Advanced De-identification & Linkage of Personal Data originating from Multiple Sources for Secondary Use

Kristof Verslype
Cryptographer (PhD.) @ Smals Research

1 April 2025

HDA
health data agency

Smals
ICT for society

Principles GDPR

Data protection by design and by default

Article 25 GDPR

”The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. ”

“Implementing technical and organizational measures, at the earliest stages of the design of the processing operations, in such a way that safeguards privacy and data protection principles right from the start.”

European Commission

Data Minimisation

Article 5, (1.c) GDPR

”Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (‘data minimisation’)”

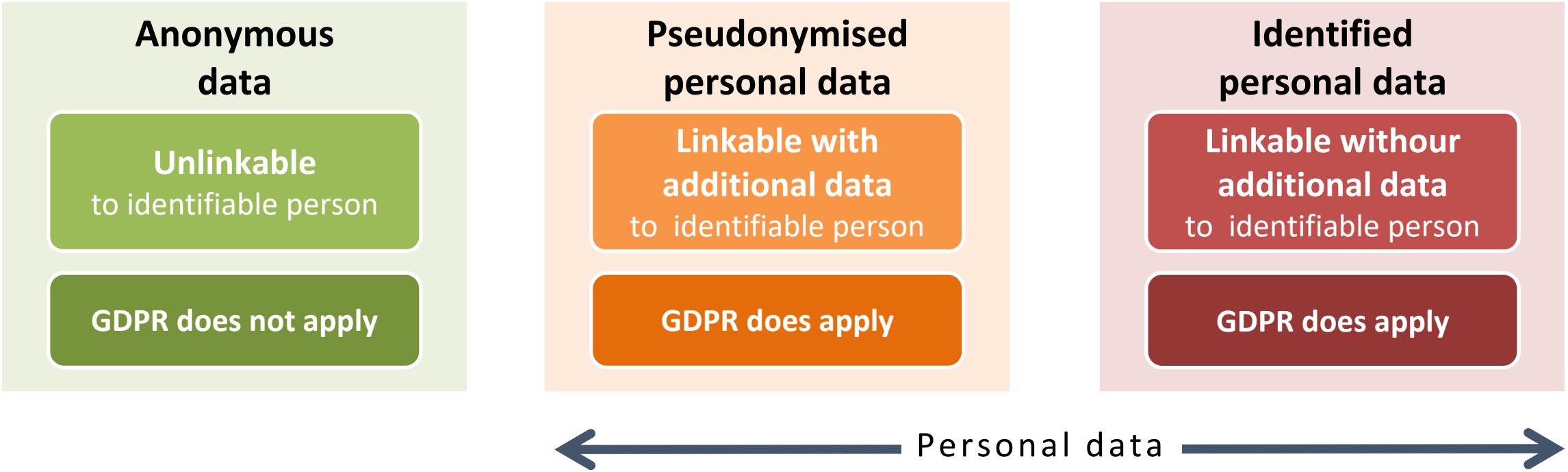
Secondary use

Article 5.1, (b,e) and 89 GDPR

*” Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to **appropriate safeguards**, in accordance with this Regulation, for the rights and freedoms of the data subject ”*

Smals ICT for society

Pseudonymisation



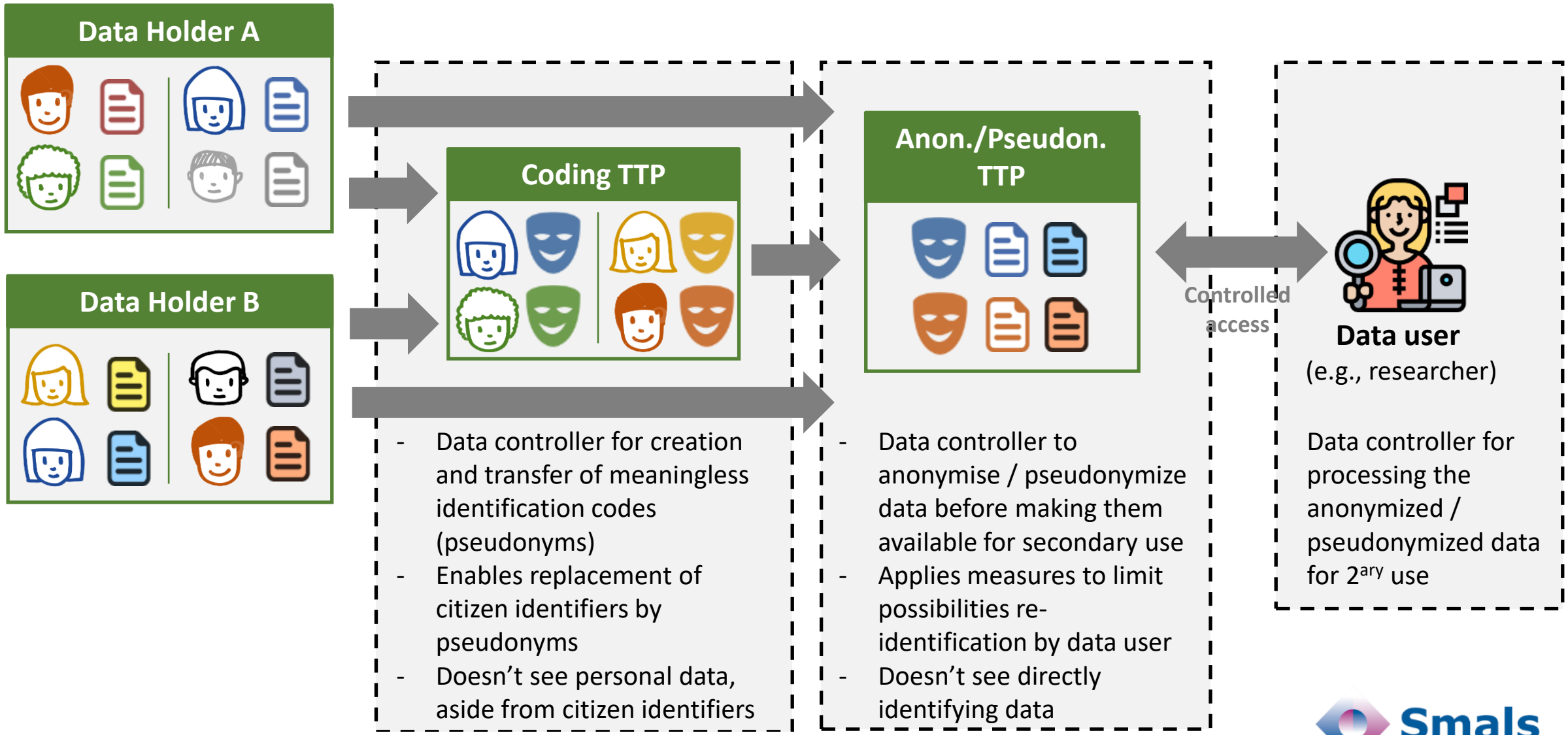
Pseudonymisation

Article 4, (5) GDPR

*"The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of **additional information**, provided that such additional information 1) **is kept separately** and 2) is **subject to technical and organizational measures** to ensure that the personal data are not attributed to an identified or identifiable natural person."*

