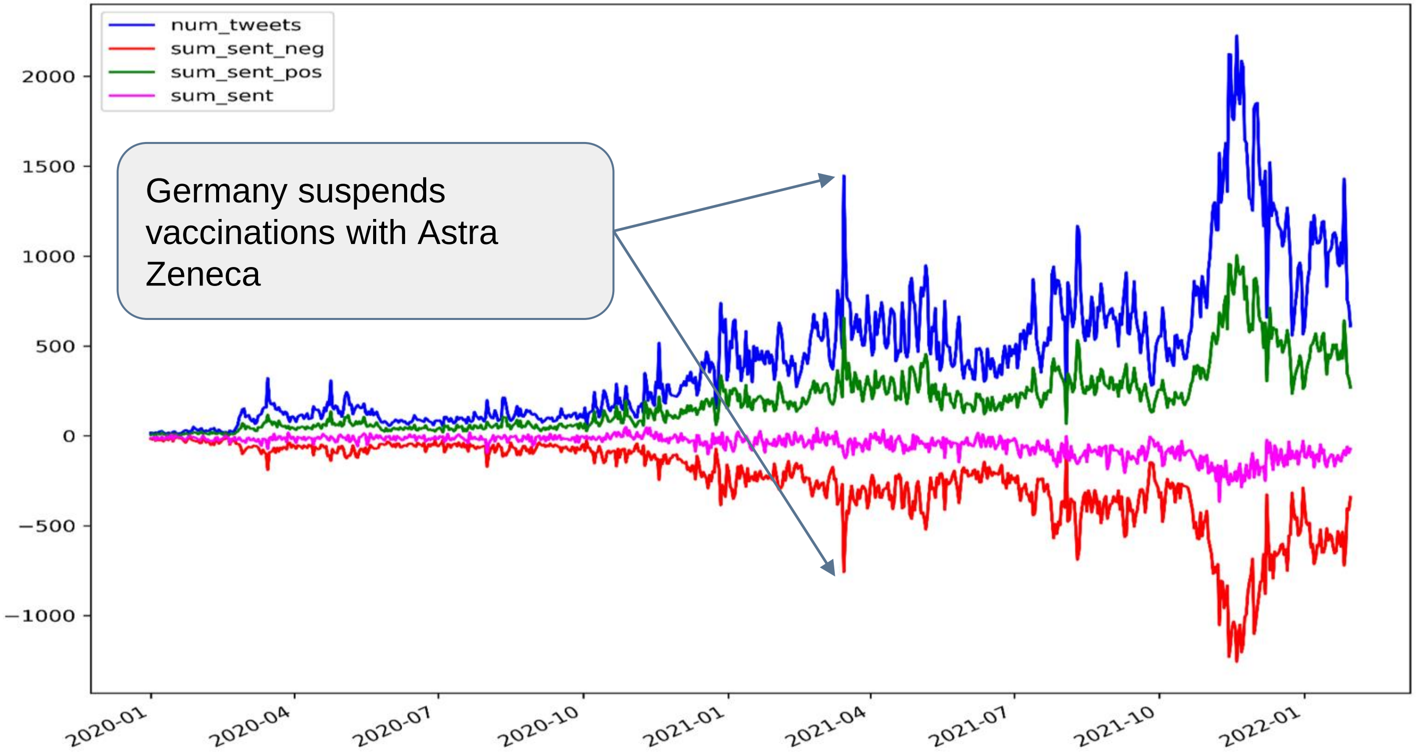
Source: FT analysis of figures from national sources and Our World in Data. Rates shown are as of November 9

RKG-based social science research using TweetsKB

Investigating Vaccine Hesitancy in DACH countries

https://dd4p.gesis.org

Twitter discourse zu “Impfbereitschaft”



Germany suspends vaccinations with Astra Zeneca

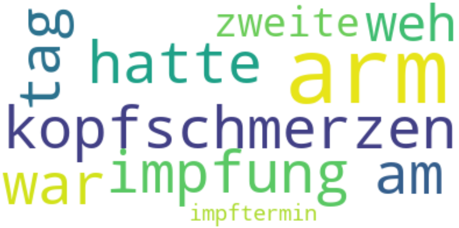
RKG-based discourse analysis using TweetsKB

