Media and Communication
2026 • Volume 14 • Article 12018
https://doi.org/10.17645/mac.12018

cogitatio

ARTICLE

Open Access Journal

Twi-XL: An Infrastructure for Cross-Media Research in the Netherlands

Danielle A. Maxwell-Fraser [1], Iris M. Baas [2], Sarah Burkhardt [1], Olga Eisele [3], Anne C. Kroon [3], Robert Jan Bood [4], Iris Geldermans [5], Sophie Ham [5], Machiel Jansen [4], Rana Klein [6], Matthijs Moed [4], Roeland Ordelman [6], Erik Tjong Kim Sang [7], Mari Wigham [6,8], Marcel Broersma [2], and Julia Noordegraaf [1]

[1] Department of Media Studies, University of Amsterdam, The Netherlands
[2] Department of Media and Journalism Studies, University of Groningen, The Netherlands
[3] Amsterdam School of Communication Research (ASCoR), University of Amsterdam, The Netherlands
[4] Scalable Data Analytics, SURF, The Netherlands
[5] National Library of the Netherlands, The Netherlands
[6] Netherlands Institute for Sound & Vision, The Netherlands
[7] Netherlands eScience Centre, The Netherlands
[8] Huygens Institute for History and Culture of the Netherlands, The Netherlands

Correspondence: Olga Eisele (o.eisele@uva.nl)

Submitted: 30 January 2026 **Accepted:** 16 June 2026 **Published:** 13 August 2026

Issue: This article is part of the issue "Open Research Infrastructures and Resources for Communication and Media Studies" edited by Silke Fürst (University of Zurich), Johannes Breuer (Center for Advanced Internet Studies / University of Duisburg-Essen), Erik Koenen (University of Bremen), Dimitri Prandner (Johannes Kepler University of Linz), Christian Schwarzenegger (University of Bremen), and Christian Strippel (Weizenbaum Institute), fully open access at https://doi.org/10.17645/mac.i504

Abstract

In the domain of media and communication research, copyright and privacy issues provide significant obstacles to achieving the strategic goals of open science. Investigating the role of different media in shaping public debates is especially difficult, as research across different media collections and platforms is not yet systematically supported by open research infrastructures on a global scale. Moreover, the private ownership of crucial communication spaces such as social media results in unreliable access opportunities for researchers and unclear conditions regarding the completeness and integrity of data where platform owners provide their own archives. This also creates obstacles for users without a technical background who are interested in gathering information, exploring, or making sense of big data. In response to these challenges, a consortium of two universities (the University of Amsterdam and the University of Groningen), two national archives (the National Library of the Netherlands and the Netherlands Institute for Sound & Vision), and SURF as a national research IT supporter created Twi-XL: Designed to facilitate cross-media research, it provides a research environment in which users of varying computational skill levels can apply

automated methods to a range of archival collections, including websites, public broadcasts, radio and podcast transcripts, and over 10 years of Dutch tweets. The societal relevance of Twi-XL lies in creating a user-friendly tool that can be used by researchers to analyze societal phenomena across social media and traditional media, capturing how public debate unfolds across interconnected media environments. We describe and demonstrate the functional logic of Twi-XL with two use cases focused on: (a) social media affordances and users' practices of news sharing, and (b) the broader public debate on sexual misconduct. Going beyond the dominance of the English language in cross-media research, our results not only showcase the advantages of developing cross-media infrastructures specific to the Dutch language, but also how Twi-XL can be utilized in qualitative and quantitative research. Embedding Twi-XL in the broader discussion on transparency, reproducibility, and accessibility, we discuss the implications and feasibility of such infrastructures, particularly for analyzing social media content.

Keywords

cross-media research; digital heritage; digital research infrastructure; public broadcasting data; social media archives; text analysis; the Netherlands

1. Introduction

In current society, online communication and social media platforms are an inherent part of information exchange, and an important factor in shaping public debate. Over the last decades, various social and cultural phenomena such as polarization, radicalization, and activist movements have been researched and understood in connection with the rise of digital media (Lorenz-Spreen et al., 2022; van Dijck et al., 2018). However, each online communication space on its own only provides a partial view of public discourse, which unfolds not only across digital but also traditional media environments (Bennett & Pfetsch, 2018). Cross-media research, examining how information moves between and across online, social, and legacy media content, is therefore a fundamental precondition to comprehensively understand the dynamics of public debate.

It is, thus, vital for a healthy democratic system that researchers and public institutions are able to monitor the evolving dynamics of mediated public debate (Freelon et al., 2025). However, access to comprehensive sets of social media data or other types of online (media) content is difficult; researchers also do not have effective tools for cross-media research at their disposal. This creates challenges for the systematic analysis of debates that emerge and evolve across different communication spaces. Social media archive access largely depends on the goodwill of platform owners, and even with archives accessible, the integrity and completeness of the data archived are not guaranteed and are hard to assess or control. The conservation of other online content is challenging mainly due to its ephemerality and sheer volume, but also rapid technological developments inducing format obsolescence, where older formats quickly become unreadable by newer software (Chapekis et al., 2024; Masanés, 2006). Thus, a major challenge in an age of digitization is how online content can be stored and made easily accessible for external users (KB nationale bibliotheek, 2022).

The Twi-XL infrastructure project (Twi-XL, 2025) is, to our best knowledge, among the first with the explicit goal of facilitating cross-media research (but see Pehlivan et al., 2025, for a Canadian example focusing on social media platforms). Twi-XL partners include a team of developers from SURF, the Dutch IT cooperative

of education and research, and the National Library of the Netherlands (KB), the Netherlands Institute for Sound & Vision (NISV), and researchers from the University of Amsterdam (UvA) and the University of Groningen (RUG) who developed use cases to ensure a meaningful development of the infrastructure. Twi-XL facilitates access to archives of Dutch Twitter (2011–2023) and Dutch webpages, both hosted at the KB, as well as a collection of public radio and television broadcasts, and podcast transcripts located at the NISV. As such, Twi-XL's aims were formulated along (a) a *deep axis*, and (b) a *broad axis*. The deep axis aimed at making an existing Dutch Twitter archive publicly accessible and queryable to allow *deep* analysis of an important Dutch social media platform. The *broad* axis aimed at comparing media collections to facilitate cross-media research.