Task division

Link and anonymise/pseudonymise personal data for secondary use, after permit has been granted

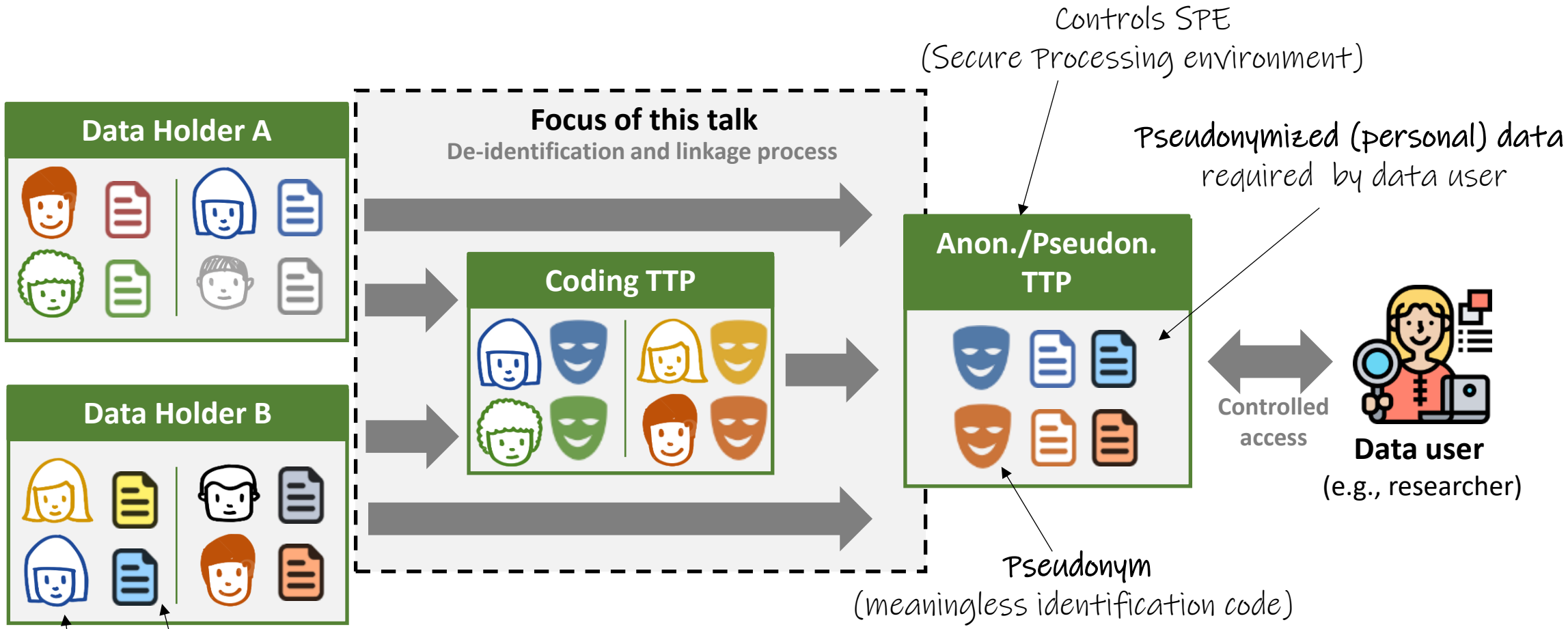


Nota Nr. 25/022 Van 27 Januari 2025 Betreffende De Algemene Principes Omtrent De Verwerking Van Persoonsgegevens, Pseudonimisering En Verdeling Van Verwerkingsverantwoordelijkheden

Smals ICT for society

Generic use case

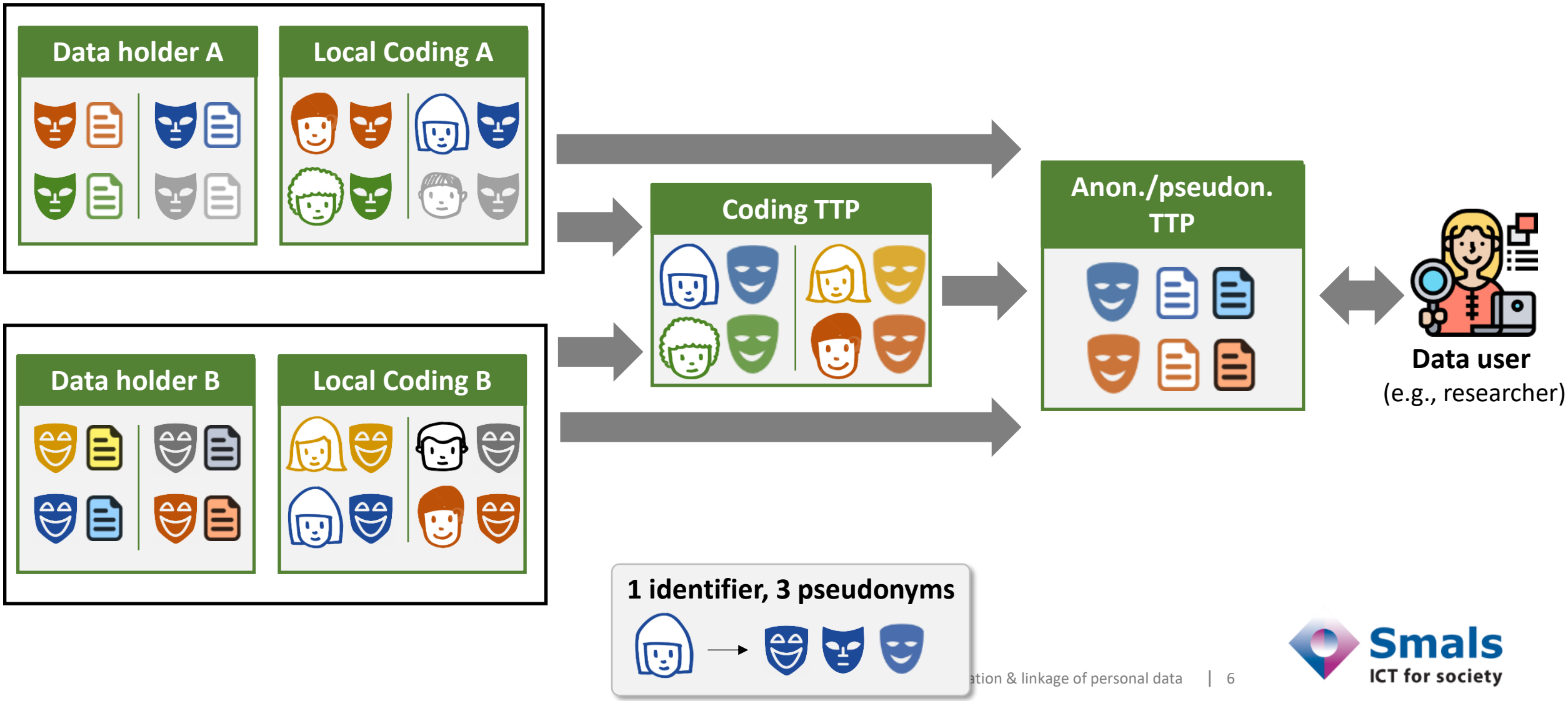
Link and anonymise/pseudonymise personal data for secondary use, after permit has been granted



Smals ICT for society

Generic use case

De-identification and linkage of personal data for secondary use, after permit has been granted



Smals
ICT for society

Questions

After the data user obtained a permit by an HDAB to access data,
A technical de-identification and linkage process should be started, resulting in a de-identified data set known by the anonymization/pseudonymization TTP in its SPE.

