Cambridge Forum on AI: Culture and Society (2026), **2**, e21, 1–10
doi:10.1017/cfc.2026.10021

Cambridge Forum

ROUNDTABLE PAPER

Lessons from archival provenance: Reframing data creator relationships in AI

Emily Maemura

School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA
Email: emaemura@illinois.edu

(Received 2 March 2025; revised 9 March 2026; accepted 15 April 2026)

Abstract

This paper presents a brief history of archival provenance and its shifting role across different paradigms of archives, considering how this past work can provide insights into current trajectories of AI. It explores what archival provenance does, or can do, differently than current imaginations of data provenance in AI. It proposes three "lessons from archival provenance" that challenge the present focus of existing approaches in AI, in particular highlighting how narrowly this work structures and models evidence, authenticity and creatorship of data. These lessons can be used to identify new opportunities for applying concepts of archival provenance by presenting a roadmap for future work where AI research might align and converge with related work in archival theory. It concludes by asserting that grounding AI datasets in this perspective of archival provenance can realize a new paradigm of "data as archives," serving to envision responsibilities to a range of data creators, determine needs for documentation of context and establish a crucial role for "data archivists."

Keywords: archives; data collection; data creators; data curation; datasets; provenance

1. Introduction

In their influential paper on fairness, accountability, transparency and ethics in machine learning (ML), Jo and Gebru (2020) propose a set of "lessons from archives," envisioning new institutional frameworks and procedures for data collection and annotation. They argue that archives and libraries, as "fields dedicated to human data collection for posterity," can contribute to the conversations around sociocultural data used in ML, especially on topics of "ethics, representation, power, transparency, and consent" (Jo and Gebru 2020, 315). But where are the archivists in these conversations? From the perspective of archivists, any "lessons from archives" should take into account the foundational role of archival provenance for archives theory and practice.

This analysis aims to generate a new set of lessons for AI data practices that revolve around archival provenance. First, I interrogate what the principle of provenance has meant for archives, and then consider how this evolving understanding might complicate current notions of evidence, authenticity and creatorship for ML training data. I develop this analysis based on a case study and close reading of recent work from the Data Provenance Initiative (DPI; Longpre et al. 2023; 2024). I close with three new "lessons," asserting that archival provenance and its evolving role in archival science can provide insights into current trajectories of AI and provide a model for ways to interrogate relationships embodied in data as (historical) sources.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (http://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

2. Characterizing archival provenance

Data, archives and records can each take on various semantic uses and connotations, and I begin by grounding this discussion in the typology of data developed by Borgman (2015). I observe that the majority of datasets used for ML training (especially for generative AI) are ultimately derived from the category she calls "records" which are produced by human activity and distinct from data such as observations of natural phenomena or derived from experimentation. Borgman's consideration of records as data sources covers a broad range of information: "documentation of government, business, public, and private activities; books and other texts; archival materials; documentation in the form of audio and video recordings," etc. (Borgman 2015, 25). I propose here that the theory and practice specifically developed by archivists and their concern with managing records can provide a useful framing, even though their perspective is rather narrower than Borgman or Jo and Gebru's conceptualization of sociocultural data as records.

For ensuring the authenticity and trustworthiness of records, archival provenance has served as the foundation of modern, Western archives from initial development in French and Dutch archival practice during the 19th century, and implemented throughout UK and US government bureaucracies in the early 20th century. The principle of provenance rests on the concept that records must be understood and interpreted in the context of their creation. In contrast to information resources in library settings which are organized through systems like subject headings and classification schemes, records within archival settings are managed and organized according to their origins, with a body of records tied to its single, original creator.[1] Archival provenance's subprinciples of *respect des fonds* and respect for original order additionally mandate that archivists do not co-mingle records from different creators and do not interfere with the structuring of records within files and folders according to the orders imposed by their creators. Archival provenance also recognizes the relationships and meaning formed between an entire body of records from the same creator, unlike understandings of provenance from other fields that focus on tracing the lineage of a single object. Preserving the relationships between records in aggregate rather than discrete units is, therefore, essential for archival provenance to ensure the evidentiary value of records and to contextualize and authenticate any given record within the creator's broader system of recordkeeping.