Vaccine Hesitancy– key topics in “safety” category

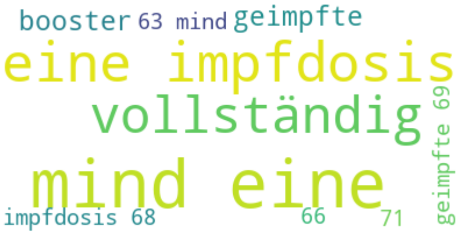
DD4P DiscourseData4Policy

https://dd4p.gesis.org

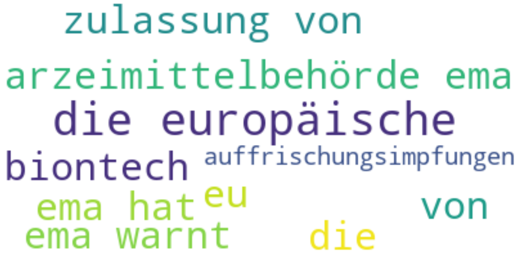
„Nebenwirkungen“



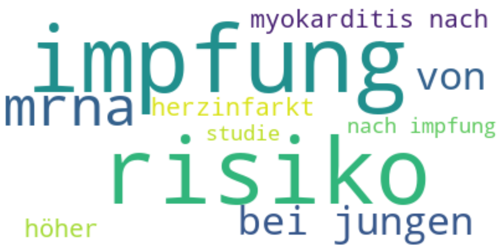
„Alter“



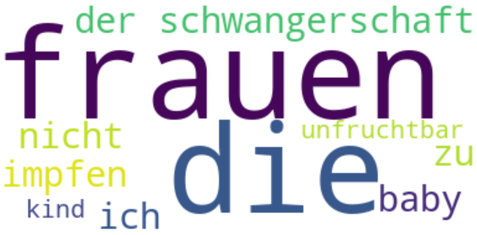
„Zulassung“



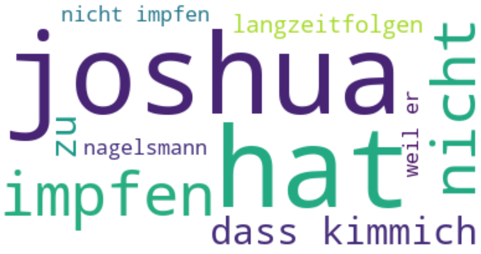
„Herzinfarkt“



„Schwangerschaft“



„Kimmich“



How about mentions of science resources on the Web?