Could we apply **Data protection by design** to further improve **data minimisation** in the de-identification and linkage process?

In particular

- Could we avoid that the coding TTP learns identifiers of involved citizens?
- Could we elegantly minimise unintended data leakage to the anon./ pseudon. TTP and/or data holders?
- Can we reduce the work done by the coding TTP?

Advanced De-identification & Linkage

of Personal Data originating from Multiple Sources for Secondary Use

eHealth Blind Pseudonymisation service

High-security service

Life

Oblivious Join

Distributed protocol

Experimental

Illustrates potential of current state of the art

Smals
ICT for society

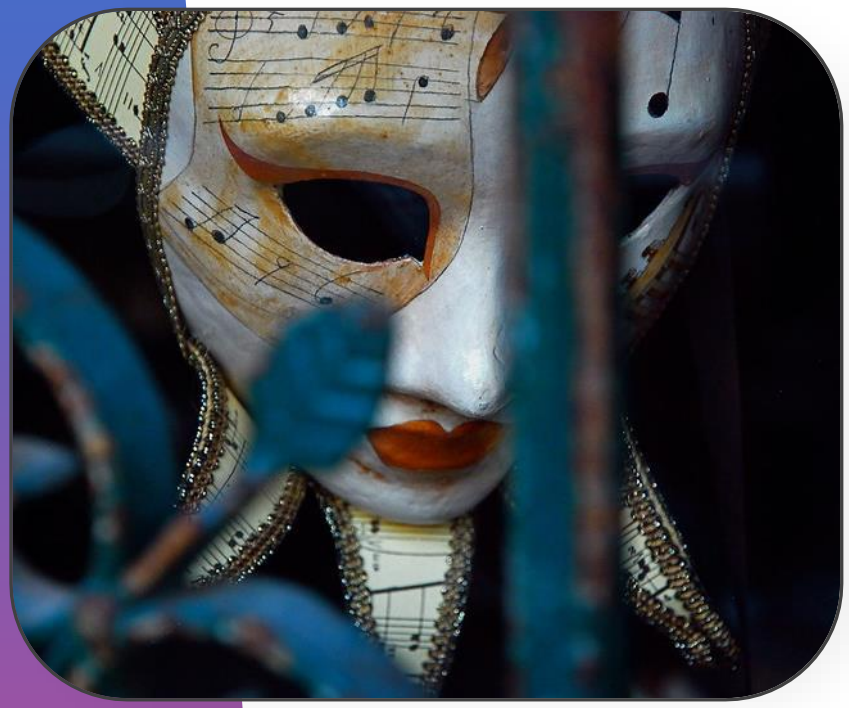
Disclaimer

This talk illustrates possibilities on the conceptual and technical level.
The goal is to inspire.

It does no do any claims regarding general feasibility!
Legal framework, procedures, governance, development costs, integration costs, operational costs, capacity limitations, … are not considered

Smals
ICT for society

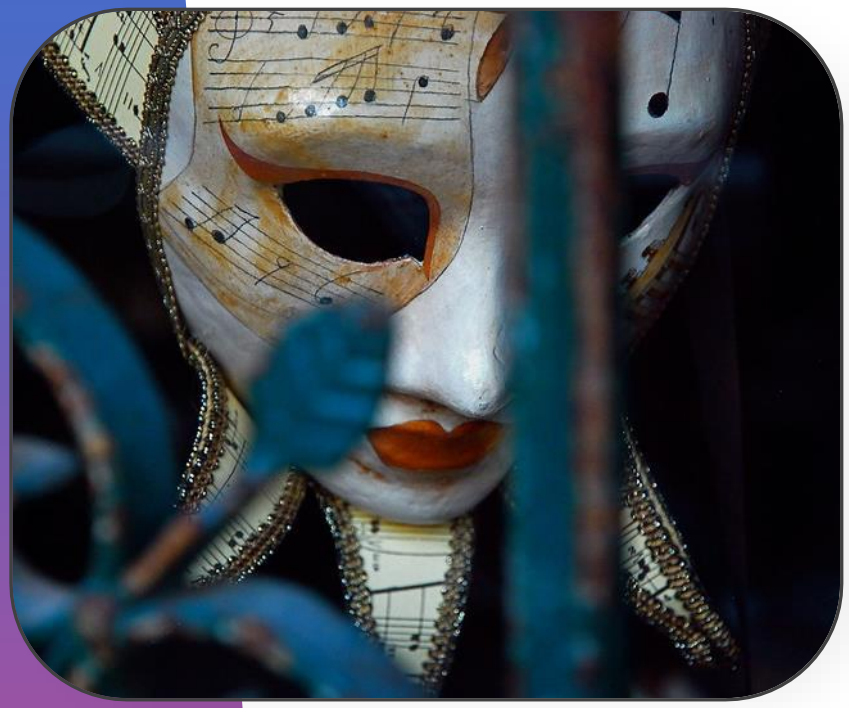
eHealth Blind Pseudonymisation



- Referral prescriptions (primary use)
- De-identification & linkage for secondary use
- Conclusion

Smals
ICT for society

eHealth Blind Pseudonymisation



- **Referral prescriptions (primary use)**
- De-identification & linkage for secondary use
- Conclusion

Smals
ICT for society

Use case 1 - Live

Referral prescription = Verwijsvoorschrift / Prescription de renvoi

What?

A certificate to start a certain treatment (e.g. physiotherapist, dieticians, speech therapists).

