InterPARES Trust AI

MA05: Personal Information (PI) Lit Review

An Annotated Bibliography

November 2024

Table of Contents

Principal Investigators: Darra L. Hofman[1] & Jim Suderman[2]

Graduate Research Assistants: Iori Khuhro[3] & Erin Gilmore[4][5]

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

Overview

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

Well, broadly speaking, they aren't providing access.

The increasingly complex and contentious nature of privacy has swung the pendulum from access more 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. 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" (Baron and Payne, 2017, p. 5). However, in not providing access to records to protect privacy, archivists are attempting to maintain an equilibrium between privacy and access. Balance, in this context, means an assessment of what action results in the least amount of harm.

The Personal Information (PI)[6] 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.

[6] The glossary of the International Association of Privacy Professionals (IAPP) notes that the terms "Personal Information" and "Personal Data" are synonymous. "Personally Identifiable Information" (PII), while not indicated as synonymous to the other two terms, likewise refers to "any information […] that can be used to distinguish or trace an individual's identity". IAPP. Glossary, https://iapp.org/resources/glossary/#paperwork-reduction-act-2 [accessed: 5.11.2024].

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 PI 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?

We were also interested in the following two questions but found little to no literature concerning:

1. What risk-based privacy protection models (any/all jurisdictions) are defined and assessed?

2. What models for parsing textual content based on legal/statutory definitions are defined and assessed? What success measures (strengths/limitations) are referenced for these models?

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

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, accurately and consistently, but some differences in writing and annotation styles will be noticeable.

Not displayed in the annotated bibliography is how we have charted "type of study" (archival/legal/computer science); jurisdiction (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 the AI model required human intervention; and novel AI model ideas into a spreadsheet for a more refined data analysis.

The articles in this annotated bibliography were initially organized under the three domains of concern: Archival Studies, Computer Science Studies, and Legal Studies. The assumption, at the time, was that the division would facilitate identifying patterns within each field; however, the categories were removed 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 professions 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.

Murphy et al. (2023), Baron and Payne (2017), Goldman and Pyatt (2015), and Yaco (2014) and lament how access is being hindered because archival institutions have no other means of dealing with PI aside from manual redaction, which consumes more resources than archivists have available to them – namely time and labour. Notably, the plight of archivists is not unique to them. Garat and Wonsever (2021), Tamper et al. (2019), Mcdonald (2019), Mcdonald et al. (2019), Mcdonald et al. (2018), Oksanen et al. (2018), Glaser et al. (2018), Dias (2017), Baron et al. (2016), and Borden and Baron (2016) all write about the same limitation of having to protect PI through manual means in a legal context.

However, it is primarily those in the legal studies who have investigated Artificial Intelligence 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.

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, 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 information is private or personal, and to whom access to data can be given, continues to remain a weighted question on the archivist's shoulders.

Lemieux and Werner 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 (2017) contend that "archivists can no longer rely on manual methods" (p. 6). AI can filter sensitive data, allowing for quicker access to records online. 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, archivists must consider how applying archival knowledge and practice –such as rich description of provenance– can ameliorate problems within AI as “[they] are critical for the protection of personal privacy now and in the future” (Henttonen, 2017, p. 86).

Another potential approach to interpreting privacy relies on the theory of “contextual integrity,” which defines privacy as a relative rather than a static concept (Nissenbaum, 2009, as cited in Bingo, 2011). 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. 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 (Henttonen, 2017).

A suggested approach involves archivists shifting their focus from the technical challenges of digital preservation and instead work on appraisal, sensitivity review, and access assisted and facilitated through AI and Machine Learning. To this effect, Moss and Gollins (2017) believe “the archive has to take what it is given, from the context in which the users have chosen to use it” (p. 6).

Discussion

Every decision an archivist makes 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.

At the same time, the lack of action taken towards protecting PI in records is not solely an issue of insufficient resources or capabilities –though the amount of manual labour and expertise that goes into redacting PI is profound– but rather a lack of strategic planning within archives to slow the steady growth of PI 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 PI are inaccessible to the public does not mean the PI 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 (*see* Bingo, 2017; Iacovino & Todd, 2007). We look forward to learning more from recent and future interviews conducted with archivists, such as Whyte and Walsh's (2024) work[7], that provide insider insight into the daily practices surrounding privacy protection which have not been documented so far in the literature.

The question, upon synthesizing the literature, is no longer whether AI can identify and then redact, anonymize or pseudonymize PI – as this annotated bilbiography will prove 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 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

Since the work of dissecting and understanding PI 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 PI 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 PI.

Some questions that need to be further explored are:

- How does AI fit with and relate to current archival thinking and practice?
- Can the results of existing studies be extrapolated onto large collections, and how do we define "large" collections in archives?

[7] 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.

Annotated Bibliography

AlEroud, Ahmed, Faten Masalha, and Ahmad A. Saifan. 2021. "Identifying GDPR Privacy Violations Usingan Augmented LSTM: Toward an AI-Based Violation Alert Systems." In 2021 IEEE Intl Conf onParallel & Distributed Processing with Applications, Big Data & Cloud Computing, SustainableComputing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 1617–24. New York City, NY, USA: IEEE. https://doi.org/10.1109/ISPA-BDCloudSocialCom-SustainCom52081.2021.00216.

Objective: The paper presents a method for identifying which GDPR article a particular privacy incident violates using text summarization and deep learning techniques.

Research Strategy/*Design:* The study was conducted using a quantitative research strategy in an experimental research design format.

Setting/Sample: The research focuses on "breaches and violations that are written in semi unstructured or unstructured natural language" (p. 1621). The experiments used 750 privacy incidents, of which many are from the GDPR Enforcement Tracker.

Method: The researchers used the Labeled Topic Modeling approach to topics, which were textual features extracted and pre-processed, associated with particular types of violations that correspond to GDPR articles. The researchers then used the labeled-Latent Dirichlet Allocation (LDA) Algorithm to train and test a Long Short-Term-Memory Neural Networks (LSTM) – deep learner – that identifies potential GDPR violations given textual descriptions.

Main Results: The use of Labeled-LDA with deep learners such as LSTM demonstrates a promising accuracy level in parsing textual content for privacy violations based on GDPR articles.

Discussion/Conclusion of Article: "The results show the need of an expanded study that utilize graph relations such as SLNs to discover relationships between different classes (violated articles), then conduct a multi-label classification on different violations"

(P.1623). The researchers also believe that the work can be expanded and be used to discover future privacy violations based on the existing privacy documentation.

Baron, Jason R., and Nathaniel Payne. 2017. “Dark Archives and Edemocracy: 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. https://doi.org/10.1109/CeDEM.2017.27.

Objective – This conference paper aims to highlight some issues of dark archives, which can be created by restricting too much information in records, and strategies to overcome them.

Research Strategy/Design – qualitative case study

Setting – The US government archives under the Obama Administration

Method – Mainly focusing on email records, the researchers conducted a descriptive study, analyzing how and why dark archives came to be despite the open government initiatives.

Main Results – Due to the tradition of manual filtering and the fact that PII is embedded in almost all public records (especially email records), it is impossible to eliminate dark archives with the current technology and systems in use. Human language is inherently ambiguous, and manual keyword searching and filtering do not work, although regular expression can help identify some PII. Machine Learning and Cloud Computing can potentially resolve this problem, as they automatically identify PII and increase the quality and quantity of retrieved data, respectively.

Discussion/Conclusion of Article – Dark archives challenge openness and transparency, the qualities records serve to protect democracy, and the lack of discussion around these qualities in a digital setting perpetuates the risk. The researchers recommend that the government work with experts from the computer science and information science field to develop methods to preserve records while protecting PII in a digital setting. They further recommend that the government should address the issue of dark archives in public forums.

Annotation – The article offers a high-level analysis of dark archives, which is yet to be addressed in the archives, records management, and information science field at the time of publication. It demonstrates the need for archivists to adopt technological solutions to ensure democracy within the digital settings.

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

In this article, Baron et al. recognizes that manual searches for records and redaction of personal information delay the FOI analysts' response time for FOI requests. They also note that privacy experts have already recommended developing AI to help with the process. The experiment outlined in this article applies classifiers to emails using machine learning technology to identify materials that can be withheld, according to FOIA regulations. The system was trained and evaluated using annotations from FOI reviewers. The results indicate that the system was very successful in identifying records that may need more attention (i.e. may need to be withheld). The system flags records that require reviewers' attention, allowing reviewers to focus on a smaller pool of records instead of all requests. It also streamlines FOI-related decisions. It is important to note that the focus on this study is on access rather than protection.

Belhi, Abdelhak, Tahani Abu-Musa, Abdulaziz Khalid Al-Ali, Abdelaziz Bouras, Sebti Foufou, Xi Yu, andHaiqing Zhang. 2019. "Digital Heritage Enrichment through Artificial Intelligence and SemanticWebTechnologies." In 2019 4th International Conference on Communication and Information Systems(ICCIS), 180–85. Wuhan, China: IEEE. https://doi.org/10.1109/ICCIS49662.2019.00039.

Belhi et al. demonstrate the ways in which Artifical Intelligence, particularly Deep Neural Networks are being used by cultural heritage institutions to tag and translate metadata for preservation in unique ways. Highlighting examples using CEPROQHA Cultural heritage Ontology, Inference Rules Engine, WordNet Lexical Database, and the OWLReady2 API, the authors demonstrate the benefits of ontological classification in simplifying and automating the classification of materials and objects in a given collection. In this way, these technologies are being used to lighten the cognitive load on archivists while using advanced image processing to ensure metadata is tagged properly.

Bingo, Steven. 2011. "Of Provenance and Privacy: Using Contextual Integrity to Define Third-Party Privacy." The American Archivist 74 (2): 506–21. https://doi.org/10.17723/aarc.74.2.55132839256116n4

Using Nissenbaum's theory of contextual integrity, Bingo defines privacy as a relative rather than a static concept. One's privacy is not always violated when a certain piece of information is shared, but when shared in an unexpected context or way. The use and dissemination of personal information can be appropriate or not based on their social settings, characterized by social norms, power structure, and internal values. With this framework in mind, Bingo further examines how contextual integrity can be applied in digital archives setting. The theory of contextual integrity emphasizes context, origin, and use of information, which makes it an ideal framework to examine the issue of third-party privacy in digital archives. As records are digitized online, the context in which records were initially made changes again, potentially violating contextual integrity.

Although many believe that item-level intervention is the only way to protect privacy in records, Bingo suggests an alternative method of evaluating and intervening the risk in records. Using the contextual integrity theory, archives can identify privacy risks by examining the contexts of the record creation, such as the creators' role and activities, during the appraisal process. In other words, the provenance reveals the context and the norms around privacy. Therefore, analyzing the provenance allows archives to evaluate the privacy risk without having to read the contents. It is important to note that Bingo acknowledges that the theory of contextual integrity does not question the social norms it operates in. This may lead to inequitable decisions, allowing certain personal information to be accessible when the creator's community may not find such a decision acceptable. Bingo, however, also urges archivists to consider the political and moral ramifications of altering the context.

For the purpose of this study, the article proposes an interesting framework to evaluate and mitigate privacy risks presented in digital collections. By framing the definition of privacy within the contextual integrity theory, the article demonstrates that archivists can examine the provenance of records rather than the contents to identify third-party privacy issues. The article, however, does not present how archives will determine the access level after identifying the risks. It also does not discuss computational approaches.

