LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 36, (2026) 1–30 | e-ISSN: 2213-056X

Relevance and Diversity – Towards Selection Criteria for Archiving the German Web

Britta Woldering

Domain Access and Engagement – Science, German National Library, Frankfurt a.M., Germany, b.woldering@dnb.de, orcid.org/0000-0002-9718-0698

Simon Gottschalk

L3S Research Center, Leibniz Universität Hannover, Hannover, Germany, gottschalk@L3S.de, orcid.org/0000-0003-2576-4640

Randi Heinrichs

Center for Digital Cultures, Leuphana University Luneburg, Luneburg, Germany, randi.heinrichs@leuphana.de

Wolfgang Nejdl

L3S Research Center, Leibniz Universität Hannover, Hannover, Germany, nejdl@L3S.de, orcid.org/0000-0003-3374-2193

Christoph Neuberger

Weizenbaum Institute, Freie Universität Berlin, Berlin, Germany, christoph.neuberger@weizenbaum-institut.de, orcid.org/0000-0003-0892-610X

Katrin Weller

Data Services for the Social Sciences, GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany, katrin.weller@gesis.org, orcid.org/0000-0003-3799-1146

Uopen Journals | http://liberquarterly.eu/ | DOI: 10.53377/lq.23531

Abstract

The motivation for this paper is the need of the German National Library to massively increase its web archiving activities as the 12,000 snapshots collected per year are considered not to be enough to reflect the German web appropriately. As collecting everything which could be defined as the German web is impossible, the concept of selecting an "exemplary diversity" is introduced and substantiated by the guiding principles of relevance and diversity. The long-term goal is to find methods to (semi-)automatise the selection process and the prerequisite for this are well-defined selection criteria and sources of relevant as well as diverse topics that can be exploited. First steps to operationalise the guiding principles are outlined and supplemented by technological challenges and requirements to support the selection process as well as in the light of the advancement of web technology. Finally, the paper proposes a mix of curational and automatised measures and sources of information that should be taken into consideration to cater for a good coverage of relevance and diversity, supplemented by technological aspects regarding alternative representations of web content in web archives to be discussed to bring the content of web archives closer to the original user experience.

Keywords: web archiving; national libraries; digitalisation; national web; selection criteria; web technology

1. Introduction

With this paper, we reflect on challenges posed to Germany's National Library (Deutsche Nationalbibliothek, DNB) in the context of the new media and publishing landscape that is shaped by digitalisation processes. The DNB's mandate is to collect media works related to Germany, i.e. published in Germany, in German language and about Germany. Since 2006, the legal deposit mandate of the DNB comprises media works in non-physical form, hence also the collection of websites. The DNB started web archiving in 2012 and currently the web archive of the DNB holds about 70,000 snapshots of about 9,000 websites, selected by librarians. This is in opposition to 17 million registered .de-domains and presumably even more websites related to Germany registered under generic domains. Therefore, there is a need for the DNB to massively increase its web archiving activities. If the collection activities are to be extended by factor 100 or even 1,000, it becomes clear that

a manually curated selection of what should be archived is impossible. The long-term goal is to find methods to (semi-)automatise the selection process. A prerequisite for this are well-defined selection criteria and sources of relevant as well as diverse topics that can be exploited to create a collection that represents the German web. As an overview of the totality of the German web reflecting the collection mandate is not at hand, a representative selection is impossible. The guiding principle the DNB has chosen to escape this dilemma, is "exemplary diversity" – a term created by Christoph Classen at a workshop on web archiving at the DNB in 2018 and describing the attempt to select a variety as broad and unbiased as possible without the aspiration to be complete or representative. A second principle is the mission of the DNB (2024) to preserve the "cultural heritage." Our research question is therefore: How can the DNB implement and technically realise the principles of "exemplary diversity" and "cultural heritage" when selecting content on the German web?

1.1. Background: The Role of Digital Transformations

Digitalisation has fundamentally changed the public sphere since around the mid-1990s. Large parts of publishing and public communication have shifted to the Internet. Media use – such as the reception of news (Behre et al., 2024; Newman et al., 2023) – is also increasingly taking place online. In contrast, older media, especially printed media, are becoming less relevant. The digital transformation not only affects traditional mass media but goes far beyond this: the whole society is mirrored in the digital public sphere including cultural, political and commercial activities. As a result of digitalisation, the public sphere is also expanding into new areas, such as everyday communication, which has become particularly visible in social media. There are several characteristics which clearly distinguish the digital public sphere from the analogue public sphere and which must be considered when it comes to preserving the digital cultural heritage.

- *Loss of memory*: Contrary to the assumption that the Internet does not forget, it is apparent that considerable parts of the offerings published in the digital public sphere are not permanently accessible but disappear (Chapekis et al., 2024). This also applies to scientific sources (Gomes et al., 2021) and journalistic offerings (Duffy & Flynn, 2021).

- *Increasing quantity*: Due to lower technological barriers and thus the broad participation of the former "audience", the amount of information increases considerably, especially on social media platforms.
- *Decreasing quality*: The ability to bypass media (disintermediation) means that there is no need for quality checks anymore before publication (Bruns, 2018). Journalism lost its role as "gatekeeper". Platform companies have little incentive to check quality (Zuboff, 2019). As a result, there are major differences in quality between digital offerings.
- *Collapse of contexts*: Boundaries are dissolving between forms of communication, including between mass communication and individual communication, between different media (convergence), between the public and private sphere or journalism and advocacy (advertising, public relations). This "collapse of contexts" (Davis & Jurgenson, 2014) makes it difficult to assign offerings to specific categories.
- *Individualisation*: The Internet expands the range of possible choice and thus promotes the individualisation of use (high-choice media environment; Van Aelst et al., 2017), both through the active choice of users and through algorithmic selection.
- *Inequality*: Despite the large number of offerings, there is a strong focus on a small number of offerings on the audience and thus a considerable inequality in the distribution of visibility and influence (Hindman, 2009).

As a consequence of the digitalisation of the publication market and the communication in cultural, political and commercial activities, the collection mandates of national libraries were extended in many countries to comprise digital media to preserve the cultural heritage in the digital age (Altenhöner & Schrimpf, 2014; Schmitt, 2011; Vlassenroot et al., 2019, 2021). This also applies to the German National Library and the law it is based on. According to the German National Library Act (Gesetz über die Deutsche Nationalbibliothek; DNBG), it is the task of the DNB to collect all media works („Medienwerke"). Media works are defined in § 3 (1) as follows: "Media works are all representations in writing, image and sound that are distributed in physical form or made accessible to the public in non-physical form" (translated by the authors). The task is not limited to distinct media. What is required is the publication, i.e. public accessibility of media works, and this includes websites and social media.

