



Knowledge graphs for sharing research data and information

Meet the Experts – GESIS online talks

*Knowledge technologies for the Social Science: Access to Social Science Data
and Services*

Benjamin Zapilko, Debanjali Biswas, 11.04.2024

Speakers



Benjamin Zapilko

- Senior researcher at GESIS in the department Knowledge Technologies for the Social Sciences
- PhD in computer science
- Research interests in knowledge graphs and knowledge engineering
- Contact: benjamin.zapilko@gesis.org



Debanjali Biswas

- Doctoral student at GESIS in the department Knowledge Technologies for the Social Sciences
- Master in computer science
- Research interests includes knowledge graphs, language models, natural language processing
- Contact: debanjali.biswas@gesis.org

Knowledge graphs for sharing research data and information

Agenda

- Knowledge Graphs for FAIR Research Data
- Research Knowledge Graphs of Scholarly Resource Metadata
- Research Knowledge Graphs of (Social Science) Research Data

Knowledge Graphs for FAIR Research Data

Sharing & Reuse of Research Data, Resources, Knowledge



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?

Sharing & Reuse of Research Data, Resources, Knowledge



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)
- **Relations and semantics** not explicit
- **Reproducibility crisis** in CS/DS/AI

FAIR principles

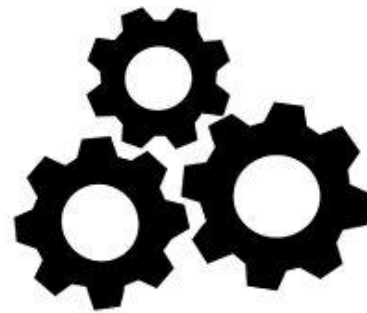
Findable



Accessible



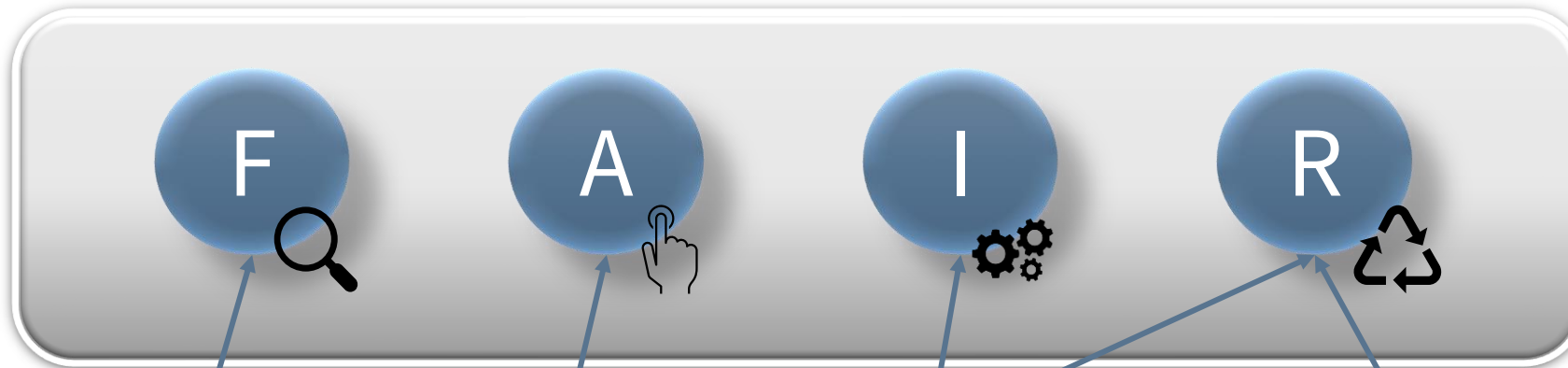
Interoperable



Reusable



Knowledge Graphs for FAIR Research Data



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

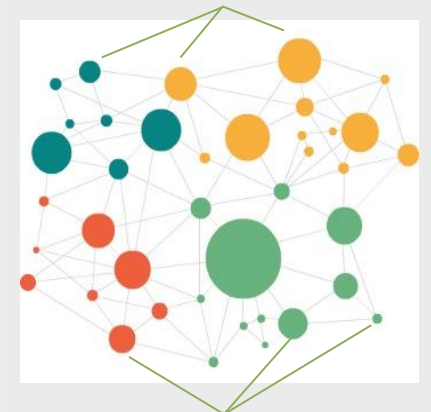
Use HTTP APIs for **data access**, reuse and linking

Using established W3C standards for data sharing (on the Web), e.g. RDF, JSON, shared vocabularies (e.g. schema.org, DCAT, DDI) to improve **data interoperability and reuse**

Making **links** between resources and concepts explicit & **machine-interpretable** (e.g. which publications cite what dataset? Which claim is supported/rejected by publication X/dataset Y?)

Resources

- Datasets
- Publications
- Code
- Software

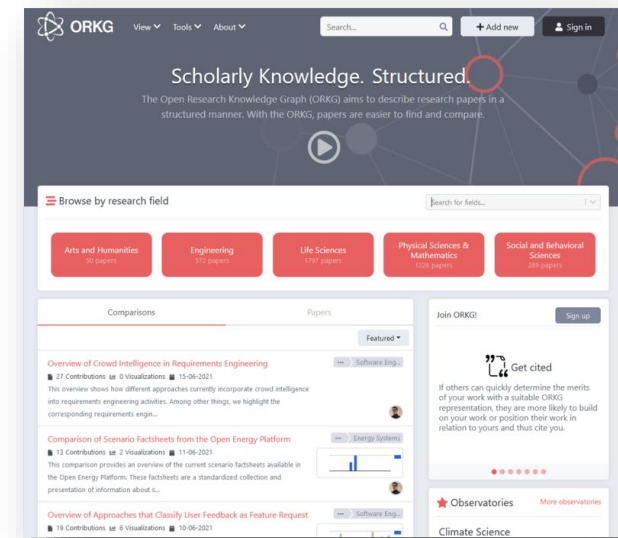


Concepts

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



Research Knowledge Graphs

- KGs of research resources (e.g., datasets, research papers, methods) and concepts (e.g., paper authors, organizations) within the scholarly domain
- Facilitating the way scientific outcomes are represented and utilized
- Facilitating the sharing of research data and information



Cologne

City in North Rhine-Westphalia

Weather

Wed 16° Thu 17° Fri 22°

weather.com


YouTube - TUI UK

Cologne TRAVEL GUIDE

4:15 25 Aug 2023

Wikipedia <https://en.wikipedia.org/wiki/Cologne>

Cologne
Cologne Cathedral · Eau de Cologne · Cologne (region) · Cologne (disambiguation)






People also ask

- Why is Cologne Germany famous?
- Is it Köln or Cologne?
- Why visit Cologne Germany?
- Is Cologne French or German?

