DC.3 Connecting PEP and DRE

Project	PRANA-DATA
Project leader	Wessel Kraaij (TNO)
Work package	
Deliverable number	DC.3
Authors	Sietse Ringers (RU)
Reviewers	
Date	May 29, 2017
Version	2
Access Rights	Public
Status	

PRANA-DATA Partners:

Portavita, TNO, Radboud Universiteit Nijmegen, Maastricht UMC+, UMCG

COMMIT/ COMMIT is a public-private research community solving grand challenges in information and communication science shaping tomorrow's society

COMMIT/ A public- private research community

This publication was supported by the Dutch national program COMMIT.

Summary

This document describes an experimental coupling between two state-of-the-art technologies for handling sensitive and personal data of research participants: PEP (Polymorphic encryption and pseudonymisation), which uses advanced encryption with distributed pseudonymisation to provide privacy friendly exchange of medical data for specific medical research purposes, and DRE (Digital Research Environment), a centralized online environment for researchers in which they can analyze, manage and safely and responsively share research data with other researchers. The coupling between these two systems allows a researcher to safely export data from PEP to DRE.

Contents

Summary 1

1 Introduction 3

2 PEP 3

3 DRE 4

4 Coupling PEP and DRE 4

5 Conclusion 5

6 Bibliography 6

1 Introduction

Because of the increasing importance of personal data in our society, when a research project requires handling personal information of research participants it is very important to handle the data responsibly, and carefully protect the privacy of the research participant as much as possible. When the appropriate protection measures are not taken, sensitive data may be accessed or manipulated by unauthorized agents, or it may leak entirely. This clearly creates a moral obligation to handle as little personal data as carefully as possible, combining with the European privacy law [1] specifying the same thing. In addition, as the awareness of these risks increases, research participants might be unwilling to participate at all if they are not ensured that their data is handled carefully and responsibly.

In the context of the "Parkinson op Maat" project [2], a large scale Parkison's Disease research project at the RadboudUMC, researchers at the Radboud University together with Verily [3] are developing PEP [4]: a system that uses advanced cryptographic techniques to protect research participant data as much as possible while it is being sent, manipulated and stored. Simultaneously, a centralized online environment for researchers called DRE [5], for Digital Research Environment, is being developed for the RadboudUMC, in which researchers can analyze, manage and safely and responsively share research data with other researchers. It makes perfect sense, then, to couple these two systems, so that research participant data is protected all the way from its creation until it is used by researchers. After introducing the PEP and DRE systems in more detail, we will describe how this coupling has been achieved.

2 PEP

The PEP system aims to endow all participants of the "Parkinson op Maat" study with just enough capabilities for them to perform their role, and no more. For example, doctors or nurses need to be able to input patient data into the system, but they need not be able to view or modify data from other participants, while researchers may need read access to all medical data but do not need to know the identities of the data subject. It is important to distinguish three levels of identification of data subjects:

- Some data can completely identify the data subject (e.g., a name);
- The data can be pseudonymized: two data records can be linked together as coming from the same data subject, but the identity of the data subject remains unknown (for example, two data records may belong to patient 7486613);
- The data can be anonymized: multiple data records cannot be linked as coming from the same data subject. In this case one says that the data records are unlinkable.

For each data record PEP internally keeps track of from which data subject it originated, but depending on the need of the data user or handler, it can ensure all three levels of identification.

The PEP system consists of the following main components:

- A storage facility, which handles only encrypted and pseudonymized data;
- An access manager, which decides who gets to see what data at which identification level;
- When data is sent from one place to another it passes through the transcryptor, which can modify the data records such that firstly, only the recipient is able to decrypt the data, and

secondly that they become unlinkable, pseudonymized or completely identifiable to the receiver, as determined by the access manager. From its own perspective all data records are fully encrypted as well as unlinkable.

Based on the partially homomorphic ElGamal encryption scheme [6], the data is always encrypted while within PEP. The access manager never gets to see the data itself; the storage facility only in pseudonymized form; and the transcryptor, which plays a critical role in the system, only in unlinkable form. Simultaneously, the system is maximally flexible due to the fact that the data can be made decryptable for certain recipients by the transcryptor as needed *after* it has been encrypted. This enables secure data flows between data actors that can be much more complex and flexible than traditional encryption could offer.

The PEP-software is at the time of writing still in development, although all key components already present and operational. It is expected to be operational by September 2017.

3 DRE

The Digital Research Environment (DRE) is a centralized online environment for researchers in which they can import, store, share, and manipulate datasets, as well as study them by integrated statistical and analytical software.

A key principle of the system is that after the data has entered DRE there is generally no need for it to be extracted from it again, and indeed DRE purposefully offers no functionality for this and even discourages it. Instead, by providing integrated statistical and analytical software, everything that a researcher would want to do with the data can be done safely within DRE itself. This central nature reduces the risk of data loss through of theft, vulnerabilities or carelessness. Additionally, it is easier to achieve compliance with (privacy) law and regulations, as there always is only one site at which this needs to be done.

