Data Protection: The TMF Approach for Medical Research Networks

Klaus Pommerening

2006 March 16

Overwiew

1. The TMF
2. Some basic principles
3. Pseudonyms
4. Models for pseudonymization
5. Results

Collections of Medical Data

For Health Care Research, Public Health and Epidemiology we would like to …

- Collect medical data from all relevant sources
  - Secondary use of electronic patient records
  - Also biological samples and genetic data
- Link them together via a personal identifier
- Build up disease registries
- Acquire follow-up data
- Evaluate the data in all possible ways
  - not always known precisely at the time of collecting
- Exchange data with other researchers and registries

German Medical Research Networks

- Health care, clinical, epidemiological research
- Research on all aspects of a specific disease
  - Examples: Paediatric Oncology, Chronic Inflammatory Bowel Disease, Rheumatism, AIDS, …
- Features:
  - Multicenter studies, central data management
  - Data pools or registries
  - Material banks (“biobanking”) and genetic data

TMF

"**Telematikplattform für die Medizinischen Forschungsnetze** des BMBF"
(Telematics Platform for the Medical Research Networks of the Research Ministry)

Goal: solve logistic, technical, and administrative problems for the research networks,
help the networks in building their infrastructure and *processing their data according to the data protection rules*.

Basic info in English:
`http://www.tmf-ev.de/site/EN/int/c_homepage.php`

Members of the TMF

- 17 “Competence Networks” in Medicine
- 12 Coordinating Centres for Clinical Trials (KKS)
- 6 Networks for Rare Diseases
- 3 Networks for Infectious Disease Epidemiology
- 6 Other Networks

$\Rightarrow$ 44 networks at present

Overwiew

1. The TMF
2. Some basic principles
3. Pseudonyms
4. Models for pseudonymization
5. Results

The Basic Rule
[of the EU D. P. Directive]

Processing of personal data is strictly forbidden.

*Member States shall **prohibit** the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and **the processing of data concerning health** or sex life*.

[Article 8 (1) of EU D. P. Directive]

Exceptions

There are exceptions (connected by logical OR)
[Article 8 (2) – (5) of EU D. P. Directive]

- Explicit informed consent by the data subject
- Protect vital interests of the data subject
- Medical treatment or prevention (for health care professionals)
- Substantial public interest (requires national law *providing for adequate safeguards*)
- …

Further Restrictions

[…] *personal data must be*: […]

(b) *collected for **specified, explicit and legitimate purposes** and not further processed in a way incompatible with those purposes*. […]

(c) *adequate, relevant and **not excessive** in relation to the purposes for which they are collected and/or further processed;* […]

(e) *kept in a form which permits identification of data subjects for **no longer than is necessary** for the purposes for which the data were collected or for which they are further processed*. […]

[Article 6 (1) of EU D. P. Directive]

⇒ Three Important Principles for Processing of Personal data

1. Specific purpose
2. Minimal possible data set (“Parsimony”)
3. Restricted time

[connected by logical AND]

Consent can be given only for 1 – 3.

Weaker rules may be possible by law or with informed consent AND additional adequate safeguards.

What about Registries or Research Projects?

Several options:

1. Use **anonymous data**
   - D. p. regulations do not apply.
   - Insufficient for many (most?) projects in Public Health or Health Care research.
2. Get **informed consent**
3. **Law**
   - E. g. for cancer registry in most federal states of Germany

Secondary Use of Medical Data from Treatment Context

Another barrier independent from d. p. laws: the professional discretion of physicians

- A physician must not give away any information she has from a particular patient.
- Protected by criminal and civil law as well as professional ethics.
- This imposes limitations on the use and processing of medical data.

However, options 1 – 3 apply as well.

Treatment Context

Barrier: Professional Discretion

Clinical Research
Health Care Research

Direct data capture

→ Export controlled by options 1 – 3 (anonymous data, informed consent, law).

Registry/
Epidemiological Research

Summary: Patients' rights

- Strict confidentiality of health data
- Even with consent only restricted purpose and time.
  - Change of purpose (e. g. secondary use of patient record) needs new consent.
- Obtaining information, revocation

Crucial question for medical research networks, registries, and biobanks:

Can the consent be "somewhat" unspecific?

National Ethics Committee: Yes, if right of revocation is preserved + additional safeguards are provided.

Summary: The German View

- Protect personal data as good as possible.
- Use anonymous or pseudonymous data wherever possible
  - even if it causes some inconvenience.
- Weakening of some d. p. procedures may be possible if additional safeguards are provided (as compensation).

Overwiew

1. The TMF
2. Some basic principles
3. Pseudonyms
4. Models for pseudonymization
5. Results

Anonymous Data

[…] *the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; …*

[Preamble (26) of EU D. P. Directive]

The German D. P. law weakens this to

… *or identifiable with disproportionate effort only*.

Data are considered anonymous when the **re-identification risk** is very low.

Pseudonymous Data

EU D. P. directive: not explicitly mentioned.

German D. P. law: Use pseudonymous data whenever possible, if you cannot attain the purpose with anonymous data.

