Tweets archivieren

Initiative, Umsetzung, Ausblick

Dr. Claus-Michael Schlesinger (Humboldt-Universität zu Berlin)
Dr. Britta Woldering (Deutsche Nationalbibliothek)

Ausgangslage

- Bedarfe
  - Literarische Texte für künftige Auswahl sichern (Passig 2021: Den Heuhaufen archivieren; Passig 2022: Rucksack oder Rechenzentrum)
  - Single Point of Failure: Twitter als Data Provider für die Wissenschaft
  - Archive und Sondersammlungen
- Transformation der Plattform
  - Unternehmensstruktur
  - Plattformfunktionen
  - Nutzer:innen
  - Archiv
    - Struktur
    - Zugang

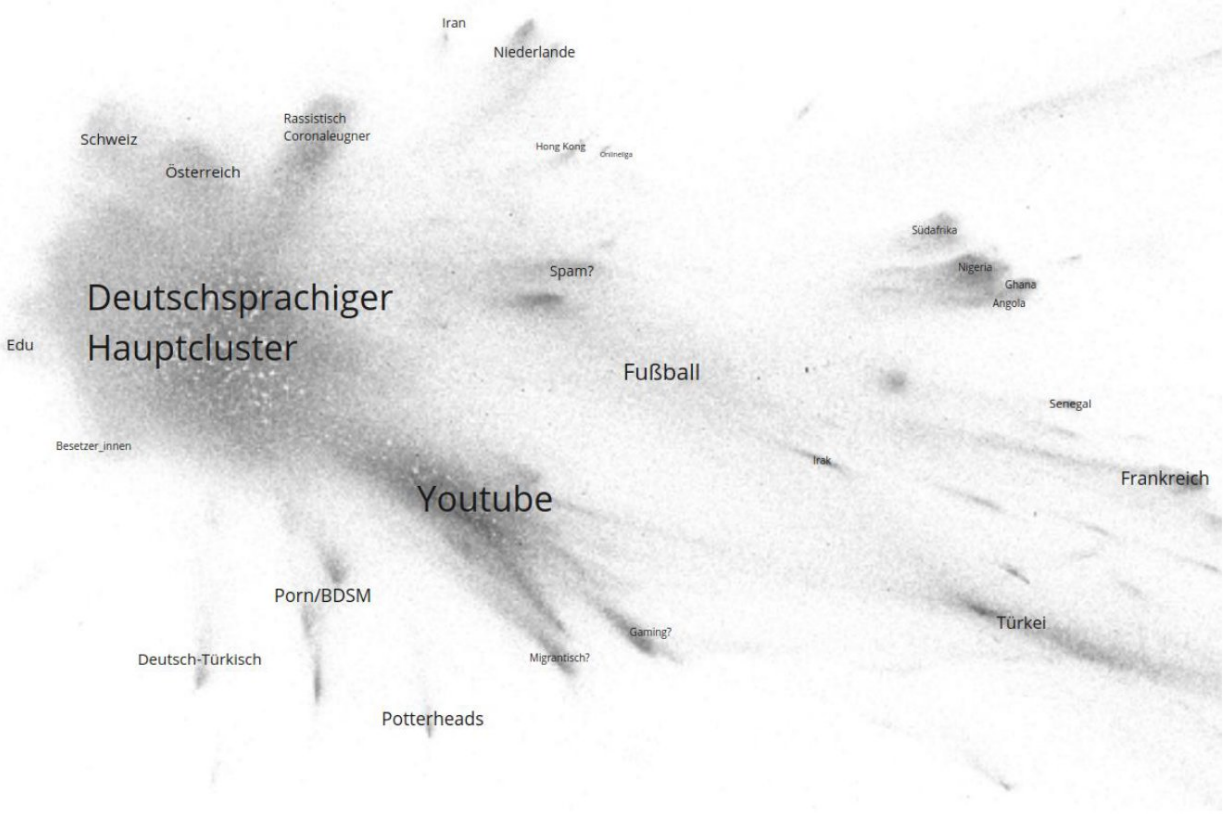
Klar archivieren, aber wer machts?