The collection mandate (DNB, 2023) applies without restriction: "Our collection mandate aims for completeness, regardless of the form of the media works or the form of publication" (translated by the authors). Completeness might be aimed at regarding distinct media like e-books or e-journals, but it cannot be achieved in the case of web archiving for several reasons. First, the totality of the "German web" is unclear as it comprises much more than the .de domain: all websites with relation to Germany, German language, culture or history as well as websites whose owner's place of business is in Germany. Besides the DENIC list,[1] it should be hard to get a complete overview of websites fulfilling these criteria, let alone to keep up with the constant change in new websites arising and websites which are taken down. Second, even after identifying a set of relevant websites, to crawl and archive all their versions over time is an impossible task since websites are constantly updated, content is added or deleted without notice and it is impossible to archive all versions (Brügger, 2018).

As completeness of the collection is impossible to achieve in archiving the German web, the DNB is confronted for the first time in its history with the question of how to define selection criteria (DNB, 2016) and to justify them (Altenhöner & Schrimpf, 2014). A key aspect of the collection mandate of the DNB is to collect "without judging content" (Deutsche Nationalbibliothek [DNB], 2017, p. 11) which does not conform with the idea of a content-related selection process. However, it can be interpreted that the selection criteria should be as neutral and consensual as possible and well justified. Guiding principles are the terms "exemplary diversity" and "cultural heritage." How these principles are to be interpreted must be reviewed permanently, also with the participation of different social groups. The meaning of these principles cannot be definitively fixed. In view of transnationally oriented media such as the web, the limitation of the collection mandate to an individual nation also requires a new, differentiated interpretation.

1.2. Scope and Structure of the Paper

The motivation for this article is the DNB's need to massively increase its web archiving activities. Before proposing our guiding principles, we provide a closer reflection on the challenges of archiving the web regarding content selection and technology in section two. Section three then addresses social media as a special case of web archiving. In section four, we introduce

our approach for operationalising the guiding principles of relevance and diversity in four dimensions. The section is completed by the technological requirements to address the challenges of the advanced web technology as well as to support a (semi-)automated selection process, complemented by aspects of transparency to be considered in the selection process. Finally, in section five, recommendations for further steps to operationalise the concept of relevance and diversity and how to cope with the abundance of the German web are given.

When full archiving is not possible, decisions about selection processes should be the starting point for building and expanding a web archive. Therefore, this paper focuses on the discussion about how to define processes for deciding what should be archived. Based on these measures could be identified and applied to enable the selection process. It is obvious that a broad expansion of a web archive bears technological challenges for the whole life cycle of web archiving as well as organisational and financial issues, but these must be addressed in a later step after first having defined content related selection criteria. The technological, organisational and financial limitations that will occur when implementing the collection concept will have an impact on the scale, but not on the basic content selection decisions. Hence, in this paper, technological issues are only discussed in the context of enabling the suggested selection criteria while the full operationalisation to address these issues must be discussed subsequently in a different context. The problems and developments regarding crawling, indexing, searching and use of web archives as well as long term preservation and legal issues (e.g., data privacy and copyright) are out of scope of this paper.

2. Challenges in Archiving the Web

The capacity of digital technology to collect, connect and save information seems to suggest unlimited access to knowledge and all-encompassing archives. However, those promises of comprehensiveness distort the understanding of digital and furthermore new web archives and more attention needs to be paid to the politics of selection and practicability. Scholars in the emerging field of critical data studies have pointed out that a rhetoric has emerged around digitalisation and datafication processes that goes far beyond their technical capacities and forgets or even obscures the work required to secure, maintain and care for these systems (Christin, 2020;

D'Ignazio & Klein, 2020; Elish & Boyd, 2017; Irani, 2015). What at first sounds straightforward, namely digitising analogue books, storing digital publications and now also archiving websites, involves its own and very specific "infra-political processes", as Thylstrup (2018) emphasises in Politics of Mass Digitization. Thylstrup et al. (2021) bring critical debates around big data into dialogue with archival theories, emphasising that it is not the supposed sense of certainty but the limitations and uncertainties that must be considered and examined to understand and think through the newly emerging regimes of cultural memory.

Building on the post-structuralist theories of the archive in the mid-20th century – including the critical positions developed in response from feminist, queer and post-colonial perspectives (Best, 2018; Butler, 2009; Hartman, 1997) – cultural studies has been discussing an archival turn, which is now receiving renewed scholarly attention in relation to digital transformations. In classical cultural studies theories, such as those of Foucault (1970) and Derrida and Prenowitz (1995), archives have been described as the places in which the order of knowledge is formed since ancient times. Archives provide the foundations and infrastructures of what can be researched, remembered or known and thus also shape what is not considered important, not collected and potentially forgotten. This must be taken into account by the institutions and practitioners responsible for these new archives, both in terms of diversity and relevance of content and the technical conditions of selecting, crawling and preserving.

2.1. Challenges Regarding Content

The DNB's mission is to archive everything that was published in Germany, in German language and about Germany. With traditional publications these could be matched to selection criteria that would work in a setting of formal publishing cultures. However, on the web, publishing does not follow formal and standardised procedures. Therefore, a new selection strategy needs to be defined that takes into account content-related criteria. Regarding print media the DNB applies formal criteria like a minimum circulation figure or a minimum volume to decide whether a publication is ingested in the collection. In the case of web archiving the selection criteria cannot be determined purely formally as the characteristics of the web as media differs significantly from print media as is shown in the list below, hence criteria must be

defined content-related. The selection must be based not only on the mission of archive libraries to preserve cultural heritage for the long term, but also on the current user demand. The question of expectations of future generations can only be answered hypothetically. These selection criteria must be justified, made transparent and be revisited regularly.

The digital public sphere has characteristics that make the systematic description and selection of web offers for archiving difficult. This results in a number of challenges for the DNB:

- *Web offers and their characteristics:* The unknown totality of web offers makes a representative selection impossible. There is no continuous recording of all offerings as in the case of press and broadcasting, for which complete directories are available.
- *Unit of selection:* It is not easy to determine how to delimit the units to be collected. External content can be embedded on a website. Providers (organisations, individuals) can operate a large number of accounts on different digital platforms at the same time (multichannel communication). The discursive context is also of interest, i.e. the connection of a large number of contributions in networks on social media (e.g. through links or hashtags) or in discussion forums (threads). Topics are dealt with in many places on the web. They spread across the boundaries of platforms (spill-over). National borders are also frequently crossed.
- *Categories*: Formats of publication continue to develop or are only vaguely defined. Predefined formats or genres would help to systematise the digital offerings (e.g. Kampes, 2021).
- *Continuous production*: The web is not a completed, stable publication like a newspaper edition; continuous updates of different parts of the web that partly appear in frequent intervals need to be handled.
- *Variability*: Offers are not uniform for the whole audience as in mass media, but are personalised to individual users. Generative AI provides individual responses on demand based on external content, lacking transparency with regard to the origin of the data used and the production process. In addition, the presentation varies depending on the used technical device.
- *Usage data*: There is a lack of valid data on the different uses of digital offers. Platforms are only willing to share data to a limited extent.

- *Multimedia*: Elements such as text, photos, graphics, videos, audio, and animation must be archived together.