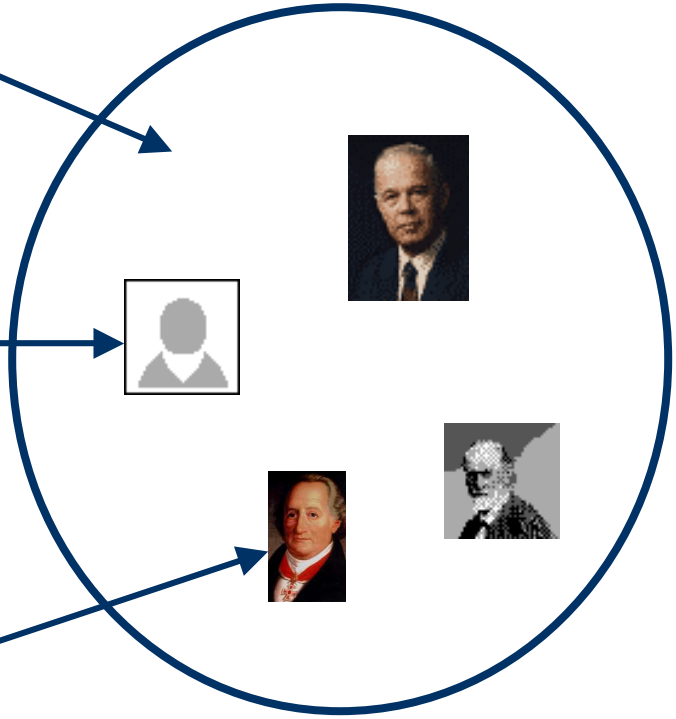
A **pseudonym** is a (random) string or number that **replaces the identity data** in a data record and gives no hint at the corresponding individual.

[See below.]

Anonymity
Reference only to
entire set of individuals

Pseudonymity
Reference to single individual
without revealing his identity

Reference to individual
The identity of the single
individual is revealed



Drawback of Anonymization

- No association between data from distinct sources
- ... or from distinct points of time.
- No way back to the patient for feedback
- ... or for recruiting suitable patients for a new research project.