Borden, Bennett B., and Jason R. Baron. 2016. "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. https://doi.org/10.1109/BigData.2016.7840978

This article by Baron and Borden, published in 2016, is concerned about the issue of access (dark archives) to presidential and federal emails that NARA will acquire. Many records would be deemed inaccessible due to PII in the records – some may be accessible after five years when people make an FOI request, but many will generally remain inaccessible for 75 years in order to protect the PII. The article first discusses legal and archival considerations that prevent access to records. It then describes some analytical toolkits that can be used to identify PII and other sensitive content. These include: technology-assisted review, social network analysis, sentiment analysis, and visual analysis. Finally, the article proposes that the archival community work with private industry "to develop a standard set of regular expressions," pilot methods used in the legal community, and test software to see if they successfully identify PII based on legislation.

Catelli, Rosario, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. “A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the DeIdentification of Italian Medical Records.” *IEEE Access* 9:19097–110. https://doi.org/10.1109/ACCESS.2021.3054479.

Introduction – Catelli et al. establish the need for publicly available de-identified electronic health records as key for ongoing medical research. Privacy legislation such as GDPR and HIPAA outline the PHI (Personal Health Information) that must be removed in order to make such records public. The authors set out with three goals: 1) the creation of a new clinical de-identification data set composed of Italian COVID-19 medical records, 2) the construction of a model with the best performing sequence labeling architecture (Bi-LSTM) for clinical de-identification of Italian medical records and finally 3) experimenting their model’s performance against BERT, another state-of-the art model for general NLP tasks including NER for de-identification.

Background and Related Works – Other anonymisation and NER tools have been successfully created for English, but these do not easily apply to other languages. The author provides other de identification anonymisation systems that have been successful for languages such as Danish, Dutch, German and French.

Materials and Methods – 115 unannotated medical records were the training dataset for this study. Researchers annotated these according to the i2b2 UT health de-identification track. This included 7 broader categories (e.g. contact, location or name) which results in 13 fields of PHI (e.g. phone, hospital, patient name). The medical records were manually annotated with the assistance of python scripts that converted PDFs into text, reverted annotations into the CONLL format followed up by a tokenizer and language model tool. Disagreement among annotations was resolved using the “Observed Agreement Index” and a Krippendorf coefficient was calculated to deduce that the annotators had reached “substantial” agreement. The authors chose to use the Bidirectional Long Short-Term Memory (Bi-LSTM) + Conditional Random Field (CRF) model as it represents the “best sequence labeling architecture recognized by scientific literature." This architecture was paired with "FastText" and "Flair" embeddings to better deal with out of vocabulary words, polysemous words, misspellings and rare words or grammatical structures.

Experimental Setup and Metrics – The authors tested their model alongside the results of the Bidirectional Encoder Representations from Transformers (BERT) model which is a "general purpose language model" which can be used for NLP tasks like NER. Evaluation methods included the standard precision, recall and F1 metrics divided into binary and the i2b2 categories and sub-categories.

Results and Discussion – The author's Bi-LSTM CRF model with the FastText and Flair embeddings provided the best f1 results at the entity and token level for the category, binary and subcategories.Their model outperformed the BERT model in all cases except one. Some categories within the experiment performed better than others such as Name versus Profession. Profession's low performance can be attributed to its lack of consistent repetition in the documents to be identified correctly. The combination of the FastText and Flair tools helped to address polysemy and handle context by working at the sub-word level. This helped to correctly identify multiple token entities for Hospital such as "Reparto di Osservazione Breve" or to correctly identify Doctor entities that possessed non-Italian last names like "Wang" or "Chunli".

Conclusions – "The Bi-LSTM+CRF architecture with the stacked embedding obtained the best results among the others. These results showed that it is desirable to adopt both contextualized and character-level language models in combination with sub-word embeddings: this way the system is capable to capture, on the one hand, the polysemy of words, their morpho-syntactic variations, rare words and/or misspelled ones and, on the other hand, the latent semantic and syntactic similarities" (p. 19107).

Annotation - Catelli et al. outline the successful creation of an NLP based NER system that can accurately identify personal health information in a language that has been historically neglected by other NER systems. Their process could be useful for other researchers looking to identify PII in their own language. Their success with the Bi-LSTM CRF model warrants it as an option to address archives based PII issues.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2022. "Archives and AI: An Overview of Current Debates and Future Perspectives." *Journal on Computing and Cultural Heritage* 15 (1): 1–15. https://doi.org/10.1145/3479010

In this survey of literature pertaining to the uses of AI in Archival studies, the authors divide the current debates and developments in the field in to four distinct categories: Theoretical and Professional Considerations, Automating Recordkeeping Process, Organizing and accessing archives, Novel Forms of Archives [created through the use of AI]. Within these four categories, the authors identify the ways in which use of Artificial Intelligence (in particular Machine Learning and Natural Language processing) is challenging and reshaping archival theory, particularly within the realm of provenance and original order.

Within the survey, Colavizza et al, highlight the occupational adaptations that are needed in order for archives to prepare for the inevitability of AI usage within Archives. Of particular interest for data privacy is the ethical, and social considerations of AI (e.g. the ways in which is can reinforce "confirmation bias" along racial and socio-economic lines within say, law enforcement databases), and how certain scholars such as Gupta, Jo, and Milligan have highlighted the need for more intentional data-design when it comes to applying AI to index community driven efforts.

Calling for further research and discussion around these ongoing changes, the authors indicate the need for more research on the limitations of AI as well as a reconceptualization of archives as "large data centers" which can potentially use a hybrid form of human/machine methods to order and preserve data. In all, AI and Data-Driven cataloguing processes are not unbiased, as the decision to use them in the first place is still one made by humans.

Cormack, Gordon V., and Maura R. Grossman. 2014. “Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery.” In *Proceedings of the 37th International ACMSIGIR Conference on Research & Development in Information Retrieval*, 153–62. Gold Coast Queensland Australia: ACM. https://doi.org/10.1145/2600428.2609601

Objective – The article seeks to compare three machine learning protocols that can be used for Technology-Assisted Review – Continuous Active Learning (CAL), Simple Active Learning (SAL), and Simple Passive Learning (SPL).

Setting/Sample – Four review tasks denoted as Matter 201, 202, 203, and 204 were derived from Topics 201, 202, 203, and 204 of the TREC 2009 Legal Track Interactive Tasks. “Four other review tasks, denoted Matters A, B, C, and D, were derived from actual reviews conducted in the course of legal proceedings.” (p.154)

Method – The researchers used Sofia-ML implementation of Pegasos SVM for the learning algorithm, with the parameters “--iterations 2000000 --dimensionality 1100000.” For each protocol (SAL, SPL, CAL), a batch size of 1000 documents was used. Each CAL, SAL, and SPL reviewed 1000 documents, and when/if indicated by the protocol, the documents were added to the training set. The process was repeated 100 times.

Main Results – The article found that the CAL protocol achieved higher recall for less effort than SAL and SPL, meaning that it is far more effective. SAL protocol, while less effective than CAL, was more effective than SPL. The article also found that the seed set selected at random yields the same or slightly inferior results than a keyword-selected seed set for CAL and SAL protocols. The entirely non-random training methods require significantly less human review effort than passive-learning methods.

Discussion/Conclusion of Article - Some people favour the random selection of samples over the non-random selection of samples because they believe that the samples selected with purpose will inherently be biased. However, such a bias is difficult to persist in CAL protocol, while the concern remains valid for SPL. Moreover, the article argues that including a single, fallible human reviewer during the training is essential, especially for CAL because the human reviewer may make a different decision based on the document's context. CAL will continue to learn from these decisions. The article notes that the objectives of SAL and CAL are different. CAL's purpose is "to find and review as many of the responsive documents as possible, as quickly as possible" whereas SAL's purpose is "to induce the best classifier possible, considering the level of training effort."

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

Objective – The article posits that the selection of pretraining data for LLMs is largely an engineering exercise and, as such, potential political and cultural impacts are given little consideration. The authors assert that the common elements in the processes of selecting pretraining data and archival appraisal and the ensuing similarities of the respective products, reveals the "implicit power" involved in the creation and use of each.

Research Strategy/Design – The research draws on literature relating to archival and LLM research, which is validated by a summary of factors drawn from four prominent pretraining datasets (WebText, The Pile, ROOTS Corpus, Dolma).

Method – The authors begin by considering pretraining datasets as collections of sociocultural material through a lens of relevant archival studies. The comparison with "traditional archives" is summarized by responses to five questions, e.g., "What impact do collections have on knowledge production?" The authors then consider several common concerns of LLM researchers with regard to selecting pretraining data and discuss how recent archival research, particularly on appraisal, might prove beneficial.

Toxic language, privacy vulnerabilities, evaluation and data contamination are the concerns addressed in the article,

Main Results – The authors report that measures to assess toxicity, privacy vulnerabilities and data contamination have known limitations. Issues commonly of concern to LLM researchers with regard to pretraining datasets include toxicity and privacy both of which are context-driven and existing practices do not adequately address that contextual element. Another common concern is the limited information generally available on the data included in a dataset which makes ensuring that data is not contaminated difficult, e.g., the presence of evaluation data that also exists in the dataset being evaluated "has implications both for rigorous evaluation and for using models for sociocultural analysis."

Discussion/Conclusion of Article – The authors conclude by advocating more 'archival like' documentation for datasets, improved finding aids to locate and describe existing datasets and assess their utility for a given purpose, and broader, community-level engagement to evaluate them.

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

Introduction:
Dias approaches the document anonymization problem through the context of a crowd sourced translation company. The company translates "real-word" documents but wants to limit any potential data exposures that are present in translated texts. He set out to create a multilingual anonymization system that could alleviate the company's problem. Dias notes the benefit of using a ML model because of their easier implementation and that most anonymization systems currently are based on ML models. He then offers a historic overview of different anonymization systems and their underlying architecture.

Metrics and Resources:
Datasets chosen by Dias to train his NER systemd included the Digital Corpus of European Parliament (DCEP) reports for the German and English in addition to CoNLL-2003, a NER testing dataset. For the Spanish and Portuguese NER's, Dias again used DCEP reports alongside a CoNLL 2002 Spanish testing dataset and "golden collection" from Segundo HAREM for the Portuguese. Some of these datasets (such as CoNLL) come pre-annotated making them a 'golden standard', yet others such as the DCEP must

be annotated by humans. Dias notes that this is a time consuming process and he created a web-browser based annotation tool, “Unannotator” to increase efficiency in this process. The DCEP corpora was chosen by the author because of the high-density of named entities mentioned in their texts. Examples of named entities for this project included organizations such as organizations, persons, and locations.

Text Anonymization:

Dias’ Text Anonymization systems contained 5 distinct modules that moved the document text through the pipeline with the end result being an anonymized document and a “table of solutions” that outlined the anonymization results such as what entities were recognized and their replacement words. Their main NER classifier used was Stanford’s CRF classifier (Stanford NER1), an open source, NLP based algorithm. Dias notes that alongside this classifier others can be used in conjunction, as done in other studies, to help improve the main NER classifier’s performance. A key addition to Dias’ system included a “Second-Pass” detection that would resolve misclassification instances where a named entity appears multiple times in a document but within different contexts. Following this second pass, the named entities pass through a Coreference resolution that helps to maintain coreferences (two linquisitc expressions that refer to the same extra-linguistic object) such as possessives (John Doe and John Doe’s) and acronyms (E.P. for European Parliament). This coreference resolution was key in helping the NER perform accurately among multiple different languages and their idiosyncrasies. Once all the named entities are recognized the text can be anonymized with 4 distinct methods as identified by Dias: suppression, tagging, substitution and generalization. Dias gives examples of each, but notes that for the most natural reading text generalization provides the best results.