Overall, the digital public sphere is a comprehensive, less structured representation of society as a whole, containing various modes of access to the world (knowledge, entertainment, culture, advertising and other forms of persuasive communication, everyday communication).

A further basic part of the definition of selection criteria is the decision whether basically everything on the (national) web may be collected or if certain categories of content must be excluded. This question is mainly relevant for dealing with potentially illegal content. This could be, for example, violence, hate speech and online harassment, or sexual content – depending on respective legal frameworks that are in place. It can be argued that everything on the web can be regarded as published, hence is part of the collection mandate. As legality of the content cannot be examined in detail before archiving, the DNB applies the common rule that it takes down illegal content and excludes it from being indexed, if appropriate, as soon as the DNB is notified of illegal content in its web archive. Generally, everything on the German web is seen as part of the historical tradition and worth to be archived.

2.2. Technological Challenges

Defining selection criteria for a national web archive is first of all content-related, but technological aspects are closely interlinked to content-related decisions, e.g. the appraisal of the relevance of a website or how to deal with personalisation and the dynamics of websites or the use of focused crawlers.

The web has seen a dramatic alteration since the first initiatives for web archiving, specifically the foundation of the Internet Archive in 1996. Consequently, Laursen and Møldrup-Dalum recognised in 2017 that "the archive separates itself increasingly from the live web the archive tries to preserve". Essential reasons for the divergence between the live web and the archived web are, on the one hand, the development of the web structure and the role of links as well as the continuous dynamisation and personalisation of the web. Both characteristics immediately affect the performances and modes of operation of classic web crawlers, whose selection strategies are only to a limited extent applicable to today's web.

Hyperlinks were the central element of the Hypertext Markup Language (HTML) and interlinked all available information in the World Wide Web (Berners-Lee, 1999). Web crawlers typically follow this paradigm and traverse hyperlinks to find potentially relevant websites. However, the role of hyperlinks has seen several developments (Helmond, 2019). Initially only being utilised as navigational elements, their distribution was soon used as an indicator of a website's importance with the PageRank algorithm developed in 1996 as a famous example. Since then, new types of hyperlinks, such as deep links and app links (Helmond, 2019), dynamic links and asynchronous triggered links, have been established. Social media have introduced platform-specific link characteristics like the usage of URL shorteners as well as the focus on intra-website links (e.g., retweets) have led to further fragmentation of the web (Helmond, 2019; Lee et al., 2009). Consequently, the paradigm that each URL uniquely identifies a page on the web, initially only queried via the GET method,[2] has been softened up.

On top of the link structure, new web technologies have made web archiving more complex. The advent of CSS, Javascript and client-side executed scripts (Ajax) had already been identified as a major challenge to web archiving in 2012 which called for new web page processing in indexing methodologies (Gyllstrom et al., 2012) and JavaScript had been responsible for 33.2% more missing resources in 2012 than in 2005 (Brunelle et al., 2016). Nowadays, websites are typically built in a modular way, consisting of several files and relying on complex server-side processes and user interaction which require server-side preservation strategies for preservation of their contents (Bočytė & de Vos, 2018).

Web crawlers have only partially been adapted to these changes of the web structure. Focused crawling techniques aim to identify web pages related to a topic of interest, but they require the identification of topic-relevant web pages, which has become increasingly inefficient and noisy as the complexity of the web has grown (Aggarwal, 2019). While Google's web crawler has allowed the use of POST requests since 2021,[3] there remains an open question of which requests are suitable for web archiving. Also, the web archive should reflect the structure of the web since users want to conduct analyses of the granularity of single websites and the structure of the web in general (Holzmann & Nejdl, 2021). Hence, new web crawlers need to drift away from the original concepts of the web and instead find new solutions that enable them to extract and store as complete information as possible from the contents of web pages.

Due to the personalisation of the web, two users can see different versions of the same web page at the same time.[4] Kelly et al. (2013) have identified different cases potentially leading to such a situation, including different website layouts (e.g., mobile vs desktop mode) and local website content based on the user's IP address or GPS position (e.g., local news and weather forecasts). Exceptional cases of personalisation are advertisements and social media: recommendations for posts and products or displayed trends can be tailored to user groups, and different user groups might be shown different content via microtargeting for political influence (Ribeiro et al., 2019).

To deal with personalised content, web crawlers can either attempt to disable any personalised information, e.g., by declaring themselves as bots in the request header, or by taking the role of different personae to retrieve different views from the same website. In the former case, the challenge lies in the definition of user profiles and the identification of websites that actually offer a relevant diversity in information regarding personalisation. When viewing a snapshot in a web archive, users should get access to metadata specifying which personalisation is behind the respective stored snapshots (Kelly et al., 2013).

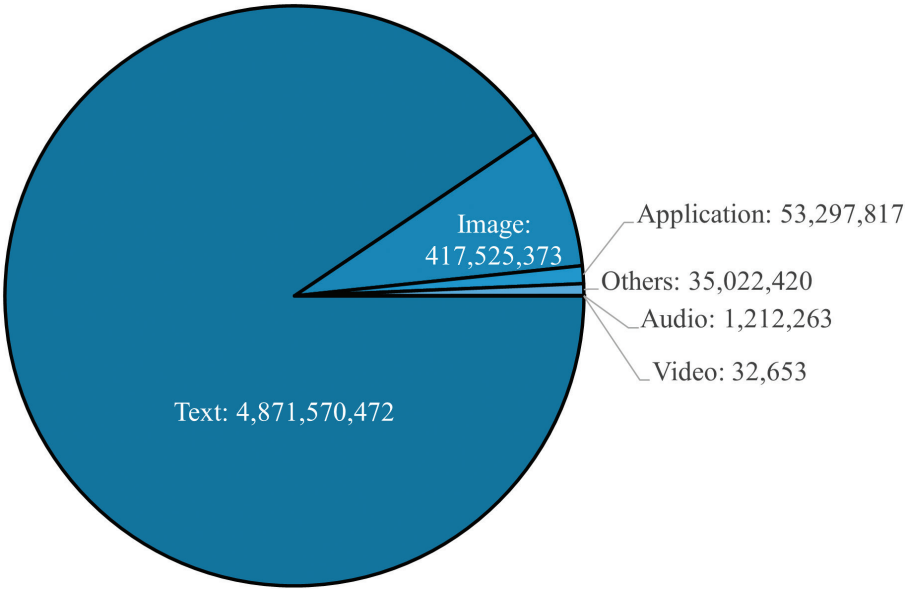
2.3. Resource Limitations

Limited resources are one of the main reasons why the web cannot be archived in its entirety. Lee et al. (2009) define three dimensions of resource requirements when creating web archives:

- Scalability: the number of web pages that a crawler can proceed – specifically, if the time and resource usage of the web crawler scales linearly with the number of pages to be crawled.
- Performance: the speed at which the crawler discovers the web.
- Resource usage: the CPU and RAM consumption during crawling. In addition, the disk space is a major requirement: the Wayback Machine of the Internet Archive consumed 57 PetaByte[5] in December 2021.