Pseudonyms

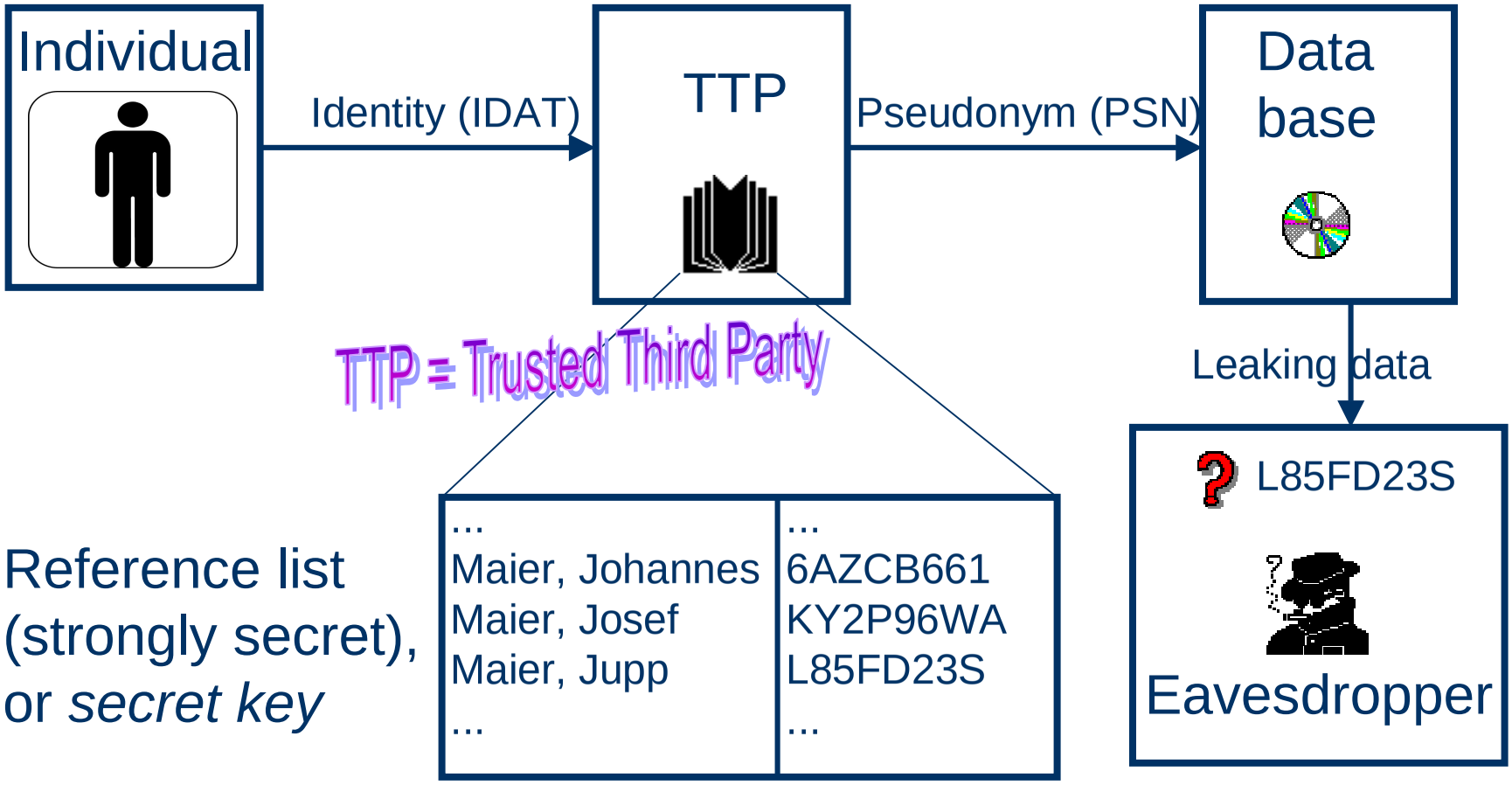
One-way pseudonyms allow

- association between data from distinct sources
- ... or from distinct points of time.

Reversible Pseudonyms additionally allow

- way back to the patient for feedback
- ... or for recruiting suitable patients for a new research project.

TTP-Generated Pseudonyms
(Basic Model)



TTP-Generated Pseudonyms

- The TTP stores a reference list or uses a cryptographic key.
- Only the TTP can reveal the pseudonym (depseudonymization).
  - Consent necessary for use of pseudonyms.
- For the attacker, pseudonymous data look like anonymous data.
  - Identifiable with disproportionate effort only.
  - The re-identification risk should be assessed.