Twi-XL is part of a large-scale effort of the Dutch government to build digital infrastructure in the Netherlands. In the domain of the social sciences and humanities (SSH), this effort is overseen by the SSH-Council (2026) and the temporary funding Platform Digitale Infrastructuur for the domain Social Sciences and Humanities (2025). Twi-XL is part of the national research infrastructure for digital humanities research (Common Lab Research Infrastructure for the Arts and Humanities [CLARIAH-NL], 2026b). It also integrates components from the national digital research infrastructure for the Social Sciences (Open Data Infrastructure for Social Science and Economic Innovations [ODISSEI], 2025). While set up in the humanities, the Twi-XL project was deliberately planned as an interdisciplinary effort, also directly involving researchers from the social sciences. Via CLARIAH-NL and ODISSEI, Twi-XL is also connected to the European Open Science Cloud. There, the functionalities of Twi-XL may connect to components developed for accessing similar collections across Europe. These include, e.g., the German Twitter archive (Mayer, 2023), the Luxembourg web archive (Bibliothèque nationale du Luxembourg, 2017), or the audiovisual collections at the French Institut national de l'audiovisuel (2026). Moreover, publicly available social media datasets have been developed, such as the TGDataset covering a large variety of Telegram channels (La Morgia et al., 2025). Such collections and datasets are stored individually and can, given data sensitivity and copyright concerns, often only be accessed on site; analysis also often requires advanced programming skills. By making the key components of Twi-XL publicly accessible via the SSH Open Marketplace (2025), Twi-XL serves as an example of a secure online environment, adaptable for cross-media analysis across Europe, including for users with only basic programming skills.

This article maps the collections accessible via Twi-XL, covering their contents, technical aspects, workflows, and ethical and privacy-related considerations. Then, the use of the infrastructure is showcased based on the two PhD projects developed within Twi-XL. The first links user identities to news sharing behavior on Dutch Twitter (based on Baas et al., 2025), showing that news sharing remains overall centered on mainstream outlets promoted by super spreaders. The second study focuses on the broader debate on sexual misconduct compared across Dutch Twitter and public television broadcasting (based on Burkhardt & Poell, 2026), revealing a strong divergence between feminist framings in public broadcasting and anti-immigration framings on Twitter. Concluding remarks highlight the contribution of Twi-XL and discuss opportunities and challenges of building digital infrastructures for cross-media research.

2. Introducing the Components of the Twi-XL Infrastructure

In the following, we first introduce the Secure Analysis Environment (SANE), which ensures that even sensitive data can be analyzed safely, accommodating privacy and copyright-related restrictions (SURF, 2024). We then

discuss the Twitter collection and the Twi-XL application programming interface (API), as well as the collections of websites and audiovisual material within the Twi-XL project (see Table 1 for an overview).

Table 1. Overview of collections accessible via the Twi-XL infrastructure.

Data controller	Collection contents	Time frame covered	Accessed via	Point of contact
KB in cooperation with SURF	Dutch Tweets (text only)	2011–2023	Twi-XL API, (Tinker or Blind) SANE	KB Datalab (KB, 2026)
KB	Websites of selected entities	1992–ongoing	Blind SANE	
NISV	Audiovisual media collection; public radio and television broadcasts and podcasts	1875–ongoing	CLARIAH Media Suite, Tinker SANE	CLARIAH Media Suite via SURF Conext; SANE via NISV Data Platform (CLARIAH-NL, 2026a; NISV, 2026)

2.1. SANE: Secure Analysis Environment

The desire to create a platform for enabling the sharing and processing of data in an efficient and secure manner gave way to the SANE project. SANE stands for Secure ANalysis Environment (van der Meer, 2023) and is, like Twi-XL, a project funded by the Platform Digitale Infrastructuur. SANE is a virtually shielded computing environment that is still in development. It enables access to and secure analysis of sensitive data, for research purposes, in a controlled and accountable way (SURF Cooperation, 2024). SANE comes in two variants: Blind SANE and Tinker SANE. Using Tinker SANE, the researcher accesses the data via a secure remote desktop that is isolated from the general web, allowing close reading and large-scale analysis of the data. In contrast, Blind SANE literally blinds the researcher, who cannot see the data but only the output of their queries. After the data provider deposits the dataset and installs tools as requested by the researcher, a secure analysis environment is launched, and access is given to the researcher. Upon completion of the analysis, the researcher submits the results for review by the data provider to ensure compliance with the agreed conditions, and upon approval, the results are released (van Dijck et al., 2018).

Overall, SANE provides a privacy-protecting, secure, and ethically grounded environment for data analysis. Aiming to create a common space for the analysis of different media collections, Twi-XL's agenda creates challenges regarding data security, covering a wide range of collections ranging from privacy-sensitive social media data to in-copyright collections of web and public broadcasting data. SANE allows us to solve these issues in a centralized and standardized manner. However, it also comes with constraints for the researcher, since the environment has to be set up entirely before users obtain access and is then (still) somewhat difficult to adapt to emerging needs in hindsight. The Twi-XL API (see following section) provides a public GitLab repository for sharing research code or feedback, but SANE lacks a central reference space to find, share, or reuse Twi-XL-related knowledge or code across use cases. In that sense, the SANE environment deliberately constrains flexibility and ad hoc adaptation, illustrating the tension between privacy and copyright protection on the one hand, and the implementation of open science principles as well as analytical flexibility on the other.

2.2. TwiNL and the Twi-XL API

The Twitter archive included in Twi-XL was originally created in the TwiNL project. Starting in 2010, the TwiNL project had the explicit purpose of creating a centralized archive of Dutch Twitter data, providing a resource for research on the Dutch language, communication patterns, and social trends. The collection contains over 4.5 billion Dutch tweets (Tjong Kim Sang, 2022). TwiNL aimed to create information about tweets written in Dutch for researchers without breaking the Twitter developer rules, as well as making it easier for users to get access to information about tweets (Tjong Kim Sang & van den Bosch, 2013). TwiNL was then included in the Twi-XL project in 2021, with the purpose of developing tools for querying it effectively to enable a deep exploration and analysis of Dutch Twitter (Twi-XL, 2025).

The TwiNL collection includes the ID and metadata of the tweet (Tjong Kim Sang, 2022). To ensure privacy of Twitter user identities, in line with the EU's General Data Protection Regulation (GDPR; see European Commission, 2026), the entire TwiNL archive is pseudonymized, i.e., no usernames or screennames are provided. In addition, the timestamp, list of hashtags, and URLs are collected. The selection of the included tweets is based on a matching process based on a list of frequent Dutch words (Tjong Kim Sang & van den Bosch, 2013). This matching process also captures languages closely related to Dutch, such as Frisian. The TwiNL collection only contains text, excluding engagement measures or audiovisual data (Tjong Kim Sang, 2022). It runs from 2011 to 2023, when the TwiNL data collection had to be halted: Under Elon Musk, X (formerly Twitter) now monetizes and controls data access, meaning there is no longer free access to the API, making data collection on the platform practically unfeasible (Gotfredsen, 2023).

To effectively query the large Twitter collection, our Twi-XL partners at SURF developed the Twi-XL API. An API enables a user to send a request to a coordinator that provides access to stored data, after which the coordinator returns the requested data. This enables Twi-XL users to search for tweets in a specified time frame based on search operators such as keywords, phrases, user handles, hashtags, or hyperlinks (Tjong Kim Sang & van den Bosch, 2013). Before Twitter's own API was closed in 2023, the Twi-XL API only gave back the tweet ID instead of the full text to ensure GDPR compliance. The tweets then had to be "re-hydrated," meaning they were re-accessed through the official Twitter API to retrieve full text and metadata such as retweets and user biographies. Since the Twitter API is now practically closed, this "re-hydration" is no longer possible. To compensate, methods for aggregate and visual analysis outputs were built into the Twi-XL API, allowing the user, for example, to count mentions of terms over time or analyze posts of specific Twitter handles (e.g., of figures of public interest, such as politicians). Raw text can also be retrieved via the API but is subject to more restrictive access policies, given that sensitive data may be contained in tweets and users may be reidentified via original tweets. While the Twi-XL API currently requires internet access, a future aim of Twi-XL is to enable the use of the API within a SANE environment; this would also allow qualitative inspection and close reading of tweets in a Tinker SANE environment.

