gesis Leibniz-Institut für Sozialwissenschaften

hhu Heinrich Heine Universität Düsseldorf

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Tales from the Inside: Experiences from 10 Years of Growing, Enriching, and Sharing a Historical Archive of Tweets

Konferenz: Archivierung sozialer Medien @ DNB

Stefan Dietze, Dimitar Dimitrov, Robert Jäschke, Sebastian Tiesler

Leibniz Gemeinschaft

Agenda

- Collecting & Archiving
- Data Sharing: Challenges & Approaches
- Summary & Outlook

Motivation

Archival perspective:

- Ensure long-term archival of volatile information from Twitter
- Independence from third-party data access / APIs

Research perspective – have tweets available for

- Training and evaluating machine learning models (e.g., NER, classification)
- Large-scale analyses (e.g., language use, trends)
- Experimenting with big data architectures (e.g., Hadoop, Elastic)

→ **Goal**: capture a representative sample of all Twitter data

Collecting

- **Objective:** collect the freely available 1% sample from Twitter's streaming API
- **Approach:**
  - Collector from TREC microblog tracks (https://github.com/lintool/twitter-tools)
  - Almost no configuration necessary; once started, collector runs indefinitely
  - Distributed setup with at least two machines running in parallel
  - Regular checks whether collector is alive; restart, if necessary
- **Result:**
  - gzip-compressed JSONL (one file per hour, ca. 100MB)


```
{"created_at":"Fri Nov 16 16:03:55 +0000 2018","id":1063462679623467008,"id_str":"1063462679623467008","text
```



```
{
  "created_at": "Fri Nov 16 16:03:55 +0000 2018",
  "id": 1063462679623467000,
  "id_str": "1063462679623467008",
  "text": "RT @markiaaan: Today Anna Busch @FontaneArchiv &amp;amp; me @UCLab_Potsdam presented our joint
research (+@peertrilcke @nrchtct @vik_bru ) in Züric…",
  "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null
```

Archiving

- **Resources**:
  - Hadoop file system (HDFS) in two clusters (40 and 7 machines, respectively)
  - Some standard servers (used mainly for collection)
- **Storage**:
  - Daily files are concatenated into one big daily file per collector
  - Copied regularly (~daily) to cluster
  - One directory per collector:

```
> hdu "/data/twitter/streams/*"
327.4 G  /data/twitter/streams/fs3
4.4 T    /data/twitter/streams/gesis
1.6 T    /data/twitter/streams/goofy
3.7 T    /data/twitter/streams/hadoop3
2.8 T    /data/twitter/streams/jerry
444.0 G  /data/twitter/streams/jerry_v2
1.7 T    /data/twitter/streams/meco
4.8 T    /data/twitter/streams/prometheus
```