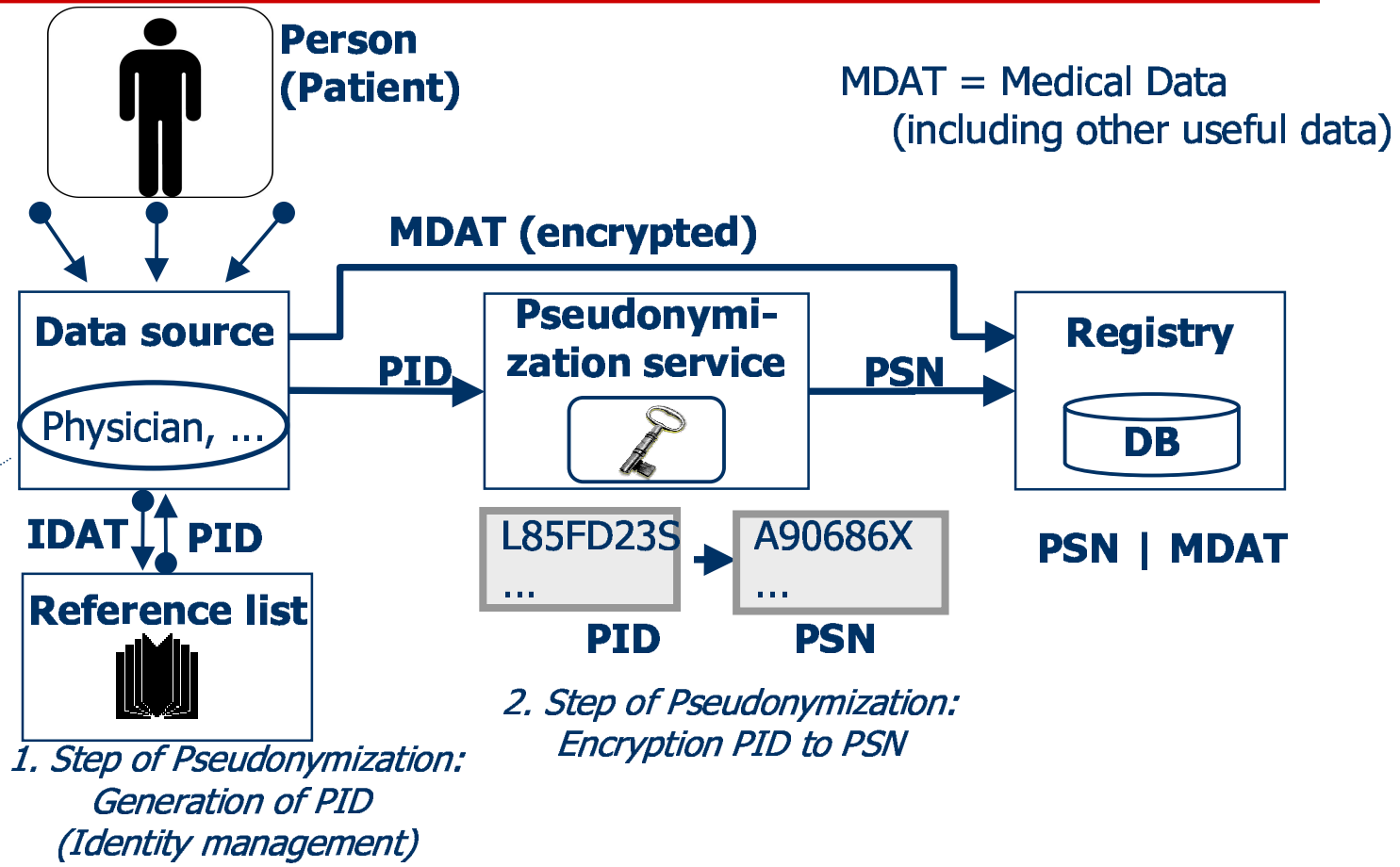
Linking Records

- Problem: Different, maybe erroneous, data sources, e. g. for follow-up data
- Is “Stefan Maier” = “Stephan Mayer”?
- This can be a problem even when universal identifiers exists (errors in data entry, homonyms, synonyms)
- Generate PID (Patient Identifier) which help of an error tolerant matching algorithm
  - *before* pseudonymization.
- Data quality assurance of records: also before pseudonymization

TTP-Generated Pseudonyms
(Extended Model)

- Separate identity management (with PID service) from pseudonymization service
- ⇒ Two TTPs:
  - PID service performs record linkage, assigns unique PID
  - PSD service encrypts PID to PSN
- Different keys for different applications mean different (unlinkable) pseudonyms.
- German view: Use even different PIDs in different networks
  - Constitution doesn't allow a "universal" ID number.

TTP-Generated Pseudonyms (Extended Model)



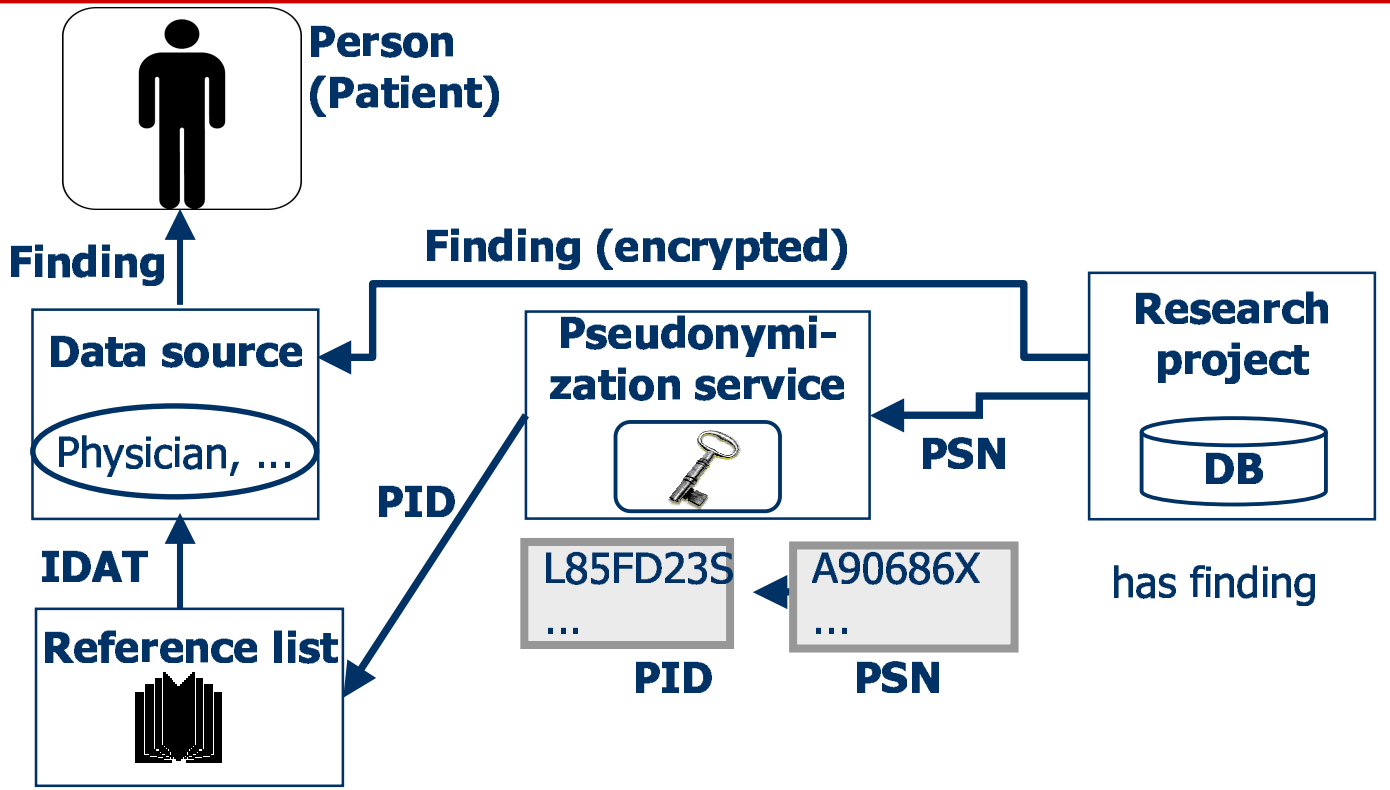
Depseudonymization

[≠ Re-identification]

Get IDAT corresponding to PSN via PSN service and PID service.