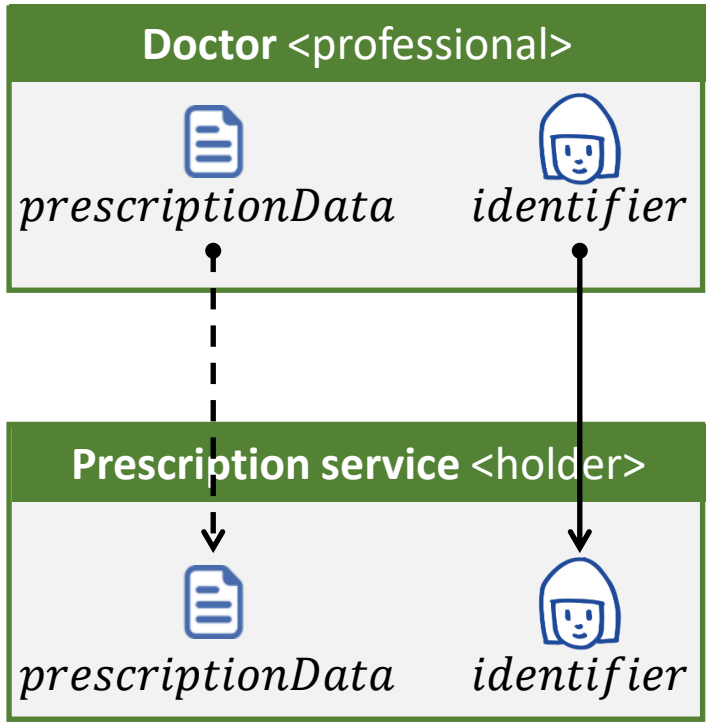
Requirements

- Prescription service should never be able to link prescription data to a citizen identifier
- No full encryption (input validation, statistics, …)

→ **Securely replace citizen identifiers by pseudonyms**

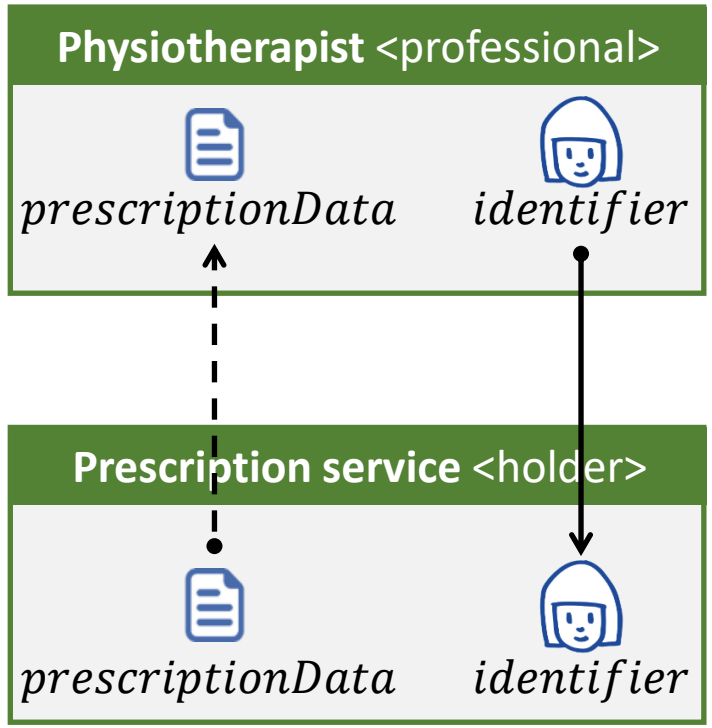
Naive scenario 1

Doctor (health professional) requests Prescription service (holder) to register prescription



Naive scenario 2

Physiotherapist (health professional) requests access to prescription for a specific citizen from Prescription service (holder)



ICT for society

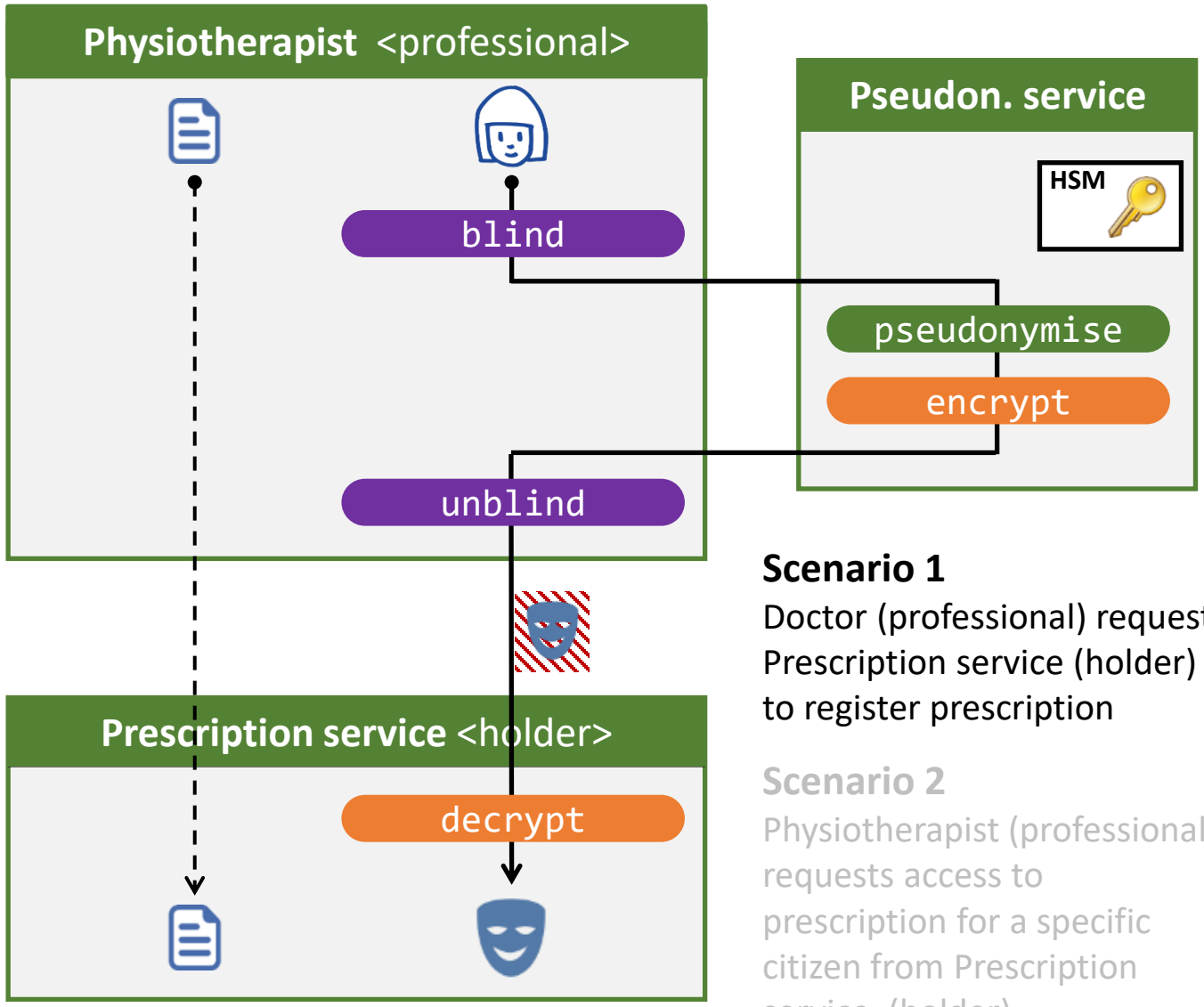
Blind Pseudon. Service
Pseudonymise

Structure blinded identifier, blinded pseudonym and final pseudonym
(AV+VXF9H5LdTe4b1 SSC7bHjp6b2enJmf plC6a3/jCR5fUHxX RSaRniYR8h7ugNqa lGvP49cZnv6lf9B7 2RUG0rA/, eSmII52CEtsZzSseU DY3YKLtSgqhq1wLPm 9ncHBzGiv1wMIxmc1 jSmpW36GhTt/s1P5s hZGhG8ncoWKSGkJDy fw=)