Feedback

Things to do

 <p>Cologne Cathedral 4.8 ★ (69K) Cathedral Website Free</p>	 <p>Museum Ludwig 4.5 ★ (7.3K) Modern art museum Website From €11.00</p>	 <p>Phantasialand 4.5 ★ (88K) Tourist attraction Website From €36.00</p>
--	---	---

More things to do →

About

Cologne, a 2,000-year-old city spanning the Rhine River in western Germany, is the region's cultural hub. A landmark of High Gothic architecture set amid reconstructed old town, the twin-spired Cologne Cathedral is also known for its gilded medieval reliquary and sweeping river views. The adjacent Museum Ludwig showcases 20th-century art, including many masterpieces by Picasso, and the Romano-Germanic Museum houses Roman antiquities. — Google

Population: 1.086 million (2019) Eurostat

Admin. region: Cologne

Dialling codes: 0221, 02203 (Porz)

District: Urban district





Elevation: 37 m (121 ft)

Founded: 38 BCE

Postal codes: 50441–51149

Feedback

People also search for

 <p>Düsseldorf</p>	 <p>Cologne Cathedral</p>	 <p>Hamburg</p>	 <p>Munich</p>
---	--	--	---

See more →

KGs in practice

© Frank van Harmelen







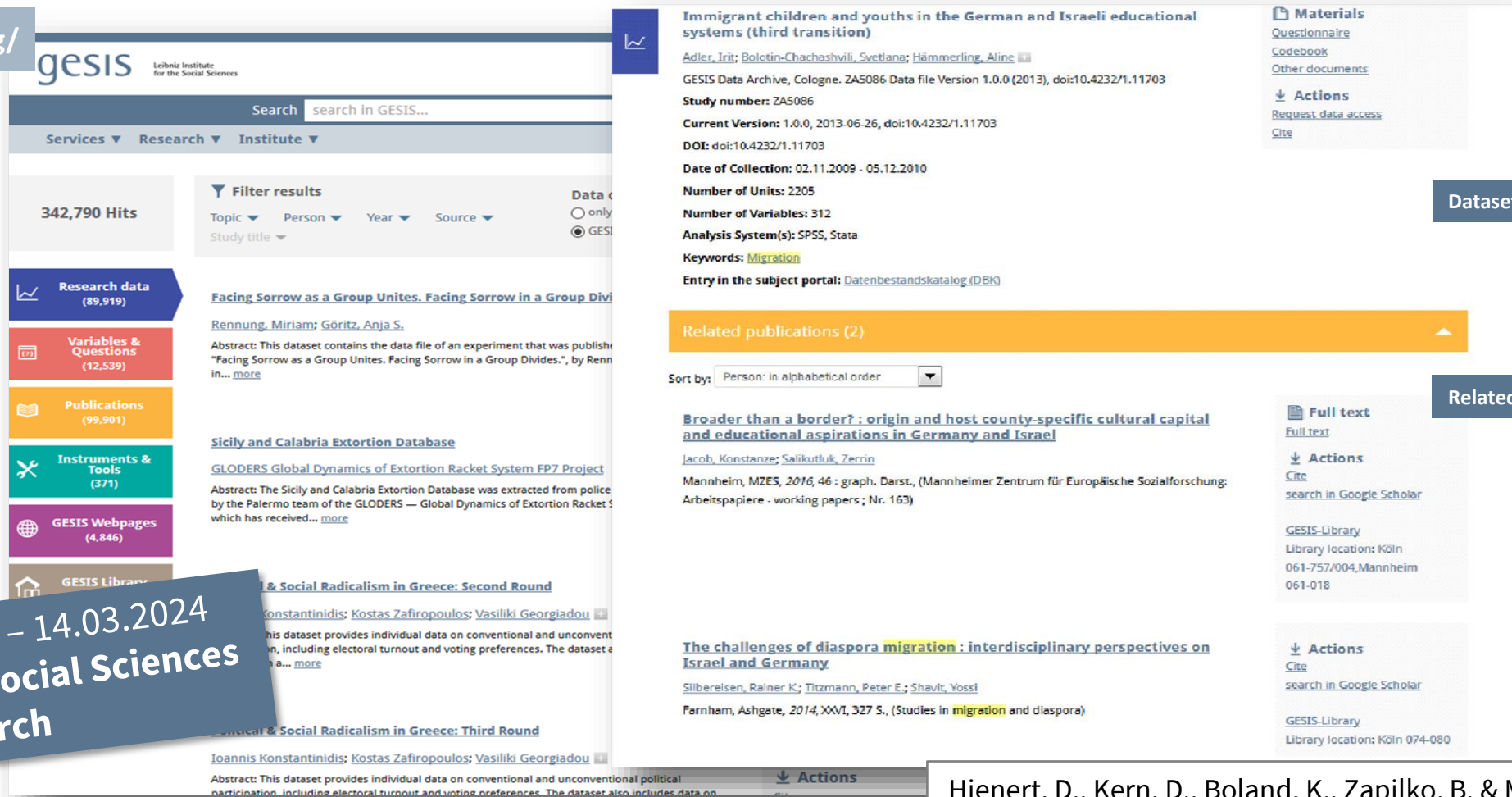




Research Knowledge Graphs of Scholarly Resource Metadata

GESIS KG: Integrated search @ GESIS

<https://search.gesis.org/>



The screenshot displays the GESIS search interface. On the left, a sidebar shows search statistics: 342,790 Hits, Research data (89,919), Variables & Questions (12,539), Publications (99,901), Instruments & Tools (371), GESIS Webpages (4,846), and GESIS Library. The main search area shows a search bar and filter options for Topic, Person, Year, and Source. The search results list several datasets, including "Facing Sorrow as a Group Unites. Facing Sorrow in a Group Divides" and "Sicily and Calabria Extortion Database". A detailed view of the "Immigrant children and youths in the German and Israeli educational systems (third transition)" dataset is shown on the right. This view includes the title, authors (Adler, Irit; Bolotin-Chachashvili, Svetlana; Hämmerling, Aline), version information, DOI, date of collection, number of units (2205), number of variables (312), analysis system (SPSS, Stata), and keywords (Migration). It also features a "Materials" section with links to Questionnaire, Codebook, and Other documents, and an "Actions" section with links for Request data access and Cite. A "Related publications (2)" section is also visible, listing works like "Broader than a border? : origin and host county-specific cultural capital and educational aspirations in Germany and Israel" and "The challenges of diaspora migration : interdisciplinary perspectives on Israel and Germany".