The KB functions as the data controller for the TwiNL collection. Users can access the TwiNL collection by contacting the KB Datalab, explaining their specific project and justification for using the archive (see also Table 1). Then, a user agreement is set up specifying conditions for research. The user must also provide evidence of ethical approval by the ethics and privacy committee of their institute or university. Thus, a clear research framework and conditions need to have been set up and officially approved. Once this process is completed, the user will receive an API token and online documentation on how to use the API; if needed, a

SANE environment will be set up. A monitoring procedure is in place to allow withdrawal of the API key in case of improper use of the API on the side of the researcher.

2.3. The Web Collection of the KB

In addition to the TwiNL collection, the KB also hosts the largest web collection in the Netherlands (KB, 2026). Starting in 2006, the KB began harvesting and archiving a collection of websites reflecting Dutch language, culture, and history. Each website is selected manually by KB researchers based on the criteria of innovative content, popularity, and relevance to the Netherlands (Geldermans, 2025). The collection includes over 23,600 website files, including the Netherlands' oldest website from 1992, and one of the first home pages in the Netherlands from 1994 (KB, 2026). The KB also harvests websites that can be relevant for future research, including topics such as news websites, the Covid-19 pandemic, the Dutch blogosphere, Chinese-Dutch relations, social criticism, and the LGBTQIA+ community (KB, 2026).

The KB's web collection is archived through the Web Curator Tool in combination with software to automatically browse and download webpages—the WebCrawler Heritix (Geldermans, 2025; Geldermans et al., 2025). The Web Curator Tool is a user-friendly application that helps non-technical users manage harvesting projects and handle metadata (Geldermans, 2025). Regarding creators' consent for the harvesting of their webpages, the KB works with an opt-out model: The KB notifies website creators when their website is harvested and asks them to notify the KB if they wish to withdraw consent. Websites are harvested once per year until the website stops running (Geldermans, 2025).

The KB has to consider a range of ethical, privacy, and technical concerns in its harvesting process. For one, harvesting contemporary websites brings particular challenges, as modern web content frequently contains highly personal information. These datasets are often too large to anonymize completely. Considering GDPR requirements, harvesting new websites requires the library to make a careful assessment of legal obligations regarding what is considered to be sensitive information, causing technical limitations. Additionally, sensitive data change over time, especially considering the political and societal developments that may increase the vulnerability of groups in Dutch society. The KB has the responsibility to remain open and flexible to how websites are harvested, stored, and made accessible. When content becomes highly sensitive, alternative access measures can be applied, such as locking the website for a set number of years, as was historically done with newspapers and archival material. In doing so, the KB aims to balance public access with the protection of individuals.

Researchers who wish to use the KB's digital collections for research do so under the same conditions as described for accessing the Twitter archive. They can request data and analyze them on their in-house workstation at the KB Datalab (see also Table 1). This allows researchers to view collections and apply text and data mining. Through the Blind SANE environment, where data remain completely hidden from users' view, researchers can now also access the KB's web collection remotely (Geldermans et al., 2025). Once the KB uploads web archive data in the SANE environment together with the requested software setup, researchers can mine the KB collection and later request an extraction of results from the administrator of the environment. Due to copyright and legal restrictions, the KB limits online access to its web archive, meaning researchers can only analyze archived content without fully visualizing a website. Nevertheless, the possibility of remotely accessing and analyzing archived content represents a significant advancement, as

the web archive was previously only available for consultation on site. Further improvements will consider making a few, much sought-after thematic collections permanently available in a standard SANE setup, which can be more rapidly provided to researchers.

2.4. The Audiovisual Media Collections of NISV

The NISV contributes to Twi-XL by providing researchers access to its audiovisual collections of public radio and television broadcasting and its curated collection of podcasts. The NISV audiovisual collection consists of approximately 70% of the Dutch audiovisual heritage: documentaries, television, radio, sound recordings, photographs, objects, books, scripts, logbooks, memorabilia, and personal papers (European Federation of Video Game Archives, Museums and Preservation Projects, 2026). NISV is tasked with preserving Dutch cultural heritage, meaning that the archived material must be produced or recorded by Dutch citizens, whether in the Netherlands or abroad (e.g., the broadcasts of the Wereldomroep, the Dutch international broadcasting organization). The NISV attempts to represent all facets of Dutch society, collecting a broad set of genres in relation to culture, religion, politics, society, history, and technology (European Federation of Video Game Archives, Museums and Preservation Projects, 2026).

From a technical perspective, the NISV uses a policy based on the Open Archival Information System. An Open Archival Information System considers how to manage archival work for long-term preservation, whilst making it accessible to users (ISO, 2025). The catalogue is supported through a metadata model created by the NISV, which connects different aspects of digital assets to the collections' descriptions, as well as administrative and location information. With the development of more advanced technologies, the NISV has started to use (semi-)automatic annotation and cataloguing techniques, i.e., automatic generation of metadata based on speech and image characteristics and automatic speech recognition (ASR) to generate transcripts (Ordelman & van Hessen, 2018). The ASR feature is available for parts of NISV's public broadcasting records and represents an alternative to relying on curated archival metadata, which provides only a partial perspective on the content. Moreover, it allows researchers to analyze the content of public broadcasts at scale, using text-based computational techniques within Twi-XL.

Outside the Twi-XL infrastructure, the NISV archives are accessible via the CLARIAH Media Suite: a virtual research environment in the Dutch national infrastructure for digital SSH (Ordelman et al., 2018). The Media Suite is accessible via SURFconext, a central authentication service developed by SURF, which is available to, e.g., researchers at Dutch universities but also employees of other institutions (SURF, 2026). It contains tools for metadata inspection, search, playback of audiovisual material, visualization, annotation, and bookmarking of items and queries in a personal workspace. As such, the Media Suite can serve as a first step in exploring the NISV archive, also supporting qualitative approaches to audiovisual analysis, but it does not enable large-scale text analysis. Within Twi-XL, a large-scale analysis of transcripts is possible via the Tinker SANE environment, extending researchers' opportunities by enabling efficient and secure analysis of much larger corpora (NISV, 2024).

For accessing NISV's audiovisual archives in a SANE environment, researchers typically contact the NISV data platform (NISV, 2026) under the same conditions as described for the KB (see also Table 1). Rather than providing direct access to the data, requested metadata and ASR files are transferred into a Tinker SANE environment. This allows direct inspection while data remain protected and subject to governance by the

data provider. NISV is one of 60 Trustworthy Digital Repositories worldwide, which have received the Data Seal of Approval (NISV, 2016), demonstrating its commitment to legal, ethical, and privacy standards.

