gesis Leibniz-Institut für Sozialwissenschaften

CAIS RESEARCH FOR THE DIGITAL AGE

„Vergänglichkeit“ als Herausforderung im Umgang mit Daten aus Online-Plattformen

Katrin Weller (GESIS & CAIS) – weller.katrin@gesis.org

Berlin, 28.03.2023, Konferenz für Sozial- und Wirtschaftsdaten

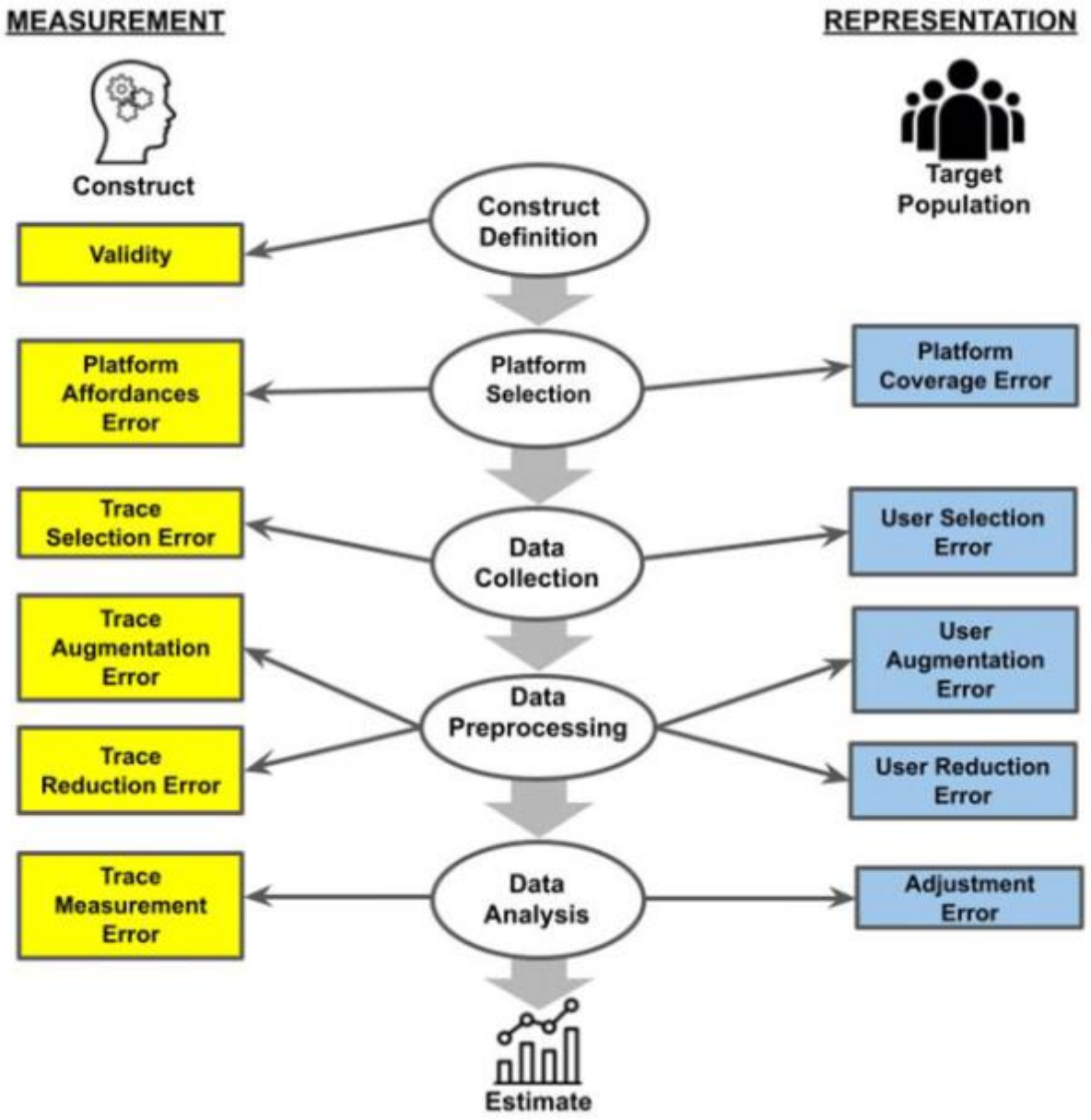
Leibniz Gemeinschaft

Daten aus Online-Plattformen

- In den Sozialwissenschaften werden Daten aus verschiedenen Online-Plattformen für die Forschung genutzt, darunter vielfach, aber nicht nur, Social Media Plattformen.
- Sie versprechen alternative und komplementäre Einblicke in menschliches Verhalten oder zu Charakteristiken und Einstellungen, u.a. auch für Bereiche, die mit klassischen Methoden der empirischen sozialwissenschaftlichen Methoden schwer zu erfassen sind.
- Dabei entstehen jedoch übergreifende und für jede Plattform spezifische Herausforderungen, welche die Datenqualität beeinflussen können.

Error Framework: TED-ON

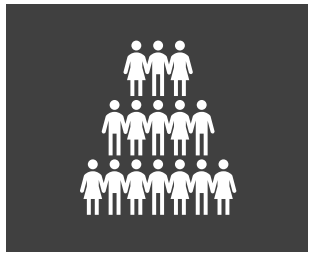
- Fokus auf Daten aus „Online Plattformen“
- Inspiriert vom Total Survey Error, TSE (Groves et al.)



Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. Public Opinion Quarterly 85 (S1): 399–422. doi: https://doi.org/10.1093/poq/nfab018.

