Research data beveiligen en delen met behulp van Polymorfe Encryptie en Pseudonimisatie (PEP)

Jean Popma
Interfacultary Hub for Security, Privacy and Data Governance
J.Popma@ru.nl 29-10-2019

About myself

- Technical support manager 1987-2000 at Radboud University
- Managing director of ICT-service centre from 2000 to 2013 at Radboud University
- Corporate Information Security Officer and acting Privacy Officer from 2013-2016 at Radboud University
- Chairman of CERT-RU 2001-2017
- SURF CyberSaveYourself steering committee 2012-2016 https://www.cybersaveyourself.nl/
- Member of the board of the Privacy by design Foundation (working on IRMA, https://privacybydesign.foundation)
- Currently Project Manager Applied Security Research working on PEP https://pep.cs.ru.nl/ at iHub, Radboud University https://www.ru.nl/ihub

Biomedical research & big data … What's the big deal?

Biomedical data are sensitive personal data.

- In case of a data breach:
  - harmful to participants/patients and
  - reputation damage for researchers
  - undermining patient's willingness to participate in future research is a show-stopper for medical research
- Legal requirements are strict
  - More so after GDPR has become fully operational (May 2018)
  - high fines / repercussions
- Ethical restrictions apply
- Professional cooperation between specialists from multiple disciplines (computer science, legal, ethics) and medical scientists is essential

Data Breach – As-A-Symptom

- Authorizations to broad
- Orphaned data
- Stolen or lost devices
- Copies traveling around
- Combination of different sources
- Identity theft
- Ransomware
- Lack of maintenance
- Hidden tracking/profiling
- Changing context
- Too many data collected
- Data reused for unintended purposes
- Fraud
- Revenge
- Spionage
- Wrongly addressed mail
- Forgotten or lost print-output
- Active hackers
- Stupidity

Privacy By Design?

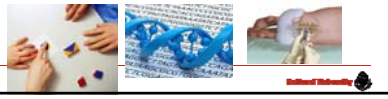
- Data Design Strategies
  - Minimise
  - Separate
  - Aggregate
  - Hide
- Proces Design Strategies
  - Inform
  - Control
  - Enforce
  - Demonstrate

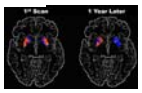
source: Jaap-Henk Hoepman @ https://arxiv.org/pdf/1210.6621.pdf

Personalised Parkinson's Project

- 850 participants suffering from Parkinson's
- Followed over a period of 2 years
- Data will be available to scientists at dutch UMC's and other research institutes (worldwide)
  - Clinical data (questionnaires, tests)
  - Biophysiological data (from blood, plasma, CSF)
  - Wearable sensor data (2 years)
  - ECG
  - fMRI
  - Genome
  - Microbiome

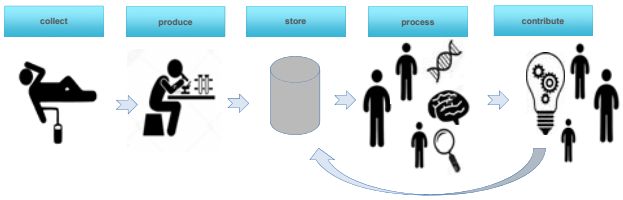
Study Total: > 0.5 PB of data





Research Data Repository

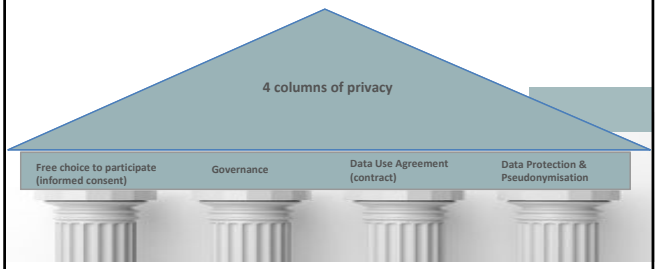
2 main functions : **share** data, and ensuring **scientific integrity**



Challenges

- Protect the privacy and personal data of participants
- Enable the use of data collected for legitimate research
- If data are shared, receiving party obtains controllership over the data
- What about subject's rights?
- What about the obligations of the primary controller?

Building Trust



Data Protection & Pseudonymisation

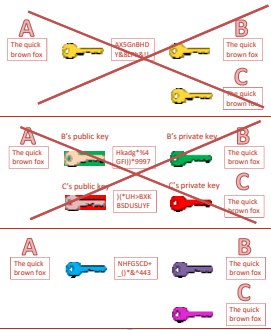
- Research data should be encrypted and pseudonymised
  - Encryption means that data cannot be read unless you have a proper key to unlock them.
  - Pseudonymisation means that all directly identifiying information is removed from the data and replaced by meaningless identifiers.
- Special system built as a data repository for these data: PEP = Polymorphic Encryption en Pseudonymisation (https://pep.cs.ru.nl) . PEP infrastructure takes care of the key- and data management