While we do not yet have exact ways to measure the size of the German web, some data may help to illustrate the potential dimensions: Figure 1 is based on the analysis of the content of .de snapshots in the Internet Archive

Fig. 1: Number of snapshots of .de URLs in the Internet Archive per file type between November 1994 and September 2013. The application data type contains file formats such as PDF and JSON.



between 1994 and 2013 performed by Holzmann et al. (2016), and highlights the challenges regarding the size of the German web offers with more than 5 billion snapshots until 2013 including 400 million images. Both numbers undoubtedly have increased substantially in recent years.

3. Online Platforms and Social Media

Finally, for preserving the web and our digital heritage, archival institutions also need to consider activities that are happening within a variety of online platforms. In the context of the web being shaped through platformisation, platforms are viewed as "(re-)programmable digital infrastructures that facilitate and shape personalised interactions among end-users and complementors, organised through the systematic collection, algorithmic processing, monetisation, and circulation of data" (Poell et al., 2019, p. 3). Online platforms exist for various domains and purposes, including notably e-commerce or entertainment. Also, a broad category of online platforms is social media.

Over the years, social media platforms have grown into a viable part of the web. They are not only being used by individuals, but also by numerous official actors and institutions, such as media, politicians and government agencies, cultural institutions, NGOs, or celebrities. Such platforms used both for personal and professional communication settings are of particular interest, contribute significantly to the information that can be found on the web, and thus are also relevant from a web archiving perspective. Their relevance for today's web landscape is not just due to size but also comes from the unique quality of their content. With their multifaceted contributions from different user communities they represent important cultural and even historical artefacts: For example, they carry the native traces of several protest movements such as the Arab Spring (Harlow & Johnson, 2011) or the Occupy Wall Street (DeLuca et al., 2012) movement, and have been used to raise awareness to societal challenges at large scale, for example with the #metoo hashtag pointing out sexual harassment.

Archiving from web platforms, however, comes with its own additional challenges for the web archiving community (Vlassenroot et al., 2021). The collaboration between the Library of Congress (LoC) in the U.S. and Twitter Inc. has been the only attempt so far to professionally preserve a social media platform in its entirety – and it did not lead to the desired success. The decision by the LoC to shift to selective collection (Library of Congress, 2017) is linked to some of the more general challenges that web archivists have to face for preserving social media platform content, e.g., capturing multimedia content shared through social media and handling outgoing links to other websites or other platform content.

Furthermore, platforms are not stable and the entire platform context may change over time. In the case of Twitter, the takeover by Elon Musk has led to numerous changes of how the platform operates including changes of the platform brand from Twitter to X. This has affected the platform's user community and how users interact on and with the platform. It has also heavily reduced the options for accessing data from the platform. The availability of APIs for accessing online platforms is frequently changing, with the closure of the previous Facebook API in 2018[6] being another impactful event for everyone aiming at collecting and preserving platform data. Partly closures of APIs as access points have inspired implementing alternative approaches such as web scraping, but these are also often technically restricted by the platforms. There is ongoing work to find new solutions for enabling data

access (e.g., at the Social Media Archive SOMAR)[7] and work on new models for making online platform data accessible for research while also respecting data protection regulations (European Digital Media Observatory (EDMO), 2022). Overall, the question of what can be accessed and thus archived from online platforms has to reflect different technical challenges, but also legal constraints (including the ones imposed by platforms through their terms of services) and ethical considerations (for example around collecting sensitive information from vulnerable groups).

Various efforts to collect and preserve social media content exist, at national libraries (Vlassenroot et al., 2021) as well as in academia ranging from individual researchers to academic institutions interested in social media content as research data (Weller & Kinder-Kurlanda, 2016). Therefore, the selection processes are often influenced by specific current research interests, rather than by specific preservation strategies and long-term perspectives.

A selection strategy for social media archiving as part of a national web archive will have to take several factors into consideration, including:

- *Platform selection*: As a first step it needs to be decided which platforms should be included in archival processes. This decision might have to be made based on practical considerations, e.g., technical feasibility. It could also be informed by the size of the user community, or the impact a platform has in a specific country.
- *Content selection*: Assuming that full coverage of an entire platform is likely not feasible, additional strategies have to be found to create subsets. This could for example be based on topics (requiring good strategies to translate a given topical interest into a search query on the platform), random data, person-centred or networked approaches.
- *Format selection:* Finally, strategies have to be defined on how to archive content and make it accessible. A fundamental question here could be to decide between purely text-based formats and multimedia content, potentially even preserving the "look and feel" of the platform.

Considering social media as part of a general web archiving strategy is desirable for preserving the cultural heritage of the social web. The sheer size of content produced on social media platforms makes selection

strategies necessary, but this is challenging due to its very ephemeral nature. Furthermore, access limitations as well as the need to quickly react to new developments (new platforms, new platform features, content changes like deletions or edits) requires very specialised setups and dedicated resources.

4. Operationalisation

In order to fulfil its collecting mission of exemplary diversity, the DNB strives for a selection based on relevance and diversity as central criteria. There is an obvious tension between relevance and diversity as selection criteria: relevance demands strict selection and the focusing of attention on a few important elements in a dimension. In contrast, the demand for diversity requires the broad representation of as many topics, actors or opinions as possible, regardless of their importance and quality. Both criteria are important for liberal democracy as well as for an archive of the German web.

4.1. Relevance and Diversity

Relevance is measured in approaches of communication studies via operationalisations of visibility and influence in the public sphere (as a concept see Schatz & Schulz, 1992). Indicators for visibility can be, for example, the usage of offers, number of citations and mentions of actors, and the number of contributions to a topic. Indicators for influence can be agenda-setting effects and opinion power ascribed to ("leading") media and actors like politicians in high positions. A special approach to define the relevance of topics is the "news value" (Galtung & Ruge, 1965).

Relevance in the public sphere can be supplemented by relevance indicators defined from the perspective of functional subsystems of society (such as politics, economy, science, art, or sports). These subsystems have their own measures of importance. For example, prominence of scientists in public often differs from their reputation in the science system itself. The system-specific relevance becomes obvious through the roles of the actors in the system's own hierarchy (e.g., the chancellor in the political system) or the results of internal recognition mechanisms (e.g., awards). Following Bourdieu (1985),

the definition of relevance at the heteronomous pole of a field is oriented toward outside influences and audience tastes, while at the autonomous pole relevance assessment follows internal standards of quality and self-defined societal goals of professions.

Other forms of social differentiation also have an impact on relevance assessments, e.g., regions, milieus, or classes. Therefore, relevance must always be defined in a distinct perspective. This can lead to a variety of different definitions of relevance. For the purpose at hand, it is important to follow the most consensual and stable relevance assessments possible.