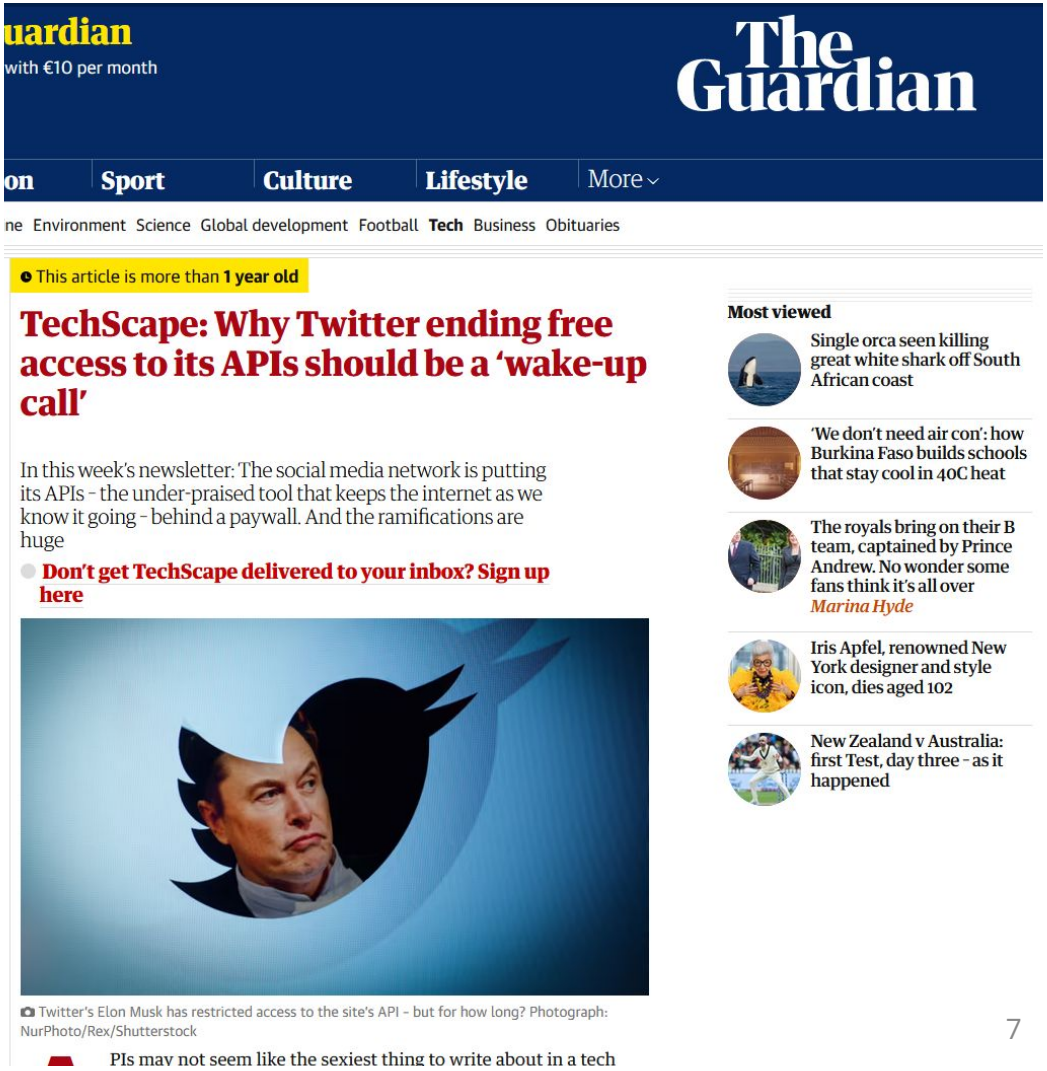
Agenda

- Collecting & Archiving
- Data Sharing: Challenges & Approaches
- Summary & Outlook

Data sharing challenges: overview

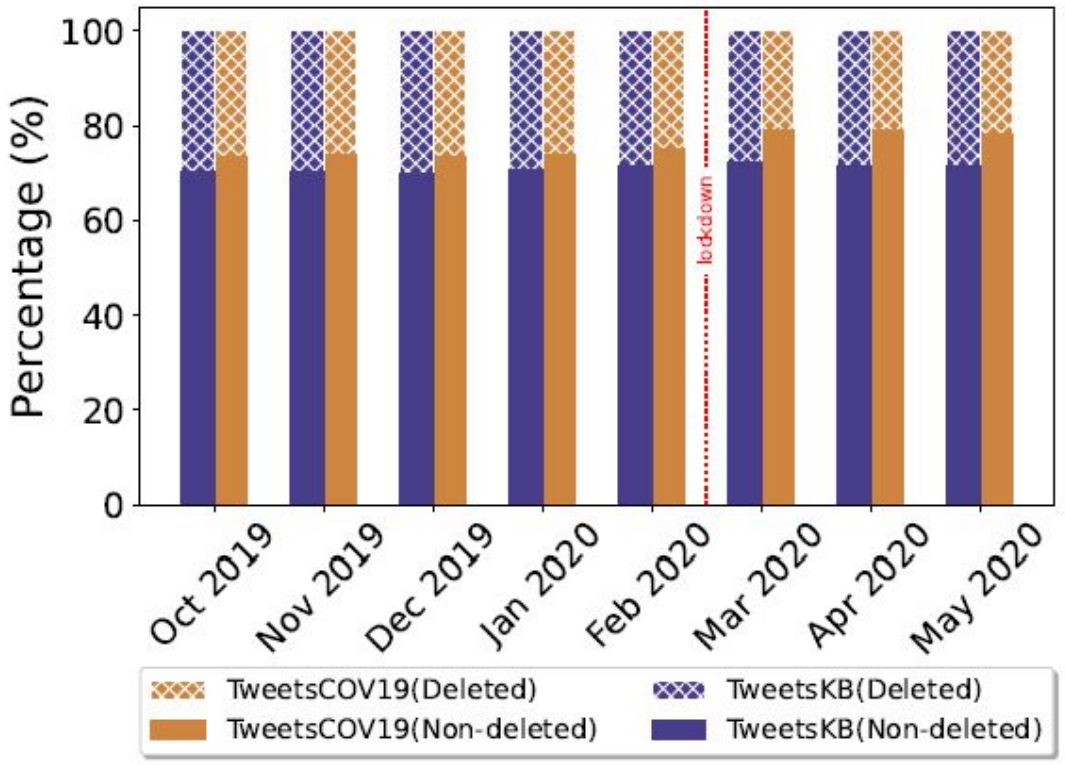
- Licensing / legal aspects: Twitter terms of service, copyright, etc.
- Ethical concerns, e.g., when information is taken out of context
- Sharing tweet IDs rather than full-text widely established practice to comply with Twitter ToC