Ephemerality?
4 Ebenen für Vergänglichkeit von Online-Plattform-Daten

1. PLATTFORM



2. NUTZENDE

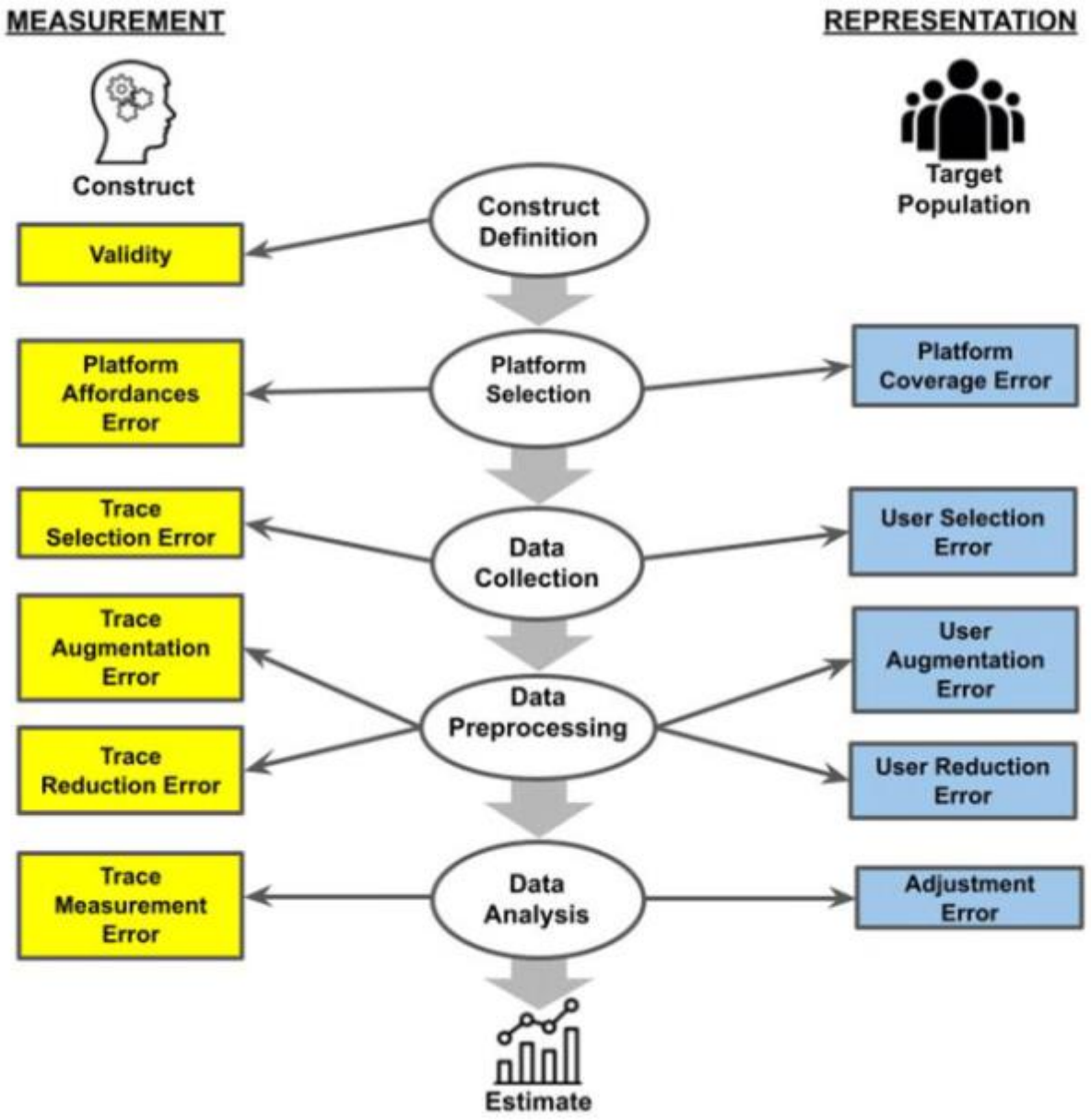
3. INHALT/DATEN

4. DATEN/ZUGANG

Betrachtet werden hierbei nicht die Meinungsänderungen / Verhaltensänderungen der potentiellen Zielgruppe (wie sie ggf. auch Untersuchungsgegenstand sein können), sondern verschiedene Arten von Veränderungen, die beeinflussen, welche Daten wir über Zielgruppen erheben können.

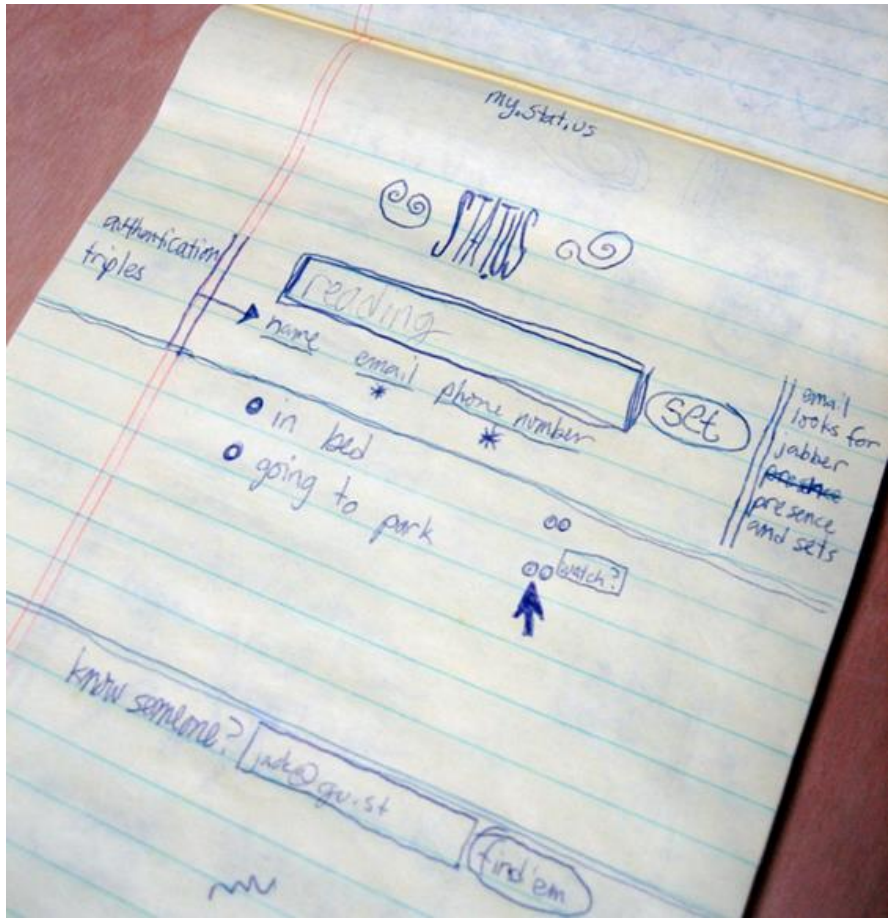
Error Framework: TED-ON

- Viele der bestehenden Herausforderungen verstärken sich durch die verschiedenen Ebenen, die von einer „Flüchtigkeit“ oder „Vergänglichkeit“ betroffen sind.
- Die Vergänglichkeits-Dimension müsste praktisch für alle Arten von Errors zusätzlich überprüft werden.



Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. Public Opinion Quarterly 85 (S1): 399–422. doi: https://doi.org/10.1093/poq/nfab018.

1. Plattformen
Beispiel: Veränderungen der Plattform Affordances



https://www.flickr.com/photos/jackdorsey/182613360

Herausforderung:
Ständig wandelnde Funktionen.
“Look & Feel” der Plattformen kann im Laufe der Zeit, aber auch basierend auf anderen Eigenschaften (Land, Gerät), variieren.