Evaluation:

Overall, Dias notes that the NER classifier performed best when “trained with corpora from the same text domain” and that the Second-pass detection after the initial NER classifier worked to increase the performance in F1 scores. This performance further increased with the implementation of the Coreference resolution module noting the efficacy of a hybrid blend of NER and language detection tools. Dias further notes that "there is no anonymization method that fits all scenarios" and that while tagging (replacing a word with a unique numeric identifier, e.g. "person123") does not result in a very readable text, "it has been shown to be one of the more acceptable solutions for anonymizing a text." Furthermore he states that indirect identifiers in a text may hint at the true identity of entities, which in the worst case results in the identification of anonymized information by an informed reader.

Conclusion:

Dias concludes that a single NER tool "provides the best precision, but not the best performance" and that the hybrid combination of tools he used yielded the best performance. Out of all the anonymization methods, generalization (replacing a named entity with a more generalized version of a word) yielded the most readable results yet performed slowly since it had to draw from a knowledge base of replacement word candidates. The random substitution method worked well but at times resulted in semantic drifts within the sentence structure.

Annotation:

Dias' offers an in-depth analysis of implementing a multilingual anonymization system and its challenges. His use of an NLP driven NER shows that AI can yield good performance at anonymization yet this improves with the addition of other tools to assist the NER. Dias also notes that best results were attained when the NER is trained with "corpora" from the same text domain as the documents in need of anonymization. He further outlines the benefits and pitfalls of varying anonymization techniques (suppression, substitution, etc.) and highlights how the most natural reading of the methods, generalization, can be the slowest due to the need for the system to interface with a knowledge base of generalized terms. Dias' system, while quite robust, was designed around a small set of broader PII fields such as person and organization. It was not clear whether his system could anonymize other terms such as gender, address, education, etc.

Farley, Laura, and Eric Willey. 2015. “Wisconsin School for Girls Inmate Record Books: A Case Study of Redacted Digitization.” *The American Archivist* 78 (2): 452–69. https://doi.org/10.17723/0360-9081.78.2.452

In this qualitative case study, Farley and Willey examine the ways Wisconsin School for Girls Inmate Record Books can be digitized for greater access while protecting the subjects’ personally identifying information. These records are protected by law since they include information about juveniles and because many parts of them constitute as medical records. They are also protected from unmediated access (i.e. all records need to be redacted) through the children’s code, protecting juveniles. Currently, the Wisconsin Historical Society is responsible for granting access to researchers.

To find out whether it is feasible to digitize the records or not, Farley and Willey digitized 100 pages of 50 inmates, redacting only name and age, the information in easily identifiable fields. After redacting these two fields only, the authors could still find PII and contextual information leading to the mosaic effect in these records. The records were not consistent enough, and the redaction failed to protect 40% of the sample. Based on this result, Farley and Willey demonstrate that redaction based on a shared pattern may not be feasible. They contend that “redacted digital representations of a limited number of the institution’s records” along with the online user agreement are viable means to digitize the collection and make them accessible to the public.

This article demonstrates that redaction purely based on an expected format is ineffective in protecting personally identifying information, especially when documents are not kept in a consistent format.

Garat, Diego, and Dina Wonsever. 2022. “Automatic Curation of Court Documents: Anonymizing Personal Data.” *Information* 13 (1): 27. https://doi.org/10.3390/info13010027

In this article, the authors discuss how they solved a bottleneck that slows down access to the data at the National Jurisprudence Database. They identified the process of the anonymization of personal information as a bottleneck. The problems of manual de-identification and anonymization are that they are unreliable and delay data publication. The authors, therefore, suggest automation or semi-automation of pre-publication tasks. It automatically de-identifies proper names and replaces them with fantasy names consistently. Assigning the same fantasy name to all references of the same identity proved to be difficult. Initial attempts with off-the-shelf NER tools proved to be ineffective, and therefore, the authors had to retrain a NER module of SpaCy using transfer learning. They also decided to use a standard unsupervised agglomerative clusting algorithm implemented in Scikit-Learn and standard distance functions between strings to resolve co-references between the named entities. For this experiment, the authors used 797 documents as a training set and 200 documents for validation. In the corpus, there were 7748 mentions for training and 2220 mentions for validation in these sets.

Garcia, Antonio, Jina Lee, Jonathan McClain, Craig Jorgensen, and John Lewis. 2018. "Protecting Sensitive Textual Information Using Information Extraction and Semantic Technologies." In *The 17th International Semantic Web Conference*. Monterey, CA, USA: OSTI.GOV. https://www.osti.gov/servlets/purl/1504816

Objective – The article presents an approach to identifying sensitive information by using semantic web ontologies, Information Extraction, and SPARQL queries. This method incorporates organizational knowledge, which is a vital context that determines sensitivity.

Research Strategy/Design – qualitative research design

Setting/Sample – The researchers used "a collection of textual information pulled from the NASA's James Webb Space Telescope website" (p.2), contriving sensitivities.

Method – The researchers created two ontologies, a JWST ontology and a sensitive information ontology, using semantic technologies. JWST ontology defines organizational information. These ontologies are extracted and mapped using ontological data and Named Entity Resolution (NER) models built with the CoreNLP system. The output of the IE system is analyzed and used to create an ontological graph. Then SPARQL can be used to query whether the document holds sensitive information or not.

Main results – The model does not automatically identify the document's sensitivity, but it can suggest whether a textual document is sensitive, helping human reviewers verify the machine's output. It also provides a reason why the machine thought the document might be sensitive.

Discussion/conclusion of the article – The researchers found this method to be a successful approach to protecting sensitive information, such as organization's trade secrets and intellectual property. They plan to continue the research with more complex relationships in text.

Annotation – It is unclear whether the method would be as straightforward with real sensitive data, rather than using contrived sensitive data extracted from a public website

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

Glaser et al. establish the need for an AI centered tool which can significantly quicken the anonymization process for German court documents. Previous manual rule based NER programs "that rely on dictionary look-ups and regular expressions" lack the capacity to accurately identify outliers or rare occurrences in named entity representations (p. 205). In using AI, the authors hypothesize that "a ML model must be specifically trained for the jurisdiction at hand and cannot be utilized as a general purpose model" (p. 206). To build their model, they used a ruled based classification algorithm that detected anonymization placeholders within a paragraph. Then the paragraph gets tokenized (broken up) and fed into a pre-trained masked language machine learning model (BERT) that “masks” (i.e. anonymizes) the candidate tokens. This anonymization model was then paired with an automatic pseudonymization method which uses an NER to replace potentially sensitive tokens. Non-recognized tokens are then compared to the anonymized document to locate undetected entities and then get tagged as ‘unknown’.

By following this process and pairing the pseudonymized documents with anonymized ones, Glaser et. all offer that this can help safely train an anonymization model that would usually require an original and anonymized document pair. For each respective evaluation set (from two different courts) the author’s methods yielded a 62-68% precision rate or a recall ranging from 79% to 52%. German court staff who anonymize documents noted that these rates were not high enough to allow unsupervised anonymization. The authors mention that to heighten their model’s performance, a specialized NER model with more reference types will be required. In sum, the author’s reiterate the need for their automatic pseudonymization system to create pairs of originals and anonymized documents which can accurately train an anonymization model. They also underscore that "contextual sensitivity classification represents an important foundation" for their work (p. 209). Overall, they note that this is a complex problem and more data is needed for an autonomous anonymization system.

The model was implemented using the Tensorflow framework using Python and the NumPy library. SpaCy was used for application to original documents.

Goldman, Ben, and Timothy D. Pyatt. 2013. “Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives.” *Library & Archival Security* 26 (1–2): 37–55. https://doi.org/10.1080/01960075.2014.913966

Goldman and Pyatt highlight the current state of privacy and PII within archival studies. The article also highlights the various standards and protocols in place at existing archival institutions that protect PII information for donors to their collections. Considering how easy it may be for PII to remain dormant in old hard drives, computers, and other digital material, the need for more rigorous understanding about the longevity of PII within these items is needed in order to properly appraise material in a way that protects privacy and ensures that privacy breaches are kept to a minimum. The author's recognize that the risk of a privacy breach can only be mitigated and not completely removed and as such, aim to push the archival community to develop their own metrics for what accounts for an "acceptable" amount of risk when it comes to the privacy of donors and patrons.

Beyond legal statutes like HIPAA and FERPA, the authors recommend archives establish processes and communication standards for identifying PII within born-digital material. These include educating researchers *and* donors on PII issues and involving IT and other technician professionals in the design and creation of archival records and record management processes. Many of these revolve around more direct communication with donors and even the subjects of research themselves, which can be difficult when dealing with, for example, the primary papers of a deceased author.

For our research purposes, this article not only puts forward a survey of issues relating to PII and risk management, but also puts forth a working definition of PII from Lee and Woods that we could hopefully develop along the way: "*any data that are personally identifying, could be used to establish the identity of the producer, establish the identity or personal details of individuals known to the producer (e.g., friends, family, and clients) or are associated with a private record (e.g., medical, employment, and education)*."

Gonzalez-Granadillo, Gustavo, Sofia Anna Menesidou, Dimitrios Papamartzivanos, Ramon Romeu, Diana Navarro-Llobet, Caxton Okoh, Sokratis Nifakos, Christos Xenakis, and Emmanouil Panaousis. 2021. "Automated Cyber and Privacy Risk Management Toolkit." *Sensors* 21 (16): 5493. https://doi.org/10.3390/s21165493

Objective – This article presents a cyber and privacy risk management toolkit named AMBIENT (Automated Cyber and Privacy Risk Management Toolkit), which automatically assesses cyber and privacy risks, and recommends mitigation measures to reduce the risks. AMBIENT was developed for the healthcare industry, but the authors contend that it can be used for any field and setting.

Research Strategy/Design – After a literature review, the authors present AMBIENT, explaining its compositions and the qualities of each module that is part of the toolkit. The authors then demonstrate how the toolkit may work in real-life scenarios by presenting a few use case scenarios.

Method – This is a descriptive study of AMBIENT.

Main Results – The authors believe AMBIENT is one of the first toolkits that integrate cyber security risk management and privacy risk management systems.

Discussion/Conclusion of Article – Developed specifically with the healthcare sector in mind, AMBIENT is a tool to support decision making, not a tool that makes decisions on behalf of the organization. It is up to the organization to act upon the recommendations provided by AMBIENT. AMBIENT automates the PIA process, quantifying the risks an organization faces. However, the toolkit does not automatically assess privacy and cyber security risks. It requires manual input from the organization to calculate the risks and recommend mitigation measures. This, in turn, means that limited or incomplete input data will result in inaccurate risk levels and recommendations.

Annotation – While the automation aspect of the toolkit may seem to suggest that it has qualities of artificial intelligence, all the inputs to calculate the risk have to be entered manually. The article also focuses on the PROTECT-P (from the NIST privacy framework), and dismisses other privacy risks. Therefore, the article may not be useful in examining the ways artificial intelligence can be used to assess privacy risks. That said, it may be an interesting framework to assess privacy risks associated with cyber security.

Gottehrer, Gail, and Debbie Reynolds. 2022. “The GDPR So Far: Implications for Information Governance,eDiscovery, and Privacy by Design.” In *The GDPR Challenge: Privacy, Technology, and Compliance in an Age of Accelerating Change*, edited by Amie Taal, First

edition. A Science Publishers Book. Boca Raton, FL London New York: CRC Press, Taylor & Francis Group.