- API shutdown in May 2023: dehydrated Twitter datasets (and research) not reproducible anymore



Data sharing challenges: volatility & decay

- Data is not persistent
- Example: deletion ratio of tweets (approx. 25% on average)



Data sharing options

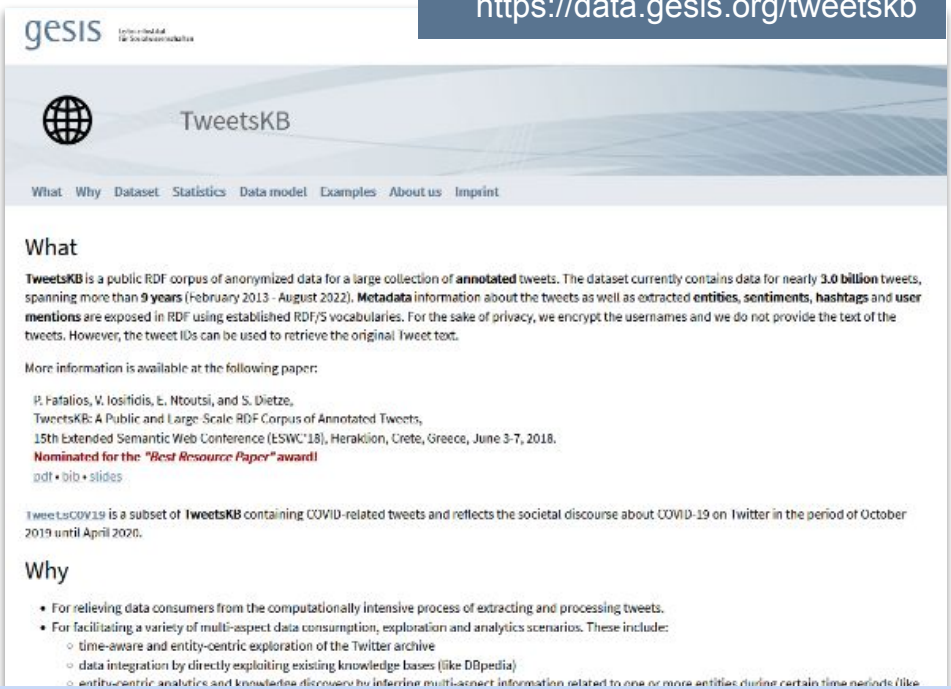
- **Sensitive data access**
  Facilitating on-prem research on data (e.g. online/offline secure data centers) or contract-based sharing of sensitive data

- **Public, non-sensitive data offers**
  Creating non-sensitive derivatives from raw data to facilitate research