Diversity as a norm means that all interests of population groups and all manifestations of society shall be represented in the public sphere, regardless of the quality, visibility or influence of these interests and manifestations. Diversity can refer to different dimensions of content, including events, topics, opinions, actors, spaces, and formats. The norm of diversity has been intensively discussed and empirically investigated in the context of media and public sphere (e.g., Hendrickx et al., 2022; Joris et al., 2020; Loecherbach et al., 2020; Napoli, 1999) to which the DNB can refer.

The operationalisation of diversity as well as of relevance depends on several normative decisions. In the first step, dimensions of diversity must be defined for both sides: for society's representation and preferences of audience groups. Then, in the second step, the different manifestations must be found in one dimension, for example, forms of everyday culture in different milieus or different opinions on a given issue. Ultimately, it is a question of capacity, determining how much diversity can be documented and archived, since the manually curated selection of diverse material is complex and time-consuming. Relevance and diversity can be operationalised meticulously and with high complexity which detracts the practical manageability and raises many questions of detail. A compromise between a documented rationale and practicability needs to be found.

4.2. Dimensions of Relevance and Diversity

Regarding relevance and diversity the following four dimensions need to be considered.

4.2.1. Dimension "Topics"

The first decision in the selection process for web archiving is about the topics. Different from print media or broadcast, there are no established thematic structures or genres for online media so far (Kampes, 2021). Kampes distinguishes online offers into information based, e-commerce, social media, search engines and games. She analysed the elements of the titles of information based online offers and condensed the following 23 top genres which are defined further in her thesis: digital, sports, lifestyle, guidance, finances, knowledge, entertainment, regional, culture, cars, health, travel, family, business, style, games, football, career, politics, forum, news, real estate, newsletter (translation by the authors).

The assumption is that information providers label their offers to be understood intuitively by their users and to be found easily by search engines. Hence, although there are no established structures or genres for online media, a de facto structure created by search engine optimisation can be defined. It is important to point out that the results of the analysis show a long tail distribution, and the above mentioned 23 genres are the top ones which are used very often while a lot more topics or genres are used much less.

Having such top topics as starting points, they have to be explored further regarding their subsystems, and finally with regard to relevance and diversity within these subsystems.

4.2.2. Dimension "Actors"

The dimension "actors" comprises individual persons and collective actors, such as organisations, social movements etc. Relevance can be defined by formal positions, roles which are connected with high visibility (attention, range, prominence), high esteem (positive rating, reputation, appreciation) or influence (power in different aspects). This can be determined by lists, rankings, statistics, experts etc.

Depending on the topic and the respective social subsystem separate taxonomies have to be used or created. For example, in politics roles

and positions are highly formalised which facilitates the selection process: top institutions on state and federal state level as well as organised interest groups. In contrast, the topic lifestyle is barely formalised and characterised by high volatility (fashions, trends) and a strong differentiation by milieus. To approach this, several subsystems need to be distinguished, and their taxonomies defined. They might be defined sociologically, i.e. lifestyle distinguished by its aspiration for singularity (Reckwitz, 2020), distinction (Bourdieu, 2001) and experience (Schulze, 1992), or by different forms of culture, e.g. popular culture. Popular culture again might be differentiated into fields like fashion, leisure activities or lifestyle products.

This shows that relevance and diversity are closely linked and that even the relevance aspect needs diversity. To gain even more diversity besides relevant actors, public and amateur actors should be considered. For example, in politics, members of the public, honorary politicians, citizens' movements or, in lifestyle, people interested in food, health, nature and environment, all kinds of leisure activities and amateur communities in this field. Further aspects of diversity can be included by considering sociodemographic characteristics, e.g., age, gender, religion, level of education, phase of life or milieu.

4.2.3. Dimension "Opinion"

For the dimension "opinion" also the aspects of relevance and diversity apply. Opinion leaders, trendsetters, influencers in a broader sense are indicators for relevance in the subsystems of the different topics and again have to be chosen in a large variety to show the whole – or at least a large – spectrum of opinions.

4.2.4. Dimension "Space"

For the dimension "space", the different spatial levels should be taken into account, i.e. transnational, international, federal, states, regional, or local aspects of a topic.

4.3. Technological Requirement and Automated Assessment

Manually curating selection processes require high personal resources. Therefore, technologies that can support or conduct the selection process will be needed to fully implement the selection based on relevance and diversity.

4.3.1. Technological Requirements

As a first step towards this goal, a set of technological requirements for the algorithms and tools to be created can be identified to address the challenges and the strategies for operationalisation discussed above:

- Technical operationalisation of selection criteria that are defined around the concepts of relevance and diversity and their interdependencies.
- Web crawling strategies that cope with the structure of today's dynamic web.
- Consideration of personalisation of content, either by targeting at the archival of non-personalised contents or by disclosing the identities adopted when archiving personalised content in the metadata of the crawl.
- Re-visit policies that define the frequency of archiving websites, which should be adaptive with respect to the domain, topics and type of a website.
- Rules to identify national contents that are in line with the specific characteristics of the targeted nation. To determine websites beyond the .de domain to be collected, subdomains, language and geo data can be factored in as shown in strategies of other national libraries (Vlassenroot et al., 2019).
- Methodologies to ensure the quality of archived websites regarding appropriateness of the sources and the preservation of their original contents and look and feel.

Facing this number of technological requirements, the concepts and methodologies behind traditional web archiving need to be questioned as key aspects of today's web can no longer be reconciled with traditional methods of web archiving: web archive content is not presented in a personalised

way, dynamic and embedded components of web pages are often not archived, and user interactions in the archived web are oftentimes not possible.

4.3.2. Automated Assessment of Relevance and Diversity

Sources of information that allow the automated identification of relevant topics include news articles (Leban et al., 2014), Wikipedia (Miz et al., 2020), social networks (Panagiotou et al., 2016) and scientific literature (Chen, 2006). However, automated topic identification based on these sources typically focuses on the detection of current trends and events (i.e., related to the criteria of news value as in Section 4.1), thus neglecting topics without strong temporal focus. Further, a critical analysis by Hanada et al. (2013) on the use of link analysis metrics such as link counts and article lengths for document selection has revealed that such metrics are more correlated to popularity than to quality and importance, hence questioning their use as unbiased selection strategies. In contrast, an approach for an unbiased and diverse selection of websites is the random selection of web contents, e.g., from the DENIC list.

With the rise of Large Language Models (LLMs) in recent years, a pressing question is whether LLMs can be used as tools to advance web archiving and address the challenges and issues raised. To date, while the use of LLMs for tasks such as web scraping has shown promising use cases (Padoan & Vinciguerra, 2024), their use for web crawling has not been fully explored (Al Galib et al., 2023). Although LLMs may struggle to identify trends and new websites on their own, we foresee their use to navigate the web for crawling and relevance assessment purposes. LLMs have recently been used to simulate human behaviour as demonstrated in Turing Experiments in different settings (Aher et al., 2023). These capabilities of LLMs have been utilised to simulate web navigation behaviour of users, targeting at solving specific tasks such as searching for weather reports (Lai et al., 2024). Adapting these approaches to the needs of web crawlers is a promising direction, allowing crawlers to emulate human-like navigation. Such LLM-based web navigation could specifically be used in focused crawling, which extracts relevant links from web pages and automatically assesses page relevance. While doing so, the LLM could even be prompted or fine-tuned to include the relevance and diversity criteria in its decision.