This traditional view of archival provenance has also been challenged and debated over the past several decades.[2] Rather than re-tread this ground, I build here upon two influential pieces that have traced the histories of provenance, beginning with Douglas (2017) who recounts the origins and evolution of the principle of provenance, exploring its shifting role over time. Douglas notes that the introduction of provenance began as an organizing principle, then shifted to a conceptual construct for arranging a group of records according to a singular creator. She then describes more recent understandings of provenance as the broader sociohistorical context that influences the creation and preservation of records. For instance, Nesmith (2006) develops the concept of societal provenance using the example of fur trader Johann Steinbruck's birchbark journal from 1802 to 1803. He argues that the journal's provenance extends beyond the individual trader and his employer, the North West Company, considering how the journal's materiality stems from Aboriginal communities' practices of writing on birchbark, and noting the additional context of Steinbruck's marriage to an Aboriginal woman. The journal further illustrates Nesmith's concept of societal provenance that includes the broader circumstances of European-Aboriginal relations at the time, as well as subsequent actors in the record's lifecycle such as the family who retained custody of the journal, and the editors and translators involved in reproducing and publishing the journal into a coffee table book in 1999. Work by Wurl additionally complicates an understanding of provenance, considering ethnicity as provenance

[1]In archives, a records creator is typically understood as a department or agency in the case of government and business archives, or an individual or family for personal archives.

[2]Most recently, this has led to scholarship such as a recent special issue devoted to new perspectives on archival provenance (Bastian et al. 2024).

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

and addressing the need for archivists to recognize and document "environments of social affiliation" that "may often overlap, intersect, and even push against each other … society truly does not sort itself out in neat corporate compartments, and as archivists, we need to learn to brace ourselves accordingly" (Wurl 2005, 71). In her own study on writers' archives, Douglas (2016) has also noted the range of actors who shape archives over time, as custodians and other interested parties – including archivists themselves – may act as "coaxers and coercers" influencing the historical record. Thus, archival provenance has evolved to account for a range of actors, frames of reference and societal influences that determine if and when a record is created, by what means, and what information is included within it.

Where Douglas traces evolving concepts of the principle of provenance, Cook (2013) delineates this evolution as a series of paradigm shifts, each defining a new role for both archives and for archivists. He observes that early and pre-modern archives emphasized the evidentiary value of records and positioned archivists as neutral, impartial actors preserving authentic sources of juridical evidence. Subsequent post-modern approaches emphasize the partiality of materials that ultimately come to reside in archives, recognizing the active role of archivists in appraisal, selection and interpretation of these records. He identifies a central tension between evidence and memory, noting shifts in practice that have archivists have variously embraced "our guardianship role, … of the archival product, the evidence, on the one hand, *versus* our interpretive or mediating role, on the other, as manifested in all of the many archival processes, the memory-making" (Cook 2013, 99–100).

Recognizing that archives are not neutral stores of evidence has led to studies of power enacted by archives and their resulting gaps and absences, highlighting where particular voices and marginalized groups are not represented in archives. For instance, Stoler (2002) addresses the colonial systems of knowledge and power relations inscribed in archives and proposes a shift from studying "archive-as-source to archive-as-subject" in order to interact with colonial archives as historical sources, which requires rethinking "what kinds of truth-claims lie in documentation" (94). Harris (2002), working within the context of South Africa and the government archives of apartheid rule, presents the metaphor of the "archival sliver" to recognize what is not included in the official historical record, through both "deliberate and inadvertent destruction by records creators and managers" as well as archivists' selection for what is preserved over time, leaving researchers with access to only "a sliver of a sliver of a sliver" (65). While Harris addresses the gaps of the documentary record as inevitable, Cook takes an optimistic view of the role of the archivist as a mediator of the documentary record, proposing that archivists can be pivotal in shaping societal memory for social justice. Cook envisions a new paradigm for the contemporary moment in which archivists work in closer collaboration with a broader set of actors from outside of professional and institutional settings to determine archival value and authenticity.

Thus, archival provenance has evolved from a focus on maintaining rigid, bureaucratic structures supporting objective evidence, to a guiding principle for more subjective interpretation of archival aggregations. Newer theorizations of archival provenance have asserted the role and value of archival materials for social identity (and reciprocally shaped by social systems and pressures) and continued to question traditional notions of "creators" who contribute to records over time, incorporating collaborative efforts of communities to engage in processes of meaning-making.