This book chapter by Gottehrer and Reynolds describes how the implementation of GDPR has affected information governance, eDiscovery, and the role of privacy in tech development. Not only is GDPR a legal requirement, but it is also a risk companies should manage. GDPR promotes the concept of Privacy by Design, which is “built on the premise that when technology is being created, developers must build in protection mechanisms for any data of EU subjects that may be affected by the new technology” (p. 104). This concept has a significant impact on eDiscovery process because the goal of Privacy by Design and the goal eDiscovery are on opposite ends. eDiscovery seeks to understand people through information, whereas Privacy by Design aims to mask an individual's identity. In order to comply with GDPR during the eDiscovery process, the authors posit that it is important to identify and clearly define data controllers and data processors under the GDPR. For instance, a vendor (the third party) handling eDiscovery would be considered a data processor and the company requesting the service would be considered a data controller. Both data processors and controllers have joint responsibilities in protecting the information, and therefore they must understand their roles and responsibilities. The authors also contend it is important for companies to keep a detailed and accurate record of how data is processed so that they can prove their compliance. Finally, companies that subcontract their eDiscovery process must understand if the third-party vendors use GDPR-compliant software. It is, therefore, important for companies to regularly review contracts with third-party vendors. The chapter then briefly discusses the difference between the GDPR and privacy-related legislation from the United States and argues that it would be infeasible for companies to completely block the information off.

Grossman, Maura R., and Gordon V. Cormack. 2013. “The Grossman-Cormack Glossary of Technology-Assisted Review.” U.S. Federal Courts Law Review. https://www.fclr.org/fclr/articles/html/2010/grossman.pdf

This article consists of an introduction to Technology Assisted Review (TAR) and the glossary of TAR. In the introduction, Facciola contends that TAR is a disruptive technology that is cheaper, performs better, and is more reliable than human reviewers, especially because words we use are often vague and can be polysemous. This is a problem in the legal field. TAR can potentially help to solve this problem. The article then provides an array of definitions of various terms one would need to know to understand Technology-assisted review.

Grossman, Maura R., and Gordon V. Cormack. 2016. "A Tour of Technology Assisted Review." In *Perspectives on Predictive Coding: And Other Advanced Search Methods for the Legal Practitioner*, edited by Jason R. Baron, Ralph C. Losey, Michael D. Berman, and American Bar Association. Chicago, Illinois: American Bar Association, Section of Litigation.

This chapter by Grossman and Cormack discusses all the components required for a tool to be considered a Technology-Assisted Review (TAR). It first establishes that search, analysis, and review are not identical, although the terms are often used interchangeably by some. TAR is specifically made for the review process. It then discusses how TAR requires control sets and should be able to comb through the search terms. It can utilize techniques such as relevance ranking, similarity search, and classifiers. TAR's machine learning mechanism can also be supervised or unsupervised, passive or active, simple or continuous. This chapter would be useful for anyone who does not know much about TAR or needs a further understanding of TAR.

Guo, Wei, Yun Fang, Weimei Pan, and Dekun Li. 2016. "Archives as a Trusted Third Party in Maintainingand Preserving Digital Records in the Cloud Environment." *Records Management Journal* 26 (2):170–84. https://doi.org/10.1108/RMJ-07-2015-0028

Objective - This study aims to find out how archives can work with private companies' digital records to protect the trustworthiness of records.

Research Strategy/Design – A qualitative Case Study of records from Tianjin Otis Elevator Co., an elevator company located in Tianjin, China

Method - The researchers conducted a descriptive study, examining the concept of public archives as a trusted third party within a cloud environment. The researchers observe and describe how Tianjin Otis Elevator Co.'s records were moved to Tianjin Municipal Archives' cloud.

Main Results - The concept of archives as a trusted third-party remain relevant even in the cloud environment. In fact, they are critical in maintaining the reliability and authenticity of digital records in the cloud.

Discussion/Conclusion of Article - The concept of archives as a trusted third party must be renewed to meet the challenges posed by the changing technological infrastructure.

Annotation - Guo et al. start the article with a history of trustworthiness as an archival concept and third parties in archives. They explain the division between creation and preservation endows record with authenticity. They then present a case study of Tianjin Otis Elevator Co., whose maintenance records were moved to the Tianjin Municipal Archive's cloud, which discharged accountability from the company and ensured the reliability and authenticity of the records. The article further argues that to ensure the authenticity of records, the role of public archives extended to the private company.

While the article does not explicitly discuss the issue of privacy within digital archives, the article raises interesting points regarding the digital curation of private records and the trustworthiness (reliability, authenticity, and accuracy) of private records in the public archive setting. The authors briefly discuss confidentiality, security, and privacy issues in the article, but they believe that these can be protected with rigorous security measures. They do not consider archives as a potential source of privacy breach even though the maintenance records hold various personally identifiable information.

Gupta, Abhishek, and Nikitasha Kapoor. 2020. "Comprehensiveness of Archives: A Modern AI-Enabled Approach to Build Comprehensive Shared Cultural Heritage." https://doi.org/10.48550/ARXIV.2008.04541

Citing the importance of more decentralized, community based archives for marginalized groups seeking to protect their cultural sovereignty and heritage from institutional and political oppression, Gupta and Kapoor observe the need for more technological approaches to the ways in which cultural heritage and knowledge is stored beyond centralized, larger archival institutions (which may themselves have a problematic history in regards to the preservation of artifacts and the intentional obfuscation of knowledge regarding women, people of color, marginalized gender identities, etc…).

Community based archives allow for a more self-affirming, intentional means of preserving a culture apart from the history it may have in relation to an oppressor. This holds true in particular for Native American cultures whose internal privacy has been repeatedly violated by governments for centuries. In this case, the need to develop more technologically sound preservation mechanisms for these community archives works as a means of protecting their privacy, identity and heritage. However, the current state of the internet, wherein search indexing algorithms tend to obscure archival projects that do not have the proper SEO (Search Engine Optimization) in mind, makes it much more difficult for these archives to gain more recognition and develop further programming and outreach within their own communities:

"While the internet has enabled a larger populace to self-document and own their own narratives, such records continue to appear online in fragmented ways – leading to low discoverability and digital marginalization. The collation efforts by archivists are thus limited by the technical ability to find fragmented context-rich records. Automated methods – web crawlers, discovery algorithms and other online nudge approaches –create real challenges for discoverability of lesser dominant records. Records become especially obsolete by online search engines and tools when they are not in the dominant languages of the internet, for example."

Henttonen, Pekka. 2017. “Privacy as an Archival Problem and a Solution.” *Archival Science* 17 (3): 285–303.https://doi.org/10.1007/s10502-017-9277-0

In this article, Henttonen first establishes that the idea of privacy is defined differently based on the interpretation of information and the context the interpretation is made. Therefore privacy violations occur when the information crosses the borders of contexts and/or social spheres. He further states that archives inherently violate privacy because archival work is a secondary use of the records. With this assumption, Henttonen examines five strategies to mitigate privacy concerns in the digital environment and their limitations. These strategies are purpose limitation, privacy self-management and the right to be forgotten, destruction, anonymization, and information safe haven approach. Henttonen demonstrates that the choice of the strategy affects the information that is preserved and how it can be used. He argues that the safe-haven approach – taking information into archival custody and preserving it there with restricted access until public access is possible – is the best strategy from an archival perspective because it prevents contextual transfer and preserves the information. Other strategies either allow contextual transfer to occur or cause a lot of information to be lost. He finally contends that the archival needs are hardly considered in these strategies and calls for records management professionals and archivists to let their concerns known as these strategies are implemented.

Hutchinson, Tim. 2017. “Protecting Privacy in the Archives: Preliminary Explorations of Topic Modeling forBorn-Digital Collections.” In *2017 IEEE International Conference on Big Data (Big Data)*, 2251–55.Boston, MA, USA: IEEE. https://doi.org/10.1109/BigData.2017.8258177

Objective – Focusing specifically on born-digital collections, the conference paper aims to explore the ways natural language processing techniques can be applied to identify sensitive contents.

Research Strategy/Design – Case study

Setting/Sample - the records of a University of Saskatchewan Associate Vice-President for Information and Communications Technology

Method – The author analyzed records using ArchExtract, a software/project created by UC Berkley's Bancroft Library and is no longer developed, to identify HR-related documents. He used the topic modeling method to identify PII, running each configuration with 10, 15, 20, 40, and 100 topics.

Main Results – It is incredibly challenging to identify sensitive documents. While topic modeling was successful for high-level classification, it was rather unsuccessful on a document level.

Discussion/Conclusion of Article – The author concludes that the training for topic modeling needs to be refined and more specific.

Annotation – The author conducted further research on some of the issues he identified. The findings from the further research were published in 2018 with the title "Protecting Privacy in the Archives: Supervised Machine Learning and Born-Digital Records."

Hutchinson, Tim. 2018. "Protecting Privacy in the Archives: Supervised Machine Learning and Born-Digital Records." In *2018 IEEE International Conference on Big Data (Big Data),* 2696–2701. Seattle, WA, USA: IEEE. https://doi.org/10.1109/BigData.2018.8621929

Objective – The conference paper describes the author's experience in developing trainings sets for supervised ML on HR-related documents, which contain PII.

Research Strategy/Design – Case study

Method – Using WEKA, Java-based open-source software for data mining, the author worked with the University of Saskatchewan Associate Vice-President for Information and Communications Technology collection, which includes approximately 2000 documents, to develop training sets for supervised machine learning. The training set categorized records into three groups: HR-personal (HR records with PII), HR-general (HR records without PII), and non-HR (all other records). The author used the Naïve Bayes (multinomial) classifier. There were a total of 16 rounds of training set development.

Main Results – When the model used two categories (generally HR and non-HR), the recall for HR was successful, but recall for non-HR was mixed. The precision for HR was uniformly high, whereas the precision for non-HR was uniformly low. When three categories were used (HR-general, HR-personal, non-HR), the recall for HR-general was decent, HR-personal was reasonable, and non-HR was mixed. Precision for HR-general was poor, HR-personal was mixed, and non-HR was excellent. None of the rounds had high precision scores for all categories. The systematic approach to creating a training set was unsuccessful. On the other hand, a manually generated training set proved to be most successful in both recall and precision.

Discussion/Conclusion of Article – The author suggests "supervised machine learning could be a viable approach for a "triage" method of reviewing collection for restrictions." The author further contends that ML would be useful to measure the privacy risk and help organizations determine the required level of access restrictions. The researcher, however, recognizes various questions this study did not answer and calls for more research.

Iacovino, Livia, and Malcolm Todd. 2007. "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 (1): 107–27. https://doi.org/10.1007/s10502-007-9055-5

Objective – It aims to answer various questions regarding the tension between the right to privacy and the right to information from a legal, ethical, and digital archival perspective.

Research Strategy/Design – Comparative qualitative study

Setting – the United States, the European Union, Australia, and Canada

Method – Mainly focusing on the European Union, the article examines privacy legislation in the EU, Australia, Canada, and the United States

Main Results – The European Union's Data Protection Directive 95/46/EC introduced a full and detailed order regarding privacy, requiring several countries to review and revamp their own privacy legislation. In the EU, the right to privacy triumphs over the right to information. In the United States, the right to information triumphs over the right to privacy. There are not as comprehensive restrictions regarding privacy in Australia, Canada, and the United States, especially in the private sector.

Discussion/Conclusion of Article – The article concludes that technical details are critical when PI needs to be preserved in digital format. It recommends an early appraisal of records to protect the privacy of those appearing on records, using a Belgian model as an example. It further recommends that the law be modified to provide a broader scope of interpretation like it is done in Italy. In Italy, archival and researcher ethics are embedded in privacy legislation. The article finally questions how data that has been de-identified for privacy purposes can be re-identified in the future for archival purposes.

Annotation – The article joins archival and legal perspectives on privacy in digital records setting, which would be helpful in this study. However, it must be noted that this article is from 2007, and its analysis of the legislations is outdated. For instance, the article focuses on the EU's general directive and individual country's legislations under the EU, but GDPR has replaced the directive.