Ergebnisse der ersten Sammlungsphase

- Dauer der aktiven Sammlungsphase: Januar bis April 2023
- Verteilte Archivierung
- ~200 Millionen Tweets
- ~5,8 Millionen Accounts
- Zeitraum 03/2006 - 05/2011
- Vernetzung (Korpora, Sammlung, Bereitstellung)

Vorarbeiten
Korpusbildung, Analyse, Qualitätskontrolle

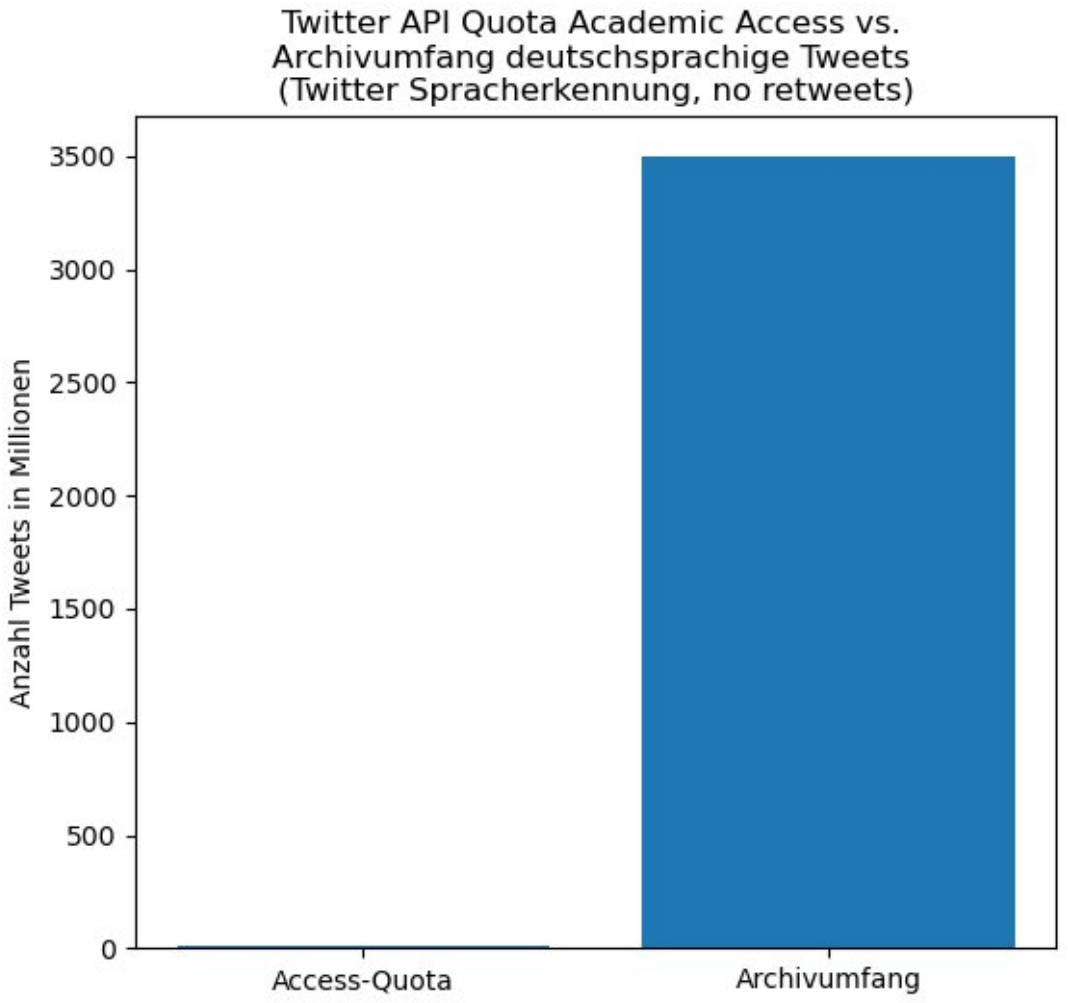
- Passig, Kathrin (2021): Den Heuhaufen archivieren, Beitrag zur Tagung "#LiteraturarchivDerZukunft" am Deutschen Literaturarchiv Marbach, https://docs.google.com/presentation/d/1CQuVkXaDvsl0psAfkCSkzBcyitxsAAIXEjhLmQOIVaY/edit?usp=sharing
- Scheffler, Tatjana (2014): A German Twitter Snapshot, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, S. 2284–2289, http://www.lrec-conf.org/proceedings/lrec2014/pdf/1146_Paper.pdf.
- Hammer, Luca (2020): Vermessung der deutschsprachigen Twittersphäre, Bachelorarbeit Universität Paderborn, 2020, https://lucahammer.com/wp-content/uploads/2021/05/Hammer-Luca-Vermessung-der-deutschsprachigen-Twittersphare-WEB.pdf.
- Pfeffer, Juergen; Mooseder, Angelina; Lasser, Jana u. a. (2022): This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API, 11.04.2023, https://doi.org/10.48550/arXiv.2204.02290.
- Fafalios, Pavlos; Iosifidis, Vasileios; Ntoutsi, Eirini u. a. (2018): TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets, in: Gangemi, Aldo; Navigli, Roberto; Vidal, Maria-Esther u. a. (Hg.): The Semantic Web, Cham 2018 (Lecture Notes in Computer Science), S. 177–190, https://doi.org/10.1007/978-3-319-93417-4_12.