Example: Twitter

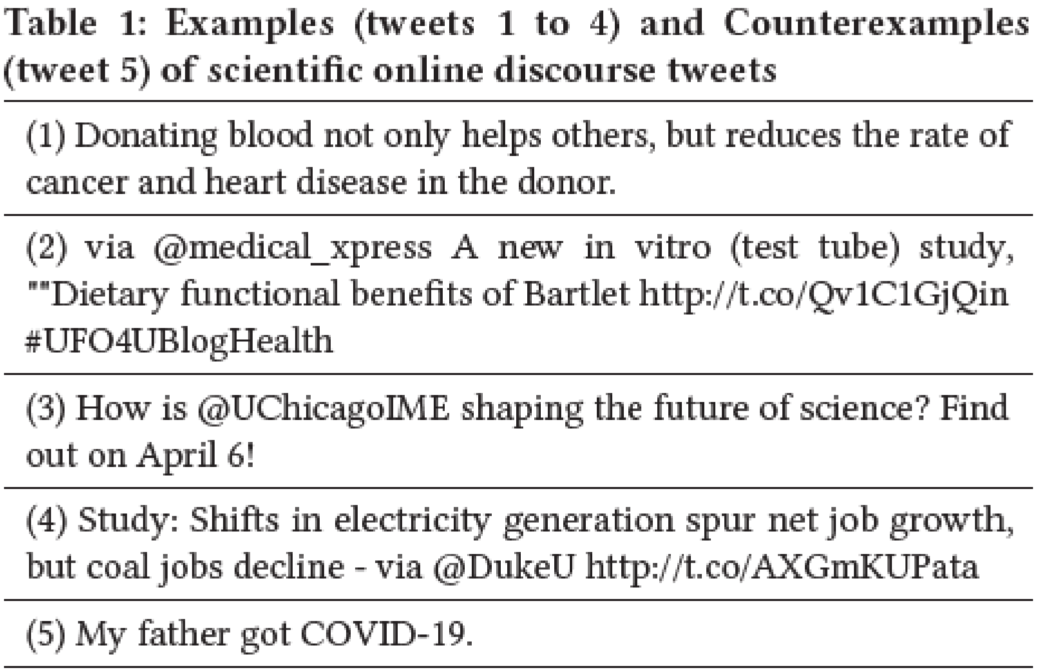


Table 1: Examples (tweets 1 to 4) and Counterexamples (tweet 5) of scientific online discourse tweets

Science claim	(1) Donating blood not only helps others, but reduces the rate of cancer and heart disease in the donor.
Science reference	(2) via @medical_xpress A new in vitro (test tube) study, ""Dietary functional benefits of Bartlet http://t.co/Qv1C1GjQin #UFO4UBlogHealth
Science relevance	(3) How is @UChicagoIME shaping the future of science? Find out on April 6!
Science reference	(4) Study: Shifts in electricity generation spur net job growth, but coal jobs decline - via @DukeU http://t.co/AXGmKUPata
No science	(5) My father got COVID-19.

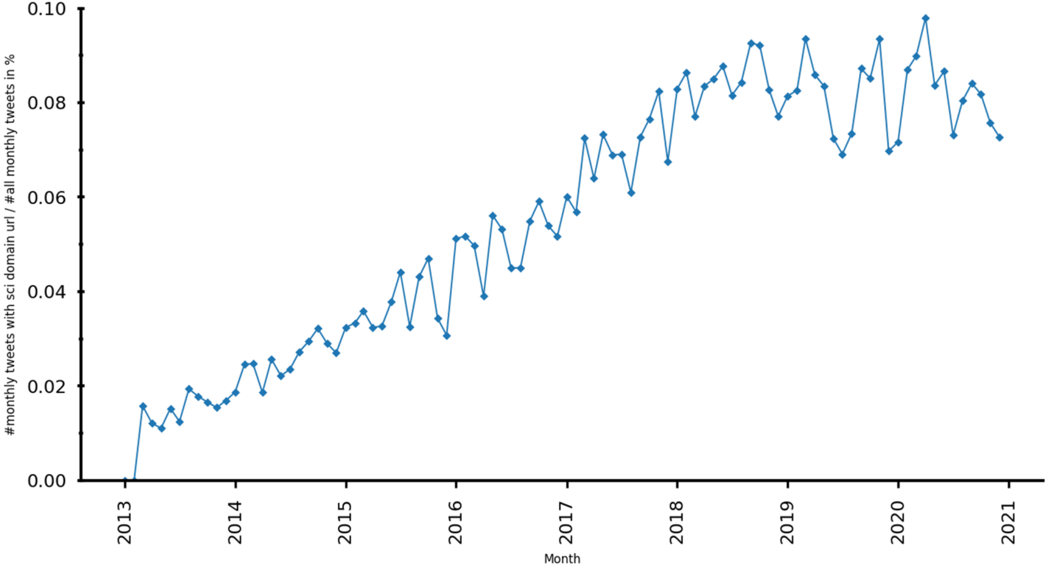
Hafid, S., Schellhammer, S., Bringay, S., Todorov, K., Dietze, S., "SciTweets - A Dataset and Annotation Framework for Detecting Scientific Online Discourse", **CIKM2022**

How about mentions of science resources on the Web?

