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

Focused Tutorial on Capturing, Enriching, Disseminating Research Data Objects

Stefan Dietze, 25.11.2022
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

Research Data Cycle

1. Plan a study & collect data
2. Find & access data
3. Process & analyze data
4. Archive & share
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?
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
- **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“)
Research KGs in Practice: integrated search @ GESIS

https://search.gesis.org/

Dataset

Related publications (2)

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

Joachim, M., 2016, 40, graph, euros, (Manneheimer Zentrum für Europäische Sozialforschung: Arbeitspapiere - working papers; Nr. 163)

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

Sibbing, Heike & Puschmann, Peter & Shalom, Yossi

Yarmek, Ashgar, 2014, 207, 27 (Studies in migration and diaspora)
From publications to machine-interpretable metadata KGs
Disambiguation of dataset & software/script citations

- 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)

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

From publications to machine-interpretable metadata KGs
Understanding scientific software/data usage

- Top adopters of data science/AI/software...

Top adopters of data science/AI/software...
...follow the worst citation habits
Building a public research knowledge graph from Twitter data

https://data.gesis.org/tweetskb

From social media to machine-interpretable research data KGs
From social media to machine-interpretable research data KGs

TweetsKB – a large-scale research KG of societal opinions

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

KTS research focused on evaluating & developing semi-supervised methods for online discourse analysis:
  - Stance detection [IJJS2020]
  - 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


https://data.gesis.org/tweetskb
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

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

<table>
<thead>
<tr>
<th>Country</th>
<th>Share (%)</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>22.1</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>1.5</td>
</tr>
</tbody>
</table>

Source: FT analysis of figures from national sources and Our World in Data. Rates shown are as of November 9.
Germany suspends vaccinations with Astra Zeneca

Twitter discourse zu “Impfbereitschaft”

Investigating Vaccine Hesitancy in DACH countries

https://dd4p.gesis.org
RKG-based discourse analysis using TweetsKB
Vaccine Hesitancy – key topics in “safety” category

https://dd4p.gesis.org
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 | Science reference
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Donating blood not only helps others, but reduces the rate of cancer and heart disease in the donor.</td>
<td>via @medical_xpress A new in vitro (test tube) study, &quot;Dietary functional benefits of Bartlet <a href="http://t.co/Qv1C1GjQin">http://t.co/Qv1C1GjQin</a> #UFO4UBlogHealth</td>
</tr>
<tr>
<td>(2) via @medical_xpress A new in vitro (test tube) study, &quot;Dietary functional benefits of Bartlet <a href="http://t.co/Qv1C1GjQin">http://t.co/Qv1C1GjQin</a> #UFO4UBlogHealth</td>
<td>(3) How is @UChicagoIME shaping the future of science? Find out on April 6!</td>
</tr>
<tr>
<td>Science relevance</td>
<td>Science reference</td>
</tr>
<tr>
<td>(3) How is @UChicagoIME shaping the future of science? Find out on April 6!</td>
<td>(4) Study: Shifts in electricity generation spur net job growth, but coal jobs decline - via @DukeU <a href="http://t.co/AXGmKUPata">http://t.co/AXGmKUPata</a></td>
</tr>
<tr>
<td>Science reference</td>
<td>No science</td>
</tr>
</tbody>
</table>

How about mentions of science resources on the Web?
Example: Twitter

- 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

Al4Sci project: understanding and classification of science discourse online (news, social Web)
How about mentions of science resources on the Web?
Example: Twitter

SciBERT classifier

Heuristic: Sci term

Sci subdomain
SciTweets dataset & classifier

- 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

<table>
<thead>
<tr>
<th>Task</th>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>binary</td>
<td>1 - Science-related</td>
<td>84.70</td>
<td>83.99</td>
<td>84.34</td>
</tr>
<tr>
<td></td>
<td>2 - Not Science-related</td>
<td>92.67</td>
<td>93.03</td>
<td>92.85</td>
</tr>
<tr>
<td>multi</td>
<td>1.1 - Scientific Claim</td>
<td>75.00</td>
<td>81.18</td>
<td>77.97</td>
</tr>
<tr>
<td></td>
<td>1.2 - Reference</td>
<td>76.19</td>
<td>77.01</td>
<td>76.60</td>
</tr>
<tr>
<td></td>
<td>1.3 - Research Context</td>
<td>81.06</td>
<td>79.65</td>
<td>80.35</td>
</tr>
</tbody>
</table>
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)
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.