3. Showcasing the Twi-XL Infrastructure

Within Twi-XL, we implemented a co-creation framework (e.g., Felt et al., 2023). This was done through two PhD projects, several workshops, and university teaching in both the humanities and the social sciences, which yielded substantive feedback for the infrastructure project and for the use of SANE in particular. The PhD researchers from UvA and RUG were closely involved in developing the Twi-XL infrastructure to make sure that it serves the needs of researchers in the SSH. In the following, two studies from these PhD projects are showcased which focused on Twi-XL's broad and deep axes: a study of link-sharing patterns across broad news outlets and political identities on Twitter, conducted using the Twi-XL API outside SANE (Baas et al., 2025; Section 3.1); and a longitudinal study on the impact of feminist perspectives on sexual misconduct discourse on Dutch Twitter and television (Burkhardt & Poell, 2026), conducted within a SANE environment (Section 3.2).

3.1. Social Media Affordances and Users' News Sharing Practices on Dutch Twitter

Baas et al. (2025) investigated who shares news on Dutch-language Twitter by analyzing how users describe themselves in their Twitter biographies and whether such self-ascribed identities can be meaningfully grouped and linked to patterns of news sharing. This was done through computational text analysis of user biographies from tweets containing news-related URLs. While the substantive contribution concerns identity and news circulation, the paper serves as a case study of how infrastructures such as Twi-XL enable large-scale social media research without dependence on platform-provided data access. However, the affordances of the infrastructure itself simultaneously structured the scope, scale, and epistemic orientation of the research, shaping both its analytical possibilities and its limitations.

The study relied on a large-scale, language-specific dataset of tweet IDs, which were retrieved via the Twi-XL API. The authors collected a stratified random sample of one million Dutch-language tweets from 2021 that contained at least one URL. The infrastructure retrieves tweets in monthly "buckets," with bucket sizes adjusted to the volume of tweets available per month. This design ensures temporal coverage across the entire year rather than overrepresenting periods of high activity. The study mainly focused on the Netherlands and Flanders, with limited spillover from other Dutch-speaking regions. Thus, the Twi-XL infrastructure enables a geographically and linguistically coherent case study of news sharing in a relatively bounded media system.

When this research was carried out, only tweet IDs were returned by the Twi-XL infrastructure, meaning all tweet IDs had to be re-hydrated. At the time of analysis, this also helped to determine which tweets were still publicly available. Only tweets that could be successfully re-hydrated, and for which user metadata (including biographies) could be retrieved, were included in the study; tweets that were no longer accessible were discarded. Following this procedure, the authors retained 392,435 tweets, meaning that over half of the originally sampled tweets could not be recovered, those which had likely been removed by the platform or users themselves. The researchers implicitly accepted this loss as a structural feature of contemporary

platform research. Twi-XL thus exemplifies how platform compliance produces datasets that are large but incomplete, and how research designs must accommodate this condition.

All URLs were extracted from tweets and subsequently compared with a list of news sources. The analysis started with any tweet containing a URL, rather than restricting the dataset to URLs from predefined news outlets. This inclusive strategy served a dual purpose. First, it aligned with the affordances of the Twi-XL infrastructure, which enables large-scale collection of URL-containing tweets without knowing or classifying tweet contents. Second, and more importantly, it allowed the researchers to assess empirically what proportion of the infrastructure's URL-traffic actually consisted of news content.

On the basis of this comprehensive, curated URL set, the authors then operationalized "news outlets" through a multi-step filtering process. Links redirecting to social media platforms, personal blogs, and webshops were excluded. Following this, the most frequently shared remaining domains were manually inspected to determine whether they could be classified as news, making their primary function to inform audiences about current events. This procedure resulted in the identification of 52 relevant news websites. A small number of sources accounted for the majority of shared news links, and therefore the analysis focused only on the 35 most shared outlets, which together represented over 70% of all news URLs in the dataset. By restricting the analysis to the top 70% of shared news URLs, the authors made an explicit trade-off between representativeness of the research and feasibility of the analysis. This choice is methodologically justified but also crucial for understanding the study's conclusions: The relative homogeneity of shared news sources across identity groups is partly a result of focusing on the infrastructural "core" of the news ecosystem, i.e., the most visible news sources, rather than its periphery.

From the full tweet set, the authors identified 11,417 unique users who shared at least one of the top 35 news sources and extracted their biographies for analysis. Those were analyzed through unsupervised natural language processing (NLP) techniques: Twitter bios are extremely short texts, and in the context of this research, they were analyzed without contextual cues such as profile images or follower networks. The authors therefore relied on text-based methods, namely contextual topic modeling and hierarchical clustering, to extract structure from sparse data. Given the large-scale dataset, computationally intensive methods such as zero-shot contextualized topic modeling trained on BERTje, a Dutch language model (de Vries et al., 2019), were employed to run the experiments in a secure research cloud. Extensive normalization was performed on the raw text, such as lemmatization, stop-word removal, and the replacement of politically or institutionally salient terms with placeholders (e.g., POLITIEK_L for left-wing political parties and POLITIEK_R for right-wing political parties). These steps are essential for enabling topic modeling on short texts.

The analysis of Twitter users who share news based on their self-ascribed identities revealed five larger categories of self-ascribed identities. As shown in Figure 1, these include politically oriented users, news creators, science and research professionals, lifestyle and culture-oriented users, and "general users," of which the latter includes all bios that could not be categorized otherwise. Right-leaning political users in particular tend to state their political affiliation in biographies, often through "anti" statements (e.g., "I am NOT a left voter"; Rogers & Jones, 2021). Moreover, the findings support the notion of identity hierarchy, where certain identities become more salient than others depending on context. This phenomenon is, thus, not only specific to journalists and political actors, as the literature suggests (Brems et al., 2017; Molyneux

et al., 2018; Sillesen, 2015), but extends to other highly active news sharers as well. So-called "supersharers" similarly combine multiple social and professional roles in their bios, indicating that identity salience on Twitter is not limited to institutional actors. Although the results align with the idea of a political "mega-identity" (Rogers & Jones, 2021), they also demonstrate that identity formation among news sharers is more heterogeneous and cannot be reduced to political or professional categories alone. In terms of limitations, only users who shared news were analyzed, and only a subset of news sources was included in the analysis, which is implicitly defined also by the boundaries of the infrastructure. Future research could combine biography analysis with tweet content or longitudinal data, using the data available within the Twi-XL infrastructure. In this sense, the paper functions not only as an empirical study but also as a proof of concept for what this infrastructure enables researchers on news sharing to explore. While the "re-hydration" of tweets is no longer viable, Twi-XL now compensates for the unavailability of the Twitter/X API by making tweets available under secure access conditions, and this case study delivered important feedback for these technical developments.