Kastenhofer, Julia, and Shadrack Katuu. 2016. "Declassification: A Clouded Environment." *Archives and Records* 37 (2): 198–224. https://doi.org/10.1080/23257962.2016.1194814

Kastenhofer and Katuu's descriptive study on declassification in a cloud environment offers a high-level overview of the concept and its application in Intergovernmental Organizations. First, they examine the idea of classification, including sensitivity classification, categories of information sensitivity (genuine national secrecy, bureaucratic secrecy, and political secrecy), and demonstrate how records are often classified based on other criteria, other than risk, due to political intentions. They then introduce types of declassification and stages of declassification, which is followed by a case study of five different intergovernmental organizations' declassification processes, in which they identify the following:

- Unclassified information is not the same as publically available information
- Declassification is an ongoing activity
- The decision to restrict information is easier than the decision to declassify
- When a record is classified, a specific reason for its classification should be indicated on the record
- When a record is classified, a potential declassification date should be indicated on the record
- Declassification takes time (pp.220-221)

Kastenhofer and Katuu conclude the article with the call for further examination on the topic.

The article does not examine the issue of PI, but it may still be useful for the purpose of our studies because it examines how sensitive records should be handled in organizations. It also briefly discusses the risk-management procedure as it explores the declassification techniques. The authors highlight that the MPLP approach can be helpful in declassifying records – conducting functional analysis of the records first to determine the risk and do a line-by-line review for those considered high-risk.

LeClere, Ellen. 2018. "Breaking Rules for Good? How Archivists Manage Privacy in Large-Scale Digitisation Projects." *Archives and Manuscripts* 46 (3): 289–308. https://doi.org/10.1080/01576895.2018.1547653

Focusing on the privacy issues around archives and digitization in particular, LeClare interrogates the notion that archives are inherently a public good within a Western-style democracy:

> The claim that archives – and by extension, digital archives – serve public interests
> within a liberal democracy is not uncontroversial. Archives in liberal democracies create a sense of accountability, transparency and access to information, but maintaining these values comes at the expense of asking marginalized groups for higher contributions for fewer benefits. This argument is also uncontroversial – access to archives has been historically controlled by privilege and power.

With this lens, LeClare uses a case study of the ongoing efforts to digitize moments within the Civil Rights Movement, wherein organizers and activists were routinely surveilled and harassed by white anti-integrationists, often with the support and cooperation of local government and law enforcement. LeClare probes the ways in which the identities and private information of individuals who fought against integration in the South were often subject to lengthy battles over their privacy as a means to "keep the peace" while the same information from black and brown activists were deemed fair game. This in turn informs the ways in which an archive detailing this involvement seeks to digitize and preserve this moment in history while protecting the rights and privacy of those whose lives were marginalized and often at risk for sabotage.

LeClare's interviews with archivists working in this field highlights the disparities around privacy rights that occur along racial and class lines. Though many in the field often argue for a more "open archive", accessible and transparent to all, more critical work needs to be done to ensure what's preserved is not built at the expense of these communities.

Lee, Benjamin Charles Germain. 2022. "The 'Collections as ML Data' Checklist for Machine Learning & Cultural Heritage." arXiv. https://doi.org/10.48550/ARXIV.2207.02960

Objective – The article draws on developments in the cultural heritage community regarding responsibly applying machine learning techniques in libraries and other cultural heritage institutions to propose a practitioner's checklist "with guiding questions and practices" for use in machine learning projects using cultural heritage data.

Research Strategy/Design – Compilation of checklist questions from existing guidance followed by application to case studies and finalization through practitioner feedback.

Method – Starting from a survey of checklists, toolkits, impact assessments, "and beyond" from in the machine learning domain, Lee produced a taxonomy to increase the comprehensiveness of guiding questions and to provide cultural heritage practitioners with a guide to existing work. Lee then selected five representative cultural heritage projects to use as case studies for developing and testing the questions making up the proposed checklist. The proto-checklist was iteratively tested against the projects and finally, seeking feedback from researchers and practitioners.

Main Results – The resulting checklist, which is included as an appendix to the article, has four sections: 1) The Cultural Heritage Collection as Data; 2) The Machine Learning Model; 3) Organizational Considerations; and 4) Copyright, Transparency, Documentation, Maintenance, and Privacy.

Discussion/Conclusion of Article – The article summarizes case study findings for each section of the checklist identifying influential sources from which checklist questions were derived. The case study findings highlight challenges of responding to the questions such as:

- complex histories surrounding the creation and curation of cultural heritage collections such as "The Real Face of White Australia" case study (first section);
- Identification of stakeholders in order to appropriately engage them in the decision to employ machine learning using a cultural heritage collection (third section).

The author acknowledges that despite the many iterations leading to its development, it is still not comprehensive and, in any case, use of a checklist "does not mean the project should not be interrogated further or documented more extensively."

Lee, Christopher A., and Kam Woods. 2012. "Automated Redaction of Private and Personal Data in Collections." In *Conference Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation.* Vancouver British Columbia, Canada: UNESCO. http://www.ils.unc.edu/callee/p298-lee.pdf

Lee and Woods outline the ways in which PII and Private data can be accumulated in archival institutions without any proper appraisal or oversight. Some of this can be attributed to the fact that properly appraising and assessing privacy vulnerabilities within records takes time and expertise that many archives are not able to lend given resource and time constraints. To this end, Lee and Woods demonstrate the scope and intentions behind the BitCurator project (Github linked here), which provides a suite of open source software tools to scrape hardware, software, and firmware for PII and presents them in a human readable format, complete with memory addresses and storage data. With this information, digital forensics tools normally used within criminal investigations can be utilized and customized for a given institution and/or collection.

These tools have undergone significant updates and revisions since Lee and Woods wrote about them here. As such, there may be some potential in measuring how they're currently used and how effective and efficient they are in saving time measuring potential privacy vulnerabilities within collections.

Lemieux, Victoria L., and John Werner. 2023. "Protecting Privacy in Digital Records: The Potential of Privacy-Enhancing Technologies." *Journal on Computing and Cultural Heritage* 16 (4): 1–18. https://doi.org/10.1145/3633477

Objective – The article is an argument for archival experimentation with Privacy Enhancing Technologies (PETs)—"a class of emerging technologies that rest on the assumption that a body of documents is confidential or private and must remain so."

Research Strategy/Design – The article is a scoping review, by which the authors aim to highlight PETs-style technologies within the cultural heritage realm. This approach is distinct from a systematic review that might have the aim of summarizing how the class of technology is being used or to comprehensively evaluable PETs.

Method – The authors sought survey articles on the use of PETs for the protection of PI from the past six years to understand how they were currently being used and which were most common. Key words drawn from the survey articles and citation chaining were then used for subsequent searches for an inclusive search of the available literature.

Main Results – The article provides a comprehensible and comprehensive overview of PETs, including the concepts on which they are based, often some history of the origins of each technique, and their limitations. The discussion section considers possible use cases of PETs in the archival domain.

Discussion/Conclusion of Article – The authors set the stage by noting 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 due to complexities of applying existing privacy protection legislation to large and often poorly described archival collections. The results of such approaches are insufficiently accurate and, in any case, fall short of the scale needed. Less human-dependent approaches, such as neural networks, likewise lack the accuracy needed meaning that trust in both the tools and the archival institutions that might use them would erode.

“PI” (for “Private Information”) is the abbreviation introduced and used in this article despite a brief discussion on the prevalence of PII (Personally Identifiable Information) in the U.S. and personal data in the European context. It is not clear why a new term was needed, particularly given that “PI” (Personal Information) is commonly used in Canada.

The authors identify several types of tools/techniques that fall into the category of PETs:

- Homomorphic Encryption (HE)
- Trusted Execution Environments (TEE)
- Secure Multi-Party Computation (SMPC)
- Differential Privacy (DP)
- Personal Data Stores (PDS)
- Privacy-Preserving Machine Learning (PPML).

They note that technologies supporting these tools include tokenization, synthetic data, blockchain and distributed ledger technologies and zero knowledge proofs.

The article notes that HE and DP approaches are the “most prevalent”. However, HE has limitations, not least its expense and slowness, while the application of DP is challenged by balancing the level of “noise” introduced into the dataset (which protects privacy) against the utility of the results. Those testing TEE were found to pair it “with some other privacy-preserving technique.” SMPC focuses on computational privacy among distributed partners, i.e., computation is enabled on a consolidated data store without partners being able to “see” each other’s data. PDS technologies enable individuals to access and control the sharing of their own PI. PPML is focused on protecting the privacy of information in machine learning models.

Noting that archives frequently have resource limitations, the authors suggest ways PETs might make archival holdings at least partially accessible, e.g., using HE to enable access to full census records before they reach a century in age. Another example is the use of PDS as a “more human-centric and participatory approach to archiving” that supports of truth and reconciliation processes or human rights cases.

The authors conclude by noting that many barriers exist before PETs may be implemented in archival functions, including an absence of clear technical standards or ethical frameworks.

Liu, Bo, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. “When Machine Learning Meets Privacy: A Survey and Outlook.” *ACM Computing Surveys* 54 (2): 1–36. https://doi.org/10.1145/3436755

Liu et al begin with an overview of Machine Learning and the various forms it can take (e.g. Supervised/Unsupervised, Centralized/Federated, etc…). They then categorize the various types of Privacy Issues related to ML, such as whether Machine Learning is the target of a privacy attack, a tool used to prevent privacy, or an attack tool used to breach user and system privacy.

There are also considerations when it comes to whether or not an attack is directed towards the Machine Learning model itself (which can house proprietary information for companies that rely on it for revenue) or say, the training data used to develop the ML-Model (which can house private user data). They post various methods by which an attacker can attempt to hijack and reverse engineer an ML model by pouring through either the outputs of the algorithm or brute-forcing its way in the ML model itself.

In terms of protection, there are a number of ML protection schemes that involve encrypting the data itself, planting false and obfuscating reference points within the training data that the ML model knows to ignore but is hidden from attackers, and many more.

In all this article gets into more granular detail about the technical aspects of privacy protection and Machine Learning. While not entirely focused on archival practice, it is a good way of highlighting some of the potential privacy challenges and opportunities ML based tools can afford archival institutions as they seek to include such tools in their work.

Liu, Xiaojing, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. "Graph Convolution for Multimodal Information Extraction from Visually Rich Documents." https://doi.org/10.48550/ARXIV.1903.11279

Liu et al. utilize a graph convolutional architecture to conduct information extraction (IE) on visually rich documents (VRDs). Based on the examples provided (purchase receipts and invoices) these would represent structured, textual records often used in a business context. The authors note that few other IE activities focus on the location of entities within a document and rather execute IE tasks on plain text documents where the information solely lies within the document's language. Their model aims to combine the more established plain text entity recognition with a graph convolutional network to incorporate the relative position of entities within a document into the IE process.

Through their experiments, Liu et al. establish the efficacy of their graph convolutional architecture when compared to the baseline BiLTSM-CRF named entity recognizer. Specifically for entities such as ‘price’, ‘tax’ and ‘buyer’ the author’s model more successfully identified these within the VRD when compared to the named entity recognizer baseline. The application of graph convolutional networks to structured documents appears to offer a method for incorporating the formal elements of a document into the entity recognition and information extraction tasks. Liu et al. help demonstrate the importance of such formal elements in computationally determining the information within a document.