- Scenarios:
  1. Feedback to patient
  2. Record linkage between different registries
  3. Recruiting patients

Feedback to Patient



Record Linkage between Different Registries (Proposal)

1. Establish a temporary TTP_{temp} as “One-Time-PID-Service”.
2. Establish a temporary database DB_{temp}.
3. Send depseudonymized data from both registries via TTP_{temp} to DB_{temp} (MDAT encrypted)
4. Link records via PID_{temp}, update.
5. Send back via TTP_{temp} and pseudonymization services.

TTP_{temp} sees IDAT and several PIDs.

DB_{temp} sees PID_{temp} and MDATs.

Cancer registry model gives alternative method of pseudonymous record linkage (control numbers).

Overwiew

1. The TMF
2. Some basic principles
3. Pseudonyms
4. Models for pseudonymization
5. Results

Scenarios for the Use of Medical Data in Research Projects

1. Single data source, one-time use
2. Overlapping data sources, one-time use
3. One-time use of data with depseudonymization
4. Pseudonymous registry
5. Central “clinical database” with multiple uses

Methods: Pseudonymization, separation of “informational powers”, templates for consent, SOPs, policies, contracts.

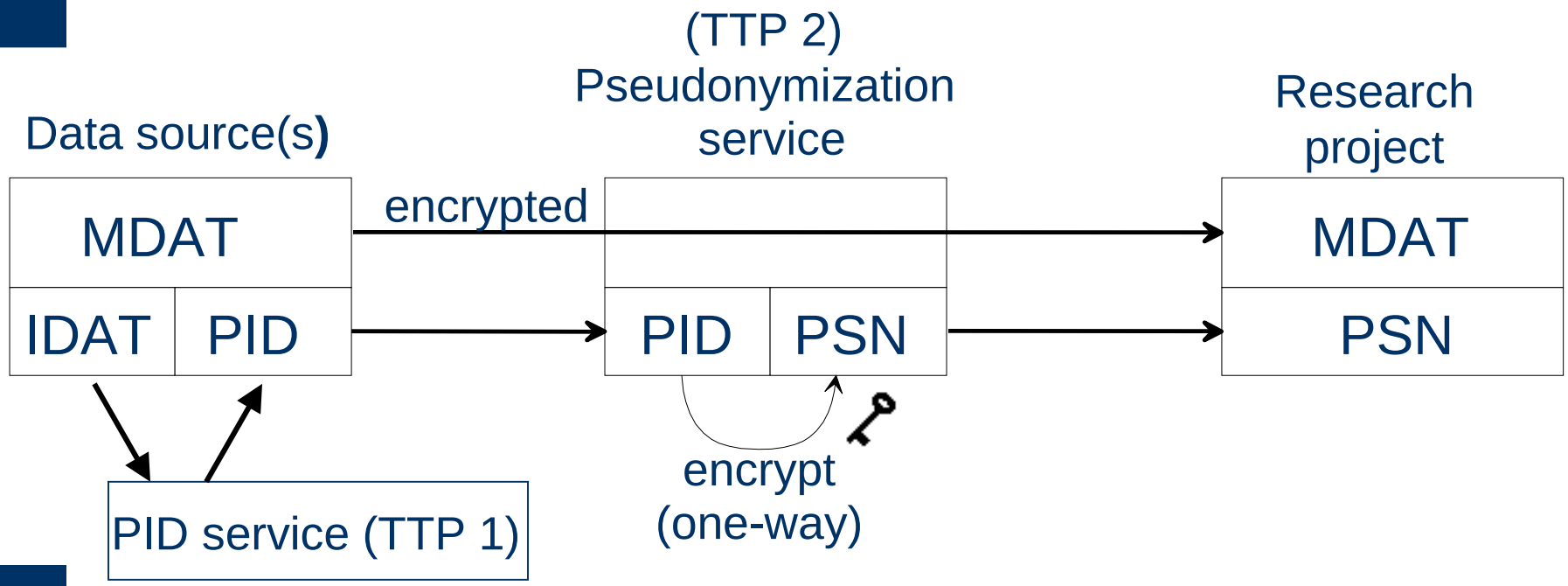
(1) Single Data Source, One-Time Use

- Typical application case for **anonymization**.
- Example: A simple statistical evaluation of epidemiological data, e. g.
  - from a registry
  - or specifically collected for this purpose.

(2) Overlapping Data Sources, One-Time Use

- Data from diverse sources must be linked together.
- Examples:
  - Record linkage from several sources,
  - Follow-up data.
- Typical application case for **one-way pseudonyms**.