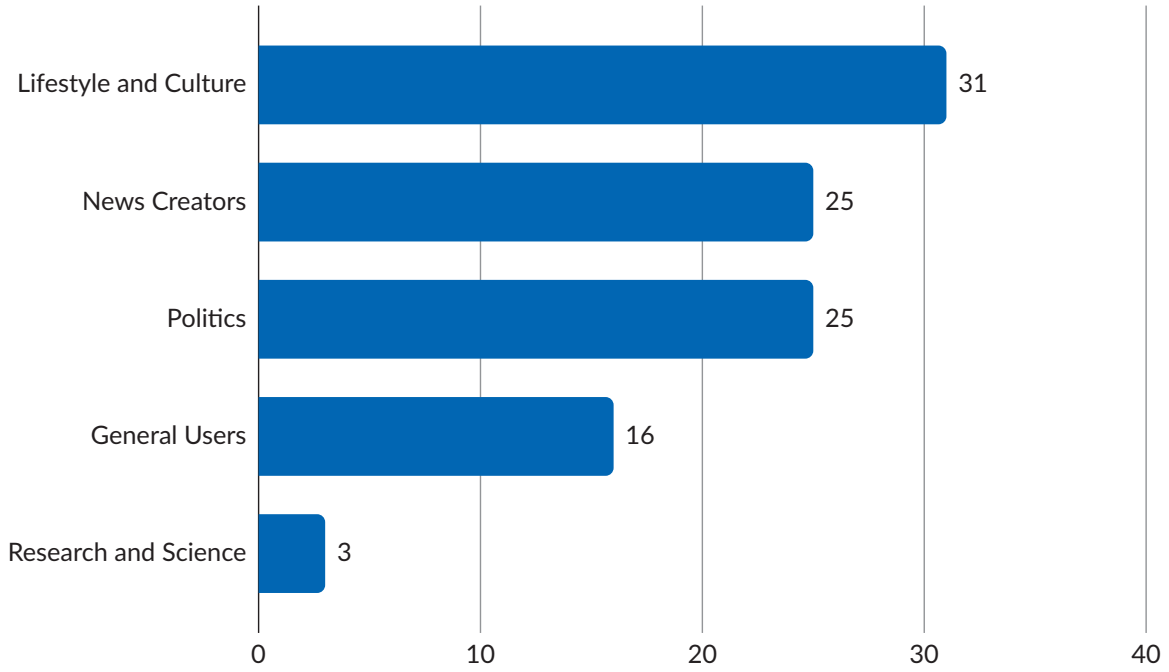


Figure 1. Five macro-categories and percentages of self-ascribed identities of Twitter users sharing news.

While this research examined how the Dutch news spectrum relates to political identities in one crucial online communication space, a cross-media approach to news-sharing behavior across platforms is a promising avenue for future research. A relevant point of reference is the study by de Keulenaar et al. (2025), which combines Dutch public television transcripts from the NISV archive with Twitter and Facebook data to analyze audience responses to immigration coverage. Whereas such cross-media analyses have typically depended on ad hoc data integration and bespoke analytical workflows, Twi-XL provides the infrastructural conditions under which comparable news-related case studies can be conducted in a more systematic and reproducible way.

3.2. Feminist Perspectives on Sexual Misconduct Discourse on Dutch Twitter and TV

Burkhardt and Poell (2026) investigated how sexual misconduct was represented and communicated in the Netherlands over time, focusing specifically on Dutch public television broadcasting and Twitter. Conceptually, the research built on feminist perspectives to investigate how public discourse of sexual misconduct has historically been shaped by different feminist waves in the Netherlands. Methodologically, Twi-XL was utilized via APIs to access and systematically organize tweets and television transcripts related to this debate. The paper finds that Dutch television adhered to a continuous feminist-leaning agenda over 35 years (1984–2018), framing the issue as an internal and complex structural concern and centering on personal testimonies of victims. However, broadcasters' feminist perspectives were constrained by their focus on the majority white Dutch population and framed within a dominant rhetoric of national health and citizen wellbeing. Dutch Twitter (2011–2018), in particular since 2015, by contrast, became a popular stage for far-right narratives that linked sexual misconduct to immigration and Islam, framing the issue as an external threat.

Besides highlighting key differences in feminist framings of sexual misconduct on Twitter and public television, the findings illustrate how Twi-XL enables researchers to investigate longitudinal and broader cross-media dynamics in terms of agenda setting and discourse. Dutch public broadcasting and Twitter were found to mediate the issue in relative isolation from each other. The #MeToo hashtag sparked explicit attention to feminist activism on Dutch television, while the hashtag and feminist activism did not appear in the dominant Twitter discourse. On television, viral social media discourse was only integrated when aligning with its established feminist-leaning agenda. At the same time, broadcasters continuously focused on the feminist debunking of perpetrator myths dating back to the feminist second wave era, rather than tackling the Islamophobic perpetrator myths that had turned hypervisible on Twitter. Such gaps in mediation not only risk further societal fragmentation but also create leeway for the observed radicalization online. However, cross-media dynamics may unfold differently for other debates, where Twi-XL allows for their systematic investigation in a national context across different issues, providing crucial insights into the politics and power structures underlying different media systems.

Methodologically, Twi-XL was used to retrieve the data for this study by querying the TwiNL database and the NISV's television collection via protected APIs. The Dutch search query was designed broadly to capture a wide and diverse discourse around sexual misconduct, using regular expressions of terms such as rape, sexual harassment, sexual intimidation, sexism, or transgressive behavior. For Twitter data, the Twi-XL API was queried for a time span between 1 January 2011 and 31 December 2018, amounting to a total of 1,251,203 tweets. For searching through television records, NISV's ASR feature was used (Ordelman & van Hessen, 2018). Querying the ASR television transcripts from 1 January 1984 to 31 December 2018 amounted to 8,378 broadcast transcripts.

Twi-XL offers a high degree of flexibility for researchers to develop computational methods in alignment with their research agenda, which proved essential for optimizing the final sample for the analysis. While this study relied primarily on a close reading and inductive coding of 35 broadcasts and 800 tweets, computational techniques were leveraged through Twi-XL to identify which records were most suitable for the analysis. For Twitter data, engagement metrics were re-calculated to retrieve the 50 most visible and 50 random tweets per year, which were then analyzed as pseudonymized text. For television data, ASR

transcripts were analyzed regarding their relative focus on the issue of sexual misconduct, which allowed the identification of the five most topic-focused broadcasts per five-year span. The selected broadcasts were then watched and coded via the CLARIAH Media Suite and its annotation tool (Melgar-Estrada et al., 2019; Ordelman et al., 2018).

Because the TwiNL collection does not store engagement metrics such as retweets or quote tweets, these were calculated retroactively. After removing identical tweets, tweet visibility was estimated through text-based similarity across retweets (marked by "RT") using Levenshtein distance. The resulting metrics only approximate visibility and do not reflect Twitter/X's inaccessible official engagement data. As TwiNL archives tweets that may have been deleted, these recalculated metrics are difficult to compare to official numbers. However, unlike official metrics, they are reproducible and transparent. Nevertheless, similar to official API data, they are not representative of a ground truth. Different text similarity measures yield varying results, underscoring that no database captures absolute truth, but Twi-XL offers transparent and reproducible frameworks which are essential for research (Rieder & Hofmann, 2020).