Example: Twitter

https://ai4sci-project.org/

- Percentage of tweets containing links to scientific articles (journals, publishers, science blogs etc)
- Uses list of > 30 K science web domains
- Data source: TweetsKB (https://data.gesis.org/tweetskb/), > 10 bn tweets archived since 2013

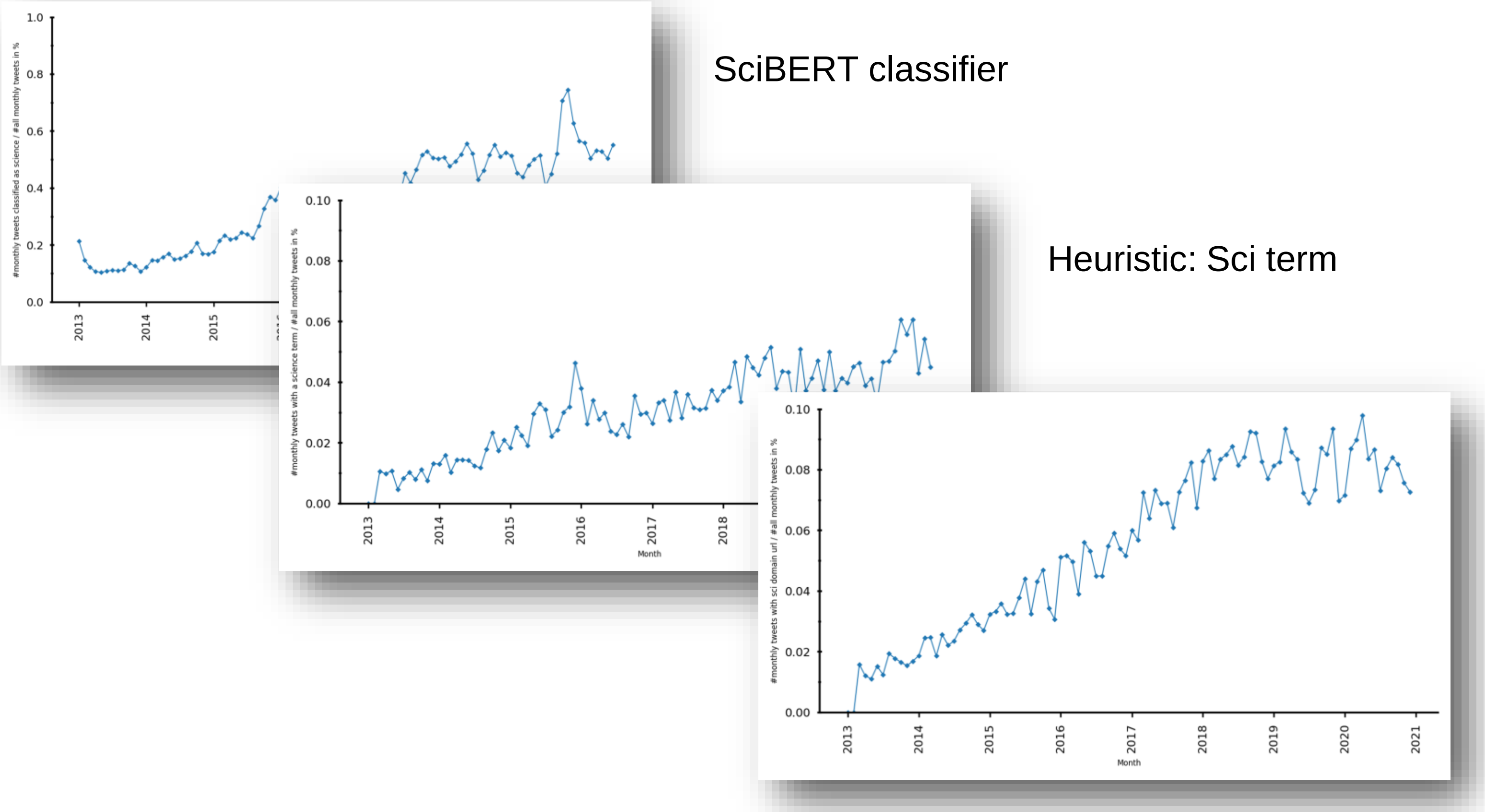


- AI4Sci project: understanding and classification of science discourse online (news, social Web)

How about mentions of science resources on the Web?