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

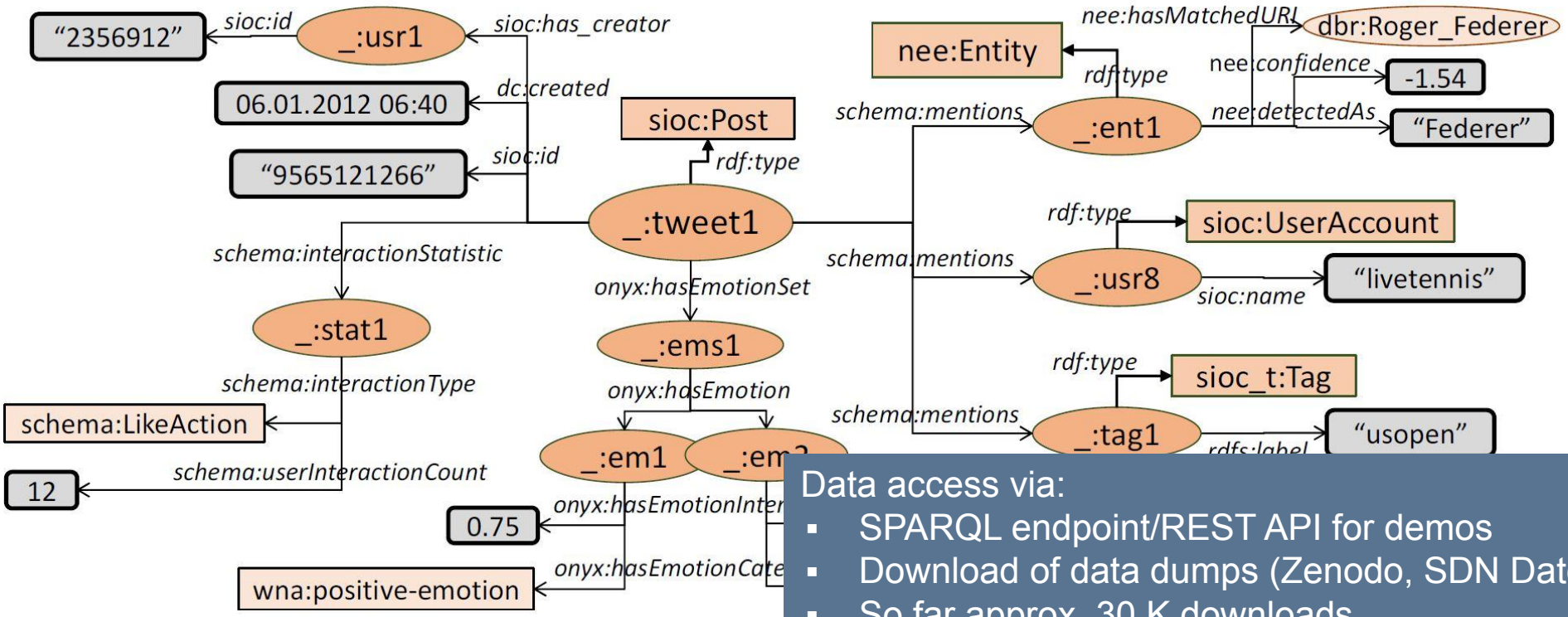
https://data.gesis.org/tweetskb

- 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

TweetsKB – knowledge graph schema & data access



Data access via:

- SPARQL endpoint/REST API for demos
- Download of data dumps (Zenodo, SDN Datorium)
- So far approx. 30 K downloads

