IHE Work Item Proposal Update De-Identification Handbook

# Proposed Work Item: Update De-Identification Handbook

Proposal Editor: Soyean Kim, CIDGOH, Kyoohyung Han, Hyojin Yoon, Samsung SDS, John Moehrke, By Light Professional IT services, co-chair Security WG HL7

Work item Editor: Soyean Kim, CIDGOH, Kay Han

Date: November 13, 2024

Version: <current version number of proposal>

Domain: De-identification, Privacy Enhancing Technologies

# The Problem

<Summarize the integration problem. What doesn’t work, or what needs to work?>

Post pandemic landscape with a significant acceleration towards digital health introduced emerging threats and evolving technology landscape for privacy best practices. With new health care data modalities (i.e. genomic) and new use cases involving new patterns of information flow (i.e. data linking involving multiple parties), implementers are tasked with decision making for responsible data sharing and use(s) where common best practices for de-identification do not yet exist.

Privacy-enhancing technologies (PETs) are based on key concepts of the information flow and can be broken down into various stages of the data life cycle. For example, Input privacy refers to the ability to process information that is hidden from you and to allow others to process your information without revealing it to them. Output privacy allows you to receive/read the output of an information flow without being able to infer further information about the input and, symmetrically, to contribute to the input of an information flow without worrying that the later output could be reverse engineered to learn about your input.

Initially published in 2014, the IHE de-identification handbook aims to provide guidance for removing individually identifiable information from healthcare data. This includes de-identification, pseudonymization, re-linking, design considerations, techniques, and risks. The intended audience is IHE Profile editors and 105 healthcare information technology implementers needing a guide for designing and implementing de-identification systems.

Targeted opportunities exist to introduce new use cases and supplement the IHE de-identification handbook as a minor version update. For example, it is currently difficult to define (i.e. threat model) and address privacy challenges for collaborative analysis involving patient data sets securely (i.e. data linking) .

The first step of collaborative analysis is data matching, which identifies rows from the same individual or device. We can apply a salted hash algorithm to Personally Identifiable Information (PII) for each dataset. However, the salt value must be kept confidential. Thus, two organizations cannot use the same salt for hashing without the assistance of a trusted third party, such as a government agency. In other words, it is impossible to perform a collaborative analysis while adhering to the guidelines.

The proposed update (minor version update) to the de-identification handbook is to review algorithms and privacy enhancing technologies and current practices. The proposal will introduce a use case of collaborative analysis using genomic data for secondary use (involving data linking) and the type of PET used for cryptographic de-identification.

The focus will be on private data matching between two datasets from two different organizations (parties) using cryptographic PETs. By applying those algorithms and technologies to the data category "Person identifying direct identifiers" in Table 3-1, either or both organizations can obtain data linking information without revealing any other personal information where PII introduces substantial challenges and perceived risk is high (sharing with a 3rd party in semi trusted or untrusted settings).

<Describe the Value Statement: What is the underlying cost incurred by the problem and what is to be gained by solving it? If possible provide quantifiable costs, or data to demonstrate the scale of the problem.>

With new health care data modalities (i.e. genomic) and new use cases involving new patterns of information flow (i.e. data linking involving multiple parties), implementers are tasked with decision making for responsible data sharing and use(s) where common best practices for de-identification do not yet exist. Use cases include but not limited to: Infectious diseases management, multi jurisdictional studies, cross border R&D and data sharing via linking (hospital to hospitals, public health lab to hospitals,etc.). Opportunities to broaden the scope of data sharing in semi trusted and/or untrusted settings based on updated threat models, attacker goals.

# Key Use Case

An example of technical workflow for tokenization from cryptographic PETs is shown below. As private key is not revealed to each other, re-identification risk is mitigated. Furthermore, two parties can collaboratively analyze their datasets using the results. In addition, we are also able to use a protocol called Circuit-PSI to exclude any information except data analysis results.