Pseudonymization for One-Time Use



MDAT = Medical Data
IDAT = Identity Data

PID = Unique Patient Identifier
PSN = Pseudonym

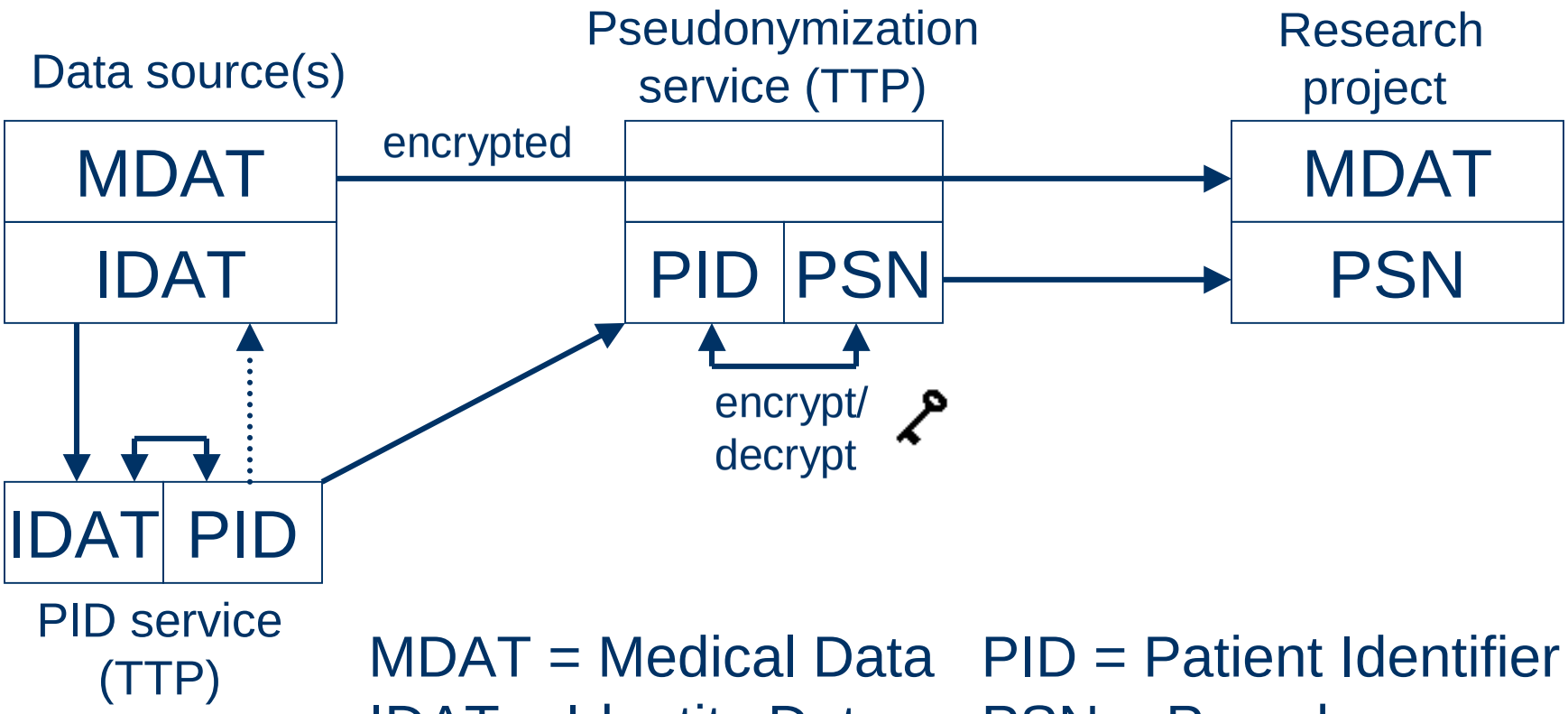
Properties of Scenario (2)

- Medical data (MDAT) are encrypted with public key of research project –
  - The TTP 2 cannot read the MDAT.
  - Only the researchers can decrypt them.
- The pseudonym (PSN) is the encrypted PID
  - With a secret key, known only to the TTP 2,
  - By a one-way procedure.
- The TTP 2 doesn't store anything (except the key).

(3) One-Time Use of Data with Depseudonymization

- Use the TTP model of scenario (2),
  - PSN service performs *reversible* encryption procedure.
- Depseudonymization involves PSN service and PID service.

Pseudonymization with Possible Depseudonymization



MDAT = Medical Data
IDAT = Identity Data
PID = Patient Identifier
PSN = Pseudonym