Example: Twitter

https://ai4sci-project.org/



SciBERT classifier

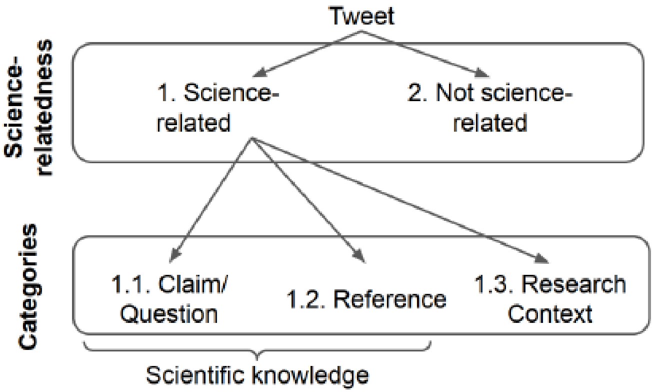
Heuristic: Sci term

Sci subdomain

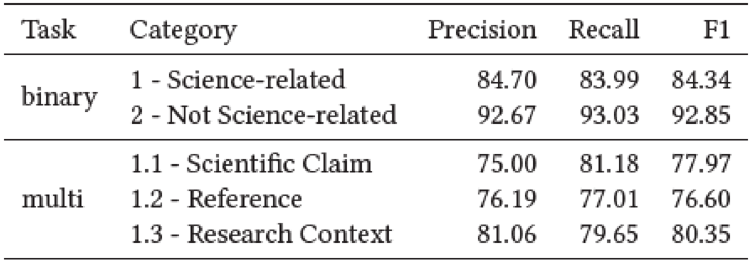
SciTweets dataset & classifier

https://ai4sci-project.org/

- Ground truth dataset, heuristics-based sampling strategy and annotation framework for testing classification models
- 1261 expert-labeled tweets across all classes/labels



- Baseline classifiers based on SciBERT transformer model (fine-tuned/tested on SciTweets)
- Ongoing: analysis of large-scale science discourse and its evolution



Task	Category	Precision	Recall	F1
binary	1 - Science-related	84.70	83.99	84.34
	2 - Not Science-related	92.67	93.03	92.85
multi	1.1 - Scientific Claim	75.00	81.18	77.97
	1.2 - Reference	76.19	77.01	76.60
	1.3 - Research Context	81.06	79.65	80.35

Hafid, S., Schellhammer, S., Bringay, S., Todorov, K., Dietze, S., *SciTweets - A Dataset and Annotation Framework for Detecting Scientific Online Discourse*, **CIKM2022**

Summary: Research KGs @ GESIS

Tools for constructing scholarly knowledge graphs

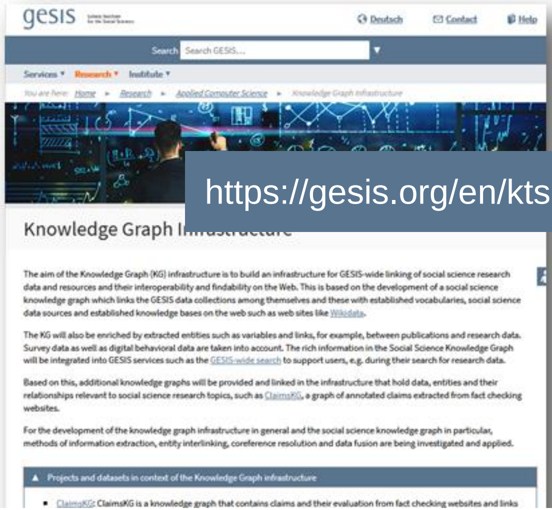
- NLP and deep learning-powered methods for extracting large-scale KGs about methods, claims, data, software involved in the scientific process

Large-scale scholarly KGs, e.g.

