Where's the Balance?: AI, Archives, and Privacy

Iori Khuhro[1], Erin Gilmore[2], Jim Suderman[3], & Darra L. Hofman[4][5]

InterPARES Trust AI

Public Symposium: Bringing AI to the Archives

Regis University | Denver, Colorado | October 25, 2024

Positionality

I, Iori Khuhro, was an uninvited guest on the traditional homeland and buffalo hunting grounds of the Arapaho, Cheyenne and Ute Nations –Denver, Colorado, United States– where I had presented this symposium paper. I had written this paper on the traditional, ancestral, and unceded territories of the xwməθkwəy̓əm (Musqueam), Sḵwx̱wú7mesh (Squamish), and səlilwətaɬ (Tsleil-Waututh) Nations – also known as Vancouver, British Columbia, Canada. Despite being a second-generation immigrant whose homeland was violently colonized and divided, I must acknowledge that I am afforded certain privileges and rights in Canada that historically have been stolen from and are currently being denied to the rightful stewards –the Indigenous peoples– of the lands that make up Turtle Island. As a student of archival studies, I am cognizant of the fact that archives and archival practices have often been used to uphold a status quo that has underrepresented or harmed marginalized –especially Indigenous– communities. It is with this understanding that I endeavour not to duplicate that harm within my research and future work, and so I urge all researchers to reflect on the lands on which they live and how their work impacts those people and their land.

Overview

[1] The University of British Columbia. ORCID: 0009-0002-6403-4149.
[2] San José State University
[3] InterPARES Trust AI. Email: suderman.mawg@gmail.com
[4] San José State University. Email: darra.hofman@sjsu.edu. ORCID: 0000-0002-1772-6268
[5] The authors would like to thank Kisun Kim (Okanagan College) and Carlos Quevedo, previous InterPARES Trust AI Graduate Research Assistants, for their contribution to this work.

How are archival institutions protecting privacy in digital records containing PII when providing access to them?

Well, generally speaking, they aren't… Providing access, that is.

— — —

Archivists have historically taken on the role of a steward – sometimes trusted, sometimes not. In the early days, they were tasked with preserving the patriotic history of their nations, but quickly found themselves balancing the idealized nation against critical social memory. Archives have evolved to make space for both the outdated and the contemporary. A good archive is always in dialogue with itself; the archivist should constantly adapt based on those discussions. It is an archivist's responsibility to maintain an equilibrium between the records and the people in hopes of providing a public good.

So, isn't access a public good? Why aren't archives providing access to digital records then?

It is not for a lack of trying. The unfortunate reality is–at the risk of sounding like a broken record–that the digital backlog is so immense that archives physically cannot get around to manually protecting entire collections filled to the brim with Personal Identifiable Information (PII) just to make them accessible. As a point of reference, the US National Archives is preserving almost 300 TB of White House emails, but "none have been systematically opened by archivists for public access, nor is there any strategic plan for doing so in the immediate future."[6]

The increasingly complex and contentious nature of privacy has swung the pendulum from access moreso to privacy; especially from the standpoint of archivists having many tools and much experience in providing access to records but with much fewer tools and experience with protecting privacy. However, in not providing access to records to protect privacy, archivists

[6]Baron and Payne, "Dark Archives and E-Democracy."

are staying faithful to work. Archivists, at their core, are stewards. Just as they balance history and the modern day, they must find a way to create an equilibrium between privacy and access. Balance, in this context, does not mean a ratio of 1:1 but rather an assessment of what action results in the least amount of harm.

The PI Lit Review Study, hoping for a solution that would assist archivists in mitigating their "hindered access" dilemma, put together an annotated bibliography that aggregated and recontextualized articles from the domains of Archival Studies, Computer Science Studies, and Legal Studies exploring the extent to which and how Artificial Intelligence (AI) tools and techniques could address or resolve privacy challenges faced by archival institutions when providing access to records containing PII.

Research Questions

The literature we analyzed primarily answered our first four research questions:

1. How are archival institutions dealing with protecting privacy in digital records containing PII when providing access to them?
2. How could AI tools and techniques contribute to the challenges faced by archival institutions in providing access to these kinds of records?
3. What are the implications of using AI tools and techniques to deal with privacy issues in records?
4. How effectively can machine learning (ML), natural language processing (NLP), and named entity recognition (NER) enable the identification and location of personal information in large digital textual collections?

Methodology

Based on our research questions, we began an iterative review of the literature. In screening for inclusion, our initial inclusion criteria included: date, peer review, type of publication, research setting, and research design.