While Cook and Douglas briefly touch on the implications of digital technologies for provenance, archival scholarship in this area has continued to develop since their writing. For instance, studies of web and social media archiving have addressed differences from traditional archives due to the scale of collection necessitating "Wild West Web Crawling" similar to what Jo and Gebru have described for ML data collection (Hegarty 2025; Maemura et al. 2018; Ogden 2022; Summers and Punzalan 2017). Jaillant (2021) and Colavizza et al. (2021) both explore how AI can be applied to existing archives workflows and models of practice to support automation of appraisal, description and enable new forms of access. Mordell (2019) considers new possibilities for an "archives as data" paradigm, following from the approach of "Collections as Data" (Padilla et al. 2019), by presenting ways that

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

computational archival science can engage more directly with "allied disciplines like digital humanities, critical information studies, and more specialized fields like critical data studies, ancestral and decolonial computing, and software studies" (159–60) for more pluralistic and inclusive approaches. Caswell (2023) presents an emphatic counterpoint in the position paper "Against Archival Collections as Data," arguing that datafication can rob archives of their context and meaning. Rather than focusing on the myriad influences of data practice and culture upon archives that have been considered for "archives as data," my analysis here takes the inverse perspective to reposition "data as archives,"[3] interpreting data's relationships and contexts via archival provenance. I explore what can be learned from the trajectory traced by archival theory's evolving conceptualization of provenance over the past two centuries, and how this might inform current work in AI to better conceptualize the origins, creators and authenticity of data upon which AI systems are built.[4]

3. Examining paradigms of AI data provenance

Addressing how archival provenance might serve to inform future work in ML and AI, I focus here on an analysis of recent work by Longpre et al. (2024) which I use as a case study and representative example of state of the art scholarship on data provenance in AI. The authors point to a range of current issues that result from the lack of data provenance, referencing their related work developing solutions through the DPI (Longpre et al. 2023). I examine how these authors frame "data provenance," specifically attending to their concepts of authenticity, evidence and creators.

In describing the need for data provenance, Longpre et al. use several phrases referencing the concept of "tracing authenticity" and the idea of tracking sources and origins of data. Specific phrases include: "source authenticity," "source lineage," "digital watermarks" and "content authenticity verification," citing work from the Coalition for Content Provenance and Authenticity, which is developing methods for detecting tampering with digital assets (Longpre et al. 2024, 5). This discursive framing of authenticity emphasizes tracing a chain of custody for data as evidence, connections to reproducibility of methods and singular technological approaches to assessing what is considered authentic. I infer here that this understanding of authenticity and evidence in relation to data provenance stems from science and computationally intensive research. For instance, within scientific workflows, data provenance has been established to track data origins or lineage to ensure authenticity and support the reliability and reproducibility of results (Moreau et al. 2008). Framing provenance (and data) in this way lends validity and credibility to the whole project of AI – provenance enshrines the value of data as scientific evidence and reinforces the products of AI as objective and neutral.

While Longpre et al. do not specifically define or characterize "creators," they emphasize the need to "protect rights" and "avoid harms" to creators. They reference artists and writers as specific creators who are "dissatisfied and disempowered" by AI and discuss broader effects on the "creative economy" (Longpre et al. 2024, 2). Citing specific lawsuits brought by creators against AI companies, the authors note how creators are concerned with "the three C's of creative rights: compensation, control and credit" (Longpre et al. 2024, 2; referencing Chayka 2023). In effect, this discourse emphasizes creators specifically as copyright holders, whose role is to grant licensing agreements for the benefit of AI progress, furthering research and innovation.

"Creators" also appear to be considered as actors distinct from "dataset creators." In the context of the DPI, Longpre et al. (2023) describe dataset creators as "institutions of the dataset authors,

[3]While "data archives" can also describe scientific data repositories and databases, the curatorial practices typically used within those contexts are not informed by training in archives, records management, and historical or humanistic methods of inquiry (see Borgman et al. 2019; Nadim 2016).

[4]I want to emphasize how the origins and conceptualizations of provenance in archives are distinct from the consideration of data provenance emerging from computer science, but note here that this work of bringing concepts from archival science to data shares an affinity with Fordyce and Jethani (2021) who take an alternate approach by extending Buneman, Khanna and Tan's (2000) framing of data provenance for critical studies of data in human and social sciences, focusing on discourse and rhetoric.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

including universities, corporations, and other organizations" (5). There is also limited discussion of secondary roles in creating data. For instance, in discussing the possibilities for science and scholarship, Longpre et al. (2024) briefly juxtapose ML outputs with data from human subject research noting "annotation decisions are more opaque, pretrained models are not always fully open, and model-based agents are neither as controllable as a human confederate nor as realistic as a study participant" (3). However, the role of other humans as annotators or other participants in the creation of data is largely unexplored in their work, and this brief gesture serves to maintain a framing based on social scientific investigations and regulatory ethics frameworks of research institutions.

