# A Social media archive for digital memory research

Costis Dallas, Ingrida Kelpšienė

Abstract: Social media is an important social and cultural interaction arena, and a growing field of social research. Acknowledging the limitations of social media platforms and institutional web archiving initiatives to fully support the needs of researchers, this chapter makes the case for a reorientation of social media archiving, drawing from critical digital curation and archival theory to define specifications for a data architecture applying knowledge graphs, and aspects of the Open Archive Information System standard, to support research on Lithuanian memory, heritage, and identity interactions on social media. Based on this experience, it discusses broader implications for web archiving and digital curation in the context of research data infrastructures.

Keywords: social media, semantic modeling, web archiving, research data archives, digital curation.

## 1. Introduction

Social media platforms have become central to contemporary social and cultural practices, profoundly informing the way individuals and communities communicate, share information, construct their identities (Papacharissi 2011), establish affiliative relationships (Baym 2010) and social networks (Garton, Haythornthwaite, and Wellman 1997). Social media practice is central in phenomena such as the memory wars in Eastern and Central Europe (Rutten 2013), political protest in the Arab spring (Tufekci 2017), Holocaust memory (Manca 2020), fake news in nationalist narratives (Bonacchi 2022), and contestations on the difficult past and heritage (Kelpšienė et al. 2023). Such practices are distinct in their reliance on the 'logic' of social media platforms.While these platforms support identity, presence, relationships, reputation, group membership, conversations, and content sharing functions for their users (Kietzmann et al. 2011), they do so by activating mechanisms of datafication, commodification, and algorithmic selection (van Dijck et al. 2018). These mechanisms pose challenges to the autonomy and agency of individuals and communities, simultaneously subverting truth, trust, and the public sphere.

The ubiquitous nature of social media, coupled with its dynamic and interactive capabilities, renders it an invaluable resource for information, communication, and social science research (Garton, Haythornthwaite, and Wellman 1997; Snelson 2016; Stoycheff et al. 2017; Stieglitz et al. 2018; Shibuya, Hamm, and Pargman 2022). However, the ephemeral nature of social media content, combined with the proprietary algorithms and platform policies governing data access, poses significant challenges for researchers. As Ben-David (2020) contends, the task of researching social media is fraught with complexities not least because of the "unarchivable" nature of platforms like Facebook, which claims the role of the "archon" of

Cotis Dallas, Vilnius University, Lithuania, konstantinos.dallas@kf.vu.lt, 0000-0001-9462-0478
Ingrida Kelpšienė, Vilnius University, Lithuania, ingrida.vosyliute@kf.vu.lt, 0000-0003-3741-9510

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Cotis Dallas, Ingrida Kelpšienė, *A Social media archive for digital memory research*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.28, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 321-340, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

contemporary records of communicative social action. Platform architecture, which privileges the constant accumulation and flow of new content, complicates efforts to capture and maintain a stable record of digital interactions for scholarly analysis. Moreover, legal and ethical considerations surrounding user privacy and data access, and alarming efforts of platforms to curtail data access for scholarly research on social media practices and their political and ethical implications (Bruns 2019) further complicate the landscape for researchers seeking to investigate social media data.

Establishing a social media archiving infrastructure that can serve the needs of scholarly research beyond the affordances of platforms thus emerges as a crucial challenge for social media researchers. In this chapter, we share our insights from establishing a data archive suitable for investigating digital memory, heritage, and identity practices on Lithuanian social media within the context of the Connective Digital Memory in the Borderlands research project.[1] We discuss how this endeavor addresses theoretical and methodological issues relevant to web archiving, digital curation, and social media research. Specifically, we explore our understanding of the challenge in designing a social media archive suitable for scholarly research; our knowledge graph approach to the semantic representation of social media data to tackle issues of intelligibility, analytical power, and theory building; how our social media open archives architecture and workflow addresses the complementary methodological challenge of reliability of collected data, and the related digital preservation requirements of integrity and authenticity. Finally, we reflect on the lessons learned from this experience regarding the specification and design of social media archives for scholarly use.

## 2. The Challenge of social media archiving for scholarly research

Numerous studies in the field of digital preservation and research data infrastructures have sought to address the challenges inherent in social media data archives: data capture and collection (Littman et al. 2018; Lomborg and Bechmann 2014; Marres and Weltevrede 2013; Pehlivan, Thièvre, and Drugeon 2021); long-term preservation and data management (Thomson 2017; Voss, Lvov, and Thomson 2017; Hemphill, Leonard, and Hedstrom 2018); ethical and legal constraints (Zimmer and Kinder-

[1] "Connective Digital Memory in the Borderlands: A Mixed-Methods Study of Cultural Identity, Heritage Communication and Digital Curation on Social Networks" is a three-and-a-half-year project undertaken by the Connective Research Group at the Faculty of Communication, Vilnius University, aiming to explore heritage, memory, and identity-related interactions on Lithuanian social media. The project received funding from the European Social Fund (project No. 09.3.3-LMT-K-712-17-0027) under a grant agreement with the Lithuanian Science Council (LMT).

Kurlanda 2017; Bruns 2019; Franzke et al. 2020; Fiesler and Proferes 2018); and, last but not least, access and use of social media archives, and their impact on scholarly research methods and practices (Thomson and Kilbride 2015; Weller 2014). Concurrently, several social media archiving tools have emerged (Borji, Asnafi, and Naeini 2022; Social Media Lab 2024), and some national libraries, national archives, and other bodies and initiatives in the field of digital preservation and curation, have been actively setting requirements, developing guidelines, and initiating wide-scope projects for institutional social media archiving on both national and international scale (Hockx-Yu 2014; Thomson and Kilbride 2015; Pehlivan, Thièvre, and Drugeon 2021; Acker and Kriesberg 2017). However, while the scholarly value of web archives in general is widely acknowledged (Brügger and Schroeder 2017; Brügger and Laursen 2019; Vlassenroot et al. 2019), there is little evidence that institutionally-based *social media* web archives are actively used by researchers. This suggests that such initiatives, unlike focused, narrowly targeted data archives produced by researchers to support their own investigations, aim for the broad objective of long-term digital preservation, rather than active research use. Indeed, as noted by Vlassenroot et al., "the use of publicly accessible national archives and large-scale social media archives in scientific studies [is] only just emerging", with the focus primarily on social media use by government entities, and in the context of natural and health emergencies (Vlassenroot et al. 2021, 118; cf. Michel et al. 2021; Milligan, Ruest, and Lin 2016).