Criterion	**Initial Requirements**	**Expanded?**
Date	2017 and subsequent; initially chosen due to the breakthroughs in AI	Yes – critical earlier publications included
Type of publication and peer review	Peer-reviewed journal articles and conference proceedings	Yes – relevant grey literature included, including white papers and reports
Research setting	Inclusive	No
Research design	Inclusive	No

Figure 1: Inclusion Criteria

We started by searching for literature from 2017 and the subsequent years because we recognized the breakthroughs happening with AI in 2016. However, critical earlier publications, specifically concerning privacy, also found their place in our review. Understanding that not all privacy, archival or computer science work happens in peer-reviewed journal articles and conference proceedings, we also expanded type of publications to include grey literature and white papers and reports. Both research setting and research design were inclusive.

Throughout the course of the study, multiple Graduate Research Assistants have graciously contributed to the annotated bibliography; they plotted out the objectives, research

questions, core concepts, research setting, research design, key findings, and implications for each article.

Not displayed in the annotated bibliography is how we have charted "type of study" (archival/legal/computer science); jurisdiction (for example, North American vs European privacy laws); privacy scope (from the very broad, such as "private user data" to very specific types of personal data, such as "email addresses, email messages, and headers"); how the study deals with privacy; success measures; whether human intervention was needed with regard to the AI model; and novel AI model ideas for future interrogation into a spreadsheet for a more refined data analysis.

We originally organized the articles under the three domains of concern: Archival Studies, Computer Science Studies, and Legal Studies. The assumption, at the time, was the division would facilitate identifying patterns within each field; however, we removed the categories because it became evident that the domains were not mutually exclusive and that there was a more overarching issue at hand: primarily, how do the professionals define and apply privacy?

Findings

The findings, discussions, and conclusions found within the articles of this annotated bibliography are vast and diverse, providing insights into privacy and AI conversations from around the world. The articles that have been aggregated and codified shatter the notion that each discipline is on its own island; the archivists grapple with the computer scientists, who grapple with the legal professionals, who grapple with the archivists. Despite the little attention they pay to one another, the findings of one discipline should have a great deal of impact on the other.

Baron and Payne[7], Goldman and Pyatt[8], Yaco[9], and Murphy et al.[10] lament how access is being hindered because archival institutions have no other means of dealing with PII aside from manual redaction, which consumes more resources than archivists have available to them – namely time and labour. However, the plight of archivists is not unique to them. Baron et al.[11], Borden and Baron[12], Dias[13], Mcdonald[14], Mcdonald et al.[15], Glaser et al.[16], Oksanen et al.[17], Tamper et al.[18], and Garat and Wonsever[19] all write about the same limitation of having to protect PII through manual means in a legal context.

However, it is primarily those in the legal studies who have investigated AI and Machine Learning (ML) as a means of overcoming these access issues. This demonstrates that as archival studies remain introspective and consider the nature of sensitivity, context, and privacy within their collections, the legal and computer science domains are already investigating and providing potential solutions to dealing with PII in more automated fashions.

It is worth noting that while computer science studies are experimenting with Machine Learning, Natural Language Processing (NLP), and Named Entity Recognition (NER) to test the efficacy of these techniques for identifying, redacting, and anonymizing PII in records, their concerns lie with the unavailability of training data sets, and success measures for their field,

---

[7] Baron and Payne, "Dark Archives and Edemocracy."
[8] Goldman and Pyatt, "Security Without Obscurity."
[9] Yaco, "Balancing Privacy and Access in School Desegregation Collections."
[10] Murphy et al., "Failure Is an Option."
[11]Baron, Sayed, and Oard, "Providing More Efficient Access To Government Records."
[12] Borden and Baron, "Opening up Dark Digital Archives through the Use of Analytics to Identify Sensitive Content."
[13] Dias, "Multilingual Automated Text Anonymization."
[14] McDonald, "A Framework for Technology-Assisted Sensitivity Review."
[15] Mcdonald, Macdonald, and Ounis, "How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review."
[16] Glaser, Schamberger, and Matthes, "Anonymization of German Legal Court Rulings."
[17] Oksanen et al., "ANOPPI: A Pseudonymization Service for Finnish Court Documents."
[18] Tamper et al., "Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens."
[19] Garat and Wonsever, "Automatic Curation of Court Documents."

including precision, recall, accuracy, and/or F1 scores, which serve as adequate measures for determining how well an algorithm identifies true and false positives or negatives. But, determining whether or not data is private, and to whom access to data can be given, continues to remain a weighted question on the archivist's shoulders.

Lemieux and Werner[20] explain –in their scoping review of privacy-enhancing technologies for archives– that despite experimentation with AI-enabled (predominantly NLP-based) approaches, effective ways to responsibly balance provision of access with protection of privacy remain elusive for archivists. This is largely due to the complexities of applying existing privacy protection legislation to large and often poorly described archival collections. The results of such approaches are insufficiently accurate; even if more accurate models are developed, current AI privacy solutions fall short of the scale needed for archival privacy management. Less human-dependent approaches, such as neural networks, likewise lack the accuracy needed at this point in time. Deploying privacy tools that are insufficiently accurate could erode trust in both the tools and the archival institutions that might use them.