Data minimisation
- Professional only sees identifiers
- Holder only sees pseudonyms
- Pseudon. service sees neither

Reduced overhead
- Direct communication between healthcare professional and prescription service
- No in-between entity

Low-intrusive side professional
- No extra keys required
- Relatively simple implementation

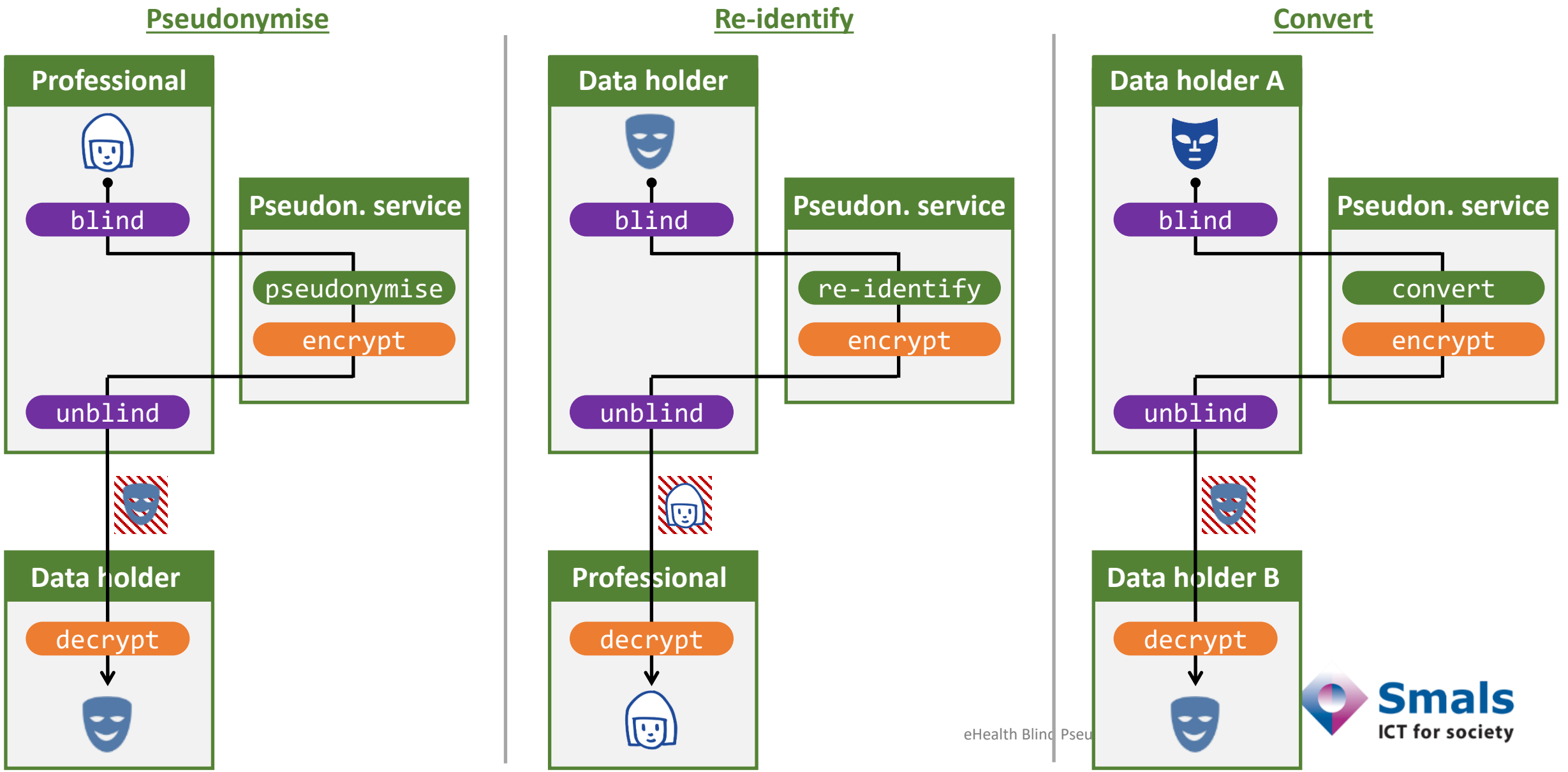


Scenario 1
Doctor (professional) requests Prescription service (holder) to register prescription

Scenario 2
Physiotherapist (professional) requests access to prescription for a specific citizen from Prescription service (holder)

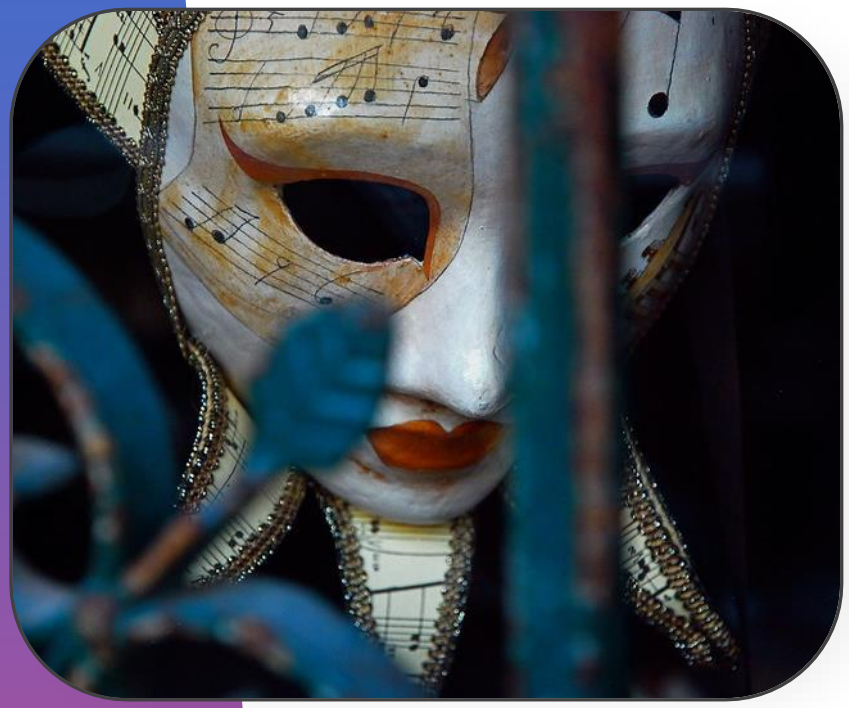
Three operations

Necessary and sufficient for a generic service



Smals
ICT for society

eHealth Blind Pseudonymisation



- Referral prescriptions (primary use)
- **De-identification & linkage for secondary use**
- Conclusion

Smals
ICT for society

Questions

Data minimisation
Can we use the blind pseudonymization service to improve data minimisation: Can we hide citizen identifiers from the coding TTP?