The risk of "epistemic failure—the inability to account for diverse theoretical, substantive and methodological perspectives in particular disciplinary traditions which require access to digital resources" is a concern for digital data repositories in general. This prompts a call for "a radical re-examination of current notions of *context* ... so that it encompasses the structure and evolution of the pragmatic references of such objects in the real world" as well as "semantic representations of the *epistemic content* of curated information objects … account[ing] for dynamically evolving semantic representations of 'things in the world' at the instance (occurrence) level as well" (Dallas 2007). Acknowledging the tension between researcher expectations and affordances of national web archives in Denmark and the UK, Jessica Ogden and Emily Maemura also underscore the need to consider "the relationship between records and the surrogate material objects they represent." To adequately access records, they had to move beyond document-centric and collection-centric views of WARC data in the archive, relying "on other representations and views into web archives at different junctures, including the use of Solr search indices and query interfaces, faceted and free text search interfaces under development, curation tools (which provided different seed- and collection-

centric views), and numerous ad-hoc and incomplete data" (Ogden and Maemura 2021, 59–60).

These broader challenges are further exacerbated by the unique nature of social media. The three complementary foci of social media research—content, interactions, and network structure—are constituted differently within the context of different platform affordances, disciplinary traditions, theoretical frameworks, and methodologies endorsed by specific research projects, casting doubts on the possibility of a 'one size fits all' approach to social media web archiving. The unit of inquiry in social media research is extremely variable, ranging from individual posts to the global graph of social media interactions. "Where to cut the network", to paraphrase Latour, is very much a theory- and researcher-laden question, and crucial contextual elements such as reactions, re-posts (retweets, shares), and comments are often absent from social media archives. Social media interactions are ephemeral and dynamic, accumulating interactions and "layers of context" after they have been collected (Acker and Kriesberg 2017), and retrieving information about deleted or altered data is often impossible (Ben-David 2016). Besides, on platforms such as Facebook, the record of each social media interaction is inherently plural and context-dependent, as different users may experience the same conversation differently based on their profile, history, and network of 'friends'—a reality that sits uncomfortably with the notion that web archives can capture a singular, fixed, and objective representation of social media.

Overall, this situation points to an onto-epistemological entanglement between the conceptualization of social media as a field of interactional communicative practice (Lomborg 2012) and almost all facets of establishing a research-capable social media archive. These facets include adopting appropriate capture strategies, ensuring authenticity and integrity of social media records, providing sufficient scope and expressiveness in social media data representation, and supporting digital methods of access and analysis. This entanglement is especially relevant within the context of datafication (Rogers 2013; Schäfer and van Es 2017), alongside the emergence of digital methods and practices tailored specifically for social media research (Edwards et al. 2013; Winters 2017; Perriam, Birkbak, and Freeman 2020; Wilson 2022). Recognizing the legitimacy of "a distinction between archiving social media data for a specific research purpose (scholar uses) and institutional archiving" (Pehlivan, Thièvre, and Drugeon 2021, 44), and drawing inspiration, among other factors, from advancements in the research use of web archives and the maturation of web archiving theory and practices over the past two decades (Brügger 2018; Helmond and van der Vlist 2019), we aim to challenge the notion that research-capable social media archives are still confined today to a form of micro-archiving previously characterized as "small scale", "here-and-now", and undertaken

by “individuals … whose technical knowledge of archiving or of the subsequent treatment is either lacking or on an amateur level” (Brügger 2005, 11).

In what follows, we address aspects of this onto-epistemological entanglement by sharing the challenges faced and decisions made in creating a social media archive capable of supporting data access, mixed-methods analysis, and theory-building within the framework of the Connective project.

### 3. Representing social media in a research-capable data archive

The Connective project is based on multiple analyses of a corpus comprising over 30,000 conversations (more than 250,000 posts and comments, authored by over 90,000 unique users) between 2016 and 2023, and sourced from Lithuanian Facebook accounts, pages, and groups, Instagram hashtag collections and accounts, and VKontakte communities. While it is not possible to assess reliably how much of the overall social media activity in Lithuania focuses on conversations about heritage and the past, such conversations clearly establish an active arena for negotiating important aspects of collective identity, values, and attitudes towards contemporary issues. Facebook data were identified and collected in summer 2022 by means of several hundred online queries, conducted under five different Facebook user accounts. Lexical expressions used in the queries are connected to 72 topics established by the research team after an initial qualitative scoping of relevant conversations on Facebook. These topics belong to nine broader themes: history, ethnoculture and language, 90s culture, religion, minorities, everyday life, war in Ukraine, objects and memory wars, and contemporary concerns. This corpus was further enriched by 75 interviews with users actively engaged in contested heritage meaning-making and identity work. Research in the Connective project, conducted by a transdisciplinary team of eleven researchers includes an integrative analysis of communicative repertoires, interactional patterns, and affiliative network structures with specific case studies of post-memory and identity construction on Lithuanian social media. These case studies explore themes such as: the remembrance of childhood in the late Soviet period and the challenges faced by youth subcultures in the 1990s; the post-memory of WWII Polish Home Army partisans in Lithuania; the self-identification of Russian Lithuanians, drawing on conceptions of the past; how invented traditions and ethnographic performance shape creative ethnicity among Lithuanians; the troubled memories of events related to the resistance and dual occupation by the Nazis and Soviets from 1939 to 1953; the ‘monument wars’ raging against public monuments and memory institutions focusing on historical figures from the Soviet era; conflicts around the use

of the 'foreign' letters *w*, *q*, and *x* in Lithuanian passports; and the complexities of remembering and the silencing of the Roma Holocaust. A second data collection season, focusing on additional queries representing topics relevant to these case studies, took place in fall and winter 2023. The research team employs diverse approaches, including corpus linguistics methods, narrative analysis, critical discourse analysis, metaphor analysis, visual methods, qualitative content analysis, and social network analysis, to identify cultural meanings embedded within social media content, group affiliations and influence in user interactions, and cultural schemas (such as stereotypes, conceptual metaphors, and deep narratives) that shape identity construction.