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

This thesis by Graham McDonald is about how information retrieval (IR) and text classification (TC) technologies can be deployed to assist with the sensitivity review of digital government documents in the UK government context. First, the thesis argues for sensitivity classification – a process of automatically identifying sensitive information as a document classification task. It then proposes a framework that builds upon it. More specifically, the thesis contends:

- automatic sensitivity classification can be effective for assisting human reviewers with the sensitivity review of digital government documents
- an effective sensitivity classifier can be learned by identifying the latent vocabulary, syntax and semantic language features of the sensitive information in a corpus
- by deploying an active learning strategy to select specific documents to be reviewed and by having a reviewer annotate, or redact, any passages of sensitive text in a document as they review, we can identify the most informative annotated terms to construct a representation of the sensitivities in a collection. (p.24)

Throughout the thesis, the author makes various proposals to make the sensitivity review of digital government documents more effective. In chapter four, the thesis proposes a framework that helps human reviewers with sensitivity review of digital government documents. It has four components: the document representation component, the document prioritization component, the feedback integration component, and the learned predictions component. In chapter five, the thesis argues that the document sanitation technique is not useful for classifying sensitive information and instead proposes that the identification of documents with sensitive information should be addressed as a text classifier task. Chapter six introduced "an enhanced sensitivity classification approach that integrates automatically generated features of sensitive information" (p.176). McDonald further argues that human reviewers need to continue to be involved in the reviewing process for the sensitivity classifier to be effective. The thesis then offers case scenarios demonstrating how their proposals can work together in chapters eight and nine.

McDonald, Graham, Craig Macdonald, and Iadh Ounis. 2020. "Active Learning Stopping Strategies for Technology-Assisted Sensitivity Review." In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2053–56. Virtual Event China:ACM. https://doi.org/10.1145/3397271.3401267

Objective – The study compares two active learning stopping strategies they suggest (Total Conf, LeastConf) against the three state-of-the-art active learning stopping strategies (StablePred, Classification Change, Min-Error) from the literature.

Research Strategy/Design – Experiential quantitative research

Setting/Sample – A test collection of 3801 government documents that have been reviewed by expert reviewers according to the UK FOI Act 2000 was used for this study. Of the test collection, 502 documents were considered sensitive and 3299 were not.

Method – The researchers used 500 documents (435 non-sensitive, 65 sensitive) "as a fixed held-out set to evaluate the effectiveness of the classifier at each iteration of the active learning process" (p. 2055). In each iteration, 20 documents were labelled. The researchers deployed "a SVM classifier with a linear kernel and $C = 1.0$" (p. 2055). For each identified strategies, they "set the threshold number of iterations to trigger the stopping strategies, $\epsilon = 3$, and our Cohen's κ threshold $\theta = 0.99$" (p.2055). The statistical significance was tested using McNemar's nonparametric test, with $p < 0.05$.

Main results – TotalConf stopping strategy results in the most sensitivity classifier in all aspects (Precision, Recall, F1, F2, BAC, auROC), followed by LeastCof. Both of the proposed methods outperformed the strategies presented in the literature. Out of the strategies from the literature, StablePred performed the best, but it is important to note that the strategy is far more aggressive than the strategies proposed by the researchers, meaning it tends to stop active learning before it reaches the optimal classifier.

Discussion/conclusion of the article – The study concludes that "a sensitivity classifier's uncertainty can be a good indicator of its effectiveness" (p.2055). Moreover, knowing when to stop the active learning process is essential because it not only makes the classifier more effective but also avoids unnecessary dictation of the reviewing order.

Mcdonald, Graham, Craig Macdonald, and Iadh Ounis. 2021. "How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review." *ACM Transactions on Information Systems* 39 (1): 1–34. https://doi.org/10.1145/3417334

Objective – The article aims to investigate how the accuracy of the automatic sensitivity classifier and the said classifier's confidence in its decisions affect the reviewers' accuracy and the reviewing speed.

Research Strategy/Design – The study used a within-subject design.

Setting and sample – The study sampled documents from a collection of 4000 UK government documents, which were all born-digital internal government communication documents. The study included 24 expert sensitivity reviewers and 7 participants with a background in politics or International Relations who are familiar with the concept of Freedom of Information. The study began with 8 participants, but one participant had to be removed as they did not demonstrate a good understanding of sensitivities in government documents.

Method – Expert sensitivity reviewers first reviewed the document samples to create the "ground truth." The study participants were given training, and those who demonstrated a good understanding of the concept were invited to identify sensitivities in the documents. Throughout the process, they were assisted by three sensitivity classification treatments (none, medium, perfect). Each participant reviewed 60 documents, 20 documents for none, medium, and perfect classification treatment batches. The participants' performances were evaluated against the ground truth established earlier by experts.

Main Results – The study found that as the effectiveness of the classifier increases, the mean participant balanced accuracy scores increase as well. It also found that providing reviewers with classification predictions had a meaningful impact on reviewing speeds. It significantly increased the wpm. The confidence level of a classification prediction had a significant effect on reviewers' performances. When the classifier's confidence was high, the reviewers' agreed more with the classifier's decision. However, the classifier's confidence had mixed results on reviewing speeds in terms of NPS. The reviewing speeds increased when the reviewers agreed with the classifier's predictions, but overall the confidence did not significantly impact.

Discussion/Conclusion of Article – The study concludes that the classifier can increase the number of documents reviewed for sensitivity without affecting the accuracy level since it would allow the government to hire less experienced reviewers. It also states that sensitivity classification prediction can have a positive, negligible, or negative impact on reviewing time. However, "assisting reviewers with sensitivity classification predictions can indeed reduce the additional reviewing overhead that arises from judging sensitive information" (p. 4:28).

McDonald, Graham, Craig Macdonald, Iadh Ounis, and Timothy Gollins. 2014. "Towards a Classifier for Digital Sensitivity Review." In *Advances in Information Retrieval*, edited by Maarten De Rijke, Tom Kenter, Arjen P. De Vries, ChengXiang Zhai, Franciska De Jong, Kira Radinsky, and Katja Hofmann, 8416:500–506. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-06028-6_48

Objective – The researchers wanted to find out whether it is possible to "improve upon a text classification baseline for identifying sensitive records." (p. 503)

Research Strategy/Design – Experiential quantitative research

Setting/Sample – The researchers used a test collection consisting of 1111 government records, of which 104 and 84 were sensitive under section 27 (International Relations) and section 40 (PI) of UK FOIA, respectively.

Method – After 17 assessors identified sensitivities of the records of the sample collection, the researchers deployed "a text classification approach, where a record is presented by a term frequency vector, over all terms in the collection" (p.504) to set a baseline. Terrior IR platform was used to extract and score text classification features. The researchers then applied other features, pCount (person name count), cRisk (country risk), nEntity (number of people in specific roles of interests), and subCof (subject confidence), extending the text classification approach. These features were identified through dictionaries, databases, sentiment analysis toolkit, etc. They used SVMLight with a linear kernel as a classifer.

Main results – The researchers found that cRisk and nEntity features improve the classifers' performance for IR sensitivity (section 27 of UK FOIA), but were not so helpful for PI sensitivity (section 40 of UK FOIA). They found that pCount performed poorly for both sections, not improving the classifiers' performances. subjConf, on the other hand, led to "the classifiers degradation for section 27" (p.505).

Discussion/conclusion of the article – The authors conclude that adding features described in the paper, namely the nEntitiy (number of people in specific roles of interest) and cRisk (risk scores for countries identified within a record)could improve the baseline of a text classification. However, "these features did not help to improve BAC for personal information sensitivities." (p.505).

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

In this article, Moss and Gollins argue that archivists have overly focused on the technical challenges of digital preservation for far too long, even though many of the challenges to digital preservation arises from describing and presenting items for use. They, therefore, suggest that archivists must focus on appraisal, sensitivity review, and access.

The authors believe "the archive has to take what it is given, from the context in which the users have chosen to use it" (p.6). From this context, the issue of sensitive information is raised. In recent years, personal information has become a commodity by analytics companies. Everyone is watched and monitored – we now live in a surveillance society. At the same time, some records will inevitably hold personally identifiable information, and they need to be used to hold people accountable. In other words, people desire to protect their privacy, but at the same time, records need to be kept indefinitely so that "the tractability of the internet can be used to make contact with people we do not know all around the world" (p.12). To solve this dilemma, there needs to be a framework that creates and governs the regulatory environment – i.e. regulations.

Sensitivity is a fraught term, and it is challenging for machines to detect the nuance and context the sensitive information is distributed in. It is, thus, critical to develop a system that will deal with the volume of materials to be reviewed. Risk management is done in the context of legislation, regulation, and reputation, weighing the costs and benefits. Risk-averse companies are more likely to close or destroy the records for fear of losing their reputation. Not only is the duty to record important, but also making them available is important. Given that risk management is handled by the corporate in this digital age, records managers and archivists need to "argue for a less risk-averse attitude to the release of information" within the organizations they serve (p.21).

Murphy, Mary O., Laura Peimer, Genna Duplisea, and Jaimie Fritz. 2015. “Failure Is an Option: The Experimental Archives Project Puts Archival Innovation to the Test.” *The American Archivist* 78 (2):434–51. https://doi.org/10.17723/0360-9081.78.2.434

Objective – The article examines the Experimental Archives Project at the Schlesinger Library, which considered and tested ideas from the non-archival field in order to innovate and speed up the manuscript processing.

Research Strategy/Design – qualitative experimental research

Setting – Schlesinger Library, Radcliffe Institute for Advanced Study, Harvard University.

Method – The authors first describe the digitization (or as authors put it, digital processing) project consisting of five experiments, followed by a brief literature review of technological experiments in the archival field. The authors establish that there are not enough experiments that focus on technology in the archives field. They then describe each experiment in length and the lessons they learned from each experiment.

Main Results – The first experiment processed items before arranging and describing the collection, inverting the common archival practices. Item level interventions were done to redact PII. A similar process followed in the second and third experiments, but the researchers used OCR to improve accessibility. The fourth experiment’s goal was to use digitization to simplify archival processing, and the researchers found that it is crucial to focus on the required minimum rather than subjective analysis. The researchers valued accessibility over authenticity. For the fifth experiment, the researchers decided to arrange and describe the digital surrogate only, without touching the original items. In both fourth and fifth experiments, digitization assistants had to manually intervene and redact all PII on digital surrogates. They operated with an assumption that this information would eventually be revealed to the public after some years.

Discussion/Conclusion of Article – The researchers learned the following: it is essential to be flexible when completing a digitization project; all collections require different processes because each collection is unique; digitization will be completed faster if there is less process involved; archivists should consider not arranging and describing the physical – original – collection for some; archivists should consider using software not necessarily developed for the archival field (such as Flikr) to provide access to researchers.

Annotation – The article provides an interesting perspective of digitization in archives. Using the main ideas from the More Product, Less Process theory, the researchers prioritized accessibility and searchability over other archival characteristics, such as reliability or authenticity. The way personally identifiable information is handled is also intriguing, for the researchers manually redacted the information using Adobe Acrobat. Unfortunately, the article does not provide the standard used to redact the information and what was considered as PII. It also fails to address how the archives will protect the PII in the physical collection as they remain untouched but accessible to researchers if requested. In other words, the security of non-redacted versions, whether they are digital surrogates or originals, is not discussed at all. It also does not discuss the potential threat of using platforms such as Flickr. Lastly, there is no discussion around artificial intelligence in the article.

Narvala, Hitarth, Graham McDonald, and Iadh Ounis. 2020. "Receptor: A Platform for Exploring Latent Relations in Sensitive Documents." In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2161–64. Virtual Event China: ACM. https://doi.org/10.1145/3397271.3401407

The article presents Receptor, a tool that can support sensitivity reviewers with searching large collections to find latent relations between documents, entities, and events. The researchers believe that these entity relationships indicate sensitivity, recognizing the relational nature of sensitive information. For this reason, Receptor is different from eDiscovery tools. Receptor also aims to "capture and visualize latent relations," (p.2162) which makes it different from ePADD or BulkReviewer. Receptor has four layers in its system – the data layer, service layer, business layer, and application layer. The most relevant layer to our current studies is the service layer, in which "the source data is passed through the information extraction pipeline to perform the following tasks: (1) extracting document attributes (2) Named Entity Recognition (3) Syntactic Dependency Parsing (4) Entity Resolution and (5) Information Enrichment using external sources" (p.2162). Receptor's functionalities include: exploratory search, profile (how a particular attribute/entity appears in the collection) generation, interactive visualization of latent relations (timeline and network), and query suggestions and automatic query generation. Receptor uses the classifiers defined in McDonald et al.'s "Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings." Receptor was implemented in python. Its NLP operations used spaCY, and its web-interface was implemented using the Django framework. See https://youtu.be/-e6m7lRIcsc for the video of Receptor.

Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57. https://ssrn.com/abstract=1450006.

Objective - The article's objective is to illustrate how data re-identification techniques have fatally compromised anonymization as the foundation of regulatory privacy protections. The author then proposes new approaches and tools to protect individual privacy.

Research Strategy/Design - The research is based on published research literature from the legal and computer science domains.

Method - The author begins by explaining the dominant role of technology-based anonymization as the foundational assumption in existing privacy regulation. He then sketches the improvements in what he terms "reidentification science," drawing on findings from the computer science domain, and observes how it defeats the aims of existing privacy regulations. The author concludes by outlining a more context-based approach to privacy regulation.

Discussion/Conclusion of Article - The article sets out the dominance of anonymization as the foundation of privacy regulation. High confidence in anonymization enabled a "release and forget" approach to sharing data, ie, if data were anonymized, it could be shared without privacy being compromed. The author observes that “simply removing personally identifiable information (PII)...is now a discredited approach.”

The power of re-identification approaches using data external to the anonymized data source and the ability to derive personal information from non-PII data such as movie ratings are shown to fatally compromise anonymization and “release and forget” sharing as an adequate approach to privacy protection, arguing that it becomes simply impossible to continually add to the types of data that potentially could be used for re-identification. With anonymization compromised, sharing data so as to benefit from its utility runs counter to privacy regulation.

The author describes a notional “database of ruin,” effectively the accretion of all information existing in all the databases enabling comprehensive re-identification. The article considers three interim approaches for evolving privacy regulation including: i) reverting from the current preventative approach to one where redress is sought for harm done; ii) waiting for a technological solution; and iii) simply banning re-identification.

The author concludes by proposing expanding regulation to types of databases and database owners, eg, “large entropy reducers”--owners of huge databases that contain “so many links between so many disparate kinds of information that they represent a significant part of the database of ruin,” such as LexisNexis and Google. The author proposes an approach to privacy regulation that is both comprehensive but also sector-specific. The approach is supported by a test to determine what regulation is needed based on five non-exhaustive factors: i) data-handling techniques; ii) private vs public release; iii) quantity [of data]; iv) motive [to re-identify]; and trust and concludes by administering the test using health information and IP addresses and internet usage information.

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

A follow up to the Tamper M. et al. article from 2018, this article introduces Automatic anonymization and annotation of legal documents (ANOPPI), a collaboration project between the Ministry of Justice in Finland (coordinator), Aalto University, University of Helsinki (HELDIG), and Edita Publishing Ltd. It aims to automate and semi-automate the anonymization of Finnish court documents to comply with the GDPR. It uses both statistics and rule-based NER for the finnish language, titled FiNER. After recognizing the named entities and their occurrences, ANOPPI pseudonymizes them. At the time of this review, the project seems to be still being developed. See: https://seco.cs.aalto.fi/mission/

Özdemir, Lale. 2019. "The Inevitability of Digital Transfer: How Prepared Are UK Public Bodies for the Transfer of Born-Digital Records to the Archives?" *Records Management Journal* 29 (1/2): 224–39.https://doi.org/10.1108/RMJ-09-2018-0040

Objective - This article aims to assess how prepared UK public bodies are for the transfer of born-digital records to the National Archives (TNA).

Research Strategy/Design - This is a qualitative study with a cross-sectional research design format.

Setting and sample - The researcher conducted a survey of 23 public bodies, including ministries, charities, and non departmental public bodies. The sample was selected through both the purposive sampling method and the random sampling method. The researcher tried to select public bodies based on the number of anticipated records to be sent to TNA. The rest were chosen randomly.

Method - An eight-question survey with open-ended questions about the current digital landscape and transfer of records was distributed to 27 UK public bodies, of which 23 responded.

Main Results - The study found that many of the UK ministries were not prepared for the transfer of born-digital records to TNA. Similarly, they were also unaware of the new risks associated with digital sensitivity because the processes for digital sensitivity of records were yet to be established.

Discussion/Conclusion of Article - The author concludes that "long-term planning for the transfer of born-digital records is yet to be undertaken and that public bodies are more likely to deal with the issue when their digital records are closer to reaching the point of transfer." (p. 224)

Annotation - It is important to note that the survey was done in 2017, and the article was published in 2018. Since then, many new technologies have emerged and there may have been new processes implemented. For the purposes of our studies, it highlights the importance of identifying new sensitivity issues in the digital landscape and the need to develop a process using TAR.

Rolan, Gregory, Glen Humphries, Lisa Jeffrey, Evanthia Samaras, Tatiana Antsoupova, and Katharine Stuart. 2019. "More Human than Human? Artificial Intelligence in the Archive." *Archives and Manuscripts* 47 (2): 179–203. https://doi.org/10.1080/01576895.2018.1502088

Rolan et al. demonstrate the emergence of AI tools and technology in the archival professional through analyzing four case studies in Australia: Public Record Office Victoria (PROV), NSW State Archives and Records (NSWSAR), National Archives of Australia (NAA), and The Australian Government Department of Finance (DoF).

PROV used eDiscovery tools, in combination with the Nuix tool, to perform a Technology-Assisted Review of their emails. Nuix applied an MD5 hash to tag identical emails as it identified over 40% of 4.6 million emails to be duplicates. PROV defined three groups to evaluate retention: Positive or valueable emails; Negative or low-value emails; and Macro where the metadata was assessed and applied as a means of understading functional context and roles of the creators. Despite false positives, Nuix identified 93% of the de-duplicated emails as holding value and only 7% were labeled non-records by the algorithm for the Negative category. PROV demonstrated a multi-layered approach to using the Nuix eDiscovery tool, claiming that the path should be:

remove duplicates first, identify low-value emails, and evaluate the metadata in the Macro approach.

NSWSAR piloted a program to use “off-the-shelf machine-learning software” as a means of applying a retention and disposal authority onto 8784 files. They wished to automate a previously manual function. The original data set contained 42,653 files; however, the team manually applied the disposition rules, leaving 12,369 files to be categorized – this was further skimmed down to 8784 files due to the format allowing for simple text extraction. NSWSAR used two machine-learning classification algorithms: Multinomial Naïve Bayes, which is a statistical model algorithm versus the Multi-Layer Perceptron, which is a form of deep learning network. The Multi-Layer Perceptron had a 84% success rate; however, human-level accuracy was not assessed in this study and cannot be compared.

The NAA and DoF initiatives have yet to develop into computational trials, they have provided conceptual models.

The first two case studies demostrate that AI tools have not yet reached a point of being able to work efficiently on an unstructured data set, and instead require a decent amount of filtering before accurately performing a task. Much of that filtering presents itself in manual, human intervention. It is important to note that the success rates of human-completed entry may be valuable as a measure of the AI’s success.

Silva, Paulo, Carolina Goncalves, Carolina Godinho, Nuno Antunes, and Marilia Curado. 2020. “Using NLP and Machine Learning to Detect Data Privacy Violations.” In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 972–77. Toronto, ON, Canada: IEEE. https://doi.org/10.1109/INFOCOMWKSHPS50562.2020.9162683

Introduction:

Authors introduce Natural Language Processing (NLP) and Named Entity Recognition (NER) as tools for monitoring and detecting PII. They test three NLP tools and note a 90% F1 score in the best cases for their models.

Background:

The authors offer an introductory overview of how NLP tools process text, how an NER (sub-task of NLP) finds and classifies entities into defined categories and how the performance of tools may differ within other environments. Authors provide additional examples of NLP tools and such tool's use of supervised learning to train machine learning models. For instance, Stanford CoreNLP uses *Conditional Random Fields* (CRF) which "perform segmentation and labeling of sequential data."

Related Work:

Silva et. al describe the use of NER in different contexts and experiments such as an NER system for biomedical data. They establish the lack of NER use for PII and posit that it is an "adequate Privacy Enhancing Technology when applied in privacy preserving data analysis."

NLP Tools:

Authors provide an overview of and outline the benefits of three NLP tools: Natural Language Toolkit (NLTK), Stanford Core NLP, and ExplosionAI's spaCy.

Experimental Approach and Data:

Authors tested the three NLP tools they described and trained NER models using different datasets. They begin by using generic data, then publicly available "contracts" that contained PII and other "mixed datasets" that combined generic data with context specific data (such as U.S. voter registration data). The datasets are divided into a 70/30 split for training and validation with the larger split (70%) going to training.

Experimental Results:

Within the generic datasets spaCy yields the best F1- score with Stanford CoreNLP only tailing by a small margin. The models when turned on publicly available data resulted in again similar results between Stanford and spaCy with approximately a 0.90 F1 score with Stanford performing marginally better. In regards to model training time, the lengthiest session was equivalent to about 100 minutes with each successive model training more quickly. The authors note however that labeling datasets took approximately 4.75 hours per a document resulting in 20 hours total per model to identify entities such as person, city, title, employment details and others.

Discussion:

The authors note a lack of available PII containing datasets to train models and the significant amount of time needed to label datasets. They suggest the use of a synthetic data generator and also utilizing some sort of online annotation tool. They also mention that they only used 68 percent of the PII entities they identified during the labeling process meaning there was a limit to the types of PII their models identified. Silva et. al also weigh the application scenarios of these models and suggest they would work well for data validation (ensuring PII is not erroneously submitted in text fields), PII discovery (alerting administrators that PII is contained within a dataset or document), permission checking (if the actual data matches the description of allowed permissions) and compliance purposes (such as if sharing a document would run afoul of GDPR). Interestingly they note that such NER tools could be used to unlawfully search for PII in data repositories for exploitation purposes.

Annotation:

Silva et. al provide a helpful overview of NERs, their place within the broader field of NLP and the process of creating an effective NER tool with already available models. In particular their difficulty in finding available datasets that can effectively train an NER and the work needed to label such a dataset hints at the difficulty of creating a single comprehensive and broadly applicable NER tool. The authors stop short of suggesting what to do with documents that have PII and how these could be anonymized to share with external parties.

Song, Liwei, and Prateek Mittal. 2020. “Systematic Evaluation of Privacy Risks of Machine Learning Models.” *arXiv*. http://arxiv.org/abs/2003.10595

Objective – This paper explores the ways in which attackers could derive private information from a membership inference attack, in which an attacker attempts to guess an input used as training data for an ML algorithm that could contain private information from say, a participant in a hospital study using the algorithm in question. They then propose granting members of a training data set a privacy risk score based on the probability of them being singled out by an attacker as an outlier ripe for picking through their privacy data.

Research Strategy/Design – Quantitative Study using Models

Method – Using a dataset from a shopping contest, researchers trained an ML algorithm to comb through and process the data as normal. Using a specially weighted predictive attack program designed around Membership Inference, researchers evaluated the output of the attack program against their own predictive model based on its effectiveness. They gave members in the data set a privacy risk score and compared the attacking algorithms chosen victims to their given privacy risk score to see if the score itself was an accurate predictor of chosen targets for membership inference attacks

Main Results – The study showed that the parameters used to apply a privacy risk score were accurate enough to predict whether a given member of a data set was a suitable target for a Membership Inference Attack. The score itself proved highly useful in single out members for further obfuscation of these attacks

Discussion/Conclusion of Article – The author’s demonstrate the effectiveness with which processing and rating training for potential privacy risks before being used as training data for an ML could help to prevent serious privacy breaches from malicious actors. They recommend their improved privacy risk metric to be used and improved upon for future iterations of ML data models going forward.