Longpre et al. (2024) also present a solution to these challenges: in their view, documenting provenance information can help prevent or overcome intellectual property disputes, as well as support developing "ethical and trustworthy" models by "sourcing training data responsibly" (2). Legal and ethical concerns are both frequently cited as central to the work of responsible AI and its aims of fairness, accountability and transparency. Yet, as is evident in the discourse from Longpre et al., tracing provenance to a single creator is seen primarily through the lens of legal concerns, rather than ethical concerns, due to the more immediate threat of regulatory repercussions or lawsuits. This orientation toward legal compliance is reinforced through the proposed structures for tracing data lineage in the DPI's implementation, which emphasizes the need to determine a singular rights holder. Notably, regulatory language also looms large in Longpre et al.'s construction of phrases and solutions such as "provenance disclosure" (2024, 9) and "audit of AI data provenance" (2023).

While it would initially appear that AI's orientation to provenance is entirely based on the premise of data sources as scientific, objective evidence, the legal and financial implications complicate matters further. Scientific understandings of data as evidence require a treatment of datasets as facts which are not protected according to US copyright law (therefore threatening the "dataset creators" noted above as institutions and corporations). In contrast, AI training data are currently positioned as a commodity threatened by the risk of copyright claims from cultural content creators that could disrupt the "AI data supply chain." The ambiguous legal status of AI datasets (for which copyright claims have not yet been tested or cemented in court decisions) results in hedging arguments for fair use exceptions recognized by US law and the related concept of "transformative use" versus what is considered a "copy" or derivative work for which licensing agreements are required. Conceptualizations of AI data provenance have, therefore, both relied upon and diverged from scientific understandings of data provenance.

4. Lessons from archival provenance

What can the wealth of past scholarship examining archival provenance offer our present moment to understand AI training data? I present here three "lessons from archival provenance" that can contribute a new perspective with which to weave together existing strands of critical data studies. In particular, archival provenance can be seen as a lens through which to view and critique concepts of evidence, authenticity and creatorship beyond the reductive positivist and legalistic framings currently embraced in AI. Complementing the parallels drawn by Jo and Gebru, I present here a roadmap for increased collaboration between AI and archives and also identify spaces where overlaps and convergences already exist.

Lesson 1: Authenticity and evidence are always partial and political; the "archival sliver" of AI training data is shaped by both sociohistorical context and sociotechnical decision-making. Thylstrup (2022) has already drawn upon critical archival studies, and similar to Stoler's work, calls for a critical dataset studies to focus on datasets as "subject" not just "source." Past work based in archival scholarship can additionally provide useful guidance and frameworks for studying data as non-neutral aggregations inflected by contemporary political concerns. This includes examining the ways datasets are shaped and whittled down across their lifecycle similar to Harris' "archival sliver," as seen in recent work analyzing and classifying a typology of "forgetting practices in data sciences"

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

(Muller and Strohmayer 2022). Hegarty (2022) also importantly explores how "biases, inequalities, and silences" in national web archives crawled from web materials result from varied social, material and technical bases, concluding that "silences should not be treated *a priori* as a 'gap' that needs to be 'filled'; rather absence can reflect people expressing agency over the contexts in which they interact" (40). Viewing this work as an extension of "societal provenance," the sociotechnical space of the web (from which much AI training data also originate) demands greater attention to the forces influencing the archival process that determines what materials are created and preserved – reinforcing yet again that this documentary record can never be perfect or complete, we can only hope gain better insights into its contingent nature. Recent work by Desai et al. (2024) already considers such convergences between AI and archival theory, applying the framing of archival appraisal to consider how power is enacted in the practices of generating pre-training datasets, through choices such as exclusion of materials with toxic language.

The shared concerns of evidence, authenticity and "truth-claims" for AI and archives also extend beyond the remains and absences within a dataset to address more fundamental questions of positionality and epistemology. Past work from science and technology studies has considered and challenged the notion of "ground-truth" in ML practice (Jaton 2021; Ratner and Thylstrup 2025). For archival studies, MacNeil (2025) recently calls for greater attention to how concepts of authenticity can be expanded and incorporate more inclusive and person-centered approaches, building upon long-standing work on pluralizing the archival profession (Douglas et al. 2022; Pluralizing the Archival Curriculum Group 2011). Considering data as a kind of archives places greater attention on who or what they are authentic to, and the need for embracing pluralistic approaches to address AI data's varied roles along the spectrum of evidence and memory.