Despite the kinks in the technology, Baron and Payne contend that, "archivists can no longer rely on manual methods" because AI can filter sensitive data, allowing for quicker access to records online.[21] Therefore, the relationship between privacy, archives, and AI is multidirectional. Simply relying on AI solutions to solve the problem of balancing privacy and access risks further entrenching known issues in both AI and archives. However, having perfected the balancing act of a steward, archivists must consider how applying archival knowledge and practice –such as rich description of provenance– can mitigate problems within AI because, as Henttonen explains in *Privacy as an Archival Problem and a Solution*, "the

[20] Lemieux and Werner, "Protecting Privacy in Digital Records."
[21] Baron and Payne, "Dark Archives and E-Democracy," 6.

application of archival practices is critical for the protection of personal privacy now and in the future."[22]

Another potential approach to interpreting privacy relies on the theory of "contextual integrity," which Nissenbaum uses to define privacy as a relative rather than a static concept.[23] One's privacy is not always violated when a certain piece of information is shared, but rather when it is shared in an unexpected context or way.[24] Henttonen suggests that since archival work is the secondary use of records, archives are –*inherently*– violating the privacy of those within the records, in which case strategies must be devised by archivists to address the ethical dilemma beyond burying records.[25]

Moss and Gollins urge archivists to shift their focus from the technical challenges of digital preservation and instead work on appraisal, sensitivity review, and access assisted or facilitated through AI and Machine Learning. The authors believe that "the archive has to take what it is given, from the context in which the users have chosen to use it."[26]

Discussion

It takes little thought to be critical of archivists; no one is more attuned to the fact that they need to be better at providing access to digital records than archivists are themselves. Instead, we must understand that they have been stewards for centuries; every decision is a compromise. Privacy and access are –*semantically*– at odds with each other, and archivists are constantly making judgements about what actions result in the least amount of harm and the most public good.

---

[22] Henttonen, "Privacy as an Archival Problem and a Solution," 86.
[23] Nissenbaum, *Privacy in context: Technology, policy, and the integrity of social life.*
[24] NIssenbaum.
[25] Henttonen.
[26] Moss and Gollins, "Our Digital Legacy: An Archival Perspective," 6.

At the same time, the lack of action taken towards protecting PII in records is not solely an issue of insufficient resources or capabilities –though the amount of manual labour and expertise that goes into redacting PII is profound– but rather a lack of strategic planning within archives to slow the steady growth of PII backlog in their collections. There needs to be a shift away from a purely compliance-based approach to a risk-based strategy that is cognizant of the fact that just because digital records with PII are inaccessible to the public does not mean the PII is protected in the digital environment. Part of an archives' strategy could involve a risk-based appraisal process which leans on provenance as a means of determining the sensitivity and privacy concerns within a collection[27] [28]. We look forward to learning more from recent and future interviews conducted with archivists, such as Whyte and Walsh's work[29], that provide insight into the daily practices surrounding privacy protection which have not been documented so far in the literature.

The question, upon synthesizing all the literature, is no longer whether AI can identify and then redact, anonymize or pseudonymize PII – as it has already been proven that it can do so for recognizable named entities, but rather, can archivists, legal professionals, and computer scientists look beyond the existing attempts to define privacy and begin to develop sufficiently rich, applied understandings of privacy to support the development of robust privacy AI solutions (and privacy for AI solutions) that enable archivists to carry the ethical burden of having to judge when access takes precedence over privacy and when privacy takes precedence over access, responsibly and effectively.

Future Research

[27]Bingo, "Of Provenance and Privacy."
[28]Iacovino and Todd, "The Long-Term Preservation of Identifiable Personal Data."
[29] Whyte, Jess, and Tessa Walsh. "'Carefully and Cautiously': How Canadian Cultural Memory Workers Review Digital Materials for Private and Sensitive Information".

Since the work of dissecting and understanding PII in records has fallen on archivists as both a legal and ethical responsibility, our future research will focus on analyzing the values and limitations of computational/technical success measures for AI models against what is considered an acceptable, humanist attempt at protecting PII within archival institutions. We also hope to conduct surveys, focus groups, and/or interviews with archivists to better understand an archives' internal processes when deciding the fate of digital records with PII.

Bibliography

Baron, Jason R., and Nathaniel Payne. "Dark Archives and E-Democracy: Strategies for Overcoming Access Barriers to the Public Record Archives of the Future." In *2017 Conference for E-Democracy and Open Government (CeDEM)*, 3–11. Krems, Austria: IEEE, 2017. https://doi.org/10.1109/CeDEM.2017.27.