The Connective project leverages social media as a data infrastructure to facilitate multifaceted data-intensive investigations into various aspects of encounters with the past and participatory practices on social networking sites (Dallas 2018; Kelpšienė 2021). We have been working with digital data to reveal how memory practices on Lithuanian Social Network Sites (SNS), mediated by contested heritage, shape cultural identities. Based on identifying "a tentative proposed set of relationships, which can then be tested for validity [and] can often help in working through one's thinking about a subject of interest" (Bates 2009, 3), we drew from activity theory (Engeström 1999) and cultural semiotics (Lotman 2005) to establish an event-centric ontology of SNS semiotic activity on heritage, memory, and identity (Kirtiklis et al., 2023), viewed as a practice of digital curation "in the wild" (Dallas 2016). Our research was empirical and data-driven, based on analyzing a large number of conversations on the history and difficult past of Lithuania on Facebook, Instagram, and vKontakte. As we grappled with establishing a corpus of social media data needed for our research, a key question emerged: what are the properties of archived information objects, and their relationships, that enable their use as epistemic objects suitable for evidence-based scholarly research?

Our data infrastructure embodies an approach to social media archiving designed to capture the multifaceted nature of online interactions. At its core is a commitment to preserving the integrity, authenticity, and intelligibility of social media data, thereby ensuring its utility for scholarly research. This section presents the design decisions that underpin our data infrastructure, focusing on two aspects: the capture of the 'experienced' manifestation of social media content, and the representation of its conceptual properties through knowledge graphs.

Recognizing that the value of social media content often resides in its presentation and the context of user interactions, the Connective project adopted a web archiving approach that accounts for capturing social media content as it was experienced by users. This entails collecting facsimile scrolling screenshots of expanded social media conversations, ensuring the

preservation of visual layout, interactive elements, and temporal sequencing of posts. This method mirrors the user's perspective, retaining the contextual cues and platform-specific nuances crucial for interpreting social media discourse. Such an approach not only maintains the visual and interactive integrity of the data, but also respects its original context, addressing a critical need for authenticity in digital archiving (Ben-David 2020).

In addition to preserving experienced data, we place emphasis on the ontological, structural, functional, and semantic aspects of social media interactions. These elements are re-envisioned through a redefinition of the notion of "significant properties" as those aspects that can warrant "the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record" (Grace, Knight, and Montague 2009, 3), essential for the archive's intelligibility and informational value in a research context. It has been noted that platforms such as Facebook, designed to serve business purposes, are not suitable to account for rich representations of social media useful for social research (Helmond and van der Vlist 2019, 18). Our approach therefore was to extract structured representations of social media data into an external research data archive for further analysis. To capture significant properties based on the content and context of social media interactions in our data we adopted the property graph data model supported by the Cypher query language as applied in the popular Neo4j graph database management system (Robinson, Webber, and Eifrem 2015). This approach allows for the mapping of a semantic schema onto neo4j labels representing key entity types involved in social media conversations, as well as their properties and relationships.

The Connective property graph schema provides for the semantic representation of three types of nodes accounting for social media interactions: Message, representing posts, comments or replies; Thread, representing the sequence of a post and following comments and/or replies; and Actor, representing social media users who post, react, or share Messages and collectivities (such as Facebook groups and vKontakte communities) where such interactions are displayed. An additional type of nodes, Topic, is used to represent descriptive thematic categorizations. Topics are organized in a taxonomy capturing the hierarchical relationships between concepts, and are connected to relevant Messages, Threads, and Actors in a way that allows plural characterizations of any node based on notions in the Topics taxonomy. The taxonomy is faceted, and therefore different Topic hierarchies can accommodate not just subject-laden thematic categorizations, such as people, events, and places mentioned in a Message, but also different kinds of analytical categories and discursive constructs identified by our analysis of the data. In addition, the schema includes a Deposit node type aimed to represent aspects of the process of data capture

of different sets of Threads, providing for provenance and preservation metadata in the archive. Different types of relationships (PART_OF, RESPONDED_TO, FOLLOWED_BY, SHARED_FROM, POSTED, REACTED_TO, CONNECTED_AS, DISPLAYS, CONTAINS, IS_ABOUT) are established to capture the compositional, presentation, and discursive structures of social media interactions as well as group membership, semiotic activities, and reactions of users.

The adoption of knowledge graphs using the Connective schema facilitates a dynamic representation of social media data, enabling the articulation of complex relationships and attributes inherent in online interactions. For instance, thematic connections between posts, the network dynamics of user interactions, and the temporal and spatial context of conversations are all encoded within the graph structure. This method aligns with an agency-oriented approach to digital curation theory and practice, which advocates for structured, queryable representations of data prioritizing their dynamic, event-centric dimension, and thus capable of accommodating the fluid and interconnected nature of social facts manifested in records (Dallas 2007). By deploying knowledge graphs, the Connective project ensures that the archive not only serves as a repository of digital artifacts, but also as a rich, navigable resource for exploring the depth and breadth of social media discourse.

Our dual-faceted approach to social media archiving—capturing both the experienced manifestation and the conceptual properties of data—reflects a comprehensive strategy that balances the need for authenticity with the demands of scholarly research. The project's infrastructure is designed with the "fitness for purpose" principle in mind, tailored to meet the specific needs of the social media research community (Dallas 2016). By preserving the experienced context of social media interactions alongside their conceptual underpinnings, the project addresses the critical challenge of representing digital social interactions in a manner that is both faithful to their original context and conducive to rigorous academic inquiry.