Dataset

Related Publications

Meet the Experts – 14.03.2024
Searching the Social Sciences
with GESIS Search

Hienert, D., Kern, D., Boland, K., Zapilko, B. & Mutschke, P. (2019) "A digital library for research data and related information in the social sciences." **JCDL '2019**

GESIS KG: Integrated search @ GESIS

- Derived from conducted online questionnaires and an observational study to reflect requirements of social scientists

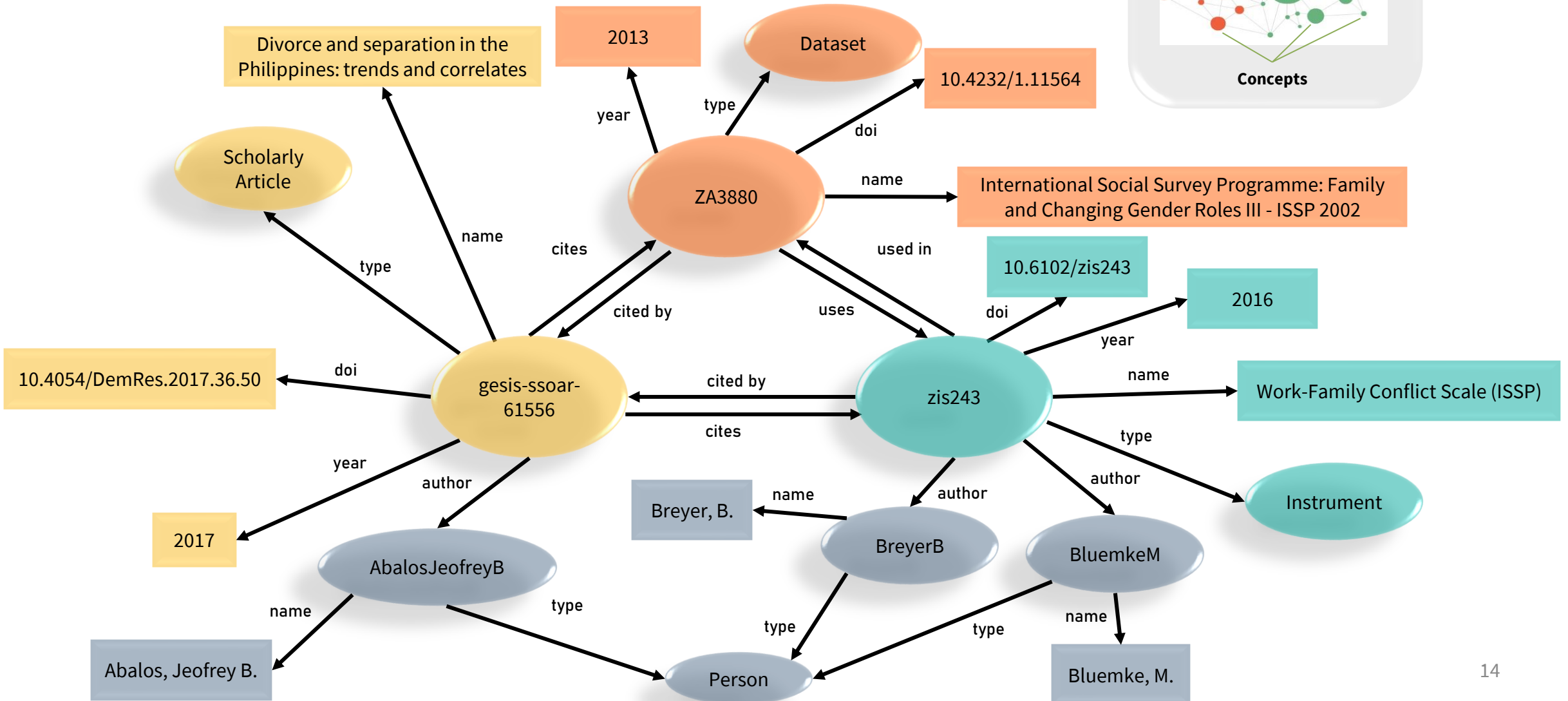
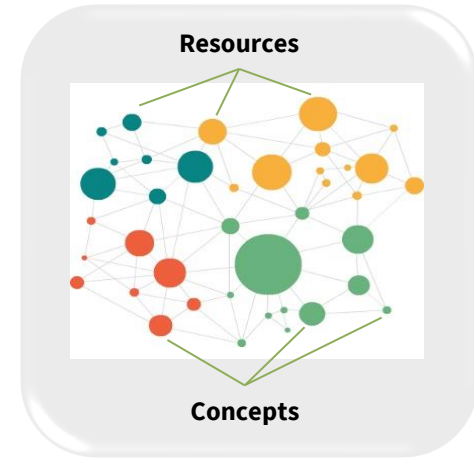
Key findings (KG related):

- Users are looking for research data mentioned in a paper
 - Dataset search suffers from missing interlinks
 - Literature search is an important part of dataset search
- Log file analysis and user study on quality of automatically generated data citations in preparation

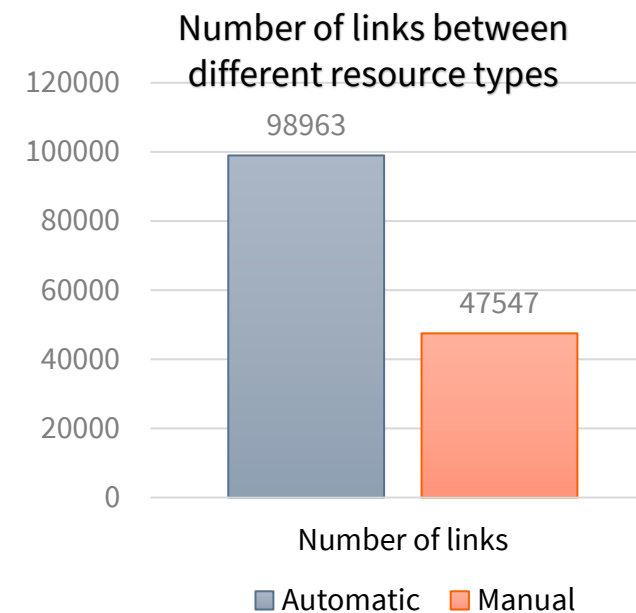
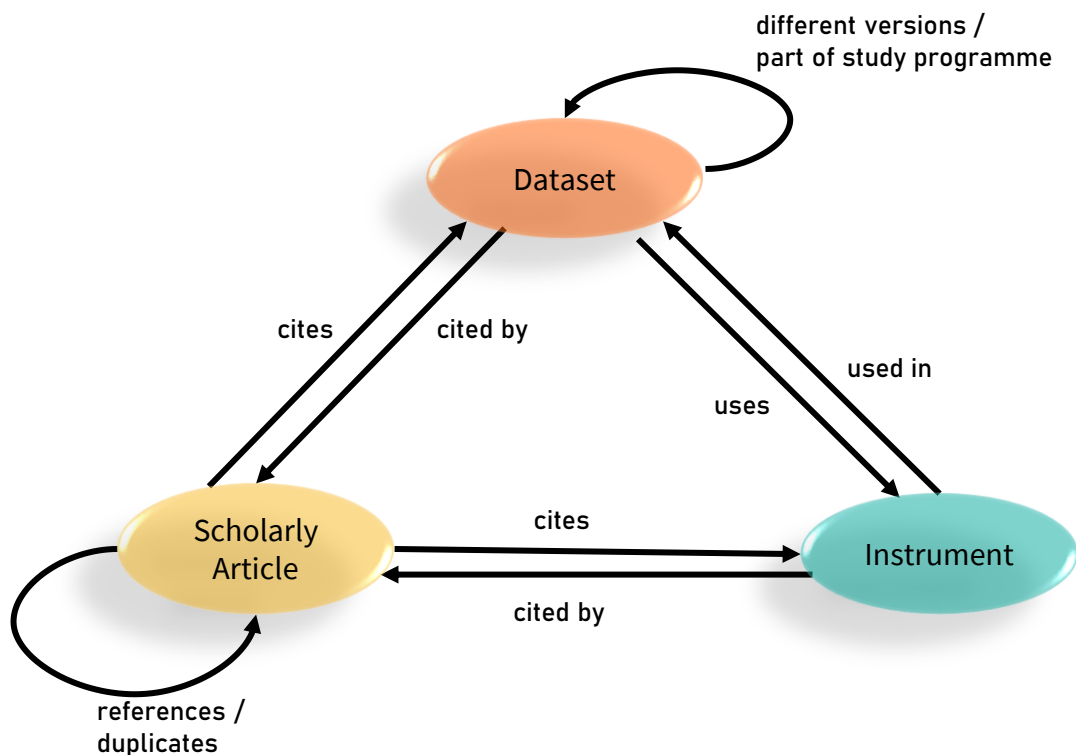
Meet the Experts – 18.01.2024
**Understanding Information-Seeking
Behavior of Social Scientists**