Lesson 2: Concerns for creators can expand beyond legal rights and licensing; archival scholarship provides a blueprint for how to broadly conceptualize and pluralize understandings of creators and their relationships to data. Work from archival theory can serve to challenge the underlying assumption in current conceptions of AI data provenance that identifying sources is beneficial to establish authority of the assembled dataset, and that the attribution to a single creator is both feasible and adequate. For instance, in discussing social justice work in archives, Punzalan and Caswell (2016) challenge the predominant legal approaches focused on individual violations of human rights, noting that this approach "ignores the realities of more subtle, intangible, and shifting forms of oppression that are also pressing social justice concerns" (8). I propose that this insight can be extended to reconsider the ethical responsibilities of AI technologies to the various actors who are related to and represented in datasets. Recent work in archives explores relationships to records beyond a single creator and can provide a counterpoint to the existing narrow focus of AI data provenance on authorship and rights holders.

As an alternative to rights-based approaches, Caswell and Cifor (2016) propose a new model of radical empathy based on feminist ethics. They define four affective responsibilities archivists must adopt, through relationships between the archivist and the following: (a) the records creator, considering if or how the creator would want to make materials available; (b) the subject of records, i.e., "those about whom records are created, often unwittingly and unwillingly," and considering how to "recover and reassert the choices of records subjects in the archival process" (36); (c) the archives user, acknowledging "the affective impact of finding – or not finding – records that are personally meaningful, and the personal consequences that archival interaction can have on users" (37); and (d) the larger community, considering "those who are not direct users of records, but for whom the use of records has lasting consequences" (38–39). Extending from this last point considering a "community" as kind of creator or relationship that archivists must responsibly foster, Brilmyer et al. (2019) explore how conceptualizations (or "imaginaries") of different marginalized communities inform community archives, and how the boundaries of the community (and the archive) shift over time. Wickner (2019) additionally considers the influence of "archivists, subjects, users, technical agents,

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

governments, corporations, communities, institutional structures, and alternative kinship ties as co-creators" (20–21) and develops a model of co-creatorship that encompasses the varied actors. Future work studying the construction of AI data might, therefore, address co-creators more directly, like the identities and positionalities of human annotators and their interpretive processes, building off work from Miceli et al. (2020; 2024; 2025) and Zhang et al. (2025). Taking inspiration from archival scholarship's focus on relationships with creators and communities represented in records, as well as custodians and actors engaged with records throughout their lifecycle, can, therefore, serve to re-emphasize the ethical dimensions of data provenance for AI.

Lesson 3: A documentation standard does not solve all provenance problems; provenance documentation is best understood as guiding an interpretive process, not as an authoritative end-product. This observation extends from the solution-oriented approach of the DPI proposal as well as a broader trend toward documentation standards for transparent, responsible and explainable AI (e.g., Bender and Friedman 2018; Gebru et al. 2021; Holland et al. 2018; Krause 2019; McMillan-Major et al. 2023; Pushkarna and Zaldivar 2022; Zelenka and Di Cara 2024). While efforts toward standardized documentation are laudable, these standards alone cannot realize the goals of ethical and responsible AI without attendant investments in the labor required to generate this documentation.[5] Rather than centering documentation products, new studies can and should focus on the processes of documentation, interpretation and contextualization of data, to better understand when and how this work can be embedded in the data lifecycle. I propose here that AI provenance documentation can take inspiration from studies of archivists' practices of description and documentation to re-conceptualize roles for "data archivists" as actors trained to engage in processes of interpreting data and their context.

For instance, Meehan (2009) outlines the various processes archivists undertake to analyze and document information about records and their creators, including sourcing evidence and making inferences. MacNeil (2005) also emphasizes the role of analysis and interpretation, drawing analogies between archival description and textual criticism. Returning to Cook (2013), his vision for a community-oriented paradigm positions archivists in the role of "mentors, facilitators, coaches, who work in the community to encourage archiving as a participatory *process* shared with many in society" (114). That vision is reflected in recent work that explores how actors outside of archival settings can be engaged in reparative or decolonizing description efforts (Haberstock 2020; Luke and Mizota 2024). Documentation requires not only standards but training and dedicated roles that support and take responsibility for interpreting and enriching datasets through insights into the processes of collection, annotation and use. Learning from archives, the role of "data archivists" in AI might be reconfigured with attendant training for individuals to interpret data broadly, engage creators and communities, and generate comprehensive description of data across its lifecycle.