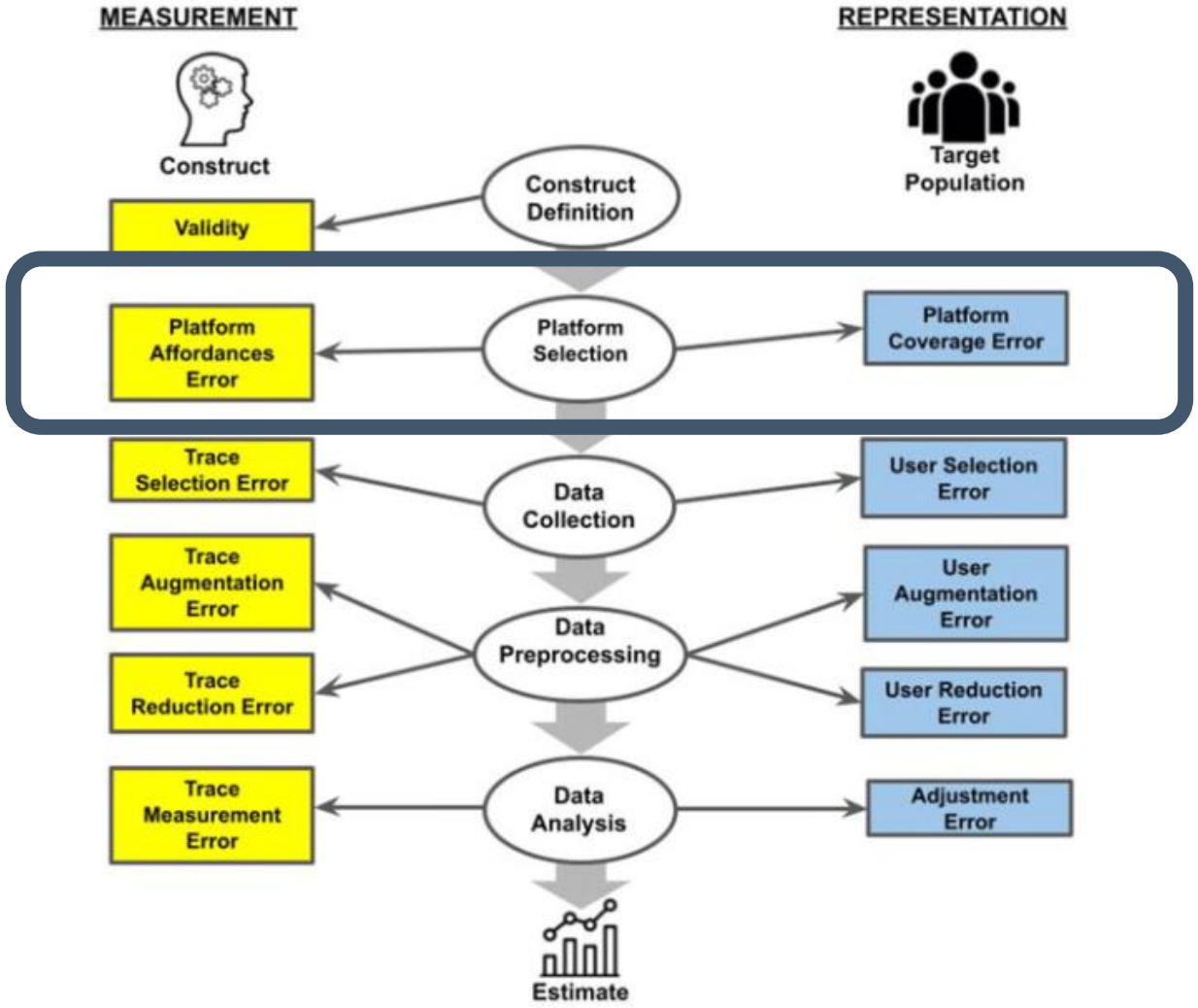
2. Nutzende: Veränderungen des Nutzungsverhaltens

https://twitter.com/chrismessina/status/223115412?ref_src=twsrc%5Etfw

Herausforderungen:

- Nutzungsverhalten ist direkt verbunden mit Änderungen der Plattform-Funktionen.
- Wenig Informationen über Demografie der Nutzenden und Entwicklung der Plattform-Population.

Ephemerality im TED-On Framework



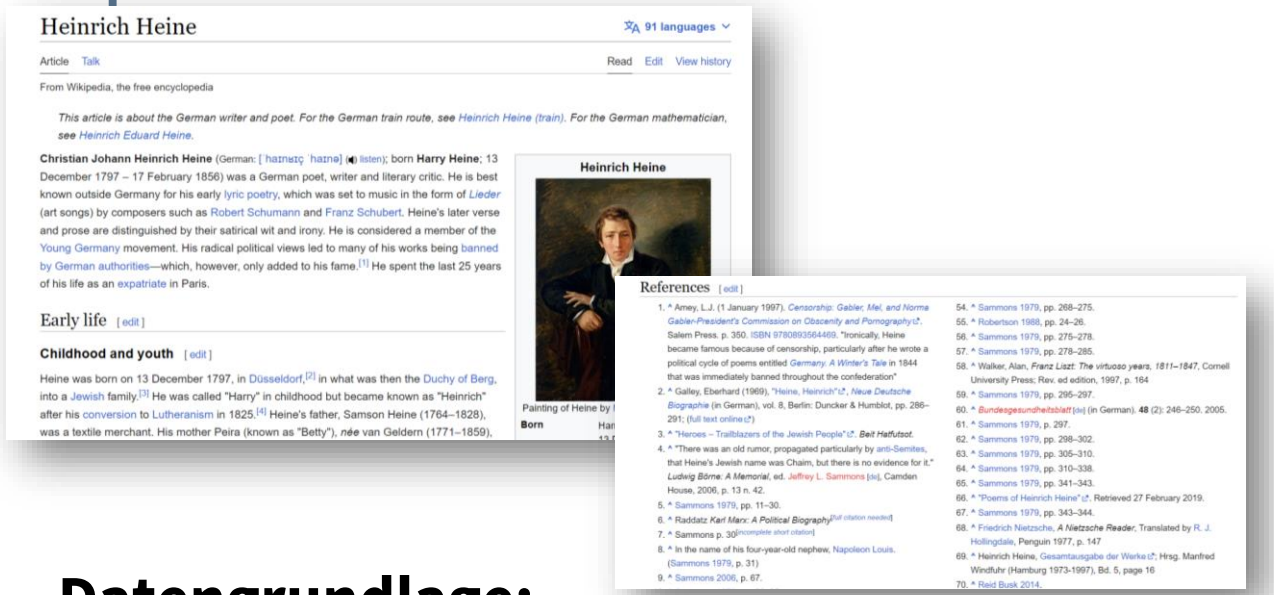
Beipspiel: Platform Selection

- **Platform Affordances** verändern sich im Laufe der Zeit.
- **Platform Coverage:** Plattform-Demografie/Nutzendenbasis verändern sich.

→ Weder Affordances noch Plattform-Populationen werden aktuell systematisch erfasst und dokumentiert. Vergleiche über verschiedene Zeitpunkte hinweg sind dadurch erschwert.

3. Inhalte/Daten
Beispiel: Referenzen in Wikipedia

Wie verändern sich Referenzen in Wikipedia im Laufe der Zeit?