Data within DRE is organized in workspaces. The owner of a workspace decides who has access to the contained data, and can share the data of a workspace with another researcher within the DRE system by giving her access to the workspace. The action of sharing data thus becomes very fast and easy, compliant, and simultaneously safe, as the data never leaves DRE.

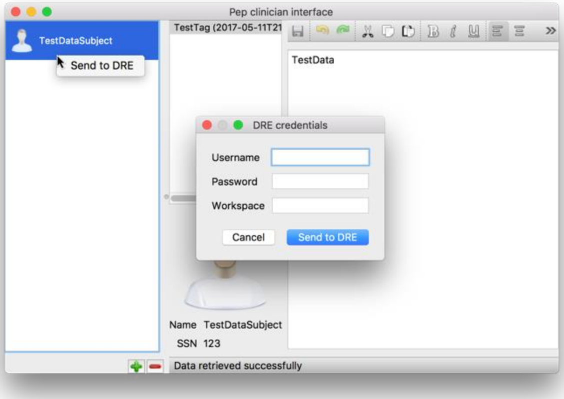
DRE has been operational since May 2017.

4 Coupling PEP and DRE

It is clear that in the interest of the security of the data and the privacy of the data subjects, DRE would be an ideal endpoint for data coming out of the PEP-system. Since PEP is not yet fully operational the connection can at this point be no more than a proof of concept, although its technical implementation has been fully realized.

The idea is that after the researcher has gone through the PEP-system to obtain the (pseudonymized) data she has access to, she can activate the PEP-DRE connection to export the data to DRE. The details are as follows:

- The researcher logs into the researcher-facing application of PEP, and selects the (pseudonymized) data she wants to export to DRE. The data is fetched from the storage facility, and the transcryptor pseudonymizes the data as well as modifying it such that it becomes decryptable for the researcher. The researcher application decrypts the data.
- The data is sent to DRE over an SFTP connection, using the username and password of the researcher at DRE;
- DRE imports the data, parses and stores it, and exposes it to the researcher for manipulation or analysis.



TestData

Name TestDataSubject
SSN 123

Data retrieved successfully

Figure 1: An experimental version of the researcher client showing the DRE connection dialog

SFTP is a standard and open technology for sending encrypted data from and to authenticated endpoints, and it is the preferred entry point for new data entering DRE. In this case, the open-source `libssh2` library [7] was used.

5 Conclusion

Both PEP and DRE demonstrate that through careful design combined with state-of-the-art cryptographical and other technical measures, it is possible to create systems that combine privacy-protecting measures with security as well as ease of use. In an ever more data-hungry society that is simultaneously becoming more aware of privacy risks, this is no luxury. The goals and functionality of PEP and DRE align to such a degree that a coupling between the two makes a lot of sense, and we believe that the coupling is an important contribution to both systems, leading to a lifecycle for sensitive data that is safe from its inception to its consumption by researchers.

At the moment, the connection between the two systems is one-way: data flows from PEP to DRE. As one of the design principles of DRE is that once data enters the system it should not be able to leave again, this makes sense, but there may be advantages to be gained from a reversed coupling. For example, if two researchers could share their data with each other from DRE through PEP, instead of purely in DRE as it is now, then the PEP system could re-pseudonymize the identifiers of the research participants. More research would be necessary to establish additional possible advantages, and to find ways in which such functionality can be implemented without compromising the security and privacy-guarantees of both systems.

6 Bibliography

[1] "General Data Protection Regulation: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data," [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679. [Accessed 12 June 2017].

[2] "Parkinson op Maat," [Online]. Available: https://www.parkinsonopmaat.nl/. [Accessed 12 June 2017].

[3] Verily, "Investigating Parkinson’s disease in a new way," [Online]. Available: https://blog.verily.com/2016/09/investigating-parkinsons-disease-in-new.html. [Accessed 12 June 2015].

[4] E. Verheul, B. Jacobs, C. Meijer, M. Hildebrandt and J. d. Ruiter, "PEP: Polymorphic Encryption and Pseudonymisation for Personalised Healthcare," *Cryptology ePrint Archive, Report 2016/411,* 2016.

[5] "DRE: Digital Research Environment," 2016. [Online]. Available: https://health-ri.org/news-events/digital-research-environment-service-be-officially-launched-during-health-ri-conference. [Accessed 12 June 2017].

[6] T. ElGamal, "A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms," in *CRYPTO '84*, 1984.

[7] "libssh2: a client-side C library implementing the SSH2 protocol," [Online]. Available: https://www.libssh2.org/. [Accessed 12 June 2017].