While broadcasts can be selected based on time spans, broadcasters, genres, featured guests, or other metadata, this study aimed to identify broadcasts that discuss sexual misconduct extensively within a national context. Many archival records, for example, are news and public affairs programs where sexual misconduct appears only briefly among other headlines. These headlines often follow standardized formats from supranational press agencies and include international coverage (Bergman, 2013). To retrieve broadcasts that engage substantively with the topic in a national context, a corpus-wide TF–IDF (term frequency–inverse document frequency) approach was applied to ASR transcripts. Using topic-related keywords, TF–IDF approximates how extensively each broadcast covers the topic relative to all other records. Broadcasts were then ranked by relative coverage, and top candidates were selected across time batches, considering further criteria such as genre diversity and a focus on national affairs. This process demonstrates Twi-XL's flexibility in optimizing the research sample, moving beyond a random selection.

Twi-XL's strength for this study lies in connecting social and traditional media to compare how feminist perspectives are amplified across platforms. The study further underscores the need to study transnational platforms such as Twitter/X within national contexts, where Twi-XL's Dutch-language focus counters anglophone bias in digital feminist activism research. In the Netherlands, scholars have documented persistent patterns of racism, ignorance, and denial in public discourse (van den Brandt et al., 2018), which Twi-XL's integrated media archives allow researchers to empirically investigate. However, the infrastructure also reveals limitations that the Twi-XL project could address. While protecting user identities is important, pseudonymization obscures which actor groups advance specific narratives, limiting insight into how social networks shape discourses and support intersectional feminist theory.

4. Conclusion: Challenges and Opportunities

This article introduced the Twi-XL infrastructure as an important advancement for enabling cross-media research on societal phenomena and public debates in a digital(ized) media environment. Twi-XL supports such research in a context where much public debate occurs on privately owned platforms that are increasingly difficult for individual researchers to access, requiring technical and legal expertise (Kamocki et al., 2022; Lomborg & Bechmann, 2014). Based on a collaboration between the UvA, RuG, KB, NISV, and

SURF, the article introduced the components of Twi-XL, including the SANE environment, the Twitter archive, the KB web collections, and the audiovisual archives of NISV. Two use cases further demonstrated how Twi-XL can be used in practice, both for large-scale computational analysis of social media and for cross-media comparisons of public discourse over time. Covering Twi-XL's broad and deep axes, these cases also highlighted the limitations of the infrastructure, strongly shaped and restricted by political, legal, and technological developments.

A first challenge is platform dependence and legal precarity. Access to social media data, and the legal conditions for collecting, storing, and sharing them—core to Twi-XL—remains controlled by a few platform companies. Changes in terms of service, API access, and pricing directly affect what data can be collected, what is feasible, and what is legally allowed (Freelon et al., 2025; Kamocki et al., 2022; Lomborg & Bechmann, 2014). In 2023, Twitter data collection had to be halted due to API and legal changes on the platform, now X, showing that publicly funded infrastructures that aim to provide access to and continuously update their collections remain dependent on commercial actors whose priorities may not align with long-term academic or public interests.

A second challenge is balancing privacy protection with analytical flexibility. Through SANE and GDPR-compliant workflows, Twi-XL enables responsible use of sensitive data, but this comes with notable constraints. All tools needed for analysis have to be considered in the setup of SANE, as additional tools are somewhat cumbersome to add in hindsight, thus limiting the technical flexibility of SANE. These constraints make the environment in its current state suitable for computational analyses that follow pre-defined professional frameworks, but less so for ad hoc experimental work and exploration. While SANE is required for analyzing tweets together with broadcast, podcast, or web data, the more flexible Twi-XL API remains valuable in supporting a more open-ended research approach.

Third, Twitter and web archives produce epistemic biases. The choices made in collecting the data and making them available to researchers to some extent limit the research questions that can be asked and foreground a version of public debate that is national, institutional, and text-based (Plantin et al., 2018). Without individual-level exposure metrics, analyses of news diets or media exposure are not possible. Visual content, memes, influencer practices, and algorithmic feeds are also underrepresented.

A fourth challenge concerns ensuring open and reproducible research under restricted data access. While Twitter data cannot be fully shared, shared environments, standardized workflows, and reusable code allow documentation and replication of analyses. SANE provides controlled access to raw data, improving transparency while reducing legal and technical burdens for individual researchers.

Finally, these challenges hold for individual collections. Yet, cross-media research introduces additional challenges, including differences in data formats, access conditions, and legal frameworks applied across datasets as well as the complexity of integrating heterogeneous sources. Infrastructures such as Twi-XL must therefore address access, but also interoperability and epistemic constraints.

Beyond these challenges, Twi-XL creates opportunities. Its main strength is enabling cross-media analysis and methodological innovation. By combining media datasets, researchers can trace how issues and frames move across social media, news websites, and broadcast media, supporting integrated analyses rather than

platform-specific studies. Twi-XL also facilitates collaboration. Centralized data collection, storage, and access reduce individual workload and allow greater focus on analysis and theory. It supports sharing of tools, scripts, and methods across projects and enables interdisciplinary research by combining computational and qualitative workflows, linking social-scientific and humanities perspectives. Twi-XL demonstrates how public research infrastructures can enable large-scale, cross-media research on public debate under platform and legal constraints. At the same time, it shows that such infrastructures shape what can be studied and how. Their long-term value depends on ongoing technical development and attention to platform dependence, privacy, and the epistemic effects of infrastructural design. Twi-XL can thus be scaled by expanding datasets, enhancing analytical tools, and building cross-national partnerships. Doing so requires careful attention to legal, ethical, and methodological challenges that shape the possibilities of cross-media research.

For the near future, Twi-XL aims to scale the infrastructure in three main ways: First, we aim to expand the type of social media collections with TikTok data in the Dutch language domain and with the repository of the documentation center for Dutch political parties (Documentatiecentrum Nederlandse Politieke Partijen, 2021). This facilitates more comprehensive investigations of how public debates trend across different platforms. It also allows a deeper understanding of the influence of platform affordances. Second, we plan to integrate new functionalities to interact with the various collections. This concerns particularly the implementation of the GPT-NL tool for chat-based analysis of the data with an open, GDPR-proof large language model; the setup of a permanent SANE environment at KB; and enabling the use of the Twi-XL API within SANE. Third, we will seek partnerships with similar collections elsewhere in Europe to facilitate the analysis of public debates as they have evolved across different European countries. In doing so, SSH scholars will be sustainably and effectively supported in critically assessing the role of both traditional mass media and social media platforms in mediating public debates.

Acknowledgments

The authors wish to acknowledge the support of the Twi-XL infrastructure project and the University of Amsterdam during the preparation of this article.