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**

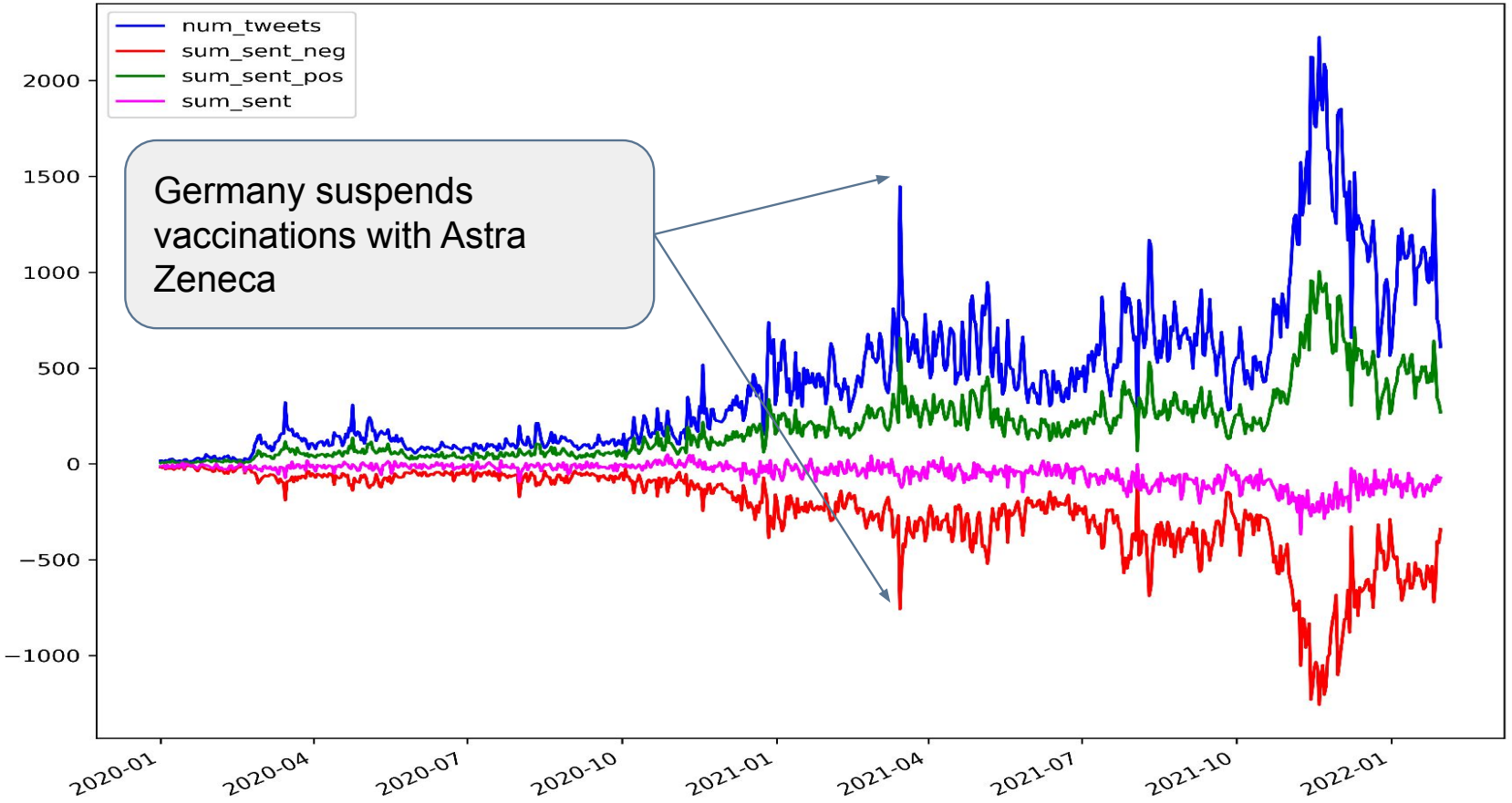
TweetsKB as social science research corpus

Investigating vaccine hesitancy in DACH countries

DD4P
DiscourseData4Policy

https://dd4p.gesis.org/

Twitter discourse zu “Impfbereitschaft”



Reflection on data sharing options

- **Public, non-sensitive data offers**
  Creating non-sensitive derivatives from raw data to facilitate research

  - Data aggregation, feature enrichment (e.g. entities, sentiments), e.g. TweetsKB
  - Provides means for analysing data without accessing sensitive information.
  - **Scales well to many users but features/data products may not be optimal for all kinds of research questions**

- **Sensitive data access**
  Facilitating on-prem research on data (e.g. online/offline secure data centers) or contract-based sharing of sensitive data

  - Requires strict output control (secure data centers) or contract-based data sharing under very strict constraints
  - **Does not scale well or requires very constrained modes of access (e.g. through predefined set of methods) but may allow users to apply their own methods (tbc)**

Summary & outlook

- Twitter archive underlying TweetsKB: largest tweet archive hosted by a public research data infrastructure (14 bn tweets, continuous data between 2013-2023)
- Data collection: easy, as long as API was available
- Lesson from API shutdown: 3rd party APIs from profit-oriented companies do not ensure reproducibility of research data
- Data sharing: hard, due to legal concerns / Twitter ToC
- Approaches: sensitive data access vs non-sensitive data offers (eg TweetsKB, TweetsCOV19)
- GESIS currently exploring both avenues to provide research data to the research community

Acknowledgements & Thanks

- Knowledge Technologies for the Social Sciences @ GESIS
  http://gesis.org/en/kts
  https://www.gesis.org/en/services/finding-and-accessing-data/gesis-web-data

- Information Processing & Analytics @ HU Berlin
  https://www.ibi.hu-berlin.de/de/institut/personen/jaeschke

- L3S Research Center
  https://www.l3s.de/