Arabzadeh et al. (2025) have demonstrated that LLMs can accurately assess the relevance of items like websites or documents, showing high alignment with human labels. However, first indications exist that LLMs may not be neutral in their assessment. For example, Ye et al. (2025) have recently demonstrated cases of LLMs' biases involved in the judgement processes, such as "the tendency to assign more credibility to statements made by authority figures, regardless of actual evidence". Better understanding these processes is critical when applying LLMs for judging the relevance and diversity criteria in the context of web archiving. Another challenge in using LLMs for relevance assessment lies in their inherent intransparency despite the need to make selection criteria and decisions transparent as discussed in the following.

4.4. Establishing High-quality and Transparent Selection and Archiving Procedures

Where archiving of the full (national) web is not possible and selection criteria need to be applied by an institution, transparency about the selection criteria that have led to the inclusion or exclusion of websites can directly impact the perceived quality of the archive. Selection criteria directly influence the composition of the collection and therefore become an integral part of the archive themselves. Only if selection criteria are documented and possible inappropriate content is defined, users will be able to comprehend why specific contents are included and what might be missing. Transparency is particularly important for sensitive topics, for example, when preparing a topical collection on LGBTQ+ websites: users of the archive could receive information on why certain websites were included. This can help to avoid "blind spots" (Donig et al., 2023) and to reflect on potential biases that may still be inherent to the collection.

Apart from transparency on selection criteria, quality-control should also respect aspects of transparency on technical dimensions. Maemura et al. (2018) define the following three elements of provenance or transparency criteria to document a specific crawl:

- *Scoping*: e.g., motivation for the collection, resulting in the focus, the seed-list, the crawl timing and configuration, inclusions and exclusions

- *Process*: How are scheduled and unscheduled events or process anomalies handled?
- *Context*: legal context, institutional setting and mandate, policies and guidelines, available resources for web archiving

Selection criteria can be presented in general for the complete crawl, or individually for each crawled website, e.g., as part of their metadata.

5. Conclusion and Recommendation

This paper was motivated by the German National Library's need to extend its web archiving activities massively and to define selection criteria which can be supported by semi-automated processes. As collecting everything which could be defined as the German web is impossible, the concept of selecting an "exemplary diversity" was introduced and substantiated by the guiding principles of relevance and diversity. Technological challenges and requirements to support the selection process as well as in the light of the advancement of web technology were described. Finally, the factors to be considered for establishing a selection strategy for social media archiving as part of a web archive were outlined.

5.1. Measures of Relevance and Diversity

A mix of curational and automatised measures and sources of information should be taken into consideration to cater for a good coverage of relevance and diversity. A selection of such measures is presented in Table 1, also summarising the automatised measures as outlined in chapter 4.3.

5.2. Representations of Web Content in an Archive

Apart from the content-related selection process guided by such measures, alternative representations of web content in web archives need to be discussed to bring the content of web archives closer to the original user experience. Hence, there is a need to question the following three aspects:

Table 1: Selection of curational and automatised measures.

	Curational	Automatised
Relevance	– Round tables with experts, e.g. scientific associations and organised groups of society – Representative population surveys regarding relevant topics in society – Thematic campaigns consulting citizens – Citizen/user suggestion scheme	Intelligent crawlers trained with large language models Exploit preselected types of offerings – News offers – News aggregators – Trend monitors – Ranking lists – Wikipedia – Hashtags on social media – Web tracking data/studies[1] – Web resources in e-publications – Scientific topics from Google Scholar, arxiv.org, …
Diversity		Random selections from – the DENIC list – Crawls of websites having a German address in their legal notice

[1]For example, the GESIS Panel Digital Behavioural Data Sample (https://www.gesis.org/gesis-panel).

- *Web crawlers*: There exists a variety of web crawlers (ranging from command line tools such as GNU Wget[8] to Heritrix[9] of the Internet Archive), tools and frameworks relevant to the creation of web archives that are explained in more detail by Vlassenroot et al. (2019). The suitability of these tools for operationalising complex selection criteria elaborated in this article must be examined.
- *Data structures*: The majority of web crawlers create or process content in the WARC file format,[10] recognised as the standard format for web archive content, particularly by national libraries. The expressivity of WARC archives in the context of the requirements, specifically with respect to the metadata and transparency requirements, needs to be assessed.
- *Interaction*: The use of web archives rarely reflects the experiences in the live web, due to incomplete, non-preserving website snapshots and incomplete crawls. To bring the content of web archives closer to the original user experience, alternative representations of web content in web archives need to be discussed.

5.3. Next Steps

As the next steps to establish a semi-automated selection process, we envision the following activities:

- The concept of relevance and diversity has to be elaborated further, based on the genres approach developed by Kampes (2021). These genres have to be investigated further regarding their relevant subtopics, actors, regional differences and sources of information and how to evaluate these results on a regular basis. Social media should be part of this approach, both as a source of information and as an object to be archived.
- To support the aspect of diversity, the web archiving institution should aim to open the selection process to various groups of persons to participate. In the table above possible groups and measures for participation are listed. These are complementing each other and should ensure that different views and aspects in the selection process are considered.
- The use of Large Language Models (LLMs) for web crawling should be further explored, in particular focused crawling as an example application of LLMs to extract relevant links from a web page and to automatically assess the relevance.

Having outlined the necessary next steps towards selection criteria for archiving the German web, we aim to elaborate them further in collaborative projects. As future work, concrete operationalisations should be explored, implemented and evaluated, also considering scalability and resource-efficiency. Balancing out the aim for completeness against the current restrictions and limitations requires solutions that facilitate the necessary selection process in a transparent and understandable way. Our concept for using relevance and diversity as main selection criteria is the proposed way of addressing this tension.

References

Aggarwal, K. (2019). An efficient focused web crawling approach. In M. N. Hoda (Ed.), *Software Engineering: Proceedings of CSI 2015* (Vol. 731, pp. 131–138). Springer. https://doi.org/10.1007/978-981-10-8848-3_13

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML '23)* (pp. 337–371).