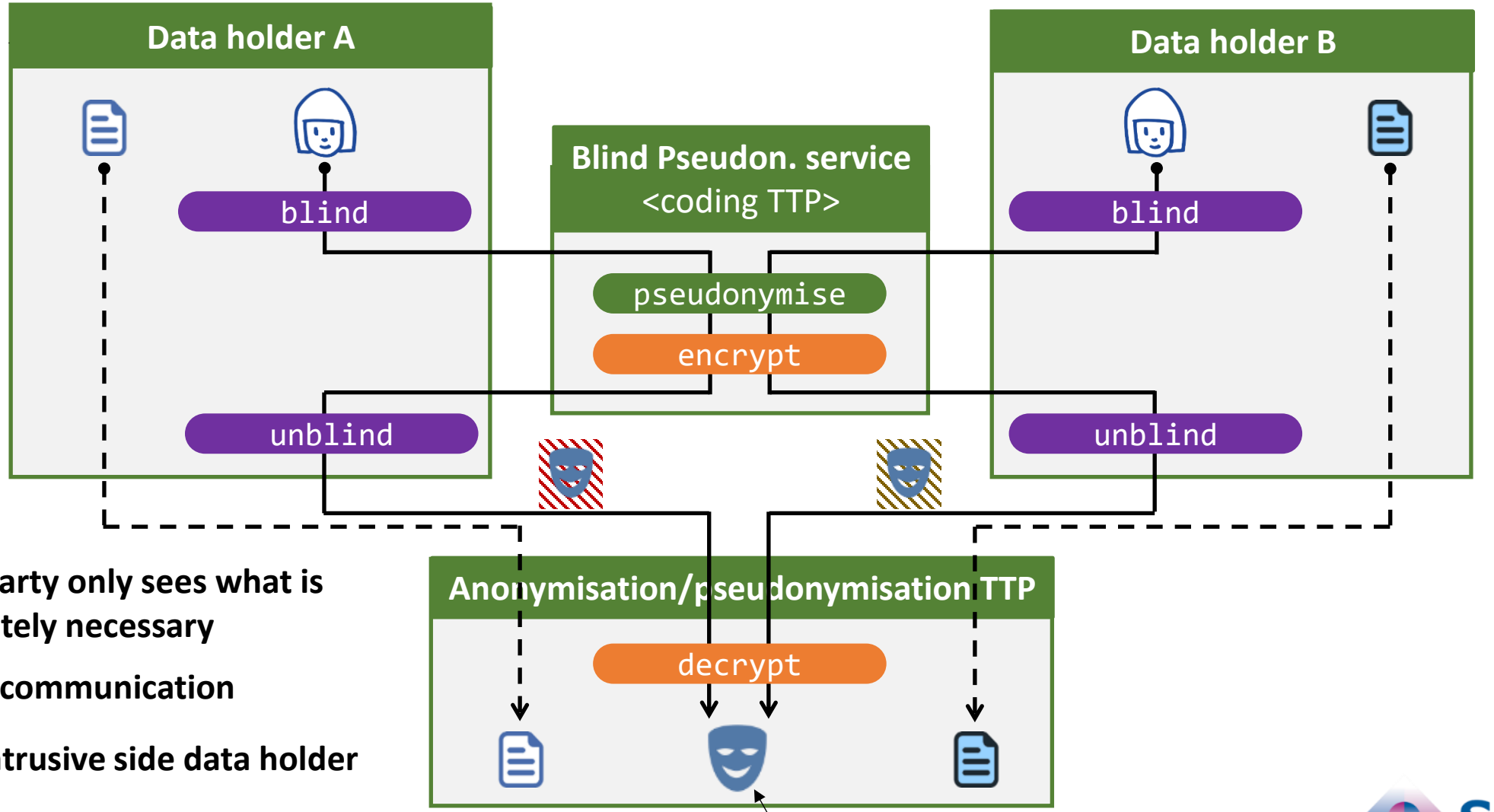
Additional data
Can the additional data, to map citizen identifiers to pseudonym and back, be kept separately with appropriate technical and organizational measures?

Linking
Is it still possible to link data originating from multiple data holders, if at least one of them uses the blind pseudonymization service?