Sovova, Olga, Miroslav Sova, and Zdenek Fiala. 2017. "Privacy Protection and E-Document Management in Public Administration." *Juridical Tribune Journal = Tribuna Juridica* 7 (2). https://www.proquest.com/scholarly-journals/privacy-protection-e-document-management-public/docview/1989835511/se-2

Objective - The paper examines legal issues regarding the right to informational self-determination and privacy protection in the digital information exchange between the public and private sectors.

Research Strategy/Design - Comparative qualitative study

Setting - The e-government in the Czech Republic

Method – After a brief analysis of legislation, the authors introduce the ways digital records with PI (both born-digital and digitized) are created and used in the Czech Republic's e-government. It offers explanatory case studies of two instances (data mailboxes and digitization of public universities), in which the public interacts with the private sector.

Main Results – The data mailboxes offer both legal and technical advantages despite its risks. It protects PII and records' authenticity. In a public university, digitization significantly improved business productivity while reducing costs. The automated document workflow has tremendous benefits, improving data quality and accuracy as human error can be removed.

Discussion/Conclusion of Article – The authors conclude that technology should be used only for legal interference with the right for informational self-determination. They also argue that automated digitization and digital records can protect privacy and ensure the authenticity of records.

Annotation – The article is written from a records management perspective rather than an archival perspective. It calls for AI to be involved in records management in order to protect records' authenticity and integrity. It also argues that automation can protect the privacy of individuals represented in records. However, the article discusses neither the type of privacy protection models involved nor the detailed automation processes. This is perhaps because the records in the case studies' settings are used for the records' primary purposes.

Tamper, Minna, Arttu Oksanen, Jouni Tuominen, Eero Hyvönen, and Aki Hietanen. 2018. "Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens." In *Law via the Internet: Knowledge of the Law in the Big Data Age*, Florence, Italy. https://seco.cs.aalto.fi/publications/2018/anonymizationservice-finnish.pdf

The article discusses the implication of GDPR on Finnish court documents and the need to anonymize the documents. Given the ineffectiveness of the manual anonymization process, the article proposes an automatic anonymization tool. The tool consists of two different software components – a web service and a user interface. The web service utilizes various natural language processing tools to identify named entities in the text and tag them with metadata. The user interface is a web-based WYSIWYG editor, which allows users to edit the changes made in the web service. At the time of publication, the tool was still in its early stages of development, and the authors published a follow-up article titled ANOPPI: A Pseudonymization Service for Finnish Court Documents.

Wu, Zongda, Jian Xie, Xinze Lian, and Jun Pan. 2019. "A Privacy Protection Approach for XML-Based Archives Management in a Cloud Environment." *The Electronic Library* 37 (6): 970–83. https://doi.org/10.1108/EL-05-2019-0127

The authors in this article point out the increasing need for a broader, cloud-based privacy system for XMl based archives. The article lays out the current state of privacy and DRM protection in current archives and discloses the vulnerabilities that can come for an internal privacy system housed either within an institution's servers or their cloud-based storage system's encryption protocol.

The new approach Wu et al. put forth is one where another server (usually working as a part of an extended microservice) houses both the privacy accreditation mechanisms needed to access a record and the encrypted record itself. The record's metadata is also encrypted so web scraping and other search-based hacking cannot access it without proper accreditation. The rest of the article is devoted to laying out the technical specifications of the proposed protocol, down to how memory is allocated and and the byte-sized management of the encryption schema created for such a service.

Yaco, Sonia. 2010. "Balancing Privacy and Access in School Desegregation Collections: A Case Study." *The American Archivist* 73 (2): 637–68. https://doi.org/10.17723/aarc.73.2.h1346156546161m8

Objective – The article aims to examine the tension between privacy and access in archives when the information in records is about ordinary people but at the same time highly political.

Research Strategy/Design – A comparative qualitative research about various archives that hold school desegregation records.

Setting and Sample – Three archives are examined in this article: Library of Virginia, a state library that generally holds records created by government institutions; Old Dominion University, a public university with collections created by a school district; and American Friends Service Committee Archives, a religious organization with records from an outreach project.

Methods – The article first examines the legal frameworks (Virginia Code, FERPA, HIPAA) that can govern the records. It then examines how each archive determines the privacy and access of those represented in school desegregation records and how the archives came to those decisions.

Main Results – The Library of Virginia's collection is open to approved researchers while records with medical records and potentially humiliating information are sealed for 75 years. Researchers must apply for permission and agree not to reveal any PI in their research to access the records. Old Dominion University returns any confidential material found in their desegregation collections to donors instead of sealing or redacting them. The university also required researchers to sign confidentiality and nondisclosure agreement. The university is currently going through digitization of the records, but at the time of the article's publication, the university is likely not to digitize the confidential records. American Friends Service Committee Archives has a uniform access policy. Collections directly not affiliated with the organization, including the desegregation records, are strictly controlled. Researchers must apply directly to archivists, and archivists must review their work before publication.

Discussion/Conclusion of Article – The article contends that there are conflicting laws regarding privacy and access. Therefore, the risk of privacy violation depends on the law and the interpretation of the said law. Archives, therefore, should create a best practices guide, which can inform their decisions.

Annotation - The article thoroughly examines the issue of privacy within archives in the United States. The article demonstrates that some archives practice risk-based models when it comes to privacy, but others do not. They instead provide universal access regardless of the contents. The article is not about digital records, but the archives' approach to privacy may be useful for the purpose of the study.

Search Terms Used:

AI in Archives
Machine Learning and Archives
Privacy and AI
Privacy and Machine Learning
Archives and Privacy Protections

Artificial Intelligence + Archives + Preservation
Risk Processing AI
Privacy Risk Model AI Tools
AI and Privacy Impact Assessment
Legal Ontologies and Taxonomies and AI
TAR, Technology Assisted Review
Jason + Privacy

Anonymization
Pseudonym
Sensitivity

Search Results:

Database	**Search Terms Used**	**Number of Results**
LISTA	SU Privacy AND SU Archive* AND Digital	13
LISTA	archiv* AND digital AND AB "personal information"	43
IEEE	Artificial Intelligence AND Archives AND Preservation	3,186
IEEE	("All Metadata":privacy) AND ("All Metadata":public archive)	34
LISTA	record* AND digital AND SU privacy	101
Jstor	((ab:(privacy) AND ab:(archiv*)) AND (digital))	16
Jstor	(((Personally Identifiable Information) AND ab:(archiv*)) AND (digital))	55
Jstor	((((privac*) AND (records management)) AND (archiv*)) AND (digital))	3105
Jstor	(("personally identifiable information") AND (archiv*) AND (digital))	131
Jstor	(("personal information") AND ("digital archives"))	44
Jstor	((privacy) AND ("digital archives") AND (authenticity))	68
Jstor	(((("sensitive information") AND (archiv*)) AND (digital)) AND (preservation))	132
Jstor	((("private information") AND (archiv*)) AND (digital)	257

	AND (preserv*))	
Jstor	((digitization) AND (privacy) AND (archival institution) AND (record))	1388
Jstor	((digitization) AND ("personally identifiable information") AND (archives) AND (record))	12
Jstor	((("digital records") AND (("personal information") OR ("personally identifiable information")) AND (archives))	372
Jstor	((archives) AND (access) AND (privacy) AND (protect*) AND (record*))	8443
ProQuest	su(privacy) AND (preservation) AND su(archives)	179
ProQuest	SU(privacy protection) AND SU(archiv*)	90
ProQuest	(privacy) AND (authenticity) AND (archiv*) AND (protect*) AND ("digital record")	294
Taylor & Francis	[Keywords: privacy] AND [All: archiv*] AND [All: digit*]	172
LISTA	risk-based AND privacy AND records	4
ACM Digital Library	[All: archiv*] AND [All: digit*] AND [All: priva*]	10453
ACM Digital Library	[Keywords: archiv*] AND [All: digit*] AND [Keywords: priva*}	9
ACM Digital Library	[All: archiv*] AND [All: digit*] AND [All: privac*]	5774
ACM Digital	[Keywords: record*] AND [All: "personal information"]	45

Library	AND [All: preservation]	
ACM Digital Library	[All: "personally identifiable information"] AND [All: archiv*] AND [All: "records management"]	5
ACM Digital Library	[All: "personally identifiable information"] AND [All: archiv*] AND [All: records]	134
ACM Digital Library	[[All: "personally identifiable information"] OR [All: "personal information"] OR [All: "pii"] OR [All: "pi"]] AND [All: "records management"] AND [[All: "artificial intelligence"] OR [All: "ai"] OR [All: "machine learning"]]	13

Jstor	((("artificial intelligence") AND (archive)) AND (preservat*))	446
Jstor	((Risk process*) AND (Artificial intelligence))	24770
Jstor	(((Risk process*) AND (Artificial intelligence)) AND (privacy))	3879
Jstor	((("risk assessment") AND (privacy)) AND ("artificial intelligence"))	263
Jstor	(("artificial intelligence") AND ("privacy impact assessment"))	4
ACM Digital Library	[All: "risk process"] AND [All: "artificial intelligence"] AND [All: privacy]	2
ACM Digital Library	[All: privacy risk] AND [All: "artificial intelligence"] AND [All: "tool kit"]	43

ACM Digital Library	[All: “privacy risk”] AND [All: "artificial intelligence"] AND [All: "tool kit"]	2
ProQuest	("privacy risk model") AND (artificial intelligence) AND (tool kit)	2
IEEE	("All Metadata":privacy) AND ("All Metadata":violation) AND ("All Metadata":machine learning)	50
IEEE	("All Metadata":privacy) AND ("All Metadata":law) AND ("All Metadata":violation) AND ("All Metadata":artificial intelligence)	7
LISS/LISTA	technology assisted review OR predictive coding AND archiv*	11
LISS/LISTA	e-discovery AND privacy	11
ACM Digital Library	[All: "technology assisted review"] AND [All: personal information] AND [All: archiv*]	12
ACM Digital Library	[All: "technology assisted review"] AND [Abstract: archiv*]	3
ACM Digital Library	[All: "technology assisted review"] AND [All: sensitiv*]	18
ProQuest	e*discovery AND record* AND privacy	37
ProQuest	"technology assisted review" AND privacy	9
ProQuest	"technology assisted review" AND ((personal information) OR (personally identifiable information))	16
ProQuest	"technology assisted review" AND sensitivity	8

ACM Digital Library	[All: anonymiz*] AND [All: sensitiv*] AND [All: court] AND [All: document] AND [Keywords: "machine learning"]	6
ACM Digital Library	[All: anonymiz*] AND [All: sensitiv*] AND [All: court] AND [Keywords: "machine learning"]	13
ACM Digital Library	[All: anonymiz*] AND [All: sensitivity] AND [All: e-discovery] AND [Keywords: "machine learning"]	124
ACM Digital Library	[Keywords: "technology assisted review"] AND [All: sensitiv*] AND [All: classification]	6
IEEE	("All Metadata":sensitiv*) AND ("All Metadata":anonymiz*)	474
IEEE	("All Metadata":privacy) AND ("All Metadata":"e discovery")	7
IEEE	("All Metadata":sensitiv*) AND ("All Metadata":records) AND ("All Metadata":anonymiz*)	90
IEEE	("All Metadata":sensitiv*) AND ("All Metadata":record*) AND ("All Metadata":public)	307