5. Conclusion: shifting paradigms for data provenance

In order for "Machine Learning to become realized as an archival science" as Taurino and Smith (2022) argue, the field must move beyond discursive references to an "archival turn" and engage more directly with archival theory, which has evolved for two centuries to consider sociocultural relations between records and creators. Reflecting on the history and evolution of archival theory and provenance, Douglas' account of the evolving conceptualization of provenance and Cook's shifting paradigms for archivist roles can serve to promote new understandings of provenance and recognize the range of creators and forces shaping the documentary record. Current work in responsible and ethical AI aims to "reduce social biases" in the data used to train models, a challenge familiar to engagement with archives as sites of memory, identity and community. Yet, the conceptualization of data provenance in AI remains rooted in an "objective evidence" paradigm and limited to tracing

[5] Gebru et al. (2021) also emphasize that documentation cannot be automated and must engage dataset creators in reflection on their own processes.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

data's lineage to a single creator. In this way, lessons from archival provenance and considerations of data as a kind of archives can help envision roles for "AI data archivists" who are responsible to a range of data creators, and broadening what contexts and processes provenance documentation must address.

Funding statement. The author(s) declare none.

Competing interests. The author(s) declare none.

References

Bastian, J. A., S. H. Griffin and J. Lowry (Eds.) 2024. "Provenance Special Issue." *Archival Science* 24 (4): 555–557. https://doi.org/10.1007/s10502-024-09467-5.

Bender, E. M. and B. Friedman. 2018. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6 (December): 587–604. https://doi.org/10.1162/tacl_a_00041.

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Borgman, C. L., A. Scharnhorst and M. S. Golshan. 2019. "Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse." *Journal of the Association for Information Science and Technology* 70 (8): 888–904. https://doi.org/10.1002/asi.24172.

Brilmyer, G., J. Gabiola, J. Zavala and M. Caswell. 2019. "Reciprocal Archival Imaginaries: The Shifting Boundaries of 'Community' in Community Archives." *Archivaria* 88: 6–48.

Buneman, P., S. Khanna and W.-C. Tan. 2000. "Data Provenance: Some Basic Issues." In *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, edited by S. Kapoor and S. Prasad, 87–93. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-44450-5_6.

Caswell, M. 2023. "Against Archival Collections as Data." *Position Statements -> Collections as Data: State of the field and future directions*, 11–12. https://doi.org/10.5281/zenodo.7888575.

Caswell, M. and M. Cifor. 2016. "From Human Rights to Feminist Ethics: Radical Empathy in the Archives." *Archivaria* 82: 23–43.

Chayka, K. 2023. "Is A.I. Art Stealing from Artists?" *The New Yorker*, February 10, 2023. https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists.

Colavizza, G., T. Blanke, C. Jeurgens and J. Noordegraaf. 2021. "Archives and AI: An Overview of Current Debates and Future Perspectives." *Journal on Computing and Cultural Heritage* 15 (1): 4:1–4:15. https://doi.org/10.1145/3479010.

Cook, T. 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.

Desai, M. A., I. V. Pasquetto, A. Z. Jacobs and D. Card. 2024. "An Archival Perspective on Pretraining Data." *Patterns* 5 (4): 100966. https://doi.org/10.1016/j.patter.2024.100966.

Douglas, J. 2016. "Toward More Honest Description." *The American Archivist* 79 (1): 26–55. https://doi.org/10.17723/0360-9081.79.1.26.

Douglas, J. 2017. "Origins and Beyond: The Ongoing Evolution of Archival Ideas about Provenance." In *Currents of Archival Thinking*, edited by H MacNeil and T Eastwood. 2nd ed. Santa Barbara, California ; Denver, Colorado: Libraries Unlimited, an imprint of ABC-CLIO, LLC, 25–52.

Douglas, J., M. Ballin and J. Lapp. 2022. "Introduction: Special Issue - Toward Person-Centred Archival Theory and Praxis." *Archivaria* 94 (Fall/Winter): 5–21.

Fordyce, R. and S. Jethani. 2021. "Critical Data Provenance as a Methodology for Studying How Language Conceals Data Ethics." *Continuum* 35 (5): 775–787. https://doi.org/10.1080/10304312.2021.1983259.

Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III and K. Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

Haberstock, L. 2020. "Participatory Description: Decolonizing Descriptive Methodologies in Archives." *Archival Science* 20 (2): 125–138. https://doi.org/10.1007/s10502-019-09328-6.

Harris, V. 2002. "The Archival Sliver: Power, Memory, and Archives in South Africa." *Archival Science* 2 (1): 63–86. https://doi.org/10.1007/BF02435631.