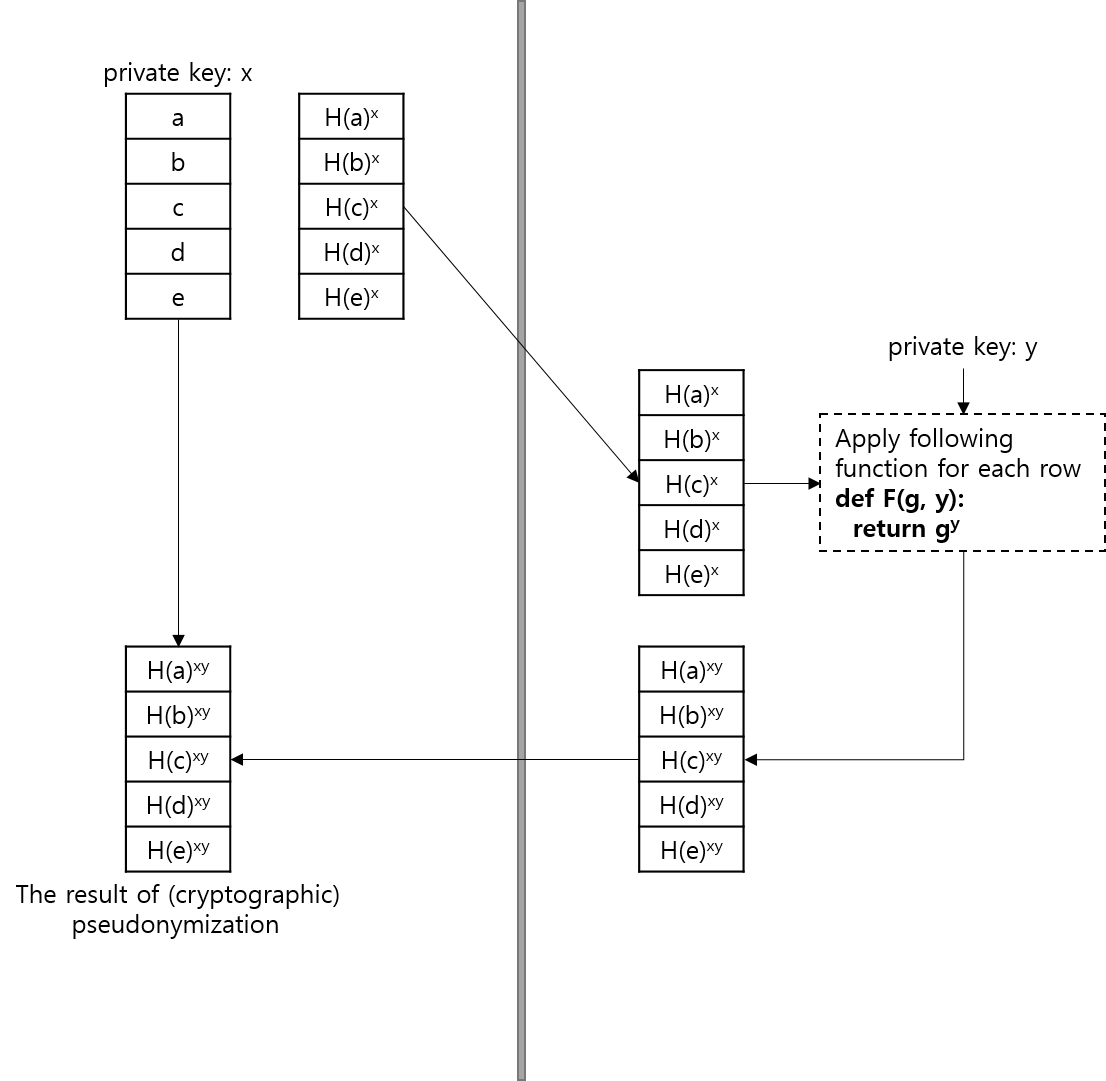


Figure 1. Cryptographic Tokenization

An example of workflow for FHIR mapping from MPC module is shown below. This workflow leverages both FHIR server and MPC modules. They can be combined to increase interoperability of health systems across multi-jurisdictional genomic data sets, which might contain sensitive identifiable data.

A diagram of a computer hardware software

Description automatically generated

Figure 2. Workflow for FHIR mapping from MPC module

An example of genomic use case workflow is shown below. This workflow illustrates how genomic reporting as a document containing antimicrobial resistant information can be stored and accessed via cryptographic de-identification by a client.

A diagram of a medical procedure

Description automatically generated

Figure 3. Workflow for Genomic report access

<Describe a short use case scenario from the user perspective. The use case should demonstrate the integration/workflow problem. Feel free to add a second use case scenario demonstrating how it “should” work. Try to indicate the people/systems, the tasks they are doing, the information they need, and where the information should come from.>

# Standards & Systems

<List existing systems that are/could be involved in the problem/solution.>

Relevant standards can be considered such as FHIR Bulk API, MolSeq, CQL, GA4GH, and/or HIE. The de-identification update may be supplemented by open source, reference implementation (with simplified version of the cryptographic tokenization for reference purposes only). The reference implementation will be tested based on the proposed use cases and will be available in the IHE International GitHub.

https://github.com/IHE

<https://f1000research.com/posters/12-893>

<If known, list specific components of standards which might be relevant to the solution.>

# Discussion

<If possible, indicate why IHE would be a good venue to solve the problem and what you think IHE should do to solve it.>

Benefit of standardizing privacy operation for emerging use cases

Establish community of practices for better governance

De-identification guide needs routine update

Keep up with others (NIST, ISO – all working on PETs, OECD, UNSD,etc.)

Needs for collaborative analysis is growing as health care is becoming more connected (IoT for home monitoring, genomics, specialized vendor products,etc.)

Current data linking practices have issues (sharing PII, sharing salt,etc.) and risks people might not be aware

<A one page proposal is preferred. Please do not exceed two pages.>