- Kern, D., & Hienert, D. (2018). Understanding the information needs of social scientists in Germany. *Proceedings of the Association for Information Science and Technology*, 55(1), 234-243.
- Krämer, T., Papenmeier, A., Carevic, Z., Kern, D., & Mathiak, B. (2021). Data-seeking behaviour in the social sciences. *International Journal on Digital Libraries*, 22, 175-195.

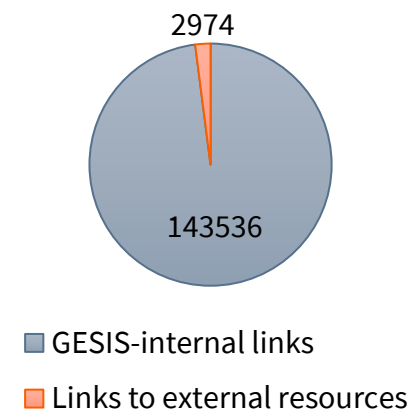
GESIS KG: Example



GESIS KG: Links between resources



Internal vs. External links

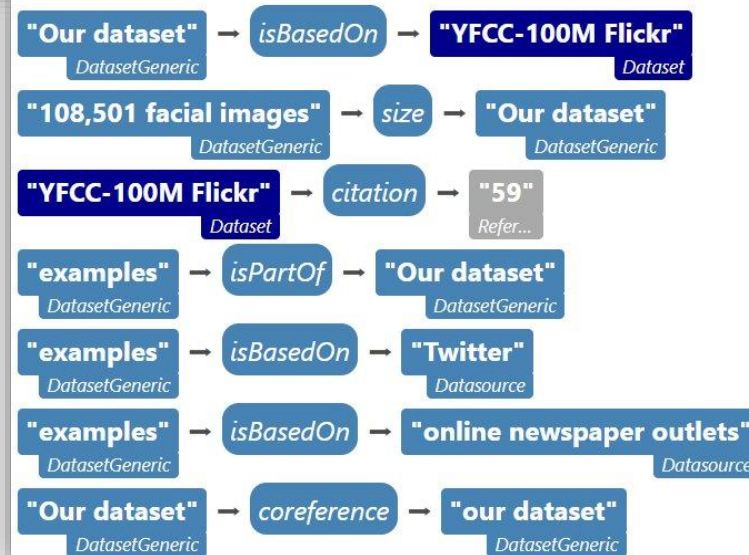


GSAP - Gesis Scholarly Annotation Project

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

- Scholarly Entity Extraction Focused on ML Models and Datasets
- For enriching Scholarly Resource Metadata KGs with relations between scholarly resources

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)

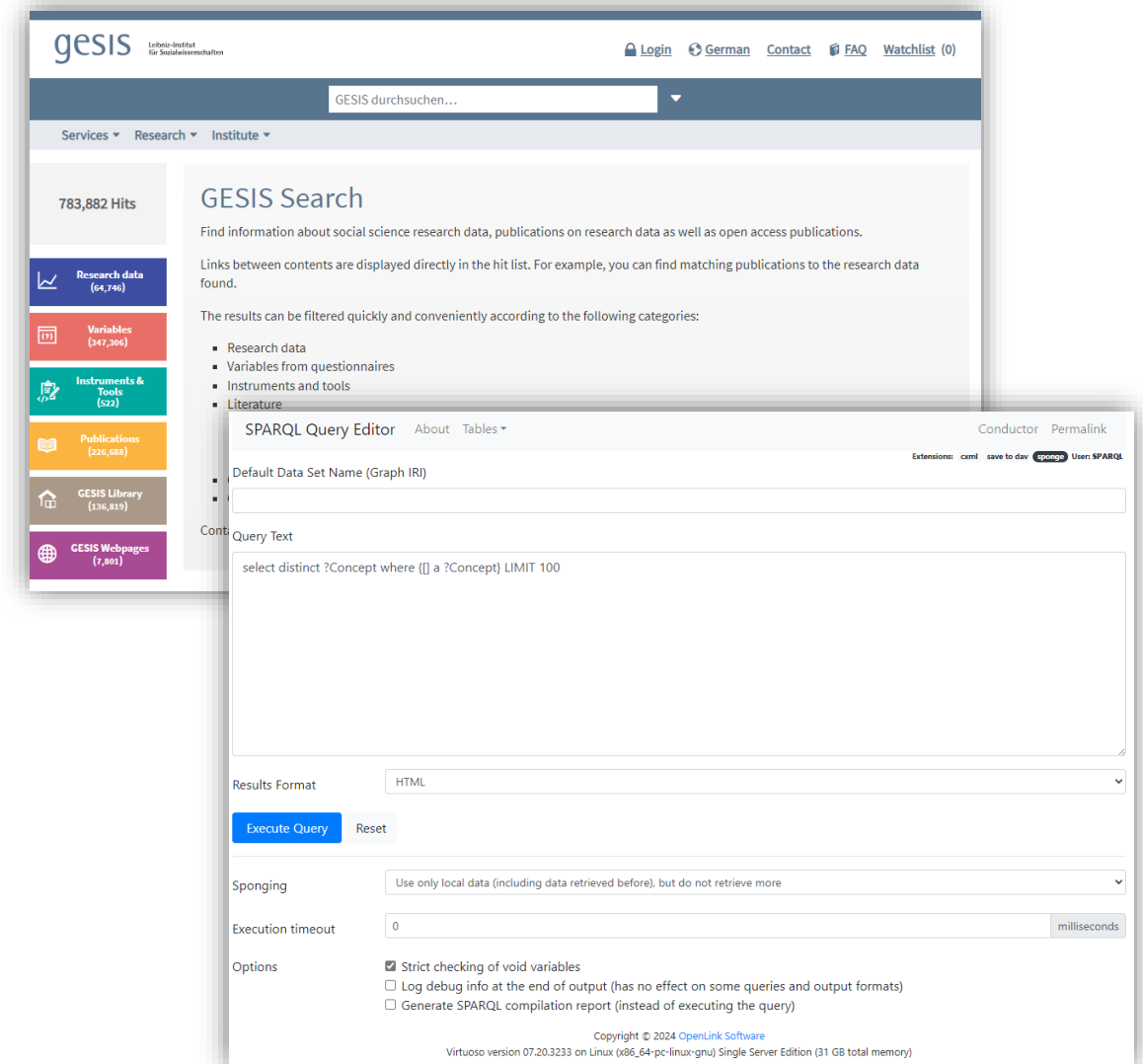


Otto, W., Zloch, M., Gan, L., Karmakar, S., & Dietze, S. (2023) GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. **EMNLP '2023**

