Protecting Privacy in Digital Records:
An Exploration of the Potential of Privacy Enhancing Technologies

Vicki Lemieux and John Werner

Abu Dhabi Symposium

February 20, 2023

Presentation Outline

1. Introduction to the Study
2. Current Approaches for Identifying and Protecting PII
3. Potential for Privacy Enhancing Technologies (PETS)
4. Survey Findings
5. Key Discussion Points

InterPARES TrustAI

1. Introduction to The Study

PII in Archival Documents

- Personally Identifiable Information (PII); Personal Health Information (PHI)
- Responsibility to provide access while protecting PII
  - Legal Obligations
- Current review techniques: slow, resource intensive and ineffective
- Potential for harm from PII: institutional mistrust and restricted archival access

InterPARES TrustAI

"Privacy in Digital Records Containing Personally Identifiable Information (PII): An Exploration of Current Status and the potential of AI Tools and Techniques" (RA06)

How could AI tools address the PII challenge?

What are the implications of using such tools in archival practice?

Researchers: Gabriela Andaur, Victoria Lemieux, Darra Hofman, Georg Geanser

InterPARES TrustAI

Rationale of Our Review

- Privacy Enhancing Technologies (PETS) as helping to address the PII in digital records challenge
- Evaluate their potential use within archives
- Can they increase accessibility of archival documents while maintaining privacy?

InterPARES TrustAI

For the next 20 minutes…

PETS

PETS = Privacy Enhancing Technologies

InterPARES TrustAI

For the next 20 minutes…

PETS

PETS = Privacy Enhancing Technologies

InterPARES TrustAI

2. Current Approaches for Identifying and Protecting PII

Manual and Risk-based approach

- Folder-level review, sampling for sensitivity
- Relies on human reviewers, not scalable
- Results in restriction of entire folder

InterPARES TrustAI

Other Automated Approaches

- Information Retrieval – use of expressions to search for sensitive strings of text

- "costly, time consuming and fragile" (McDonald 2019, p. 43)

InterPARES TrustAI

AI Based Approaches

- Machine Learning (ML) and Natural Language Processing tools
- Topic Modeling (Hutchinson 2017)
- Named Entity Recognition
- Work by Franks, Support Vector Machine and other supervised ML (2022)
- Recent approaches (e.g., US Gov FOIA requests) employ multiple ML techniques and a modular approach to achieve improved, but not perfect results (Mitre, 2023)

InterPARES TrustAI

Current Approaches for Identifying and Protecting PII DO NOT focus on . . .

- PII in models used for ML/AL/NLP identification of PII in archival documents

Current solutions focus exclusively on the PII in archival documents, but protecting models from leaking PII also needs attention

InterPARES TrustAI

3. Potential of Privacy Enhancing Technologies (PETS)

Privacy Enhancing Technologies

- Also known as privacy preserving technologies, privacy-enhanced computation or privacy enablers
- Allow useful search or analytical results that enable data use without revealing its content
- Protect data in use or processing– including model training

InterPARES TrustAI

Current PETS Applications

- Used primarily in health and financial sectors, enable use of sensitive data while maintaining privacy
- Currently few uses in archives
- PETS assume inherent sensitivity, applicable to large archival holdings where manual review of content is difficult

InterPARES TrustAI

Scoping Review on PETS

- Searched Proceedings on Privacy Enhancing Technologies (PoPETS), ACM Surveys, and IEEE databases
- Compile list of available PETS and examples of their application
- Capture "scope" of available literature versus comprehensive survey

InterPARES TrustAI

4. Survey Findings

Homomorphic Encryption (HE)

- Encryption method that allows for computable functions over encrypted data without altering data's form or characteristics
- Partially, Somewhat and Fully HE
- First Fully Homomorphic Scheme introduced in 2009
- Can be used to encrypt something like sensitive medical data and then used to train a ML model

InterPARES TrustAI

Trusted Execution Environment (TEE)

- Function by creating a trusted processing enclave within a computer–protects data in use within the TEE
- Most Common model: Intel's SGX, some by other hardware manufacturers like AMD
- Vulnerable to side channel attacks, difficult to predict and prevention focused on known attacks
- Typically paired with other PETS

InterPARES TrustAI

Secure Multiparty Computation (MPC)