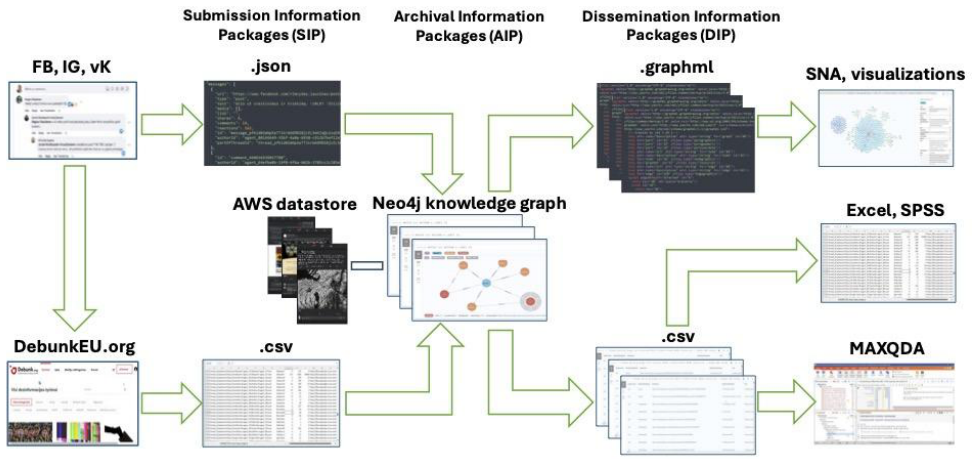
In summary, the representation of social media interactions in the Connective project's data infrastructure embodies an approach to social media archiving that upholds the authenticity of digital interactions while providing a robust framework for their analysis. Through the use of facsimile screenshots and knowledge graphs, the project captures the richness of social media discourse, ensuring its integrity, authenticity, and intelligibility for future research endeavors.

## 4. An Open archives social media data curation architecture

The Connective project's social media data curation architecture (Fig. 1) transcends the stages of the widely accepted Digital Curation Centre (DCC)

digital curation lifecycle and elaborations aiming to extend the model beyond the ingestion to disposal cycle (Higgins 2008; Constantopoulos et al. 2009). Drawing on the principles of the records continuum model (Upward, McKemmish, and Reed 2011), the architecture is designed to accommodate the dynamic and evolving nature of social media data, ensuring that the archive remains relevant and accessible to researchers across various stages of their inquiry.

Figure 1. Connective social media archive data architecture




The data representation and analysis workflow supported by our architecture begins with a multifaceted approach to data selection, guided by research questions that are served by both query-based and collection-based strategies of data capture. This dual approach allows for the targeted gathering of data pertinent to specific research questions within focal case studies, while also accommodating broader sweeps of content for scoping and exploratory analysis of the full archive. Inspired by OAIS, the ISO-standard reference model for an Open Archival Information System (CCSDS 2012), our data architecture. maps the processes, information structures, and systems used to appraise, capture, ingest, enrich, and provide research access to social media information in the Connective project, while also clearly separating representations of data in the archive between Submission Information Packages, Archival Information Packages, and Dissemination Information Packages. This provides for a reliable mechanism to ensure the integrity and authenticity of collected data, while also providing for the needs of a designated community of researchers

without sacrificing analytical power and the dynamic enrichment of data with additional properties and relationships as researchers "exercise the archive" (Dallas 2016).

*Submission Information Package (SIP).* Adopting the concept of the Submission Information Package (SIP) from the Open Archival Information System (OAIS) model, the project curates an initial collection of data that includes both 'raw' JSON streams and facsimile screenshots of social media pages. The JSON stream captures decoded textual and media content as it appears on the platform, preserving the 'raw' data in its most unfiltered form. Concurrently, the screenshots serve to document the experienced manifestation of the content, retaining the layout, design elements, and interactive features integral to interpretation. This combination of data formats provides a comprehensive snapshot of social media interactions, capturing both the underlying data structures and the user-facing presentation of content, and ensuring their integrity and authenticity.

*Archival Information Package (AIP).* The Archival Information Package (AIP) in the Connective social media archive represents a significant advancement in the curation process, where the raw submission data undergoes semantic decomposition and mapping into a knowledge graph. This transformation facilitates a structured representation of social media content, encompassing the structural and semantic elements of interactions and community structures. The knowledge graph not only captures the content and dynamics of social media interactions, but also incorporates provenance metadata detailing the circumstances of data capture. This aspect is particularly crucial in platforms like Facebook, where content visibility can vary based on user profiles and where data may be altered or deleted over time. By documenting the provenance of data, the AIP ensures the traceability and reliability of archived content, providing researchers with essential context for their analyses. Crucially, the representation of archived content by mapping originally captured data as knowledge graphs offers information relevant for research to be searched, categorized, and enriched algorithmically in ways that would have been impossible were the data retained only in their originally ingested format.

*Dissemination Information Package (DIP).* The Dissemination Information Package (DIP) is tailored to meet the diverse needs of researchers, offering both predefined and customizable access methods to the archived data. Researchers can interact with the neo4j knowledge graph through predefined or arbitrary Cypher queries facilitating the exploration of specific themes or patterns within the data. Additionally, scripts can generate filtered and ordered views of the knowledge graph, preparing data for export and further processing in analytical tools such as MaxQDA and various social network analysis software tools. This flexible dissemination

strategy ensures researchers can access and utilize archived data in ways that align with their specific research objectives and methodologies.

In essence, the Connective project's social media data curation architecture embodies an approach integrating the principles of open archives and digital curation to meet the complex demands of social media research. By navigating the challenges of data variability, provenance, and accessibility, the project establishes a robust framework for the long-term preservation and analysis of social media interactions, paving the way for insightful explorations of digital culture and communication in future research.