- KGs about scholarly use of software & research data
(e.g. **SoftwareKG**: 1.8 M disambiguated software mentions extracted from 3 M publications, https://data.gesis.org/softwarekg/)
- Web mined KGs of social science research data, e.g. public opinions, claims and attitudes expressed on social media
(e.g. **TweetsKB**: > 10 Bn semantically annotated tweets, sentiments, https://data.gesis.org/tweetskb)

Semantic Search powered by KGs and related tools

- RKG-powered search across scholarly publications, datasets, methods and their relations (e.g. **GESIS Search**, https://search.gesis.org)


gesis
Deutsch
Contact
Help
Search
Search GESIS...
Services ▾
Research ▾
Institute ▾
You are here: Home ▸ Research ▸ Applied Computer Science ▸ Knowledge Graph Infrastructure

https://gesis.org/en/kts

Knowledge Graph I

The aim of the Knowledge Graph (KG) infrastructure is to build an infrastructure for GESIS-wide linking of social science research data and resources and their interoperability and findability on the Web. This is based on the development of a social science knowledge graph which links the GESIS data collections among themselves and these with established vocabularies, social science data sources and established knowledge bases on the web such as web sites like Wikidata.

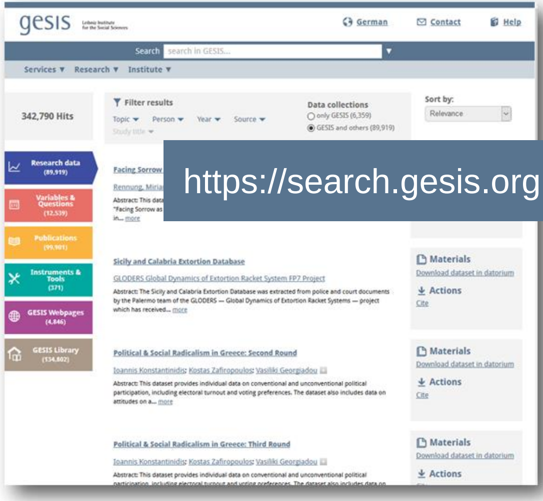
The KG will also be enriched by extracted entities such as variables and links, for example, between publications and research data. Survey data as well as digital behavioral data are taken into account. The rich information in the Social Science Knowledge Graph will be integrated into GESIS services such as the GESIS-wide search to support users, e.g. during their search for research data.

Based on this, additional knowledge graphs will be provided and linked in the infrastructure that hold data, entities and their relationships relevant to social science research topics, such as ClaimsKG, a graph of annotated claims extracted from fact checking websites.

For the development of the knowledge graph infrastructure in general and the social science knowledge graph in particular, methods of information extraction, entity interlinking, coreference resolution and data fusion are being investigated and applied.

▲ Projects and datasets in context of the Knowledge Graph infrastructure

- ClaimsKG: ClaimsKG is a knowledge graph that contains claims and their evaluation from fact checking websites and links



342,790 Hits

Data collections
○ only GESIS (6,359)
● GESIS and others (89,919)

https://search.gesis.org

Facing Sorrow

Rennung, Miria

Abstract: This data
"Facing Sorrow as
in... more

Sicily and Calabria Extortion Database

GLODERS Global Dynamics of Extortion Racket System FP7 Project

Abstract: The Sicily and Calabria Extortion Database was extracted from police and court documents by the Palermo team of the GLODERS — Global Dynamics of Extortion Racket Systems — project which has received... more

Political & Social Radicalism in Greece: Second Round

Ioannis Konstantinidis; Kostas Zafiropoulos; Vasiliki Georgiadou

Abstract: This dataset provides individual data on conventional and unconventional political participation, including electoral turnout and voting preferences. The dataset also includes data on attitudes on a... more

Political & Social Radicalism in Greece: Third Round

Ioannis Konstantinidis; Kostas Zafiropoulos; Vasiliki Georgiadou

Abstract: This dataset provides individual data on conventional and unconventional political participation, including electoral turnout and voting preferences. The dataset also includes data on

Outlook: shared tasks on scholarly information extraction

Enganging with the community to advance progress in RKGs & scholarly IE

Creating large training/testing corpora and run shared tasks for

- Software / code detection and disambiguation
- Leaderboard extraction / task-dataset-metric detection (TDM)
- Dataset mention detection & disambiguation
- Machine learning model detection & disambiguation
- Research field classification

More to be announced soon.

@stefandietze
http://stefandietze.net

gesis Leibniz-Institut für Sozialwissenschaften

BERD@NFDI