Luca Hammer, Follow-Netzwerk deutschsprachige Twittersphäre (Accounts mit vier oder mehr deutschsprachigen Tweets), in: Hammer 2020, S. 41 (Abb. 10), Ausschnitt.

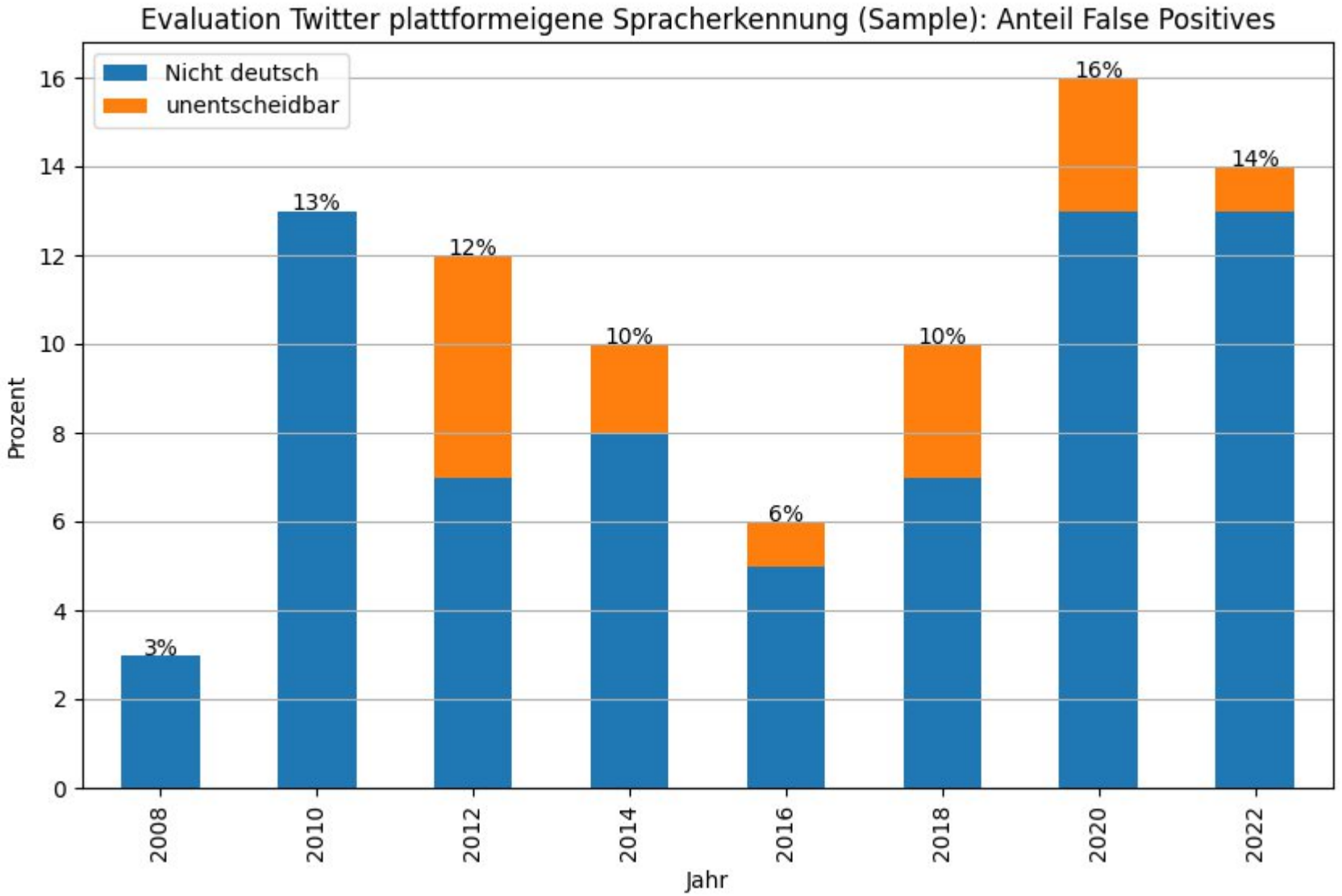
Begrenzter Zugriff und Operationalisierung *on speed*