Hegarty, K. 2022. "Representing Biases, Inequalities and Silences in National Web Archives: Social, Material and Technical Dimensions." *Archives & Manuscripts* 50 (1): 31–45. https://doi.org/10.37683/asa.v50.10209.

Hegarty, K. 2025. "Web Archives after Platformization: Reading Social Media Collections along the Archival Grain." *Information, Communication & Society* 28 (12): 2175–2190. https://doi.org/10.1080/1369118X.2024.2420033.

Holland, S., A. Hosny, S. Newman, J. Joseph and K. Chmielinski. 2018. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards." arXiv. https://doi.org/10.48550/arXiv.1805.03677.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

Jaillant, L., edited by. 2021. *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitised Archival Collections*. 1. Auflage. Digital Humanities Research 2. Bielefeld: Transcript.

Jaton, F. 2021. "Assessing Biases, Relaxing Moralism: On Ground-Truthing Practices in Machine Learning Design and Application." *Big Data and Society* 8 (1): 205395172110135. https://doi.org/10.1177/20539517211013569.

Jo, E. S. and T. Gebru. 2020. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316. FAT*'20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372829.

Krause, H. 2019. "An Introduction to the Data Biography - We All Count." January 21, 2019. https://weallcount.com/2019/01/21/an-introduction-to-the-data-biography/.

Longpre, S., R. Mahari, A. Chen, N. Obeng-Marnu, D. Sileo, W. Brannon, N. Muennighoff, N. Khazam, J. Kabbara, K. Perisetla, X. Wu, E. Shippole, K. Bollacker, T. Wu, L. Villa, S. Pentland and S. Hooker. 2023. "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI." arXiv. https://doi.org/10.48550/arXiv.2310.16787.

Longpre, S., R. Mahari, N. Obeng-Marnu, W. Brannon, T. South, K. I. Gero, A. Pentland and J. Kabbara. 2024. "Position: Data Authenticity, Consent, & Provenance for AI are All Broken: What Will It Take to Fix Them?" In *Proceedings of the 41st International Conference on Machine Learning*, 32711–32725. PMLR. https://proceedings.mlr.press/v235/longpre24b.html.

Luke, S. M. and S. Mizota. 2024. "Instituting a Framework for Reparative Description." *Archival Science* 24 (3): 481–508. https://doi.org/10.1007/s10502-024-09435-z.

MacNeil, H. 2005. "Picking Our Text: Archival Description, Authenticity, and the Archivist as Editor." *The American Archivist* 68 (2): 264–278. https://doi.org/10.17723/aarc.68.2.01u65t6435700337.

MacNeil, H. 2025. "Authenticity as a Travelling Concept: From Heritage Conservation to Archives." *Archival Science* 25 (3): 28. https://doi.org/10.1007/s10502-025-09497-7.

Maemura, E., N. Worby, I. Milligan and C. Becker. 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.

McMillan-Major, A., E. M. Bender and B. Friedman. 2023. "Data Statements: From Technical Concept to Community Practice." *ACM Journal on Responsible Computing*. https://doi.org/10.1145/3594737.

Meehan, J. 2009. "Making the Leap from Parts to Whole: Evidence and Inference in Archival Arrangement and Description." *The American Archivist* 72 (1): 72–90. https://doi.org/10.17723/aarc.72.1.kj672v4907m11x66.

Miceli, M., A.-A. Dinika, K. Kauffman, C. S. Wagner, L. Sachenbacher, A. Hanna and T. Gebru. 2025. "Methodological Considerations for Centering Workers' Epistemic Authority in AI Research." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8 (2): 1698–1710. https://doi.org/10.1609/aies.v8i2.36667.

Miceli, M., M. Schuessler and T. Yang. 2020. "Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2): 115:1–115:25. https://doi.org/10.1145/3415186.

Miceli, M., P. Tubaro, A. A. Casilli, T. L. Bonniec, C. S. Wagner and L. Sachenbacher. 2024. "Who Trains the Data for European Artificial Intelligence?" European Parliament ; The Left. https://hal.science/hal-04662589.

Mordell, D. 2019. "Critical Questions for Archives as (Big) Data." *Archivaria* 87 (Spring): 140–161.

Moreau, L., P. Groth, S. Miles, J. Vazquez-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan and L. Varga. 2008. "The Provenance of Electronic Data." *Communications of the ACM* 51 (4): 52–58. https://doi.org/10.1145/1330311.1330323.