Meet the Experts – 11.07.2024
Introduction to scholarly information extraction

GESIS KG: Availability

- Integrated as backend in the GESIS Search portal
- Public release of the GESIS KG planned for 2024, including
 - SPARQL endpoint
 - HTTP API
 - RDF dump for download



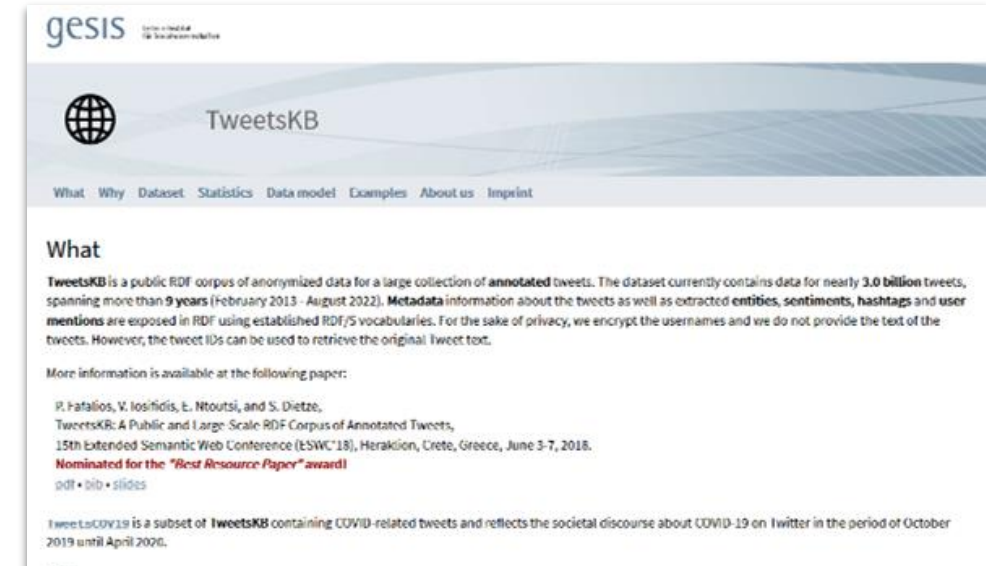
The image shows two overlapping screenshots from the GESIS website. The top screenshot is the GESIS Search portal, displaying 783,882 hits and a sidebar with categories: Research data (64,746), Variables (347,306), Instruments & Tools (522), Publications (226,688), GESIS Library (136,819), and GESIS Webpages (7,801). The bottom screenshot is the SPARQL Query Editor, showing a query text area with the query: `select distinct ?Concept where {[] a ?Concept} LIMIT 100`. The editor also includes options for Results Format (HTML), Sponging (Use only local data), Execution timeout (0 milliseconds), and various options like Strict checking of void variables.

Research Knowledge Graphs of (Social Science) Research Data

TweetsKB – a non-sensitive large-scale archive of societal discourse

<https://data.gesis.org/tweetskb>

- Subset of 3 billion prefiltered tweets (English, spam detection through pretrained classifier)
- Sharing of tweet metadata (time stamps, retweet counts etc), hash tags, user mentions and dedicated features that capture tweet semantics (no actual full texts/user IDs)
- Features include:
 - Disambiguated mentions of **entities**, linked to Wikipedia/DBpedia (“president”/“potus”/”trump” => dbp:DonaldTrump)
 - **Sentiment** scores (positive/negative emotions)
 - **Geotags** for a small subset



Feature	Total	Unique	% with >= 1 feature
Hashtags:	1,161,839,471	68,832,205	0.19
Mentions:	1,840,456,543	149,277,474	0.38
Entities:	2,563,433,997	2,265,201	0.56
Sentiment:	1,265,974,641	-	0.5

P. Fafalios, V. Iosifidis, E. Ntoutsis, and S. Dietze, *TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets*, **ESWC'18**.

RKG-based social science research using TweetsKB

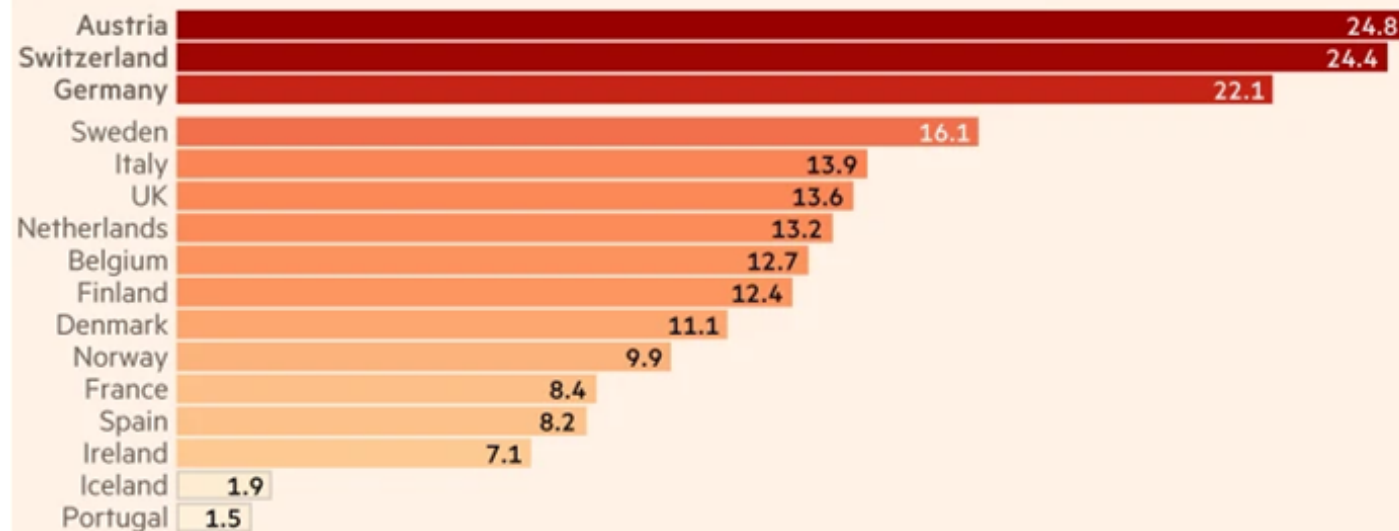
Investigating Vaccine Hesitancy in DACH countries