Al Galib, A., Mehedi, M. H. K., & Rasel, A. A. (2024). Large scale web crawling and distributed search engines: techniques, challenges, current trends, and future prospects. In N. H. Zakaria, N. S. Mansor, H. Husni, & F. Mohammed (Eds.), *Computing and Informatics: 9th International Conference (ICOCI '23)*, Kuala Lumpur, Malaysia, September 13–14, 2023, Revised Selected Papers (pp. 17–29). Springer Nature. https://doi.org/10.1007/978-981-99-9589-9_2

Altenhöner, R., & Schrimpf, M. (2014). Lost in tradition?: Systematische und technische Aspekte der Erwerbung von Internetpublikationen in Archivbibliotheken. In A. Schüller-Zwierlein & M. Hollmann (Eds.), *Diachrone Zugänglichkeit als Prozess: Kulturelle Überlieferung in systematischer Sicht* (pp. 297–328). De Gruyter Saur. https://doi.org/10.1515/9783110311846.297

Arabzadeh, N., & Clarke, C. L. (2025). Benchmarking LLM-based relevance judgment methods. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3194–3204). https://doi.org/10.1145/3726302.3730305

Behre, J., Hölig, S., & Möller, J. (2024). *Reuters Institute Digital News Report 2024: Ergebnisse für Deutschland*. Verlag Hans-Bredow-Institut. https://doi.org/10.21241/ssoar.94461

Berners-Lee, T. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Harper.

Best, S. (2018). *None like us: Blackness, belonging, aesthetic life: Theory Q*. Duke University Press. https://doi.org/10.1215/9781478002581

Bourdieu, P. (1985). The market of symbolic goods. *Poetics*, *14*(1–2), 13–44. https://doi.org/10.1016/0304-422X(85)90003-8

Bourdieu, P. (2001). *Distinction: A social critique of the judgement of taste*. Routledge.

Brunelle, J. F., Kelly, M., Weigle, M. C., & Nelson, M. L. (2016). The impact of JavaScript on archivability. *International Journal on Digital Libraries*, *17*(2), 95–117. https://doi.org/10.1007/s00799-015-0140-8

Bruns, A. (2018). *Gatewatching and news curation: Journalism, social media, and the public sphere*. Peter Lang.

Brügger, N. (2018). *The archived web: Doing history in the digital age*. MIT Press.

Butler, B. (2009). 'Othering' the archive - From exile to inclusion and heritage dignity: The case of Palestinian archival memory. *Archive and Museum Informatics*, *9*(1), 57–69.

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

Christin, A. (2020). *Metrics at work: Journalism and the contested meaning of algorithms*. Princeton University Press.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Davis, J. L., & Jurgenson, N. (2014). Context collapse: Theorizing context collusions and collisions. Information, *Communication & Society*, *17*(4), 476–485. https://doi.org/10.1080/1369118X.2014.888458

DeLuca, K. M., Lawson, S., & Sun, Y. (2012). Occupy wall street on the public screens of social media: The many framings of the birth of a protest movement. *Communication, Culture and Critique*, *5*(4), 483–509. https://doi.org/10.1111/j.1753-9137.2012.01141.x

Derrida, J., & Prenowitz, E. (1995). Archive fever: A freudian impression. *Diacritics*, *25*(2), 9–63. https://doi.org/10.2307/465144

Deutsche Nationalbibliothek. (2016). *Strategischer Kompass*. https://d-nb.info/1112299254/34

Deutsche Nationalbibliothek. (2017). *Zum Sammelauftrag der Deutschen Nationalbibliothek*. https://www.dnb.de/SharedDocs/Downloads/DE/Ueber-uns/zumSammelauftragDNB.pdf?__blob=publicationFile&v=4

Deutsche Nationalbibliothek. (2023). Sammelauftrag. https://www.dnb.de/DE/Professionell/Sammeln/sammeln_node.html

D'Ignazio, C., & Klein, L. (2020). *Data feminism*. MIT Press.

Donig, S., Eckl, M., Gassner, S., & Rehbein, M. (2023). Web archive analytics: Blind spots and silences in distant readings of the archived web. *Digital Scholarship in the Humanities*, *38*(3), 1033–1048. https://doi.org/10.1093/llc/fqad014

Duffy, C., & Flynn, K. (2021, September 10). Some of the most iconic 9/11 news coverage is lost. Blame Adobe Flash. *CNN Business*. https://edition.cnn.com/2021/09/10/tech/digital-news-coverage-9-11/index.html

Elish, M. C., & Boyd, D. (2017). *Situating methods in the magic of big data and artificial intelligence*. SSRN. https://ssrn.com/abstract=3040201

European Digital Media Observatory (EDMO). (2022). *Report of the European Digital Media Obervatory's Working Group on platform-to-researcher data access*. (EDMO WP-Content). European Digital Media Observatory. https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

Foucault, M. (1970). *The order of things: An archaeology of the human sciences*. Pantheon Books.

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba und Cyprus crises in four foreign newspapers. *Journal of Peace Research*, *2*(1), 64–90. https://doi.org/10.1177/002234336500200104

Gomes, D., Demidova, E., Winters, J., & Risse, T. (Eds.). (2021). *The past web: Exploring web archives*. Springer. https://doi.org/10.1007/978-3-030-63291-5

Gyllstrom, K. A., Eickhoff, C., de Vries, A. P., & Moens, M.-F. (2012). The downside of markup: Examining the harmful effects of CSS and javascript on indexing today's web. In *The Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)* (pp. 1990–1994). The Association for Computing Machinery. https://doi.org/10.1145/2396761.2398558

Hanada, R., Cristo, M., & Pimentel, M. D. G. C. (2013). How do metrics of link analysis correlate to quality, relevance and popularity in Wikipedia? In *Proceedings of the 19th Brazilian Symposium on Multimedia and the web (WebMedia '13)* (pp. 105–112). https://doi.org/10.1145/2526188.2526198

Harlow, S., & Johnson, T. J. (2011). The Arab Spring | Overthrowing the Protest Paradigm?: How The New York Times, Global Voices and Twitter Covered the Egyptian Revolution. *International Journal of Communication*, *5*, 1358–1374.

Hartman, S. (1997). *Scenes of subjection: Terror, slavery, and self-making in Nineteenth-Century America*. Oxford University Press.

Helmond, A. (2019). A historiography of the hyperlink: Periodizing the Web through the changing role of the hyperlink. In *The SAGE Handbook of Web History* (pp. 227–241). Sage Publications Ltd.

Hendrickx, J., Ballon, P., & Ranaivoson, H. (2022). Dissecting news diversity: An integrated conceptual framework. *Journalism*, *23*(8), 1751–1769. https://doi.org/10.1177/1464884920966881

Hindman, M. S. (2009). *The myth of digital democracy*. Princeton University Press.