Funding

The Twi-XL infrastructure project (2021–2025) was funded by the Platform Digitale Infrastructuur for the domain Social Sciences and Humanities (PDI-SSH) established via the SSH-Council by the Dutch Government. Publication of this article in open access was made possible through the institutional membership agreement between the University of Amsterdam and Cogitatio Press.

Conflict of Interests

The authors declare no conflicts of interests.

References

Baas, I. M., Broersma, M., Caselli, T., & Esteve Del Valle, M. (2025). Who is sharing the news? Identity construction of news sharers on Dutch language Twitter. *First Monday*, *30*(5). https://doi.org/10.5210/fm.v30i5.13789

Bennett, W. L., & Pfetsch, B. (2018). Rethinking political communication in a time of disrupted public spheres. *Journal of Communication*, *68*(2), 243–253. https://doi.org/10.1093/joc/jqx017

Bergman, T. (2013). Liberal or radical? Rethinking Dutch media history. *Javnost – The Public*, *20*(3), 93–108. https://doi.org/10.1080/13183222.2013.11009122

Bibliothèque nationale du Luxembourg. (2017). *Luxembourg web archive*. Luxembourg Web Archive. https://www.webarchive.lu

Brems, C., Temmerman, M., Graham, T., & Broersma, M. (2017). Personal branding on Twitter: How employed and freelance journalists stage themselves on social media. *Digital Journalism*, *5*(4), 443–459. https://doi.org/10.1080/21670811.2016.1176534

Burkhardt, S., & Poell, T. (2026). Feminist television, racist Twitter? Feminist perspectives on the mediation of sexual misconduct in the Netherlands. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448251409211

Chapekis, A., Bestvater, S., Remy, E., & Rivero, G. (2024). *When online content disappears*. Pew Research Center. https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears

Common Lab Research Infrastructure for the Arts and Humanities. (2026a). *A research environment for digital humanities*. CLARIAH Media Suite. https://mediasuite.clariah.nl

Common Lab Research Infrastructure for the Arts and Humanities. (2026b). *What does CLARIAH offer?* https://clariah.nl

de Keulenaar, E., Poell, T., Helmond, A., Rieder, B., & Van Gorp, J. (2025). Computational cross-media research: Tracing divergences between normative Dutch television and social media discourses on the 'refugee crisis' (2013–2018). *Convergence: The International Journal of Research into New Media Technologies*, *31*(5), 1606–1628. https://doi.org/10.1177/13548565241258956

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). *BERTje: A Dutch BERT model*. arXiv. https://doi.org/10.48550/arXiv.1912.09582

Documentatiecentrum Nederlandse Politieke Partijen. (2021). *Over het DNPP*. University of Groningen. https://www.rug.nl/research/dnpp/over-het-dnpp

European Commission. (2026). *Application of the GDPR*. https://commission.europa.eu/law/law-topic/data-protection/information-business-and-organisations/application-gdpr_en

European Federation of Video Game Archives, Museums and Preservation Projects. (2026). *Institute for Sound and Vision (NL)*. https://efgamp.eu/institute-for-sound-and-vision-netherlands

Felt, U., Öchsner, S., Rae, R., & Osipova, E. (2023). Doing co-creation: Power and critique in the development of a European health data infrastructure. *Journal of Responsible Innovation*, *10*(1), Article 2235931. https://doi.org/10.1080/23299460.2023.2235931

Freelon, D., Monzer, C., Jeon, G., Moy, C., & Williams, N. (2025). The post-API age of social media data access: Past, present, and future. *The ANNALS of the American Academy of Political and Social Science*, *715*(1), 16–37. https://doi.org/10.1177/00027162251372557

Geldermans, I. (2025). *Historical growth of the KB web archive*. KB Lab. https://lab.kb.nl/dataset/historical-growth-kb-web-archive

Geldermans, I., de Gruijter, M., Ham, S., & Wiedenhof, F. (2025). *Twi-XL & SANE: New ways of exploring the KB web collection*. Zenodo. https://doi.org/10.5281/zenodo.15401634

Gotfredsen, S. G. (2023, December 6). Q&A: What happened to academic research on Twitter? *Columbia Journalism Review*. https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php

Institut national de l'audiovisuel. (2026). *Bienvenue sur data.ina.fr. L'INA met l'IA au service de l'exploration des médias*. https://data.ina.fr

ISO. (2025). *Space data system practices—Reference model for an open archival information system (OAIS)* (ISO 14721:2025). https://www.iso.org/standard/87471.html

Kamocki, P., Hannesschläger, V., Hoorn, E., Kelli, A., Kupietz, M., Lindén, K., & Puksas, A. (2022). Legal issues related to the use of Twitter data in language research. In M. Monachini & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2021* (pp. 68–75). Linköping University. https://doi.org/10.3384/ecp1897

KB nationale bibliotheek. (2022, January 31). *Webarchivering bij de KB vandaag de dag (deel 2/3) | KB nationale bibliotheek* [Video]. YouTube. https://www.youtube.com/watch?v=lTK7eW_M14E

La Morgia, M., Mei, A., & Mongardini, A. M. (2025). TGDataset: Collecting and exploring the largest Telegram channels dataset. In *KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (pp. 2325–2334). Association for Computing Machinery. https://doi.org/10.1145/3690624.3709397

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*(4), 256–265. https://doi.org/10.1080/01972243.2014.915276

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), 74–101. https://doi.org/10.1038/s41562-022-01460-1

Masanés, J. (2006). *Web archiving*. Springer. https://doi.org/10.1007/978-3-540-46332-0

Mayer, I. (2023, March 8). Vier Milliarden Tweets fürs Archiv. *Süddeutsche Zeitung*. https://www.sueddeutsche.de/panorama/twitter-archiv-deutsche-nationalbibliothek-1.5764923

Melgar-Estrada, L., Koolen, M., Beelen, K., Huurdeman, H., Wigham, M., Martinez-Ortiz, C., Blom, J., & Ordelman, R. (2019). The CLARIAH Media Suite: A hybrid approach to system design in the humanities. In *CHIIR '19: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 373–377). Association for Computing Machinery. https://doi.org/10.1145/3295750.3298918

Molyneux, L., Holton, A., & Lewis, S. C. (2018). How journalists engage in branding on Twitter: Individual, organizational, and institutional levels. *Information, Communication & Society*, *21*(10), 1386–1401. https://doi.org/10.1080/1369118X.2017.1314532

National Library of the Netherlands. (2026). *Web collection*. https://www.kb.nl/en/onderzoeken-vinden/bijzondere-collecties/webcollectie

Netherlands Institute for Sound & Vision. (2016, September 8). *Sound and Vision first national audiovisual archive in the world to receive the Data Seal of Approval* [Press release]. https://nieuws.beeldengeluid.nl/135622-sound-and-vision-first-national-audiovisual-archive-in-the-world-to-receive-the-data-seal-of-approval

Netherlands Institute for Sound & Vision. (2024). *Project SANE*. https://labs.beeldengeluid.nl/projects/sane

Netherlands Institute for Sound & Vision. (2026). *Sound & vision data*. Sound & Vision Data Platform. https://data.beeldengeluid.nl

Open Data Infrastructure for Social Science and Economic Innovations. (2025). *About ODISSEI*. https://odissei-data.nl/about-odissei

Ordelman, R., Melgar, L., Van Gorp, J., & Noordegraaf, J. (2018). Media suite: Unlocking archives for mixed media scholarly research. In I. Skadiņa & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018* (pp. 21–25). Linköping University Electronic Press.