<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



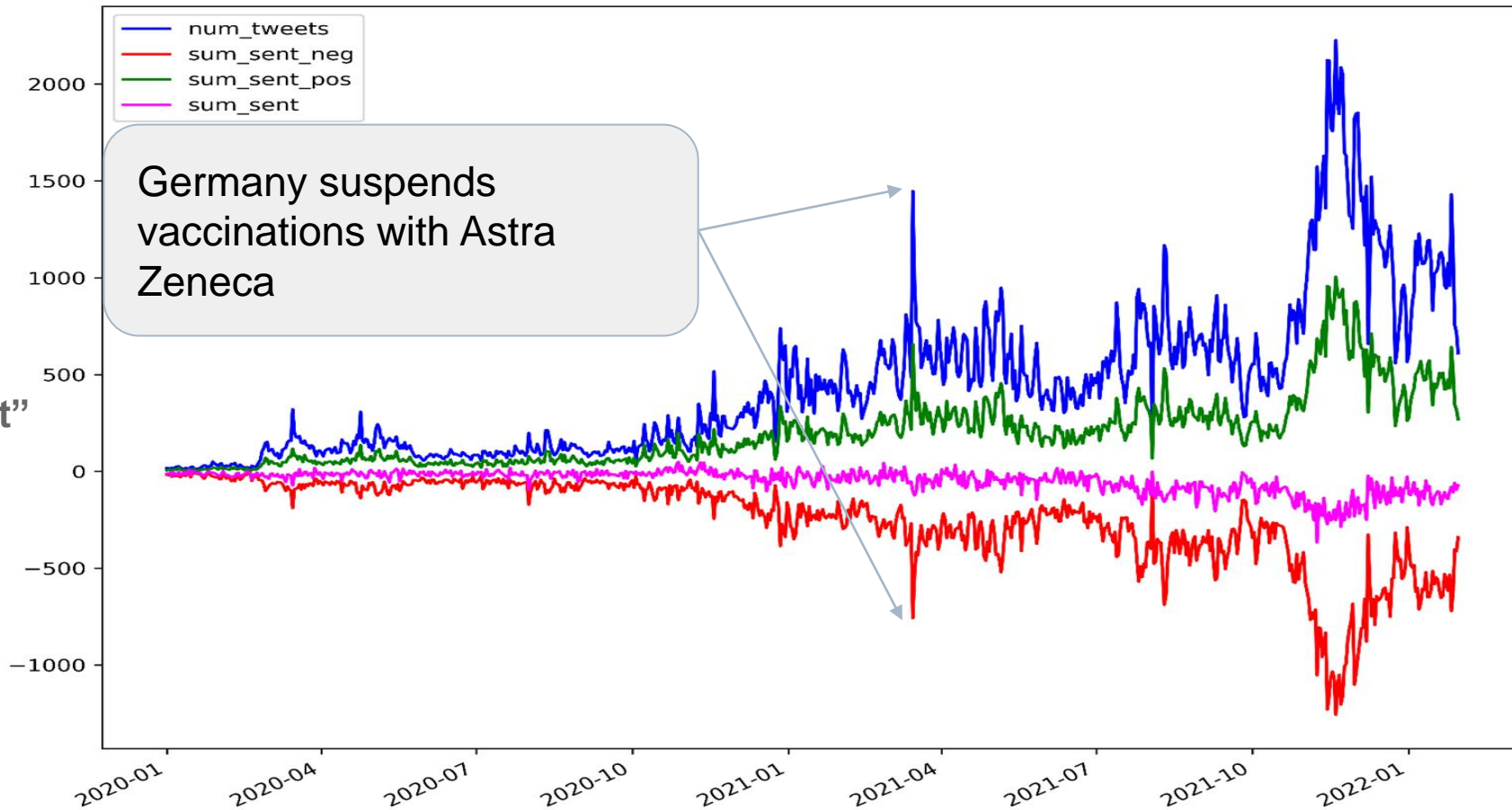
RKG-based social science research using TweetsKB

Investigating Vaccine Hesitancy in DACH countries



<https://dd4p.gesis.org>

Twitter discourse to “Impfbereitschaft”

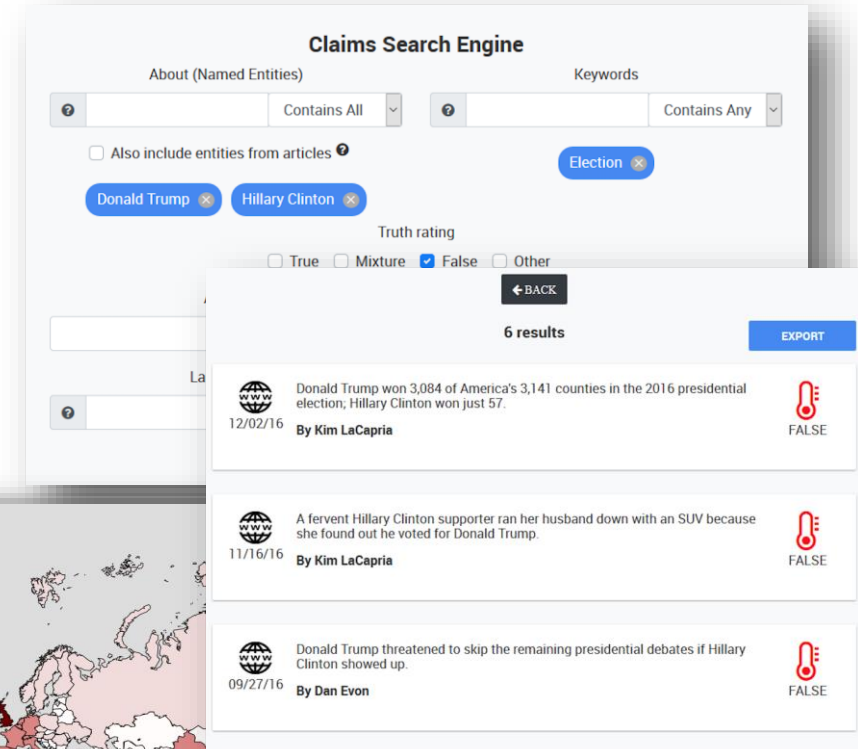
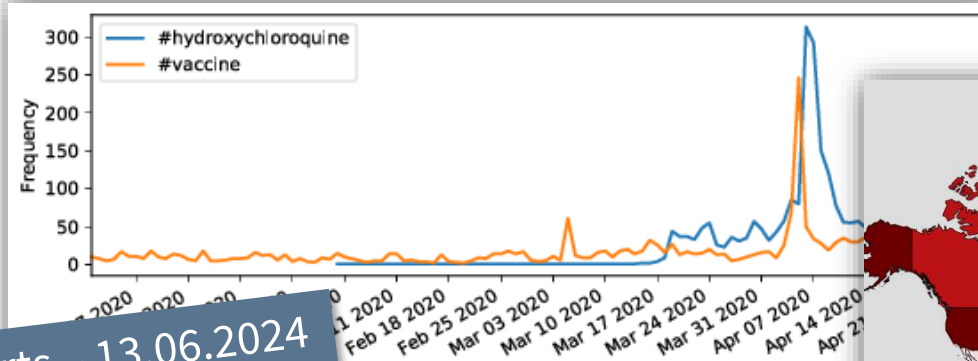
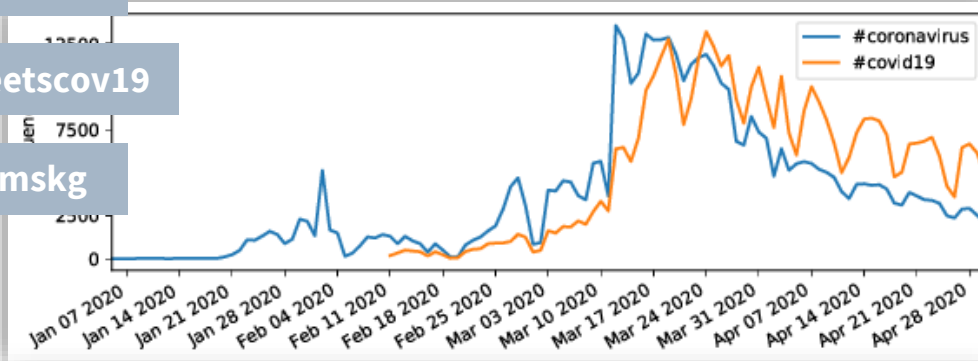