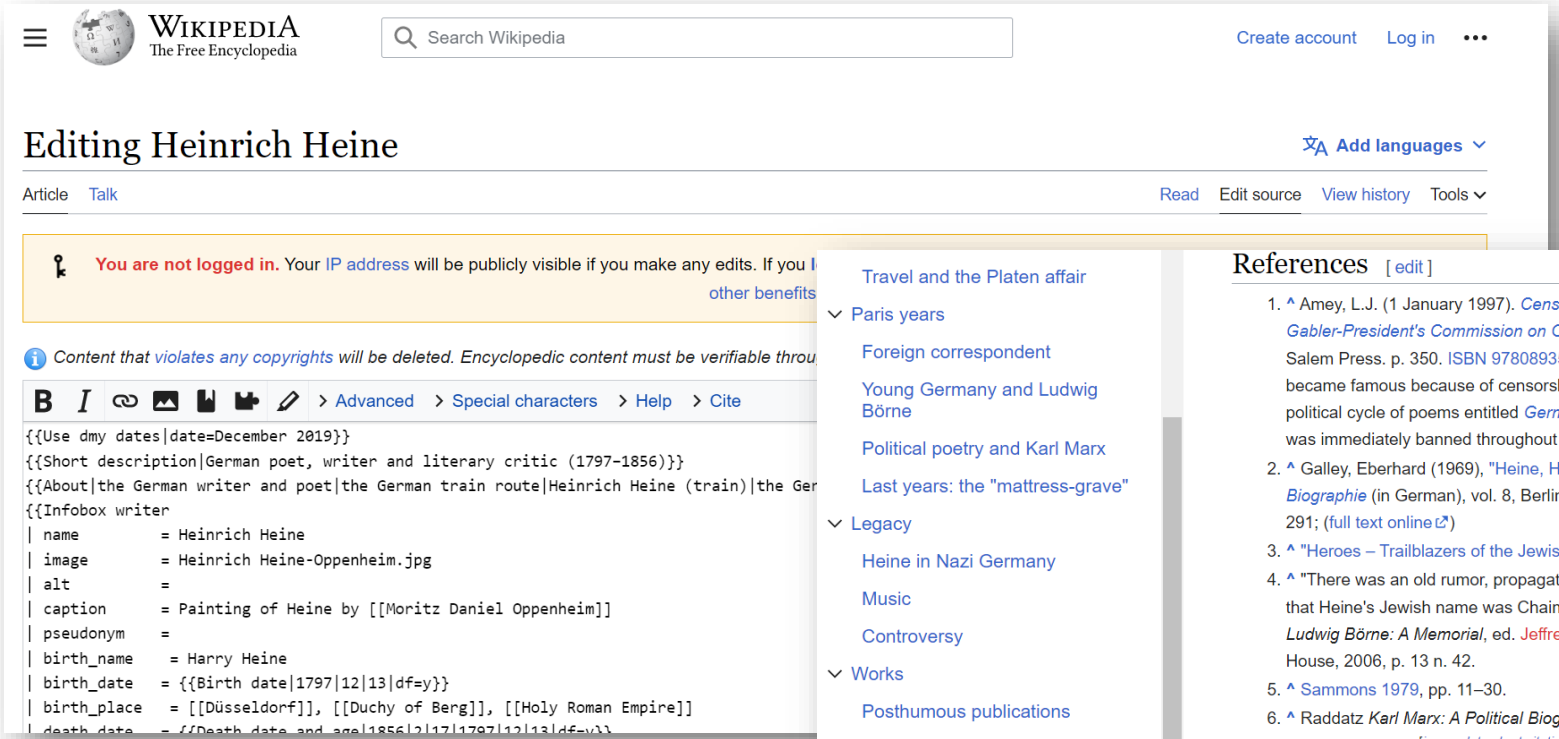


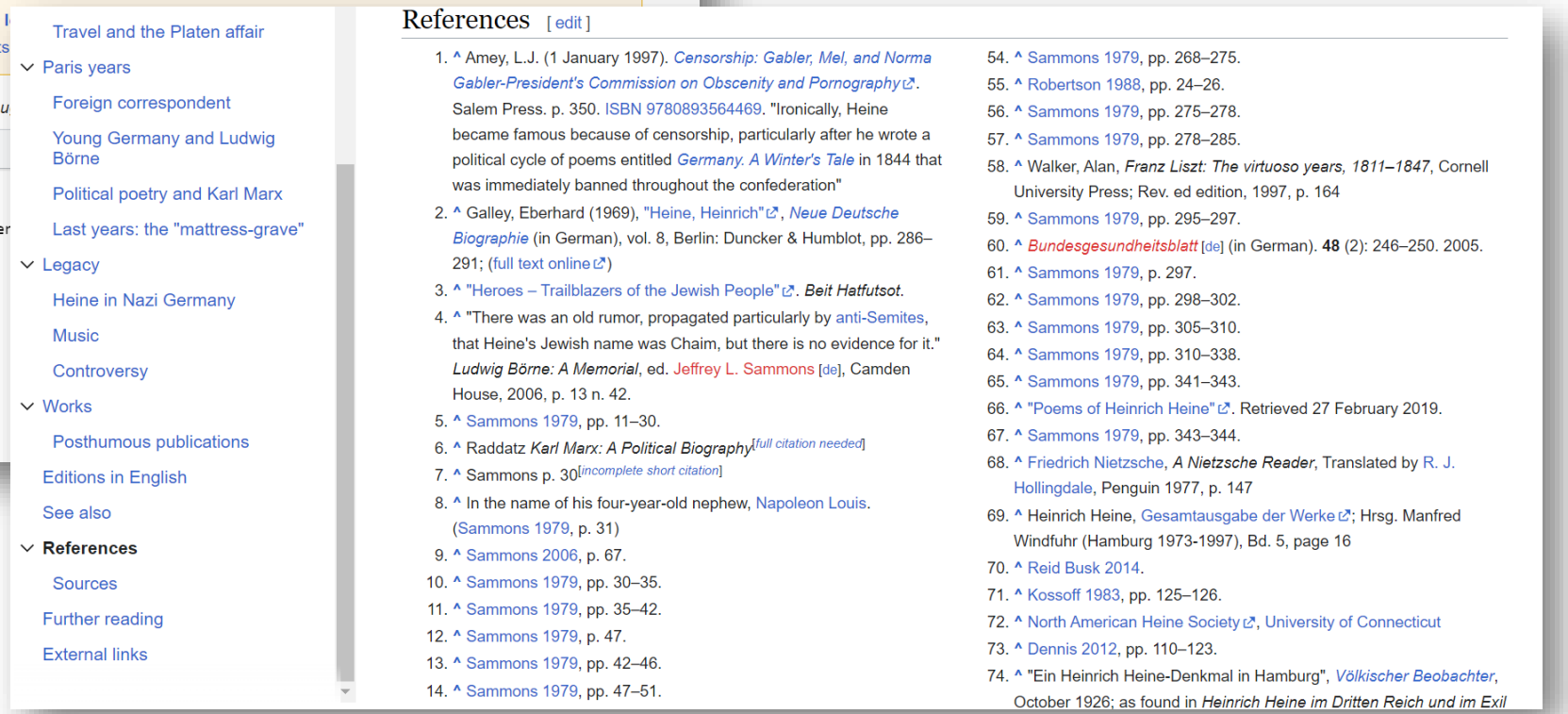
Datengrundlage:
Komplette “Edit Histories” der Referenzen der englischsprachigen Wikipedia (bis 2019).