## 5. Curation/creation: Exercising the social media archive

The innovative approach adopted by the Connective project in developing its social media data curation architecture epitomizes a dynamic interplay between archival preservation and the active generation of new knowledge. Drawing from an agency-oriented digital curation approach (Dallas 2007), the project's methodology underscores the inclusion of evolving representations of social media interactions within Derived Archival Information Packages (CCSDS 2012), dynamically enriched with additional properties and relationships as researchers work with the data. This evolving nature of AIPs facilitates the incorporation of plural insights and analyses conducted by researchers, effectively transforming the archive into a living entity that grows in informational depth and breadth over time.

Using the Connective data infrastructure, we can search for complex lexical patterns in messages, and extract dictionaries and repertoires of themes. We can query the archive asking questions such as “which thematic categories of messages attracted on average the highest number of comments?”, useful to support a mixed-methods investigation, combining qualitative with quantitative analysis on the full corpus of conversations in the archive. We may also identify attribution and predication discursive constructs in social media interactions by initiating a collocation analysis based on dictionary-based lexical queries on the text of Messages, for example, to ask: “which identifications related to heroism are made for persons who have been identified as anti-Soviet partisans?” Applying communicative nexus theory (Laužikas and Dallas forthcoming), we can also use the knowledge graph to explore the circulation of agency between historical referents identified as Topics, people and organizations identified as Actors, and communicative acts identified as Messages, for example, asking questions such as: “when, and by whom, was the idea that Peter Cvirka was a Soviet collaborator first asserted on Facebook, and how did this idea circulate across the network?”

The Connective social media archive is distinguished by the combination of semantically rich representations of social media interactions as knowledge graphs with an open archives data architecture which, while ensuring integrity and authenticity of ingested data, allows their knowledge enhancement as researchers produce plural identifications and relationships through analysis and interpretation. This approach has some important advantages for a scholarly research data infrastructure.

*Knowledge graphs as dynamic repositories.* Central to this dynamic enrichment of the AIPs is the use of knowledge graphs, which serve as the backbone for representing the complex web of social media interactions. The introduction of a faceted Topic node taxonomy, grounded in an ontology that captures the semiotic and epistemic dimensions of online discourse (Kirtiklis et al. 2023), allows for the mapping of diverse aspects identified through axial coding, such as referents, cultural models, and affiliative relationships. This ontological approach enables researchers to embed within the knowledge graph new objects representing refined characterizations and relationships, thereby expanding the archive's conceptual structure.

*Facilitating advanced social media analysis.* This expanded conceptualization of the AIPs is instrumental in supporting advanced scenarios of social media analysis and interpretation. For instance, in analyzing social media discourse surrounding Lithuanian partisans, researchers in the Connective project can identify critical semiotic and discursive constructs such as attributional relationships (whereby individuals are ascribed certain qualities or roles), predicational statements (which articulate actions or states attributed to subjects), and pervasive metaphors (which structure understanding through conceptual mappings). The dynamic nature of the knowledge graph enables these constructs to be identified, categorized, and represented within the AIPs, transforming abstract analytical concepts into tangible elements of the curated research data archive.

*Enriching grounded theory-building:* By enabling the representation of such constructs within Derived AIPs, the Connective project's archival infrastructure not only supports the initial identification of these elements, but also their utilization in iterative cycles of analysis. Researchers can leverage the enriched Derived AIPs to question, refute, or reinforce grounded theories about the phenomena under investigation. For example, the attribution of heroism to historical figures in social media discourse can be examined in the light of prevailing cultural models and metaphors identified within the archive, providing an evidence-based understanding of collective memory construction and identity negotiation in digital memory.

This capability to dynamically enrich the AIPs with new knowledge graph objects and to utilize these enhancements in subsequent analyses

exemplifies the archive's potential role as an active participant in the research process. Rather than serving merely as a repository of static data, the archive becomes a collaborative platform where the boundaries between curation and creation blur, fostering a symbiotic relationship between archival practices and scholarly inquiry.

*Fit for purpose: Serving research needs.* The specification of a research-capable social media archive, as exemplified by the Connective project, underscores the importance of aligning archival practices with the specific needs of the research community. The ability of the archive to adapt to and incorporate the evolving insights of researchers is a testament to its fitness for purpose. By providing a flexible and expandable infrastructure that accommodates the analytical endeavors of scholars, the project ensures that the archive remains not only relevant, but indispensable to the exploration of complex social media phenomena.

In conclusion, the Connective project's approach to social media archiving, characterized by its dynamic and interactive AIPs, introduces a new direction for research-capable archival systems. By enabling the continuous enrichment of the archive with new knowledge and insights generated through scholarly analysis, the project demonstrates the critical role of archives in advancing our understanding of digital culture and communication. This model affirms that for a social media archive to be truly 'fit for purpose' it must be designed to serve the evolving needs of researchers, facilitating the discovery, analysis, and interpretation of social media interactions in ways that account for the dynamic nature, plurality, and context-dependency of social knowledge.

## 6. Reflections on web archiving, digital curation, and research infrastructures

The landscape of web archiving has traditionally focused on long-term preservation, fixity, and authenticity, often neglecting the dynamic and evolving nature of digital content. Driven by institutional memory organizations, web archiving initiatives have emphasized the creation of static repositories to safeguard digital heritage for future generations. While invaluable for preserving the digital record, this approach often overlooks the active use, enrichment, and transformation that characterize the lived experience of digital research ecosystems. Similarly, the field of digital curation has faced its own set of challenges. Initially vibrant and innovative, digital curation research risks stagnation due to its close ties to institutional and preservation-centric initiatives. The prevalent 'preservation vault' lifecycle model often fails to recognize the importance of knowledge representation and the specific needs of designated user communities, such as, notably, researchers. This narrow, custodial perspective on curation

overlooks the researchers' active role in curating social media records, infusing them with meaning through their analytical and interpretive work.