- Simply defined: a distributed computing task executed in a secure manner
- Private inputs ☐ concealed computation ☐ correct output
- Variety of tools and protocols: Secret Sharing, Oblivious Transfer Protocol, Yao's Garbled Circuits, Private Set Intersections; all designed to obfuscate inputs and offer accurate output
- Applications in Law Enforcement, Medical Image Sharing, distributed ML

InterPARES TrustAI

Differential Privacy (DP)

- Adds a controlled amount of randomness (i.e. "noise") to a dataset
- Prevents reconstruction of data by external parties, hides whether an individual is in or out of a dataset
- Created as a response to k-anonymous datasets that leaked information about individuals; proposed in 2006
- Highly varied in its adoption and techniques – over 200 identified in Desfontaines and Pejó's survey (2020)
- Used in medical fields, unstructured data (video and audio), Internet of Vehicles, and any other large, public datasets

InterPARES TrustAI

Personal Data Stores

- "systems that provide individuals with access and control over data about them, so that they can decide what information they want to share and with whom . . ." (Royal Society, 2019)
- Closely linked to self-sovereign identity (SSI) solutions executed through a consumer facing application
- Individuals then use verifiable credentials as "proof" to any sort of identity inquiries like date-of-birth, asset ownership, etc.
- Derived from Zero Knowledge Proofs

InterPARES TrustAI

Privacy Preserving Machine Learning (PPML)

- Leverages PETS to protect privacy in ML models
  – HE, MPC, DP
- ML models contain essential information about the training set which is often sensitive
- Federated Learning – frequently PPML adjacent but needs PETS to protect information
- Significant use in medical field given sensitive and siloed data

InterPARES TrustAI

Synthetic Data

- Artificially generated data made for public access
- Generated algorithmically (by a model) to simulate real-world events or transactions
- Include techniques like: Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs), Variational Auto-encoders (VAEs)

InterPARES TrustAI

Hybrid Use of PETS

- PETS are frequently used together to provide a broad range of protection
- Application of each depends on the context of the data, participants, expected outcome and so forth

InterPARES TrustAI

5. Discussion Points

PETS still represent an emerging class of technologies yet offer significant potential in addressing the PII problem

Default to broadly protecting information in archival holdings with known presence of PII

InterPARES TrustAI

With a wide array of PETS available organizations can identify which ones address their needs

Helpful in creating collaborative training structures where each party wants to ensure privacy of their data

Several of the techniques can be used to ensure that ML and AI models do not leak PII or protect them from privacy attacks

Personal Data Stores to support "participatory access to archives."

Individuals allow access to records, either jointly with archives or independently

Such access could support reconciliation efforts while preserving record subject's autonomy

InterPARES TrustAI

Synthetic Data represents a valuable tool to develop and test the effectiveness of PETS

Synthetic Data can still possibly leak information; tradeoff between privacy and inferential utility of data generated

InterPARES TrustAI

We are submitting a paper to ACM's Journal of Computing and Cultural Heritage. We welcome any feedback you can provide.

THANK YOU

InterPARES TrustAI

References

Desfontaines, D., & Pejó, B. (2020). SoK: Differential privacies. Proceedings on Privacy Enhancing Technologies. https://petsymposium.org/popets/2020/popets-2020-0028.php

Franks, J. (2022). Text Classification for Records Management. Journal on Computing and Cultural Heritage (JOCCH), 15(3), 1-19.

Hutchinson, T. (2017). Protecting privacy in the archives: Preliminary explorations of topic modeling for born-digital collections. 2017 IEEE International Conference on Big Data (Big Data), 2251–2255. https://doi.org/10.1109/BigData.2017.8258177

McDonald, G. (2019). A framework for technology-assisted sensitivity review: Using sensitivity classification to prioritise documents for review [PhD, University of Glasgow]. https://eleanor.lib.gla.ac.uk/record=b3341257

mitre.org/news-insights/impact-story/mitre-tool-simplifies-freedom-information-act-requests

The Royal Society. (2019). Protecting privacy in practice: The current use, development and limits of Privacy Enhancing Technologies in data analysis. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Protecting-privacy-in-practice.pdf?la=en-GB&hash=48A28CDF4FB012663652BE671CFFED08

Sloyan, V. (2016). Born-digital archives at the Wellcome Library: Appraisal and sensitivity review of two hard drives. Archives and Records, 37(1), 20–36. https://doi.org/10.1080/23257962.2016.1144504

InterPARES TrustAI