Zagovora, O., Ulloa, R., Weller, K., & Flöck, F. (2022). 'I updated the <ref>': The evolution of references in the English Wikipedia and the implications for altmetrics. Quantitative Science Studies 3 (1): 147–173.

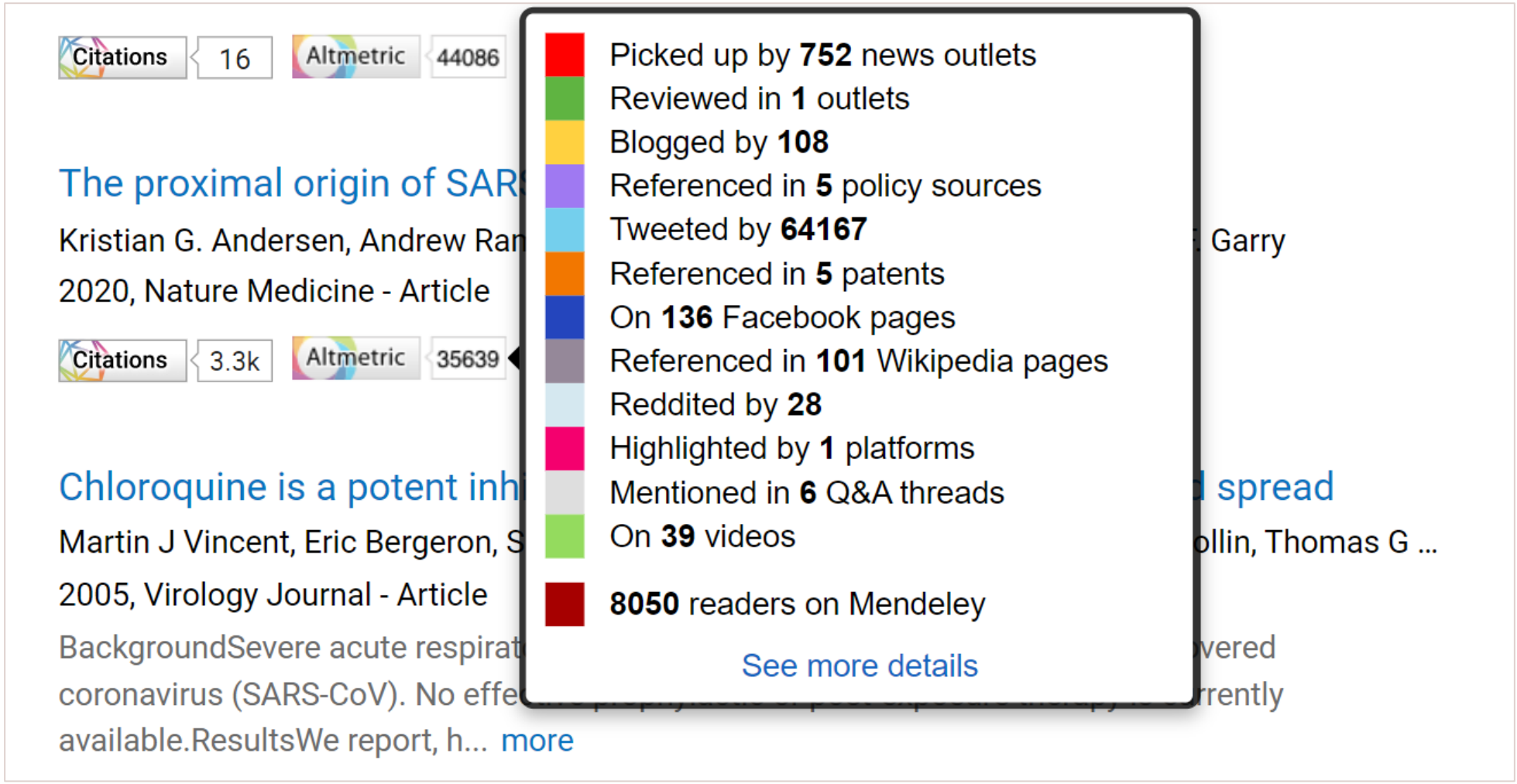
gesis Leibniz-Institut für Sozialwissenschaften CAIS RESEARCH FOR THE DIGITAL AGE

3. Inhalte/Daten
Beispiel: Referenzen in Wikipedia





3. Inhalte/Daten
Beispiel: Referenzen in Wikipedia



https://app.dimensions.ai/

Beispiel: Referenzen in Wikipedia

Dataset creation

Search for <ref> … </ref>

Considering article histories

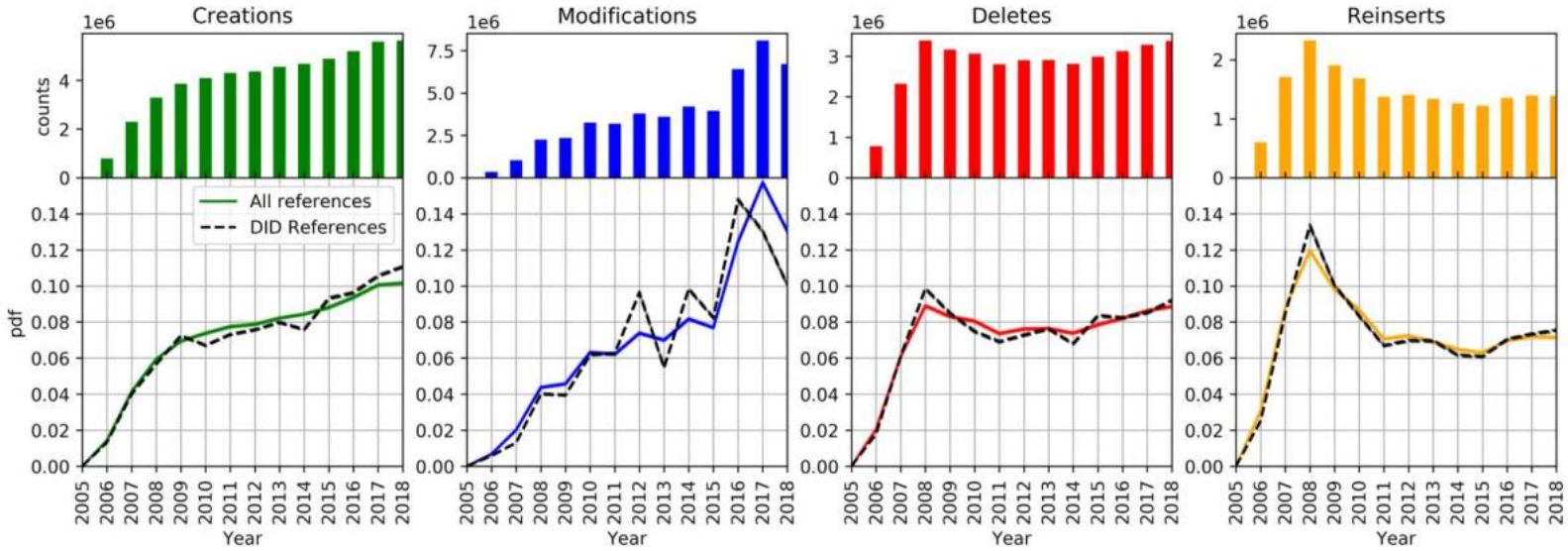
Dataset description

Dataset contains:

- 6,073,708 articles (English Wikipedia)
- 55,503,998 references
- with 164,530,374 actions (creations, modifications, deletions, reinsertions)

“Lost data”?

Around 31.8% of references had been deleted (without further reinsertions) between 2005 and 2019.



Distribution of actions over time.

3. Inhalte/Daten

Gelöschte und modifizierte Inhalte

- Wikipedia ist ein spezieller Fall durch die besondere Weise, wie editiert, gelöscht und wieder eingefügt wird.
- Gleichzeitig ist Wikipedia auch besonders transparent in Bezug auf vorgenommene Edits – das fehlt bei anderen Plattformen meist komplett.
- Andere Beispiele:
  - Gelöschte oder in bestimmten Ländern nicht verfügbare Videos auf YouTube.
  - Twitter accounts die von Twitter oder von den Nutzenden gelöscht wurden, oder auf “privat” umgestellt wurden. Twiter user handles, die von anderen Personen übernommen wurden.

gesis Leibniz-Institut für Sozialwissenschaften

CAS RESEARCH FOR THE DIGITAL AGE

4. Daten/Zugang
Veränderung der Zugänglichkeit

Computational research in the post-API age

Deen Freelon

University of North Carolina at Chapel Hill

Forthcoming in *Political Communication*

Keywords: API, computational, Facebook, Twitter, social media

2018-08-20

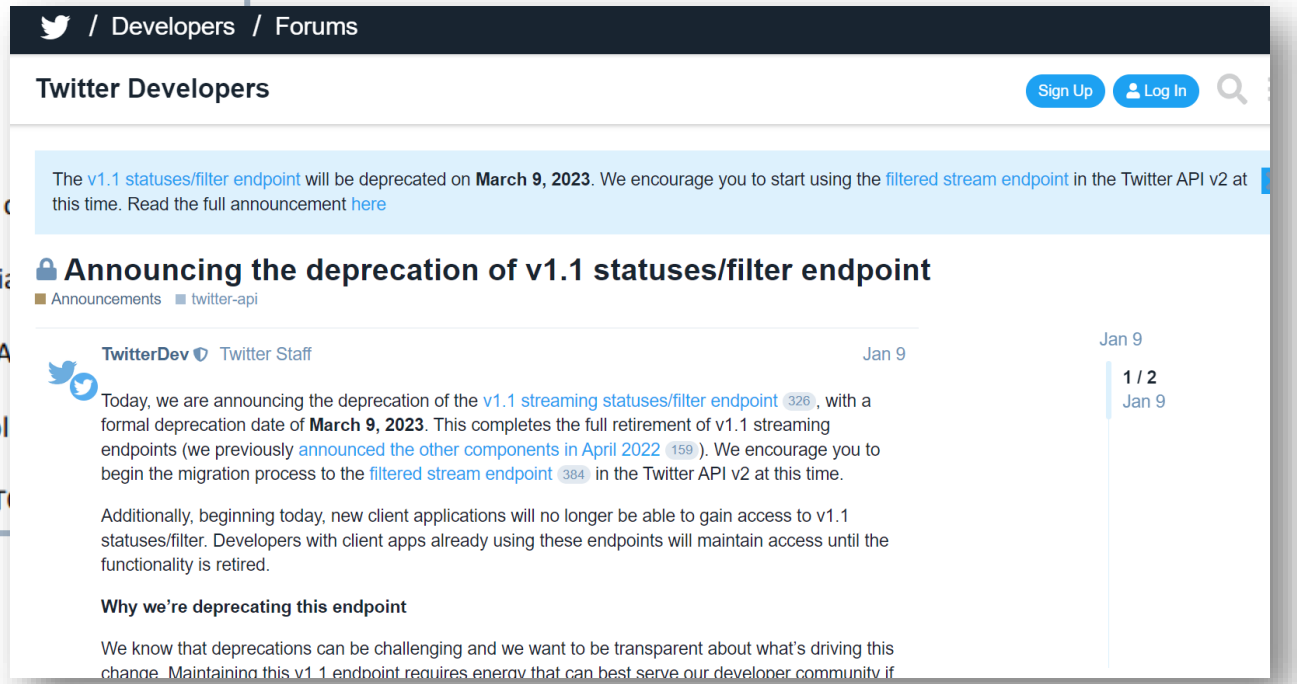
On April 4, 2018, the post-API age reached a milestone. On that day, Facebook

its Pages API, which had allowed researchers to extract all posts, comments, and associa

from public Facebook pages (Schroepfer, 2018). This decision followed the company's A

of its public search API, which provided searchable access to all public posts within a rol

window (Facebook, n.d.). The closure of the Pages API eliminated all terms of service (T



4. Daten/Zugang
Beispiel: Geocodierte Tweets

Wie können Twitter Daten für die wissenschaftliche Nachnutzung archiviert werden?

Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness

Pfeffer, Jürgen; Morstatter, Fred

. Datenfile Version 1.0.0, https://doi.org/10.7802/1166

Abstract: This dataset consists of IDs of geotagged Twitter posts from within the United States. They are provided as files per day and state as well as per day and county. In addition, files containing the aggregated number of hashtags from these tweets are provided per day and state and per day and county. This data is organized as a ZIP-file per month containing several zip-files per day which hold the txt-files with the ID/hash information. Also part of the dataset are two shapefiles for the US counties and states and Python scripts for the data collection and sorting geotags into counties.