Ordelman, R., & van Hessen, A. (2018). Speech recognition and scholarly research: Usability and sustainability. In I. Skadiņa & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018* (pp. 167–172). Linköping University Electronic Press.

Pehlivan, Z., Park, S., Abrahams, A. S., Desblancs-Patel, M., Steel, B. D., & Bridgman, A. (2025). *Building a media ecosystem observatory from scratch: Infrastructure, methodology, and insights*. arXiv. https://doi.org/10.48550/arXiv.2506.10942

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Platform Digitale Infrastructuur for the domain Social Sciences and Humanities. (2025). *About PDI-SSH*. https://pdi-ssh.nl/en/about-pdi-ssh

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1535

Rogers, N., & Jones, J. J. (2021). Using Twitter bios to measure changes in self-identity: Are Americans defining themselves more politically over time? *Journal of Social Computing*, *2*(1), 1–13. https://doi.org/10.23919/JSC.2021.0002

Sillesen, L. B. (2015, March 2). Analyzing journalists' Twitter bios. *Columbia Journalism Review*. https://www.cjr.org/analysis/analyzing_journalists_twitter_bios.php

SSH-Council. (2026). *SSH-Council of the Netherlands*. https://sshraad.nl/en/ssh-council

SSH Open Marketplace. (2025). *SSHOC and the EOSC*. https://marketplace.sshopencloud.eu/about/sshoc-eosc

SURF. (2024). *SANE: Trusted research environment for analysing sensitive data*. https://www.surf.nl/en/themes/research-infrastructure/sane-secure-environment-for-analysing-sensitive-data

SURF. (2026). *SURFconext: Secure access everywhere with one set of credentials*. https://www.surf.nl/en/services/identity-access-management/surfconext

SURF Cooperation. (2024, February 7). *Introduction to SANE (Secure ANalysis Environment)* [Video]. YouTube. https://www.youtube.com/watch?v=qxuA6GOc5Do

Tjong Kim Sang, E. (2022). *Explaining the dataset: Differences between TwiNL and the Twitter API* [Powerpoint presentation]. https://ifarm.nl/erikt/talks/slides-20221121-twixl.pdf

Tjong Kim Sang, E., & van den Bosch, A. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal*, *3*, 121–134.

Twi-XL. (2025). *About*. https://twi-xl.humanities.uva.nl/?page_id=88

van den Brandt, N., Schrijvers, L., Miri, A., & Mustafa, N. (2018). White innocence: Reflections on public debates and political-analytical challenges. An Interview with Gloria Wekker. *DiGeSt. Journal of Diversity and Gender Studies*, *5*(1), 67–82. https://doi.org/10.11116/digest.5.1.4

van der Meer, L. (2023). *Sharing sensitive data using SANE*. Zenodo. https://doi.org/10.5281/zenodo.10069901

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

About the Authors

Danielle A. Maxwell-Fraser is a junior researcher in the Twi-XL project. She graduated from the research master's in social science at the University of Amsterdam in 2025 with a thesis focusing on understanding the process of decolonizing education through critical ethnography.

Iris M. Baas is a PhD candidate in the Twi-XL project at the University of Groningen, studying public debate and polarization on Twitter. She holds a bachelor's degree in international studies and a master's degree in digital humanities. Her research interests include political extremism, polarization, and activism in relation to news online.

Sarah Burkhardt is a PhD candidate in the Twi-XL project in the Department of Media Studies at the University of Amsterdam. In her PhD thesis, she investigates cross-media discourse of feminist concerns in the Netherlands and engages critically with the development of computational methods for studying public discourse.

Olga Eisele is an assistant professor at the Amsterdam School of Communication Research (ASCoR). She is the coordinator for the Twi-XL project. Her research focuses on political (crisis) communication of different organizations in society, with a focus on the European Union as well as (computational) text analysis.

Anne C. Kroon is an associate professor at the Amsterdam School of Communication Research (ASCoR). She focuses on the role of algorithms in recruitment and hiring, and the content and consequences of (biased) media representations of minorities, drawing extensively on computational methods.

Robert Jan Bood is an advisor in the Scalable Data Analytics team at SURF, the Dutch IT cooperative for education and research. He develops and designs solutions for researchers to store, process, and analyze data.

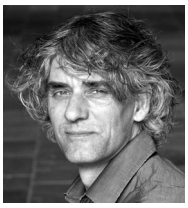
Iris Geldermans is a researcher at the National Library of the Netherlands, specializing in web archiving. Her work revolves around building workflows for archiving (digital) data as well as analyzing and researching archived data.

Sophie Ham is curator of digital collections at the National Library of the Netherlands. Her focus is on the selection, preservation, and dissemination of digital heritage, particularly born-digital publications, such as websites, blogs, and social media.

Machiel Jansen is heading the Scalable Data Analytics team at SURF, the Dutch IT cooperative for education and research. At SURF, he oversees the development and design of solutions for researchers to store, process, and analyze data.

Rana Klein is an AI developer at the Netherlands Institute for Sound & Vision. She develops machine learning models, pipelines, and benchmarking tools. Rana holds a master's degree in logic from the University of Amsterdam.

Matthijs Moed is an advisor in the Scalable Data Analytics team at SURF, the Dutch IT cooperative for education and research. He holds a master's degree in artificial intelligence from Maastricht University.



Roeland Ordelman is a senior researcher focusing on multimedia retrieval at the University of Twente (PhD, 2003) and manager in research and development at the Netherlands Institute for Sound & Vision. He is also a member of the CLARIAH-NL management board with a focus on infrastructure.

Erik Tjong Kim Sang is a senior research software engineer at the Netherlands eScience Centre, working on a broad variety of computational linguistics topics. He graduated in electrical engineering from Delft University and holds a PhD in computational linguistics from the University of Groningen.

Mari Wigham was involved in Twi-XL as a data engineer at the Netherlands Institute for Sound & Vision. She is currently a digital archivist at the Huygens Institute for History and Culture of the Netherlands.

Marcel Broersma is professor of media and journalism studies at the University of Groningen. His expertise is in journalism, media studies, communication studies, and history, focusing on the transformation of journalism, the innovation and diffusion of forms, and styles of journalism.

Julia Noordegraaf is professor of digital heritage in the Department of Media Studies at the University of Amsterdam. She is the principal investigator of the Twi-XL project and chair of the CLARIAH-NL supervisory board. Her research focuses on the preservation and reuse of audiovisual and digital heritage.