- Begrenzter Zugriff auf das Twitter-Archiv vs. sehr viele Tweets
  - Quota
  - Zeit
- Operationalisierung: Tweets nach Bedarfen, Zeit und im Sinne des gesetzlichen Auftrags der DNB
  - accountorientiert
  - **sprachorientiert**
  - ~~geolocation~~ (Scheffler 2013)



Twitter Spracherkennung

- Nicht transparent
- Unzureichend dokumentiert
- Eigene Stichprobe: ~10 Prozent False Positives
- False Negatives schwer überprüfbar (in Arbeit)



- **Sample**: je 100 Tweets bi-yearly (2008-2022)
- **manuelle Annotation** auf Basis der Texte
- nicht deutschsprachig: eindeutige Zuordnung zu einer anderen Sprache
- unentscheidbar: keine ausreichend eindeutigen Marker, z.B. nur Hashtags, nur Medien

Annotation und Visualisierung zus. mit Adam Jorczick (Universität Stuttgart)

Verteilte Archivierung

Anzahl TN/Accounts	Archivierungs dauer
1	30 Jahre
10	3 Jahre
20	1,5 Jahre
50	7 Monate
350	<1 Monat

Archivumfang

- ~3.5 Milliarden Tweets
- 03/2006 - 02/2023

Koordination

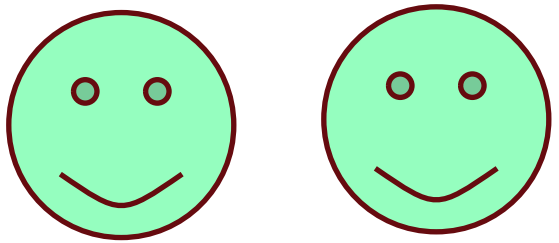
- Definition von Batches mit je 1 Million Tweets
- Reservierung von Batches für lokalen Download
- Auslösen eines serverseitigen Downloads (Tokenspende)