Primärforscher/in, Institution: Pfeffer, Jürgen; Carnegie Mellon University | Morstatter, Fred; Arizona State University

Mitwirkende, Institution, Rolle: Zenk-Möltgen, Wolfgang ;GESIS - Leibniz Institute for the Social Sciences;Contact Person

Studiennummer: SDN-10.7802-1166

DOI: 10.7802/1166

Publikationsjahr: 2016

Downloads

Datensätze

Aktionen

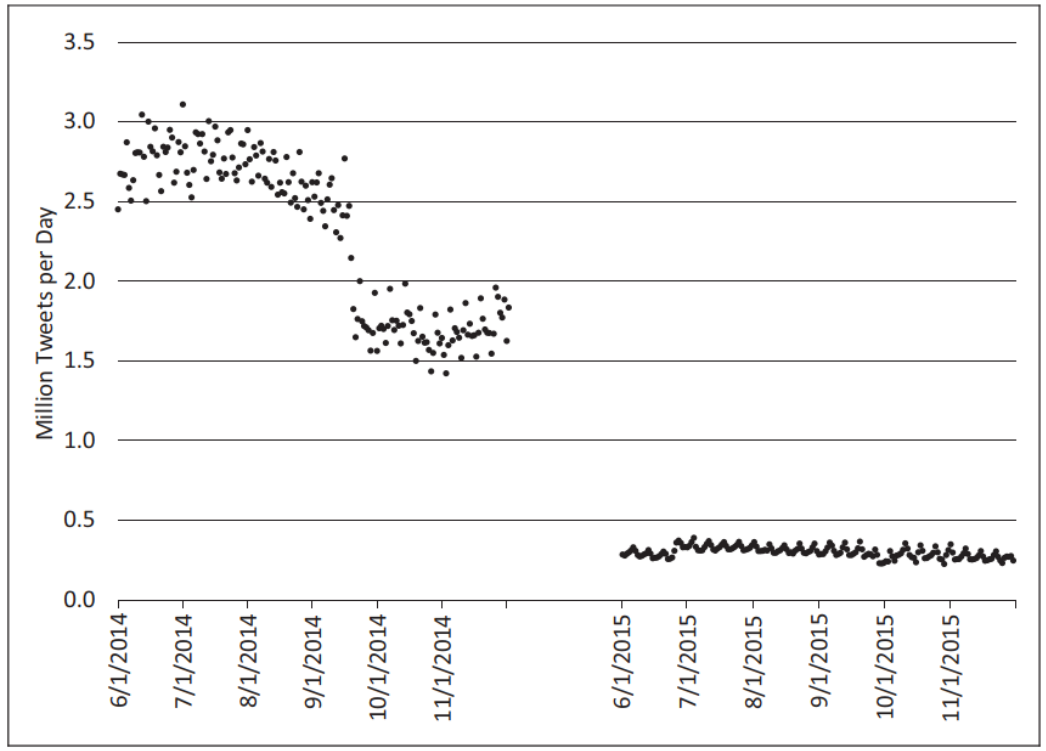
Merken

Zitieren

Datengrundlage:
Geo-codierte Tweets (mehrmonatige Sammlung von Tweets mit Geocodes in den USA aus 2014 und 2015)

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. Big Data & Society, 4(2).

4. Daten/Zugang
Beispiel: Geocodierte Tweets



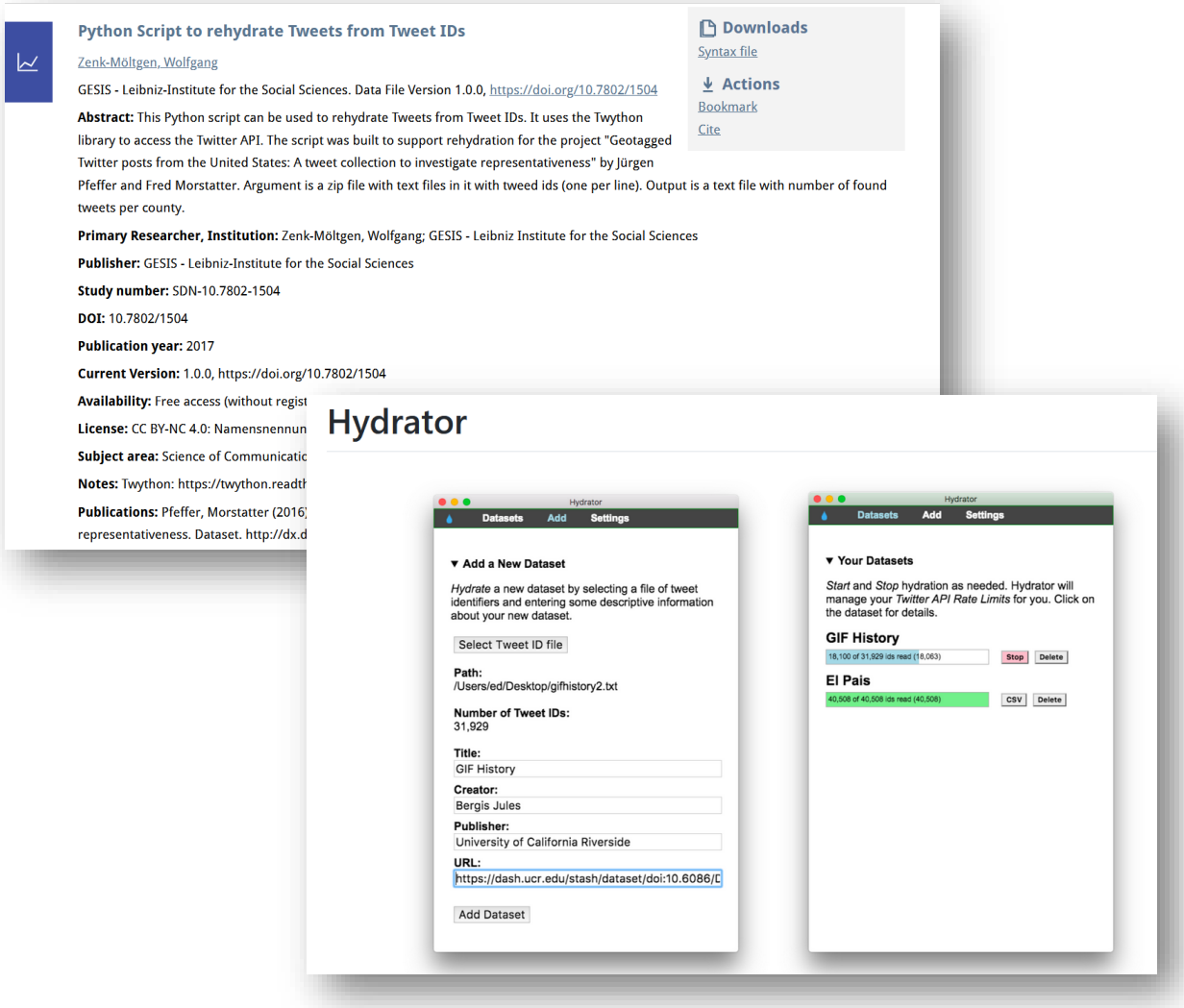
Anzahl gesammelte Tweets pro Tag

Veränderte Plattform-Affordances

Im April 2015 änderte Twitter die Art, wie Nutzende ihre Location mitteilen können.

Dies führte zu Veränderungen im Nutzungsverhalten aber auch zu veränderten Optionen bei der Datensammlung.