Research Knowledge Graphs of (Social Science) Research Data

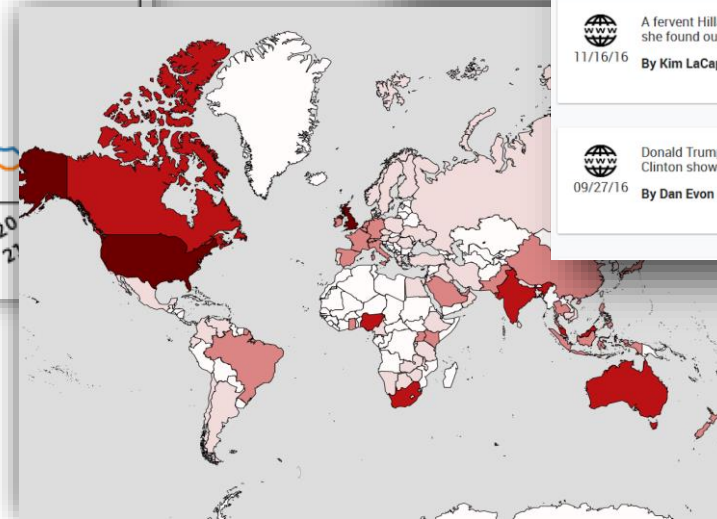
<https://data.gesis.org/tweetskb>

<https://data.gesis.org/tweetscov19>

<https://data.gesis.org/claimskg>

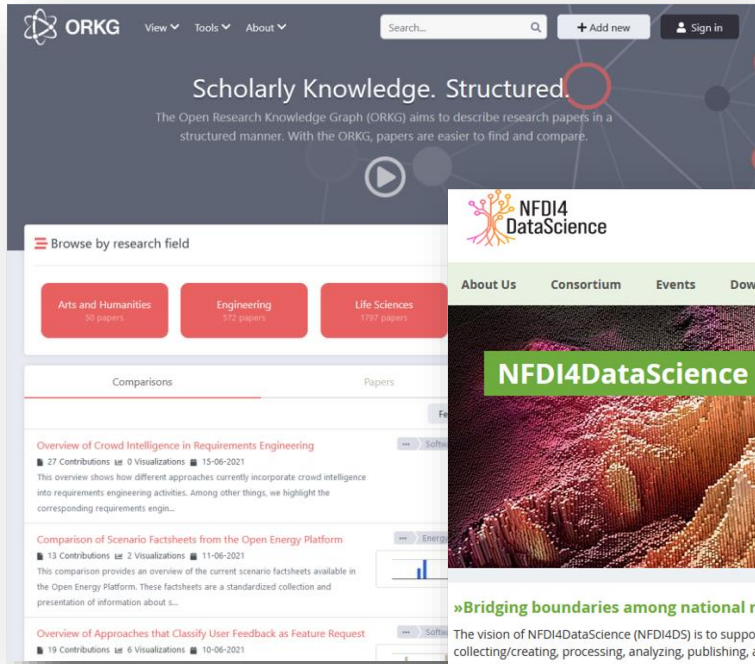


The screenshot displays the 'Claims Search Engine' interface. It includes search filters for 'About (Named Entities)' and 'Keywords', both set to 'Contains All'. The 'Keywords' field contains 'Election'. There are buttons for 'Donald Trump' and 'Hillary Clinton'. The 'Truth rating' section has radio buttons for 'True', 'Mixture', 'False' (which is selected), and 'Other'. A 'BACK' button is visible. Below the filters, it shows '6 results' and an 'EXPORT' button. The results list includes three entries, each with a date, author name, and a 'FALSE' truth rating icon.



Meet the Experts – 13.06.2024
Preserving and analysing
large-scale Twitter data

Research Knowledge Graphs: Initiatives



ORKG View Tools About Search... + Add new Sign in

Scholarly Knowledge. Structured.

The Open Research Knowledge Graph (ORKG) aims to describe research papers in a structured manner. With the ORKG, papers are easier to find and compare.

Browse by research field

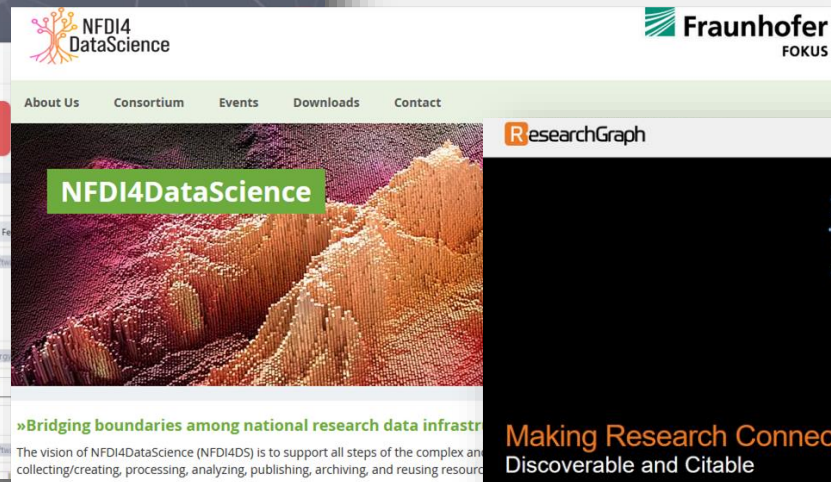
- Arts and Humanities 50 papers
- Engineering 372 papers
- Life Sciences 1797 papers

Comparisons Papers

Overview of Crowd Intelligence in Requirements Engineering
27 Contributions 0 Visualizations 15-06-2021