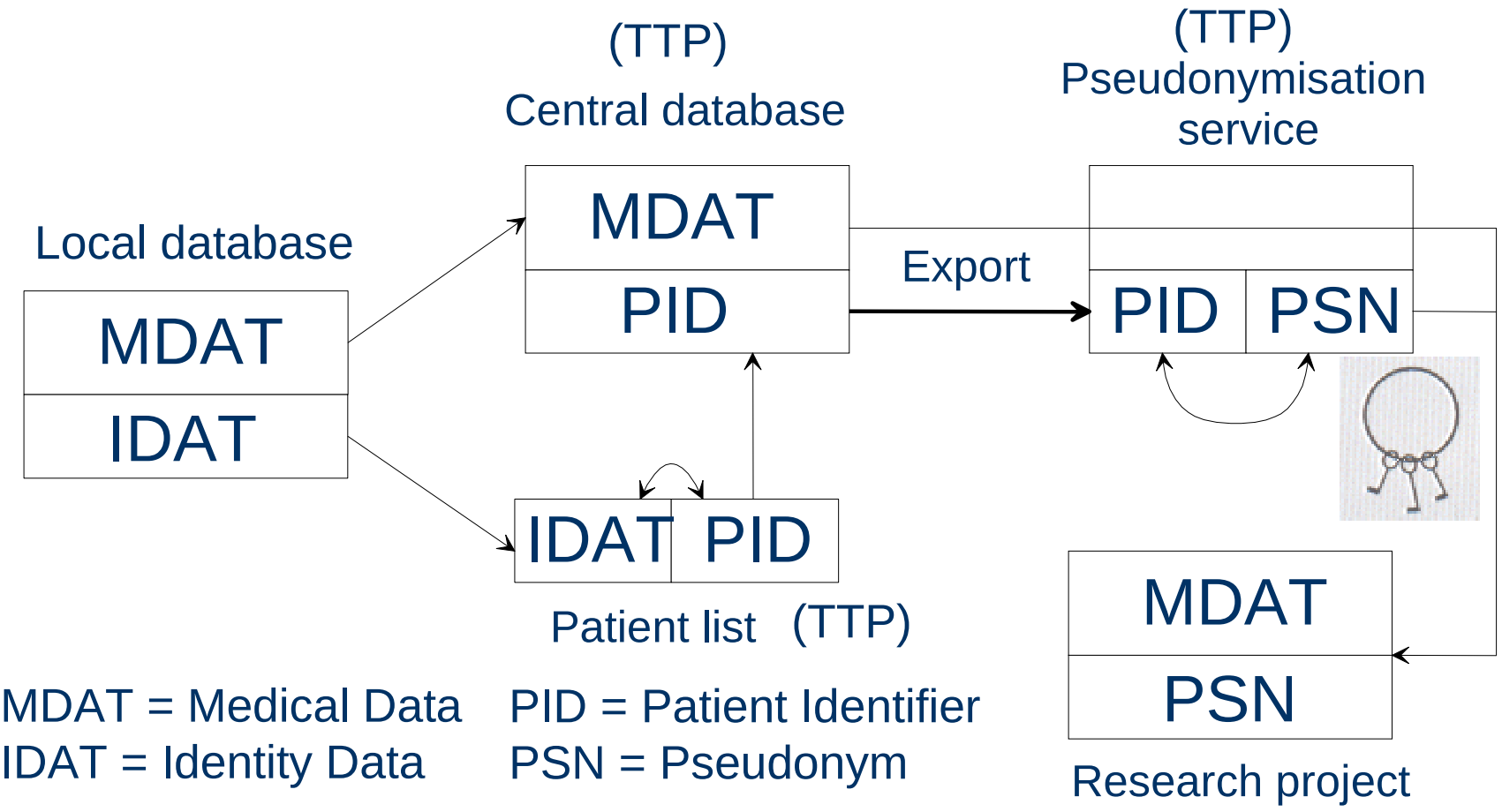
(4) Pseudonymous Registry

- Same procedure as in (3),
  - But the research project builds a (disease specific) registry.
- Long term data accumulation needs special organisational and technical security measures.
  - Protect registry from unauthorized access.
  - Control re-identification risk.
- Quality management of data should precede pseudonymization.
  - Yet another TTP service.

(5) Central Clinical Data Base with Multiple Uses

- Data pool = central “clinical” data base.
  - Access for treating clinicians.
  - No identity data in DB, only PIDs.
  - Access by temporary tokens.
  - Implemented as (yet another) TTP service.
- No online access by researchers.
  - Researchers get exported data set (anonymized or pseudonymized) as in (1) or (3).

TTPs for Central Clinical Data Base



Properties of Scenario (5)

- Advantages:
  - Better support for long-term observation of patients with chronic diseases.
  - Individual feedback of research results easy.
  - Fits well into EHR architecture or into multicenter study management.
- Drawback:
  - Sophisticated communication procedures.
  - More TTPs and secret keys involved.
  - Less useful for registries
    (But can be data source for registry.)

The TMF approach

- Generic Data Protection Conception
  - Scenario (5) as “Model A”.
  - Scenario (4) as “Model B”.
- Revision is work in progress
  - combine models A and B
  - criteria for appropriateness
- Extension to Biobanking in preparation.
- Vagueness of purpose compensated by choices in patients’ consent and by clearly defined organizational framework