To address the multifaceted challenges posed by digital curation in the social media context, our approach draws on an alternative, pragmatic perspective on digital curation (Dallas 2016). This perspective prioritizes descriptive attention to digital curation as it occurs empirically "in the wild", involving a diverse range of curating actors, activities and objects of curation, over prescriptive rules within the custodial realm of archival professional work. It highlights the importance of embracing multiple viewpoints and relationships among records and their contextual frameworks, rather than adhering to rigid and formulaic approaches. It also recognizes that records are not fixed but evolve across time as knowledge and interpretive frameworks change. This aligns with the need to account for memory and identity as relevant paradigms for archival theory, advocated by Terry Cook's fundamental critique of the narrowness of dominant approaches to the design and professional practices of archives (Cook 2013). It also resonates with the records continuum approach (McKemmish 1997; Upward, McKemmish, and Reed 2011), which emphasizes that the responsibilities of archivists and data managers extend beyond the moment of record creation, encompassing the plural nature of interpretations and the need to consider diverse perspectives, particularly within the realm of online cultures.

This approach acknowledges that social media archives are not static entities but are shaped by ongoing interactions, evolving meanings, and the diverse voices and experiences of users. It is enabled by a comprehensive framework that acknowledges the dynamic nature of social media archives and the challenges posed by diverse perspectives and modes of knowing. Building upon critical digital curation and archival theory, our framework extends beyond the traditional understanding of a social media archive as merely capturing and managing fixed records in a preservation vault. Instead, it welcomes plural interpretations of social media records within the archive, recognizing the perspectival concerns of various stakeholders, and being open to the ethical and political dimensions inherent in decolonial and indigenous ways of knowing, as well as the complexities arising from the dynamics of online communication and the logics of social media platforms.

In the realm of digital research infrastructures, a tension persists between large-scale, often multinational or national projects that mimic traditional archival and library functions, and the more granular, practice-oriented data management activities of individual researchers and projects. The latter, crucial for shaping the epistemic potential and interpretive frameworks of research, remains relatively unexplored in discussions on research infrastructures. This gap highlights the need for a deeper understanding of how data models, taxonomies, and curation practices directly influence the

trajectory and outcomes of scholarly inquiry. The Connective project's approach to creating a social media data archive tailored for scholarly research represents a promising convergence of these domains. By prioritizing the significant properties of social media interactions that are vital to researchers, the project shifts attention from mere preservation to active engagement with the data. This perspective acknowledges the challenges of authenticity and integrity in social media data, while recognizing that the meaningfulness of information suitable for research in the human sciences hinges on what Ian Hacking (1999, 123) refers to as "interactive kinds"—those that are epistemically constructed and dynamically shaped by research activities.

Knowledge representation and semantic technologies, such as knowledge graphs, provide a powerful toolkit for rethinking social media web archiving. These technologies enable a more nuanced and flexible representation of social media data, facilitating the identification, annotation, and linking of content in ways that align with the conceptual and operational needs of social media research. This approach not only enhances the accessibility and usefulness of archives for researchers, but also fosters a curatorial effect through which scholarly engagement actively contributes to the construction and evolution of the research data landscape.

In this regard, the approach adopted by the Connective project may serve as a promising model for integrating web archiving, digital curation, and research infrastructures in a manner that is responsive to the dynamic and interactive nature of social media research. By fostering dialogue between these fields and placing the research process at the forefront as a data curation practice, the project lays the groundwork for a more engaged and reflective approach to the archiving and analysis of social media content. This paradigm shift holds the potential to enrich the field of social media research by opening up new avenues for exploring the complex interplay of digital interactions, cultural practices, and societal discourses in the online sphere.

# References


Acker, Amelia, and Adam Kriesberg. 2017. “Tweets May Be Archived: Civic Engagement, Digital Preservation and Obama White House Social Media Data.” *Proceedings of the Association for Information Science and Technology* 54, 1: 1–9. https://doi.org/10.1002/pra2.2017.14505401001.

Bates, Marcia J. 2009. “An Introduction to Metatheories, Theories, and Models”. *Library and Information Science* 11, 444: 275–97.

Baym, Nancy K. 2010. *Personal Connections in the Digital Age*. Cambridge: Polity.

Ben-David, Anat. 2016. “What Does the Web Remember of Its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top-Level Domain.” *New Media & Society* 18, 7: 1103–19. https://doi.org/10.1177/1461444816643790.

———. 2020. “Counter-Archiving Facebook.” *European Journal of Communication* 35, 3: 249–64. https://doi.org/10.1177/0267323120922069.

Bonacchi, Chiara. 2022. *Heritage and Nationalism: Understanding Populism through Big Data*. UCL Press. https://doi.org/10.2307/j.ctv1wdvx2p.

Borji, Samaneh, Amir Reza Asnafi, and Maryam Pakdaman Naeini. 2022. “A Comparative Study of Social Media Data Archiving Software.” *Preservation, Digital Technology & Culture* 51, 3: 111–19. https://doi.org/10.1515/pdtc-2022-0013.

Brügger, Niels. 2005. *Archiving Websites: General Considerations and Strategies*. Aaarhus: Center for Internetforskning.

———. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge, Mass.: MIT Press.

Brügger, Niels, and Ditte Laursen. 2019. *The Historical Web and Digital Humanities: The Case of National Web Domains*. Routledge.

Brügger, Niels, and Ralph Schroeder, eds. 2017. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press. https://doi.org/10.14324/111.9781911307563.

Bruns, Axel. 2019. “After the ‘APIcalypse’: Social Media Platforms and Their Fight against Critical Scholarly Research.” *Information, Communication & Society* 22 (11): 1544–66. https://doi.org/10.1080/1369118X.2019.1637447.

CCSDS. 2012. “Reference Model for an Open Archival Information System (OAIS).” Recommended Practice. Washington, DC: Consultative Committee for Space Data Systems (CCSDS).