Baron, Jason R., Mahmoud F. Sayed, and Douglas W. Oard. "Providing More Efficient Access To Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege." arXiv, November 13, 2020. http://arxiv.org/abs/2011.07203.

Bingo, Steven. "Of Provenance and Privacy: Using Contextual Integrity to Define Third-Party Privacy." The American Archivist 74, no. 2 (September 1, 2011): 506–21. https://doi.org/10.17723/aarc.74.2.55132839256116n4.

Borden, Bennett B., and Jason R. Baron. "Opening up Dark Digital Archives through the Use of Analytics to Identify Sensitive Content." In 2016 IEEE International Conference on Big Data (Big Data), 3224–29. Washington DC,USA: IEEE, 2016. https://doi.org/10.1109/BigData.2016.7840978.

Dias, Francisco. "Multilingual Automated Text Anonymization." Instituto Superior Técnico, 2016. https://scholar.tecnico.ulisboa.pt/records/W-m-zXqhZ-Ck1jjk7_oa9h_JsK3fev6LIvK-.

Garat, Diego, and Dina Wonsever. "Automatic Curation of Court Documents: Anonymizing Personal Data." Information 13, no. 1 (January 10, 2022): 27. https://doi.org/10.3390/info13010027.

Glaser, Ingo, Tom Schamberger, and Florian Matthes. “Anonymization of German Legal Court Rulings.” In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 205–9. São Paulo Brazil: ACM, 2021. https://doi.org/10.1145/3462757.3466087.

Goldman, Ben, and Timothy D. Pyatt. “Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives.” Library & Archival Security 26, no. 1–2 (July 2013): 37–55. https://doi.org/10.1080/01960075.2014.913966.

Henttonen, Pekka. “Privacy as an Archival Problem and a Solution.” Archival Science 17, no. 3 (September 2017): 285–303. https://doi.org/10.1007/s10502-017-9277-0.

Iacovino, Livia, and Malcolm Todd. “The Long-Term Preservation of Identifiable Personal Data: A Comparative Archival Perspective on Privacy Regulatory Models in the European Union, Australia, Canada and the United States.” Archival Science 7, no. 1 (March 2007): 107–27. https://doi.org/10.1007/s10502-007-9055-5.

Lemieux, Victoria L., and John Werner. “Protecting Privacy in Digital Records: The Potential of Privacy-Enhancing Technologies.” Journal on Computing and Cultural Heritage 16, no. 4 (December 31, 2023): 1–18. https://doi.org/10.1145/3633477.

McDonald, Graham. “A Framework for Technology-Assisted Sensitivity Review: Using Sensitivity Classification to Prioritise Documents for Review,” 2019. https://doi.org/10.5525/GLA.THESIS.41076.

Mcdonald, Graham, Craig Macdonald, and Iadh Ounis. “How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review.” ACM Transactions on Information Systems 39, no. 1 (January 31, 2021): 1–34. https://doi.org/10.1145/3417334.

Moss, Michael, and Tim Gollins. “Our Digital Legacy: An Archival Perspective.” Journal of Contemporary Archival Studies 4 (2017). http://elischolar.library.yale.edu/jcas/vol4/iss2/3.

Murphy, Mary O., Laura Peimer, Genna Duplisea, and Jaimie Fritz. “Failure Is an Option: The Experimental Archives Project Puts Archival Innovation to the Test.” The American Archivist 78, no. 2 (September 2015): 434–51. https://doi.org/10.17723/0360-9081.78.2.434.

Nissenbaum, Helen, ed. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, Calif: Stanford Law Books, 2010.

Oksanen, Arttu, Minna Tamper, Jouni Tuominen, Aki Hietanen, and Eero Hyvönen. “ANOPPI: A Pseudonymization Service for Finnish Court Documents.” In Legal Knowledge and Information Systems, 322:251–54. Frontiers in Artificial Intelligence and Applications, 2019. https://doi.org/10.3233/FAIA190335.

Tamper, Minna, Arttu Oksanen, Jouni Tuominen, Eero Hyvönen, and Aki Hietanen. “Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens.” via the Internet, Florence, Italy, 2018. https://seco.cs.aalto.fi/publications/2018/anonymization-service-finnish.pdf.

Whyte, Jess, and Tessa Walsh. 2024. “‘Carefully and Cautiously’: How Canadian Cultural Memory Workers Review Digital Materials for Private and Sensitive Information”. Partnership: The Canadian Journal of Library and Information Practice and Research 19 (1):1-26. https://doi.org/10.21083/partnership.v19i1.7180.

Yaco, Sonia. “Balancing Privacy and Access in School Desegregation Collections: A Case Study.” The American Archivist 73, no. 2 (September 2010): 637–68. https://doi.org/10.17723/aarc.73.2.h1346156546161m8.