Holzmann, H., & Nejdl, W. (2021). A holistic view on web archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The past web: Exploring web archives* (pp. 85–99). Springer. https://doi.org/10.1007/978-3-030-63291-5

Holzmann, H., Nejdl, W., & Avisheh, A. (2016). The Dawn of today's popular domains: A study of the archived German Web over 18 years. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. https://doi.org/10.1145/2910896.2910901

Irani, L. (2015). Difference and dependence among digital workers: The case of amazon mechanical turk. *South Atlantic Quarterly*, *114*(1), 225–234. https://doi.org/10.1215/00382876-2831665

Joris, G., De Grove, F., Van Damme, K., & De Marez, L. (2020). News diversity reconsidered: A systematic literature review unravelling the diversity in

conceptualizations. *Journalism Studies*, *21*(13), 1893–1912. https://doi.org/10.1080/1461670X.2020.1797527

Kampes, C. F. (2021). *Angebotsfragmentierung online: Empirische Analysen struktureller Differenzierung von Medienangeboten und Medienanbietern im Online-Medienmarkt* [Doctoral dissertation, Universität Düsseldorf]. Universität Düsseldorf Docserv. https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=58308

Kelly, M., Brunelle, J. F., Weigle, M. C., & Nelson, M. L. (2013). A method for identifying personalized representations in web archives. *D-Lib Magazine*, *18*(11–12), 1–11. https://doi.org/10.1045/november2013-kelly

Lai, H., Liu, X., Iong, I. L., Yao, S., Chen, Y., Shen, P., Yu, H., Zhang, H., Zhang, X., Dong, Y., & Tang, J. (2024). AutoWebGLM: A large language model-based web navigating agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)* (pp. 5295–5306). https://doi.org/10.1145/3637528.3671620

Laursen, D., & Møldrup-Dalum, P. (2017). Looking back, looking forward: 10 years of web development to collect, preserve and access the Danish web. In N. Brügger (Ed.), *Web 25: Histories from the first 25 years of the World Wide Web* (pp. 207–228). Peter Lang.

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: learning about world events from news. In *Companion: Proceedings of the 23rd International Conference on World Wide Web (WWW '14)* (pp. 107–110). https://doi.org/10.1145/2567948.2577024

Lee, H.-T., Leonard, D., Wang, X., & Loguinov, D. (2009). IRLbot: Scaling to 6 billion pages and beyond. *ACM Transactions on the Web*, *3*(3), 1–34. https://doi.org/10.1145/1541822.1541823

Library of Congress. (2017). *Update on the Twitter Archive at the Library of Congress*. [White paper]. Library of Congress - Blogs. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf

Loecherbach, F., Moeller, J., Trilling, D., & van Atteveldt, W. (2020). The unified framework of media diversity: A systematic literature review. *Digital Journalism*, *8*(5), 605–642. https://doi.org/10.1080/21670811.2020.1764374

Maemura, E., Worby, N., & Milligan, I. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology*, *69*(10), 1223–1233. https://doi.org/10.1002/asi.24048

Miz, V., Hanna, J., Aspert, N., Ricaud, B., & Vandergheynst, P. (2020). What is trending on Wikipedia? Capturing trends and language biases across Wikipedia editions. In *Companion Proceedings of the Web Conference 2020 (WWW '20)* (pp. 794–801). https://doi.org/10.1145/3366424.3383567

Napoli, P. M. (1999). Deconstructing the diversity principle. *Journal of Communication*, *49*(4), 7–34. https://doi.org/10.1111/j.1460-2466.1999.tb02815.x

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Nielsen, R. K. (2023). *Reuters Institute Digital News Report 2023*. Reuters Institute and University of Oxford. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Panagiotou, N., Katakis, I., & Gunopulos, D. (2016). Detecting events in online social networks: Definitions, trends and challenges. In S. Michaelis, N. Piatkowski, M. Stolpe (Eds.), *Lecture Notes in Computer Science: Vol. 9580. Solving large scale learning tasks. Challenges and algorithms* (pp. 42–84). Springer, Cham. https://doi.org/10.1007/978-3-319-41706-6_2

Padoan, L., & Vinciguerra, M. (2024). ScrapeGraphAI [Online post]. https://github.com/ScrapeGraphAI/Scrapegraph-ai

Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4), 1–13. https://doi.org/10.14763/2019.4.1425

Reckwitz, A. (2020). *The society of singularities*. Polity.

Ribeiro, F. N., Koustuv, S., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., Goga, O., Gummadi, K. P., & Redmiles, E. M. (2019). On microtargeting socially divisive Ads: A case study of Russia-Linked Ad campaigns on Facebook. In Association for Computing Machinery (Ed.), *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19)*, January 29–31, 2019, Atlanta, GA, USA (pp. 140–149). ACM. https://doi.org/10.1145/3287560.3287580

Schatz, H., & Schulz, W. (1992). Qualität von Fernsehprogrammen: Kriterien und Methoden zur Beurteilung von Programmqualität im dualen Fernsehen. *Media Perspektiven*, *11*, 690–712.

Schmitt, T. M. (2011). *Cultural governance: Zur Kulturgeographie des UNESCO-Welterberegimes*. Franz Steiner Verlag.

Schulze, G. (1992). *Die Erlebnisgesellschaft: Kultursoziologie der Gegenwart*. Campus Verlag.

Thylstrup, N. B. (2018). *The politics of mass digitization*. MIT Press.

Thylstrup, N. B., Agostinho, D., Ring, A., D'Ignazio, C., & Veel, K. (Eds.). (2021). *Uncertain Archives: Critical Keywords for Big Data*. MIT Press.

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551

Bočytė, R., & de Vos, J. (2018). *Server-side Preservation of Dynamic Websites* [White paper]. Universiteit van Amsterdam. https://publications.beeldengeluid.nl/pub/633/Juli2018_Server-Side-Preservation-of-Dynamic-Websites_R_Bocyte.pdf

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, *1*, 185–111. https://doi.org/10.1007/s42803-019-00007-7

Vlassenroot, E., Chambers, S., Lieber, S., & Michel, A. (2021). Web-archiving and social media: An exploratory analysis. *International Journal of Digital Humanities*, *2*, 107–128. https://doi.org/10.1007/s42803-021-00036-1

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. In W. Nejdl, W. Hall, P. Parigi, & S. Staab (Eds.), *Proceedings of the 2016 ACM Web Science Conference (WebSci '16)*, Hannover, Germany, May 22–25, 2016 (pp. 166–172). ACM. https://doi.org/10.1145/2908131.2908172

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P., Chawla, N. V. & Zhang, X. (2025, April). *Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge*. [Published as a conference paper at the International Conference on Learning Representations (ICLR) 2025]. ICLR-2025-justice-or-prejudice-quantifying-biases-in-llm-as-a-judge-Paper-Conference.pdf. https://doi.org/10.48550/arXiv.2410.02736

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Notes

[1] DENIC provides the Domain Name System (DNS) for the .de domain.

[2] https://developer.mozilla.org/en-US/docs/Web/HTTP/Basics_of_HTTP/Evolution_of_HTTP.

[3] https://developer.mozilla.org/en-US/docs/Web/HTTP/Basics_of_HTTP/Evolution_of_HTTP.

[4] This applies especially to content which requires an account to be accessible, e.g. content behind paywalls or in Facebook groups.

[5] https://archive.org/web/petabox.php.

[6] https://techcrunch.com/2018/07/02/facebook-rolls-out-more-api-restrictions-and-shutdowns/.

[7] https://socialmediaarchive.org/.

[8] https://www.gnu.org/software/wget/.

[9] https://github.com/internetarchive/heritrix3/wiki.

[10] https://github.com/internetarchive/heritrix3/wiki.