Cook, Terry. 2013. “Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms.” *Archival Science* 13 (2–3): 95–120. https://doi.org/10.1007/s10502-012-

9180-7.
Dallas, Costis. 2007. “An Agency-Oriented Approach to Digital Curation Theory and Practice.” In *The International Cultural Heritage Informatics Meeting Proceedings*, edited by Jennifer Trant and David Bearman. Toronto: Archives & Museum Informatics. http://www.archimuse.com/ichim07/papers/dallas/dallas.html.
———. 2016. “Digital Curation beyond the ‘Wild Frontier’: A Pragmatic Approach.” *Archival Science* 16 (4): 421–57. https://doi.org/10.1007/s10502-015-9252-6.
———. 2018. “Heritage Encounters on Social Network Sites, and the Affiliative Power of Objects”. In *Culture and Perspective at Times of Crisis: State Structures, Private Initiative and the Public Character of Heritage*, edited by Sophia Antoniadou, Ioannis Poulios, George Vavouranakis, and Pavlina Raouzaiou, 116–31. Oxford: Oxbow Books.
Edwards, Adam, William Housley, Matthew Williams, Luke Sloan, and Malcolm Williams. 2013. “Digital Social Research, Social Media and the Sociological Imagination: Surrogacy, Augmentation and Re-Orientation.” *International Journal of Social Research Methodology* 16 (3): 245–60. https://doi.org/10.1080/13645579.2013.774185.
Engeström, Yrjö. 1999. “Activity Theory and Individual and Social Transformation.” In *Perspectives on Activity Theory*, edited by Yrjö Engeström, Reijo Miettinen, and Raija-Leena Punamäki-Gitai, 19–37. Learning in Doing. Cambridge; New York: Cambridge University Press.
Fiesler, Casey, and Nicholas Proferes. 2018. “’Participant’ Perceptions of Twitter Research Ethics.” *Social Media + Society* 4 (1): 205630511876336. https://doi.org/10.1177/2056305118763366.
franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Ess, and Association of Internet Researchers. 2020. “Internet Research: Ethical Guidelines 3.0 Association of Internet Researchers.” Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf.
Garton, Laura, Caroline Haythornthwaite, and Barry Wellman. 1997. “Studying Online Social Networks.” *Journal of Computer-Mediated Communication* 3 (1): 0–0. https://doi.org/10.1111/j.1083-6101.1997.tb00062.x.
Grace, Stephen, Gareth Knight, and Lynne Montague. 2009. “Investigating the Significant Properties of Electronic Content over Time (InSPECT) – Final Report.” London: King’s College London. https://significantproperties.kdl.kcl.ac.uk/inspect-finalreport.pdf.
Hacking, Ian. 1999. *The Social Construction of What?* Cambridge, Mass: Harvard University Press.
Helmond, Anne, and Fernando N. van der Vlist. 2019. “Social Media and Platform Historiography: Challenges and Opportunities.” *TMG–Journal for Media History* 22 (1). https://doi.org/10.18146/tmg.434.
Hemphill, Libby, Susan H. Leonard, and Margaret Hedstrom. 2018. “Developing a Social Media Archive at ICPSR.” In *Proceedings of Web Archiving and Digital Libraries (WADL’18)*. New York: ACM. http://deepblue.lib.umich.edu/handle/2027.42/143185.
Hockx-Yu, Helen. 2014. “Archiving Social Media in the Context of Non-Print Legal Deposit”. In. Lyon. https://library.ifla.org/id/eprint/999/.
Kelpšienė, Ingrida. 2021. “Participatory Heritage: A Multiple-Case Study of Lithuanian Grassroots Cultural Heritage Communities on Facebook.” Doctoral Dissertation, Vilnius, Lithuania: Vilnius University. https://doi.org/10.15388/vu.thesis.181.
Kelpšienė, Ingrida, Donata Armakauskaitė, Viktor Denisenko, Kęstas Kirtiklis, Rimvydas Laužikas, Renata Stonytė, Lina Murinienė, and Costis Dallas. 2023. “Difficult Heritage on Social Network Sites: An Integrative Review.” *New Media & Society* 25 (11): 3137–

64. https://doi.org/10.1177/14614448221122186.
Kietzmann, Jan H., Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. 2011. "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media." *Business Horizons* 54 (3): 241–51.
Kirtiklis, Kęstas, Rimvydas Laužikas, Ingrida Kelpšienė, and Costis Dallas. 2023. "An Ontology of Semiotic Activity and Epistemic Figuration of Heritage, Memory and Identity Practices on Social Network Sites." *SAGE Open* 13 (3): 1–25. https://doi.org/10.1177/21582440231187367.
Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. 2018. "API-Based Social Media Collecting as a Form of Web Archiving." *International Journal on Digital Libraries* 19 (1): 21–38. https://doi.org/10.1007/s00799-016-0201-7.
Lomborg, Stine. 2012. "Researching Communicative Practice: Web Archiving in Qualitative Social Media Research." *Journal of Technology in Human Services* 30 (3–4): 219–31. https://doi.org/10.1080/15228835.2012.744719.
Lomborg, Stine, and Anja Bechmann. 2014. "Using APIs for Data Collection on Social Media." *The Information Society* 30 (4): 256–65. https://doi.org/10.1080/01972243.2014.915276.
Lotman, Juri. 2005. "On the Semiosphere." Translated by Wllma Clark. *Σημειωτική-Sign Systems Studies*, 1: 205–29.
Manca, Stefania. 2020. "Bridging Cultural Studies and Learning Science: An Investigation of Social Media Use for Holocaust Memory and Education in the Digital Age." *Review of Education, Pedagogy, and Cultural Studies* 43 (3). https://www.tandfonline.com/doi/abs/10.1080/10714413.2020.1862582.
Marres, Noortje, and Esther Weltevrede. 2013. "Scraping the Social?: Issues in Live Social Research." *Journal of Cultural Economy* 6 (3): 313–35. https://doi.org/10.1080/17530350.2013.772070.
McKemmish, Sue. 1997. "Yesterday, Today and Tomorrow: A Continuum of Responsibility." In *Proceedings of the Records Management Association of Australia 14th National Convention, 15–17 Sept. 1997*. Perth, Western Australia: RMAA. http://www.infotech.monash.edu.au/research/groups/rcrg/publications/recordscontinuum-smckp2.html.
Michel, Alejandra, Jessica Pranger, Friedel Geeraert, Sven Lieber, Peter Mechant, Eveline Vlassenroot, Sally Chambers, Julie Birkholz, and Fien Messens. 2021. "WP1 Report: An International Review of Social Media Archiving Initiatives." Report. https://orfeo.belnet.be/handle/internal/7741.
Milligan, Ian, Nick Ruest, and Jimmy Lin. 2016. "Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses." In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 107–10. Newark New Jersey USA: ACM. https://doi.org/10.1145/2910896.2910913.
Ogden, Jessica, and Emily Maemura. 2021. "'Go Fish': Conceptualising the Challenges of Engaging National Web Archives for Digital Research." *International Journal of Digital Humanities* 2 (1–3): 43–63. https://doi.org/10.1007/s42803-021-00032-5.
Papacharissi, Zizi, ed. 2011. *A Networked Self: Identity, Community and Culture on Social Network Sites*. New York: Routledge.
Pehlivan, Zeynep, Jérôme Thièvre, and Thomas Drugeon. 2021. "Archiving Social Media: The Case of Twitter." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 43–56. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_5.
Perriam, Jessamy, Andreas Birkbak, and Andy Freeman. 2020. "Digital Methods in a Post-API Environment." *International Journal of Social Research Methodology* 23 (3): 277–