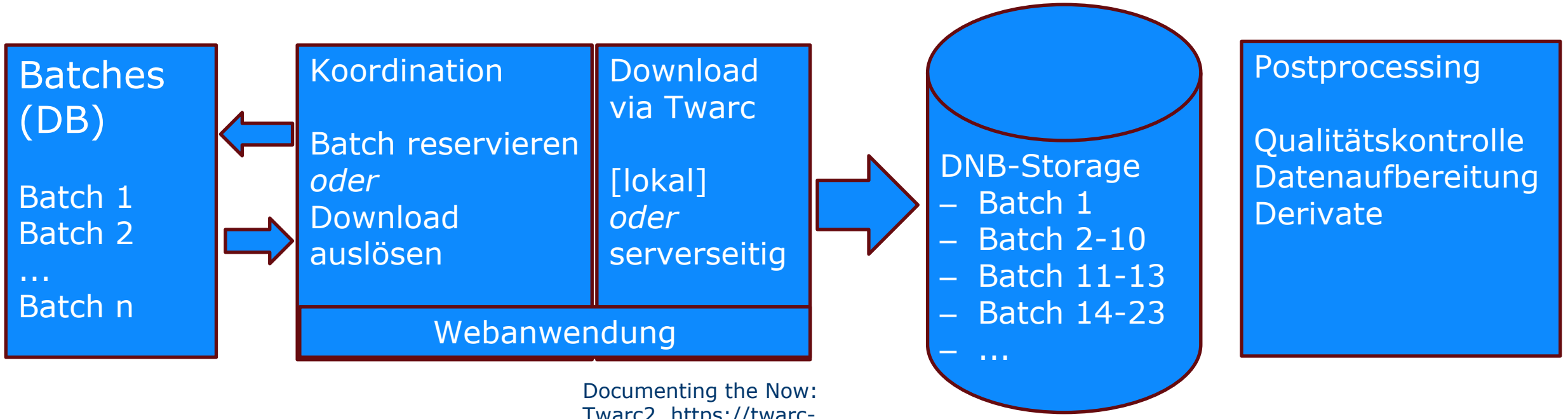
Tweets archivieren mit Twarc/Twarc2

- Für Archivierung entwickelt von *Documenting The Now* (docnow.io)
- Python Package und CLI-Programm
- Genaue Kontrolle
- vollständige Daten und Metadaten
- Logging/Prozessmetadaten
- stabil

Documenting the Now: Twarc2, https://twarc-project.readthedocs.io/en/latest/

Verteilte Archivierung





Documenting the Now: Twarc2, https://twarc-project.readthedocs.io/en/latest/

Benutzungsverwaltung, Support, Koordination, Dokumentation

Was kommt ins Archiv?

Alles*

Was kommt ins Archiv

*

- Fokus auf eine Menge von einzelnen Tweets
- Texte
- Medien
- Metadaten
  - Tweetstatistik: Anzahl der Retweets, Likes zum Zeitpunkt der Archivierung
  - Conversation-ID
  - Hashtags
  - [alle verfügbaren Metadaten]
- Keine Friends/Followers

Ausblick

Archivieren

- Strategiewechsel: vom API-Crawl zu Browser Automation?
- Weitere Korpora zur Sammlung hinzufügen
- Medien archivieren
- Vernetzung: Eine Tweet-Registry für die Wissenschaft?

Bereitstellen

- Vollständiger Datensatz im Rahmen der DH-Calls der DNB (ab 2024)
  - On-premise auf DNB-Infrastruktur
  - Nicht öffentlich
- Derivate
  - Anonymisiert
  - öffentlich

Vielen Dank!

Infos: https://dnb.de/twitterarchiv
Kontakt: twarchiv@dnb.de

DOI: 10.5281/zenodo.8006204

Britta Woldering (DNB)
Peter Leinen (DNB)
Tobias Steinke (DNB)
Claus-Michael Schlesinger (HU)
Mona Ulrich (DLA, SDC4Lit)