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

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

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
Research Knowledge Graphs of Scholarly Resource Metadata
GESIS KG: Integrated search @ GESIS

https://search.gesis.org/

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

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

GESIS KG: Links between resources

Number of links between different resource types

- Automatic: 98,963
- Manual: 47,547

Internal vs. External links

- GESIS-internal links: 143,536
- Links to external resources: 29,744
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).

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
Research Knowledge Graphs of (Social Science) Research Data
TweetsKB – a non-sensitive large-scale archive of societal discourse

- 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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Total</th>
<th>Unique</th>
<th>% with &gt;= 1 feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hashtags</td>
<td>1,161,839,471</td>
<td>68,832,205</td>
<td>0.19</td>
</tr>
<tr>
<td>Mentions</td>
<td>1,840,456,543</td>
<td>149,277,474</td>
<td>0.38</td>
</tr>
<tr>
<td>Entities</td>
<td>2,563,433,997</td>
<td>2,265,201</td>
<td>0.56</td>
</tr>
<tr>
<td>Sentiment</td>
<td>1,265,974,641</td>
<td>-</td>
<td>0.5</td>
</tr>
</tbody>
</table>

P. Fafalios, V. Iosifidis, E. Ntoutsi, 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

<table>
<thead>
<tr>
<th>Country</th>
<th>Share of population aged 12+ that has not had any Covid vaccine dose (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>24.8</td>
</tr>
<tr>
<td>Switzerland</td>
<td>24.4</td>
</tr>
<tr>
<td>Germany</td>
<td>22.3</td>
</tr>
<tr>
<td>Sweden</td>
<td>16.1</td>
</tr>
<tr>
<td>Italy</td>
<td>13.9</td>
</tr>
<tr>
<td>UK</td>
<td>13.6</td>
</tr>
<tr>
<td>Netherlands</td>
<td>13.2</td>
</tr>
<tr>
<td>Belgium</td>
<td>12.7</td>
</tr>
<tr>
<td>Finland</td>
<td>12.4</td>
</tr>
<tr>
<td>Denmark</td>
<td>11.1</td>
</tr>
<tr>
<td>Norway</td>
<td>9.9</td>
</tr>
<tr>
<td>France</td>
<td>8.4</td>
</tr>
<tr>
<td>Spain</td>
<td>8.2</td>
</tr>
<tr>
<td>Ireland</td>
<td>7.1</td>
</tr>
<tr>
<td>Iceland</td>
<td>1.9</td>
</tr>
<tr>
<td>Portugal</td>
<td>1.6</td>
</tr>
</tbody>
</table>

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

Germany suspends vaccinations with Astra Zeneca

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

Meet the Experts – 13.06.2024
Preserving and analysing large-scale Twitter data
Research Knowledge Graphs: Initiatives
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/](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](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](https://search.gesis.org))
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](http://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!