90. https://doi.org/10.1080/13645579.2019.1682840.
Robinson, Ian, Jim Webber, and Emil Eifrem. 2015. *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media.
Rogers, Richard. 2013. *Digital Methods*. Cambridge, Massachusetts: The MIT Press.
Rutten, Ellen. 2013. "Why Digital Memory Studies Should Not Overlook Eastern Europe's Memory Wars." In *Memory and Theory in Eastern Europe*, edited by Uilleam Blacker and Alexander Etkind, 219–31. New York: Palgrave Macmillan. http://link.springer.com/chapter/10.1057/9781137322067_11.
Schäfer, Mirko Tobias, and Karin van Es, eds. 2017. *The Datafied Society. Studying Culture through Data*. Amsterdam: Amsterdam University Press. http://en.aup.nl/books/9789462981362-the-datafied-society.html.
Shibuya, Yuya, Andrea Hamm, and Teresa Cerratto Pargman. 2022. "Mapping HCI Research Methods for Studying Social Media Interaction: A Systematic Literature Review." *Computers in Human Behavior* 129: 107131.
Snelson, Chareen L. 2016. "Qualitative and Mixed Methods Social Media Research: A Review of the Literature." *International Journal of Qualitative Methods* 15 (1): 1609406915624574. https://doi.org/10.1177/1609406915624574.
Social Media Lab, Toronto Metropolitan University. 2024. "Social Media Research Toolkit." *Social Media Lab* (blog). January 2024. https://socialmedialab.ca/apps/social-media-research-toolkit-2/.
Stieglitz, Stefan, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. "Social Media Analytics – Challenges in Topic Discovery, Data Collection, and Data Preparation". *International Journal of Information Management* 39 (Complete): 156–68. https://doi.org/10.1016/j.ijinfomgt.2017.12.002.
Stoycheff, Elizabeth, Juan Liu, Kunto A. Wibowo, and Dominic P. Nanni. 2017. "What Have We Learned about Social Media by Studying Facebook? A Decade in Review." *New Media & Society* 19 (6): 968–80. https://doi.org/10.1177/1461444817695745.
Thomson, Sara Day. 2017. "Preserving Social Media: Applying Principles of Digital Preservation to Social Media Archiving." In *Researchers, Practitioners and Their Use of the Archived Web*, 1–13. London: School of Advanced Study, University of London. https://doi.org/10.14296/resaw.0007.
Thomson, Sara Day, and William Kilbride. 2015. "Preserving Social Media: The Problem of Access." *New Review of Information Networking* 20 (1–2): 261–75. https://doi.org/10.1080/13614576.2015.1114842.
Tufekci, Zeynep. 2017. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. New Haven; London: Yale University Press.
Upward, Frank, Sue McKemmish, and Barbara Reed. 2011. "Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures." *Archivaria* 72 (January): 197–237.
Vlassenroot, Eveline, Sally Chambers, Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Alejandra Michel, and Peter Mechant. 2019. "Web Archives as a Data Resource for Digital Scholars." *International Journal of Digital Humanities* 1: 85–111.
Vlassenroot, Eveline, Sally Chambers, Sven Lieber, Alejandra Michel, Friedel Geeraert, Jessica Pranger, Julie Birkholz, and Peter Mechant. 2021. "Web-Archiving and Social Media: An Exploratory Analysis." *International Journal of Digital Humanities* 2 (1): 107–28. https://doi.org/10.1007/s42803-021-00036-1.
Voss, Alex, Ilia Lvov, and Sara Day Thomson. 2017. "Data Storage, Curation and Preservation." In *The SAGE Handbook of Social Media Research Methods*, edited by Luke Sloan and Anabel Quan-Haase, 161–76. SAGE Publications Ltd London. https://www.torrossa.com/gs/resourceProxy?an=5018794&publisher=FZ7200#page=190.

Weller, Katrin. 2014. “What Do We Get from Twitter—and What Not? A Close Look at Twitter Research in the Social Sciences.” *Knowledge Organization* 41 (3): 238–48. https://doi.org/10.5771/0943-7444-2014-3-238.

Wilson, Steven Lloyd. 2022. *Social Media as Social Science Data*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781108677561.

Winters, Jane. 2017. “Coda: Web Archives for Humanities Research–Some Reflections.” In *The Web as History: Using Web Archives to Understand the Past and Present*, edited by Niels Brügger and Ralph Schroeder, 238–48. London: UCL Press.

Zimmer, Michael, and Katharina Kinder-Kurlanda. 2017. *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts*. New York, NY: Peter Lang.