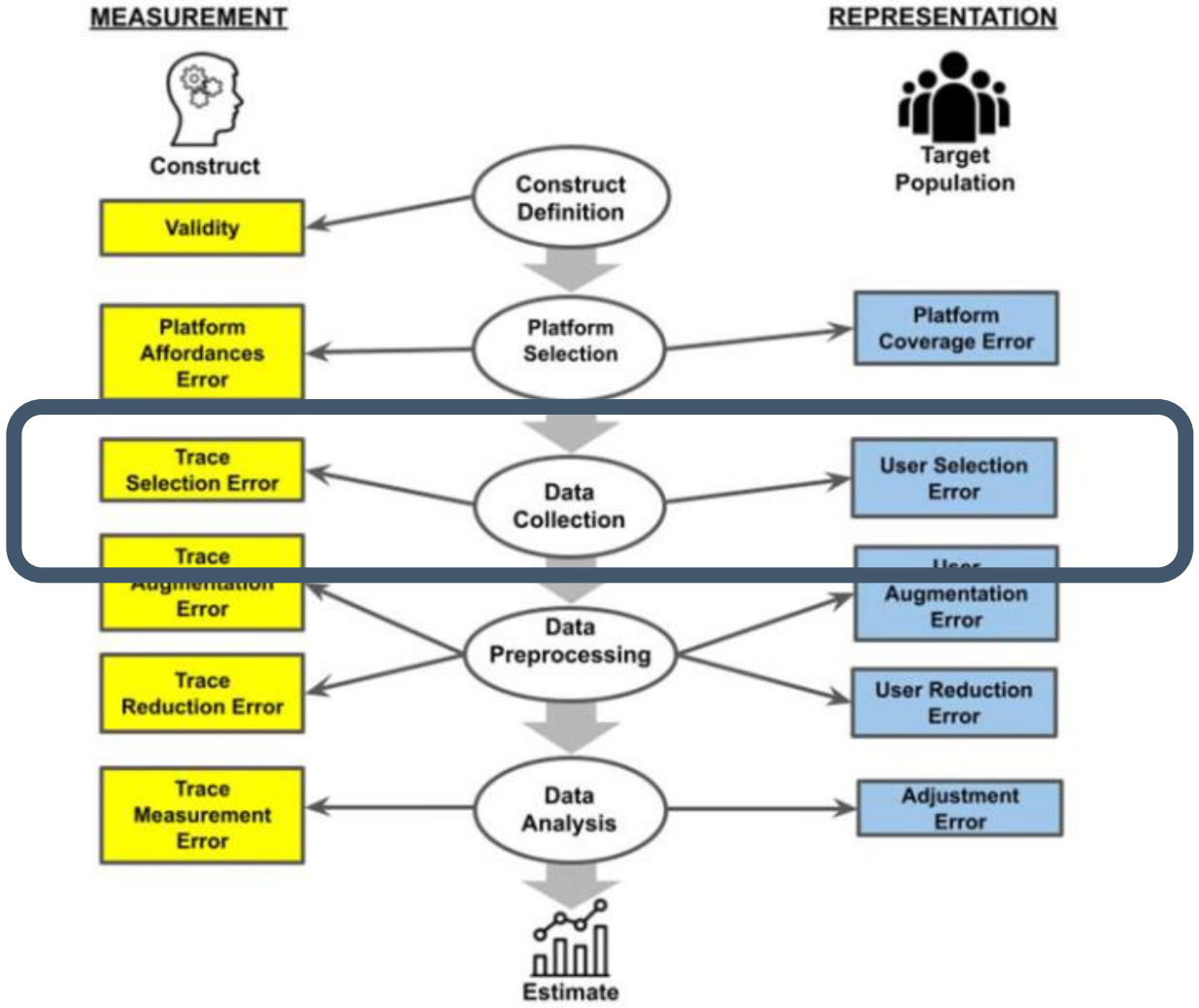
4. Daten/Zugang
Beispiel: Geocodierte Tweets



Zugang zu archivierten Daten

Mit Bezug auf die Twitter Terms of Services wurde entschieden, nur die Tweet IDs zu archivieren und zugänglich zu machen. Die vollständigen Tweets müssen dann für die Nachnutzung neu abgerufen (“rehydriert”) werden. Das ist jedoch für zwischenzeitlich gelöschte Inhalte nicht möglich. Die Änderungen an der Twitter API werden weitere Auswirkungen haben.

Ephemerality im TED-On Framework

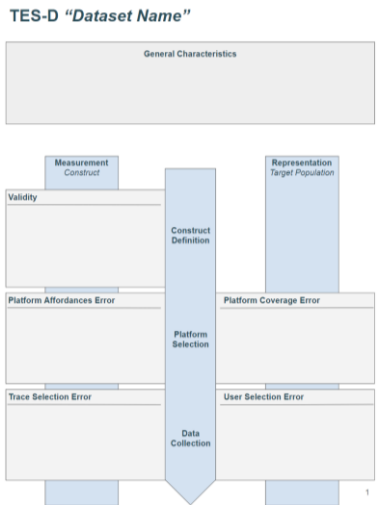


Beispiel: Data Collection

- **Trace Selection**: Der Fehler, dass nicht alle relevanten Inhalte gesammelt warden können, kann sich im Laufe der Zeit durch veränderte Zugangsoptionen (aber auch durch Löschungen) noch verstärken.
- **User Selection:** manche Gruppen von Nutzenden sind potentiell wahrscheinlicher von gelöschten Inhalten betroffen (Selbstlöschung oder Löschung durch die Plattform).

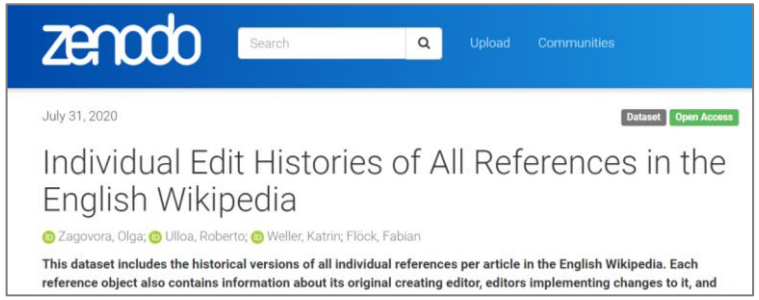
Lösungsansätze?

Dokumentation?



Total Error Sheets for Datasets (TES-D).
Work in progress bei GESIS

Benchmark-Datasets?



Vernetzung?

- RatSWD
- DGPuK
- Web Archiving
- …

EDMO Institute for Data, Democracy & Politics

Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access

Vielen Dank für die Aufmerksamkeit

Für Fragen / Anregungen / Rückmeldungen:

Dr. Katrin Weller

Leiterin Team „Digital Society Observatory“
Computational Social Science
GESIS Leibniz-Institut für Sozialwissenschaften
katrin.weller@gesis.org

Co-Leitung Team „Research Data and Methods“
Center for Advanced Internet Studies (CAIS)
katrin.weller@cais-research.de