Comparison of Scenario Factsheets from the Open Energy Platform
13 Contributions 2 Visualizations 11-06-2021

Overview of Approaches that Classify User Feedback as Feature Request
19 Contributions 6 Visualizations 10-06-2021



NFDI4DataScience Fraunhofer FOKUS

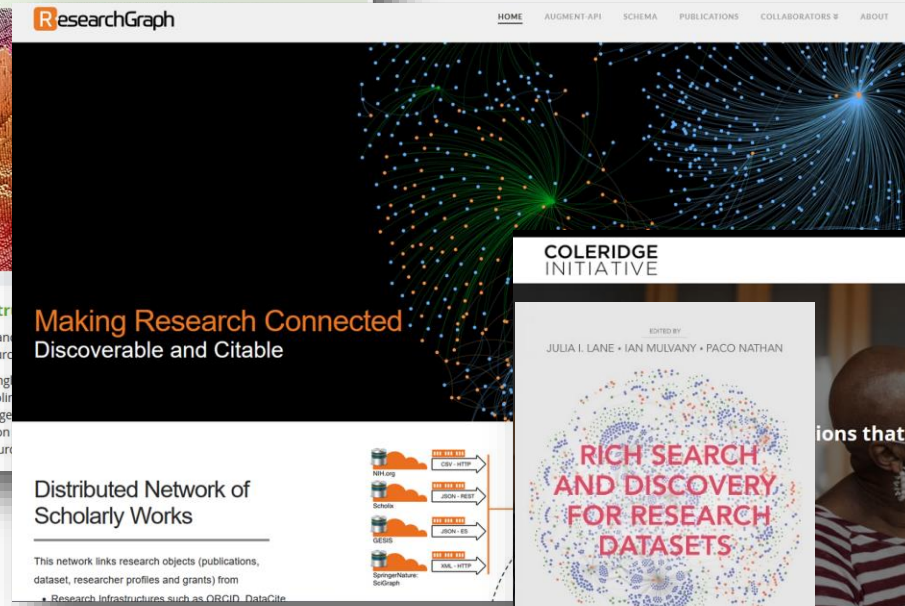
About Us Consortium Events Downloads Contact

NFDI4DataScience

»Bridging boundaries among national research data infrastr...

The vision of NFDI4DataScience (NFDI4DS) is to support all steps of the complex an collecting/creating, processing, analyzing, publishing, archiving, and reusing resourc

The past years have seen a paradigm shift, with computational methods increasing approaches, leading to the establishment and ubiquity of Data Science as a disciplin Science. Transparency, reproducibility and fairness have become crucial challenge complexity of contemporary Data Science methods, often relying on a combination promote fair and open research data infrastructures supporting all involved resourc integrated approach.

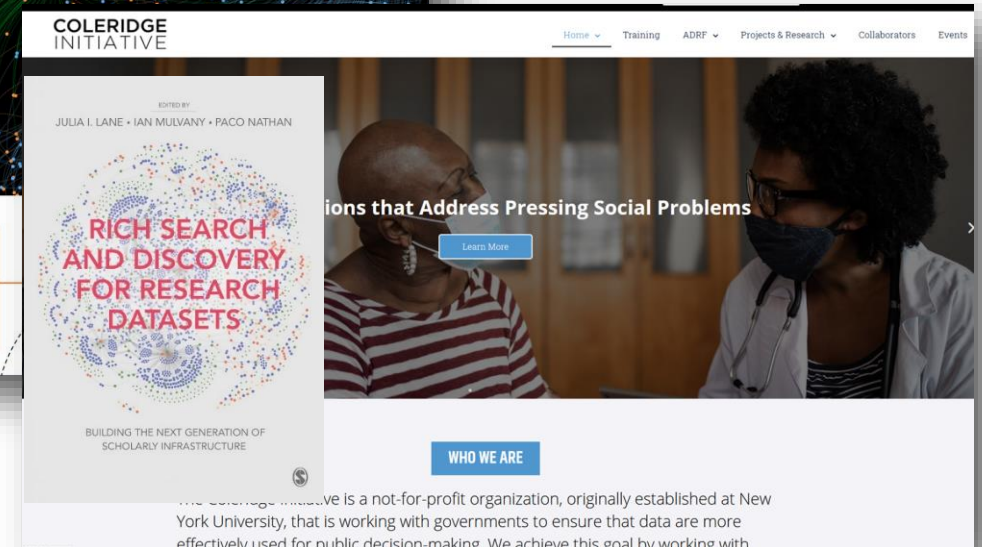


ResearchGraph HOME AUGMENT-API SCHEMA PUBLICATIONS COLLABORATORS # ABOUT

Making Research Connected Discoverable and Citable

Distributed Network of Scholarly Works

This network links research objects (publications, dataset, researcher profiles and grants) from Research Infrastructures such as ORCID, DataCite



COLERIDGE INITIATIVE Home Training ADRF Projects & Research Collaborators Events

Edited by JULIA I. LANE • IAN MULVANY • PACO NATHAN

RICH SEARCH AND DISCOVERY FOR RESEARCH DATASETS

Building the next generation of scholarly infrastructure

Who We Are

...the Coleridge Initiative is a not-for-profit organization, originally established at New York University, that is working with governments to ensure that data are more effectively used for public decision-making. We achieve this goal by working with

Research KGs @ Knowledge Technologies for the Social Sciences (KTS)

- **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>)



The screenshot shows the website for the Knowledge Technologies for the Social Sciences (KTS) department at GESIS. The page features a navigation menu with 'Services', 'Research', and 'Institute' options. A search bar is located at the top right. The main content area includes a header with the text 'Knowledge Technologies for the Social Sciences (KTS)' and a large image of a person interacting with a digital display showing various data visualizations like bar charts and line graphs. Below the header, there is a paragraph describing the department's role in developing digital services and research data infrastructures. A prominent URL, <https://gesis.org/en/kts>, is displayed in a large, semi-transparent box. At the bottom, there are several buttons for 'GESIS-Search', 'Research Labs', 'gesisDataSearch', and 'Knowledge Graph Infrastructure', followed by a list of service categories such as 'Information & Data Retrieval', 'Information Extraction & Linking', 'FAIR Data', 'Data & Services Engineering', 'Big Data Analytics', and 'Human Information Interaction'.

Expert contact & GESIS consulting



Contact: you can reach the speaker/s via e-mail:

benjamin.zapilko@gesis.org

debanjali.biswas@gesis.org

<https://gesis.org/en/kts>

GESIS Consulting: GESIS offers individual consulting in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle.

Please visit our website www.gesis.org for more [detailed information](#) on available services and terms.

Upcoming talks

- 16.05.2024: Opportunities and challenges of Large Language Models for the social sciences
- 13.06.2024: Preserving and analysing large-scale Twitter data
- 11.07.2024: Introduction to scholarly information extraction

- Please visit our meet-the-experts website:
<https://www.gesis.org/en/services/sharing-knowledge/consulting-and-guidelines/meet-the-experts>

Thank you for participating!