Smals ICT for society

De-identification a link process to make data available for secondary use

Data holders know data under citizen identifier

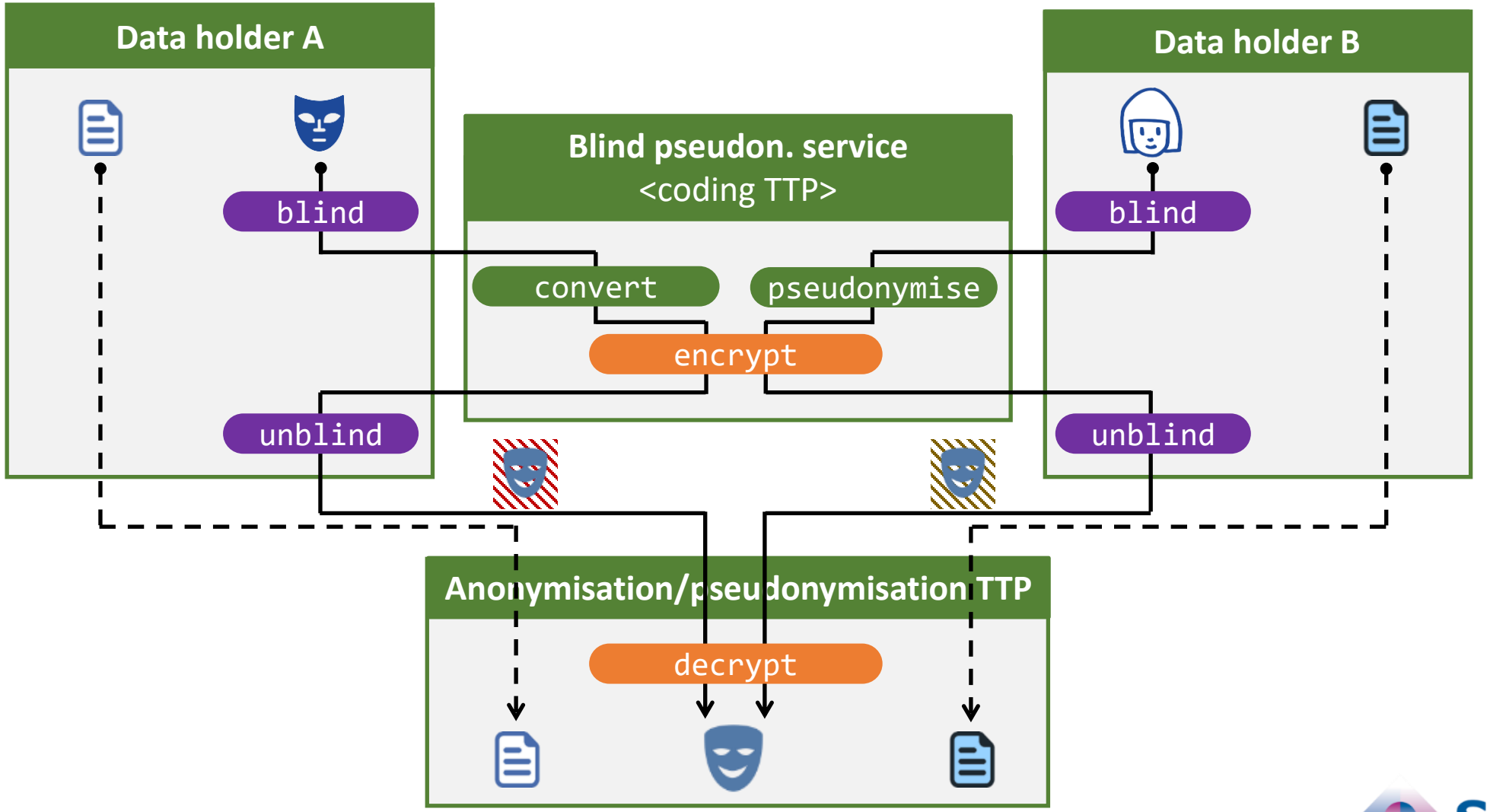


- ✓ **Each party only sees what is absolutely necessary**
- ✓ **Direct communication**
- ✓ **Low-intrusive side data holder**
- ✓ **Reuse of existing infrastructure**
  (BUT: limited capacity!)

Smals ICT for society

De-identification a link process to make data available for secondary use

Data holder A already uses the blind pseudo service to protect data in production environment



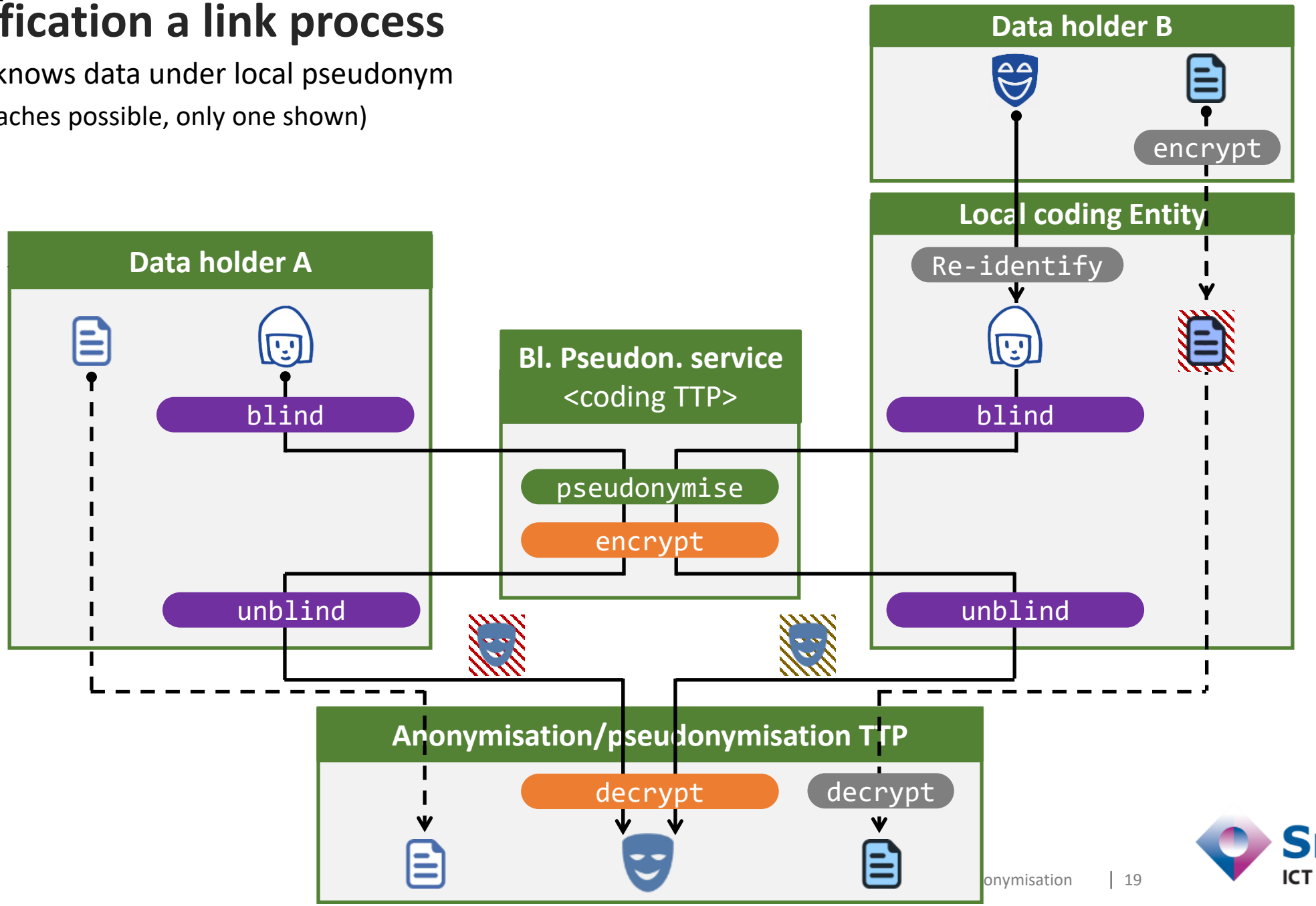
Smals ICT for society

Use case 2 – Proposal

De-identification a link process

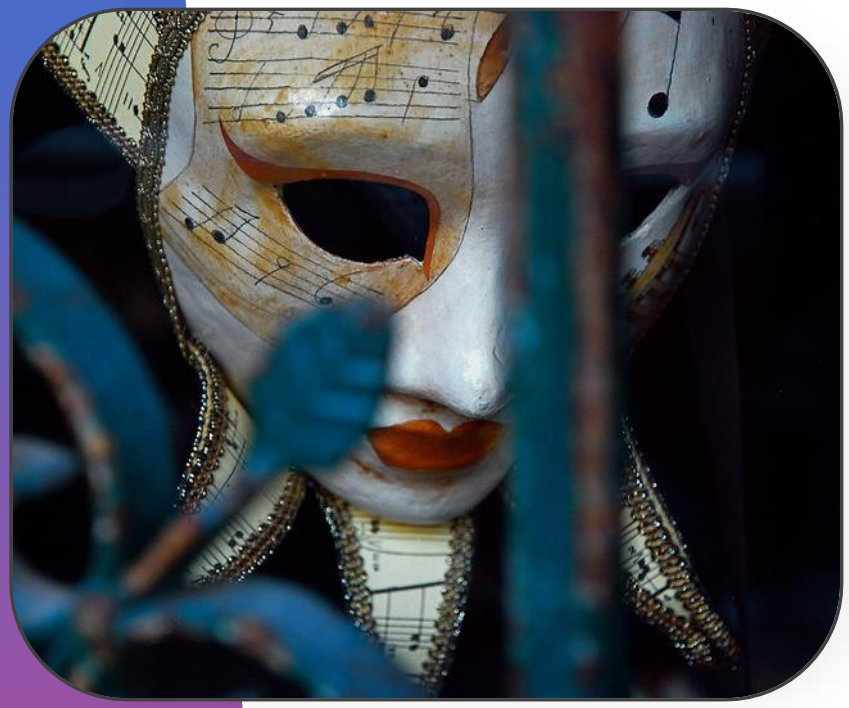
Data holder B knows data under local pseudonym