Encryption

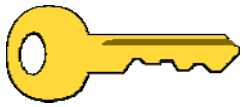
Shared secret

Public/private key

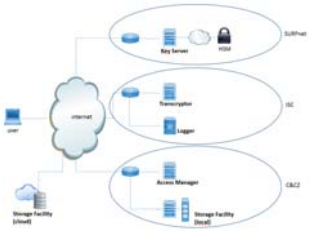
Polymorphic



The key to encryption is the key

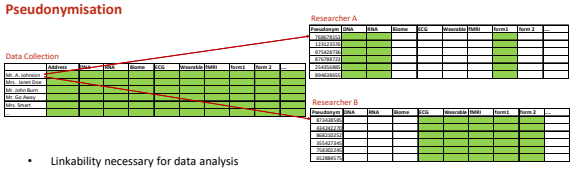


Distributed management of cryptographic keys



- User has client Software (encryption/decryption, up/download)
- 3 trusted parties take care of key management. Developers have no knowledge of keys.
- Keys in Hardware Security Modules (tamperproof)
- Storage Facility can be anything (public cloud)

Pseudonymisation

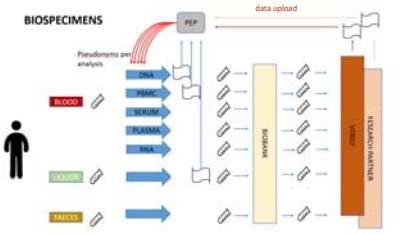


- Linkability necessary for data analysis
- Data should be pseudonymised first (no identifiers, timestamps etc. as part of the data)
- Personalized pseudonyms for each sharing
- Pseudonomous data is not Anonymous data
- Reversibility is a requirement

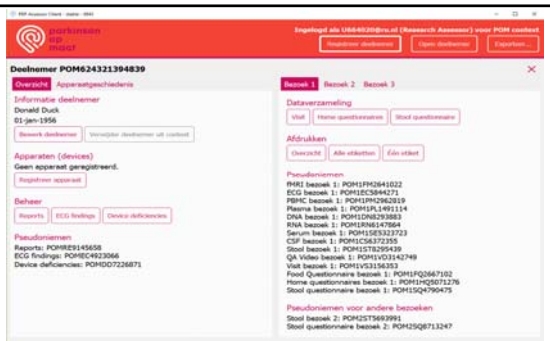
Polymorphic Encryption and Pseudonymisation (PEP) In a Data Repository

- Polymorphic encryption close to the data source
- Unique keys for different data and different users can be generated a posteriori, at the time access is granted.
- Decryption in the target processing environment
- All data streams are based on unique pseudonyms
- Unique and persistent pseudonyms are cryptographically generated for each user (user group) obtaining data from the repository

Pseudonymisation during data collection / analysis phase



- All datastreams based on unique pseudonyms
- When uploaded, these pseudonyms are translated to a polymorphic pseudonym, so data from different sources can be linked.
- Same idea for data from other sources (wearables, MRI, ECG etc.)



Pseudonymisation for data use by researchers

1. Research project **requests** data of certain types regarding subjects that meet certain criteria, based on a research proposal
2. Request is **evaluated** by a review board.
3. In case of positive **decision**:
   - Required columns for specific subjects are **authorized** to User
   - User **requests** data from PEP
4. User gets:
   - **Access** (download) to only these authorized data
   - **Pseudonyms** personalized to user(group),
   - A unique **key** is constructed to decrypt these (and only these) data
5. In the user's authorized processing environment (like DRE):
   - Data are **decrypted** for further processing and analysis
   - Derived data can be encrypted and **uploaded** back to the repository

Scientific Integrity

- Data can not be erased from the repository
- Dat can become invalid at a certain point in time
- Participants can withdraw consent
- Shareable data may change over time
- Historic queries (for reproduction of earlier studies) must be possible

Results

- During the data collection phase samples cannot be linked to the same subject by different labs involved in production of the data.
- During transport (to and from respository) and storage data are protected by encryption at all times
- Key management is distributed over multiple parties: no single party can get access to the data (storage can be anywhere)
- Authorization is fine grained, and pseudonyms are personalized for individual researchers (research groups)
- Historical queries to facilitate replication studies are possible

Privacy By Design revisited

- Data Design Strategies
  - Minimise √
  - Separate √
  - Aggregate X
  - Hide √
- Proces Design Strategies
  - Inform √
  - Control √
  - Enforce √
  - Demonstrate √

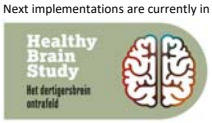
Bron: Jaap-Henk Hoepman @ https://arxiv.org/pdf/1210.621.pdf

Summary

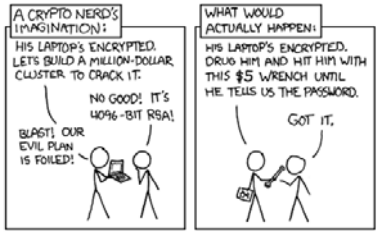
- Privacy risks are not reduced to 0 …
  - Data can be self-identifying (video/photo material, genome data etc.)
  - Conspiring scientists can link data based on content.
  - Enforcement and control of data use agreements is crucial but very difficult to do.
- Technology is not snake oil.
- Big Data, Cloud & Privacy can be combined if Privacy & Security are part of the design, at legal, organizational and technical levels **combined**.

Where do we go from here?

- First PEP pilot should demonstrate useability of the technology.
  It is a proof of concept (not for the Parkinson's project though)
- All software will be published as open source for reasons of transparancy and public value.
- Next implementations are currently in progress



- After pilot phases, a more general PEP system will be built and released
- Transfer of technology to Privacy by Design Foundation for further development and support



More? http://pep.cs.ru.nl