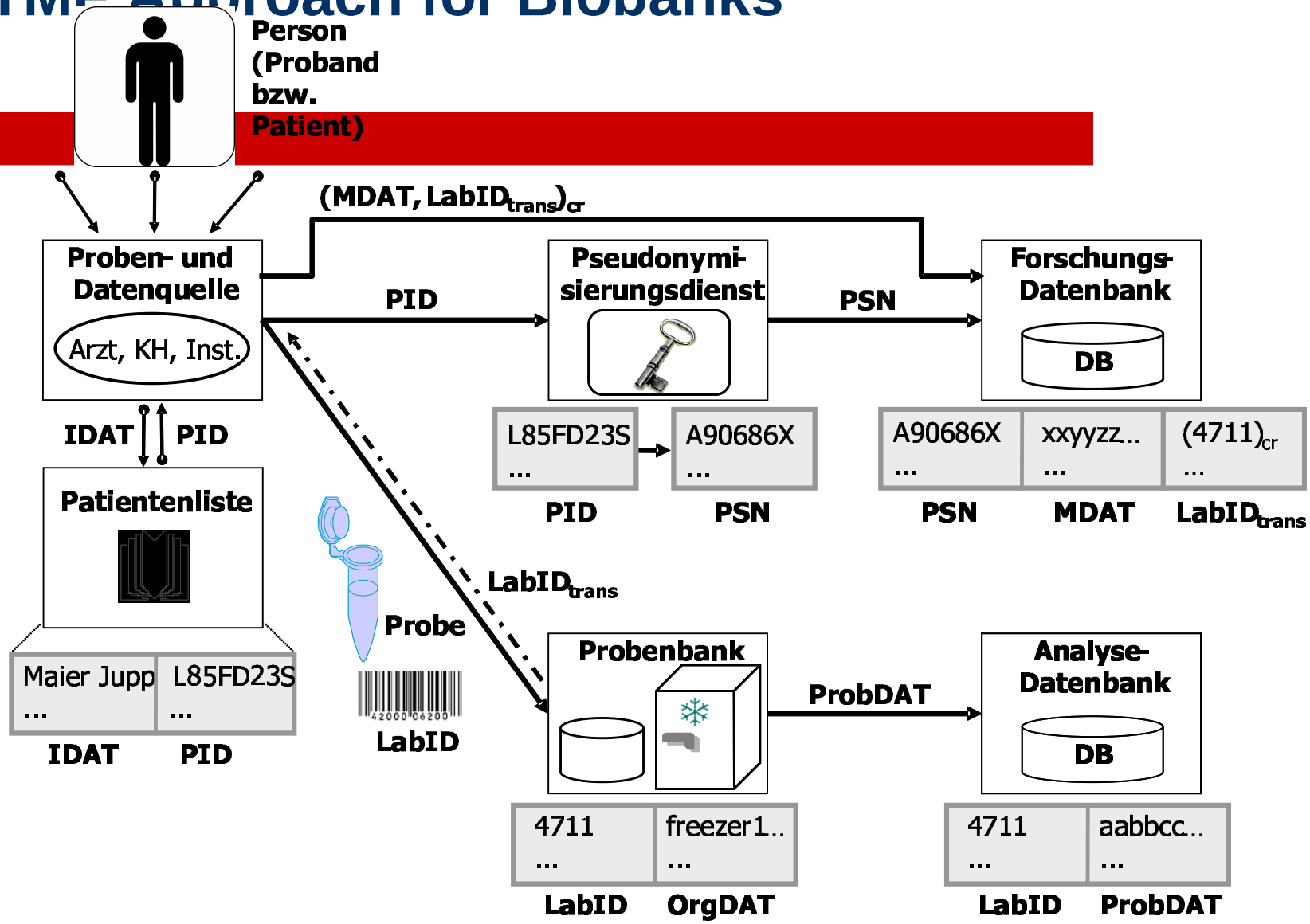
Biomaterial and Genetic Information

- Part of the human body
- Information about the person
- Information about relatives
- Information about a group of individuals (for example ethnic minorities)

Approach:

- Separate samples (and analysis results) from data.
- Use another set of identifier (LabID) and pseudonym ($LabID_{trans}$) for material samples.

TMF Approach for Biobanks



Security Infrastructure

... includes

- Network and server security
- PKI (cryptographic tools)
- SSL (secure communication via web)
- Smart cards (or “soft certificates”)

Note: *Good security needs redundancy*.
(How much is appropriate?)

Overwiew

1. The TMF
2. Some basic principles
3. Pseudonyms
4. Models for pseudonymization
5. Results

Current Status of the TMF Approach (I)

- TMF models A and B [(5) and (4)] approved by the German Data Protection Commissioners
  - (Arbeitskreis Wissenschaft der Datenschutzbeauftragten des Bundes und der Länder)
- Scenario (2) in routine use since 2002 in a health care research project of the TMF.
- Scenario (5) implemented in some networks.
  - KN CED (Chronic Inflammatory Bowel Disease).
  - Further implementations in progress.

Current Status of the TMF Approach (II)

- Scenario (4) adapted by several research networks
  - Implementations in progress.
- TMF offers software tools for the TTP services.
- Corresponding policies, sample contracts, forms for patient's consent available from TMF (free for members) [in German, international publication in preparation].

Discussion of the TMF Approach

- The TMF model architecture (variants A and B) provides ways for building registries and data pools for public health, epidemiological and health care research, that
  - conform to the German and European data protection rules,
  - respect the patients' rights,
  - and cover a wide range of situations.
- The pseudonymization scenarios look complex, but once established, work transparently.

Discussion of the TMF Approach

- The TTP services should be implemented as “communication nodes” or “communication servers”
  - That perform their duties automatically and silently, in particular perform minor transformations of data
  - But are under control by independent TTPs (in different organizations)
  - And follow predefined rules.
- The TTP services are technical tools that help enforce policies.