Muller, M. and A. Strohmayer. 2022. "Forgetting Practices in the Data Sciences." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19. CHI'22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3491102.3517644.

Nadim, T. 2016. "Data Labours: How the Sequence Databases GenBank and EMBL-Bank Make Data." *Science as Culture* 25 (4): 496–519. https://doi.org/10.1080/09505431.2016.1189894.

Nesmith, T. 2006. "The Concept of Societal Provenance and Records of Nineteenth-Century Aboriginal–European Relations in Western Canada: Implications for Archival Theory and Practice." *Archival Science* 6 (3): 351–360. https://doi.org/10.1007/s10502-007-9043-9.

Ogden, J. 2022. "'Everything on the Internet Can Be Saved': Archive Team, Tumblr and the Cultural Significance of Web Archiving." *Internet Histories* 6 (1–2): 113–132. https://doi.org/10.1080/24701475.2021.1985835.

Padilla, T., L. Allen, H. Frost, S. Potvin, E. R. Roke and S. Varner. 2019. "Final Report — Always Already Computational: Collections as Data." https://zenodo.org/record/3152935.

Pluralizing the Archival Curriculum Group. 2011. "Educating for the Archival Multiverse." *The American Archivist* 74 (1): 69–101. https://doi.org/10.17723/aarc.74.1.hv33964712745684.

Punzalan, R. L. and M. Caswell. 2016. "Critical Directions for Archival Approaches to Social Justice." *The Library Quarterly* 86 (1): 25–42. https://doi.org/10.1086/684145.

Pushkarna, M. and A. Zaldivar. 2022. "The Data Cards Playbook: A Toolkit for Transparency in Dataset Documentation." Google Research Blog. November 17, 2022. https://ai.googleblog.com/2022/11/the-data-cards-playbook-toolkit-for.html.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.

Ratner, H. F. and N. B. Thylstrup. 2025. "Citizens' Data Afterlives: Practices of Dataset Inclusion in Machine Learning for Public Welfare." *AI and Society* 40 (3): 1183–1193. https://doi.org/10.1007/s00146-024-01920-4.

Stoler, A. L. 2002. "Colonial Archives and the Arts of Governance." *Archival Science* 2 (1–2): 87–109. https://doi.org/10.1007/BF02435632.

Summers, E. and R. Punzalan. 2017. "Bots, Seeds and People: Web Archives as Infrastructure." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 821–834. ACM Press. https://doi.org/10.1145/2998181.2998345.

Taurino, G. and D. A. Smith. 2022. "Machine Learning as an Archival Science: Narratives behind Artificial Intelligence, Cultural Data, and Archival Remediation." https://ai-cultures.github.io/papers/machine_learning_as_an_archiva.pdf.

Thylstrup, N. B. 2022. "The Ethics and Politics of Data Sets in the Age of Machine Learning: Deleting Traces and Encountering Remains." *Media, Culture & Society* 01634437211060226. https://doi.org/10.1177/01634437211060226.

Wickner, A. 2019. "Recognizing Co-Creators in Four Configurations: Critical Questions for Web Archiving." *Journal of Contemporary Archival Studies* 6: 1. https://elischolar.library.yale.edu/jcas/vol6/iss1/23.

Wurl, J. 2005. "Ethnicity as Provenance: In Search of Values and Principles for Documenting the Immigrant Experience." *Archival Issues* 29 (1): 65–76. https://doi.org/10.31274/archivalissues.10940.

Zelenka, N. and N. Di Cara. 2024. "Data Hazard Labels V1.1." Accessed March 1, 2025. https://datahazards.com/labels.html

Zhang, B. Z., T. Yang, M. Miceli, O. L. Haimson and M. Thomas. 2025. "The Making of Performative Accuracy in AI Training: Precision Labor and Its Consequences." In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19. CHI'25. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3706598.3713112.

Emily Maemura is an assistant professor in the School of Information Sciences at the University of Illinois Urbana-Champaign. Her research focuses on data practices and the activities of curation, description, characterization and re-use of archived web data. She is interested in approaches and methods for working with archived web data in the form of large-scale research collections, considering diverse perspectives of the internet as an object and site of study.

Cite this article: Maemura, E. 2026. Lessons from archival provenance: Reframing data creator relationships in AI. *Cambridge Forum on AI: Culture and Society*, **2**, e21, 1–10. https://doi.org/10.1017/cfc.2026.10021

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 20:29:04, subject to the Cambridge Core terms of use.