(Multiple approaches possible, only one shown)



Smals
ICT for society

eHealth Blind Pseudonymisation



- Referral prescriptions (primary use)
- De-identification & linkage for secondary use
- **Conclusion**

Smals ICT for society

Evaluation

✓ **Data minimisation**
- ❖ Professional only sees identifiers
- ❖ holder only sees pseudonyms
- ❖ Pseudon. service sees neither

✓ **Protection of additional data**
- ❖ Cryptographic key
- ❖ Stored by blind pseudon. service in HSM (kept separately)

✓ **Appropriate measures**
- ❖ Reuse of existing infrastructure
- ❖ Acceptable technical complexity for data holders & A/P TTP

✓ **Linking**

Elegant way to deal with data holders that use blind pseudon. service

Status
- ✓ Live
- ○ Not yet used for de-identification and linkage processes (Feasibility to be checked)

Smals ICT for society

Advanced De-identification & Linkage

of Personal Data originating from Multiple Sources for Secondary Use

eHealth Blind Pseudonymisation service

High-security service

Life

Oblivious Join

Distributed protocol

Experimental

Illustrates potential of current state-of-the art

Smals
ICT for society

Oblivious Join

- Problem statement
- Concept
- Proof-of-Concepts
- Required extensions
- Conclusion

Smals
ICT for society

Oblivious Join

- **Problem statement**
- Concept
- Proof-of-Concept
- Required extensions
- Conclusion

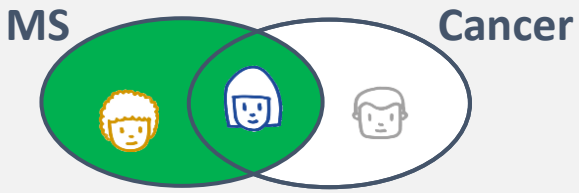
Smals ICT for society

Concrete case

Research question

Do MS patients who take medications with the molecule teriflunomide or alemtuzumab have an increased cancer risk compared to MS patients treated with other medications?

Involved citizens

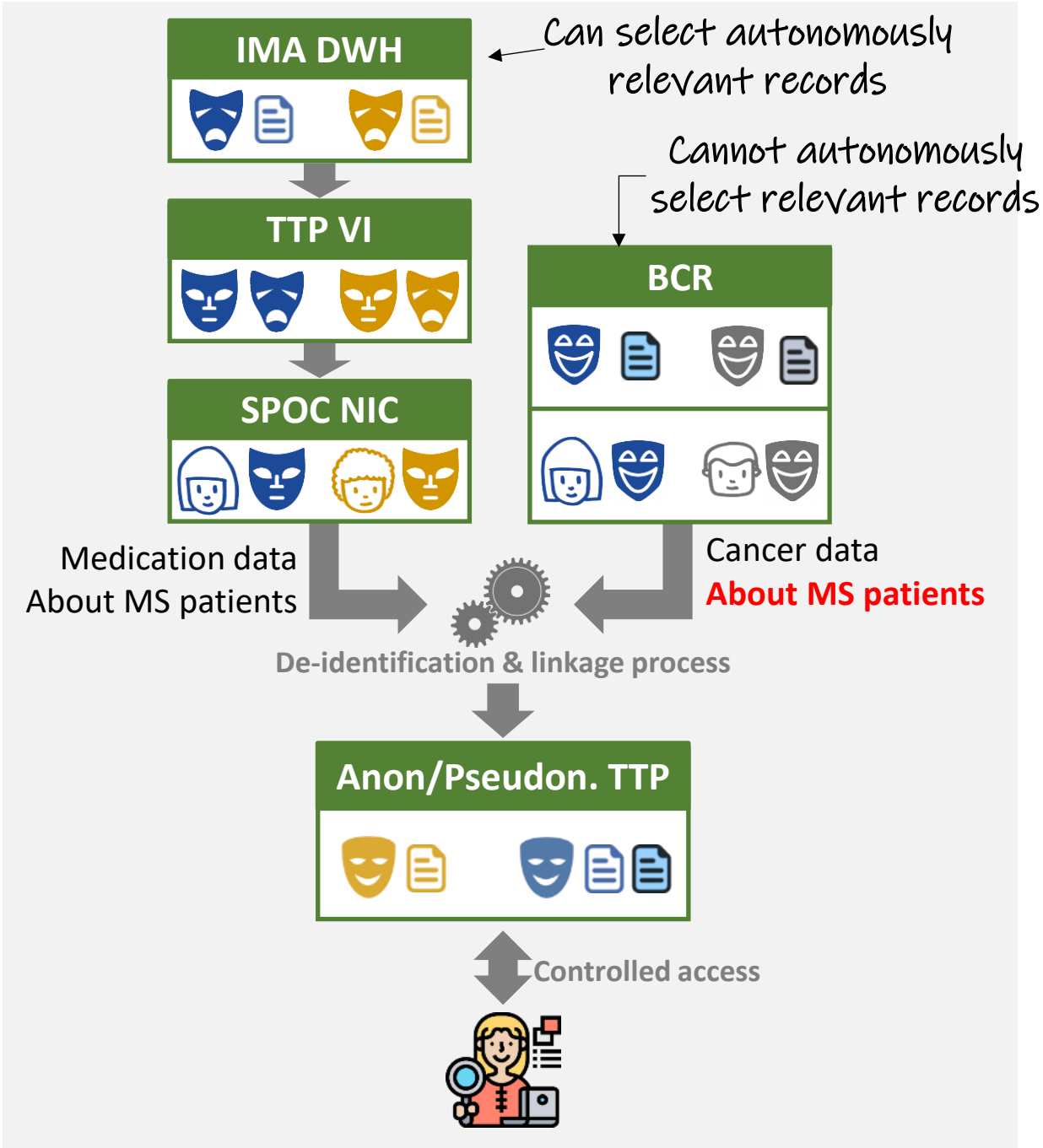


Personal data hidden from SPE

Personal data disclosed to SPE

Data minimisation challenge

How can BCR deliver only records about MS patients without learning who has MS?



Current practice

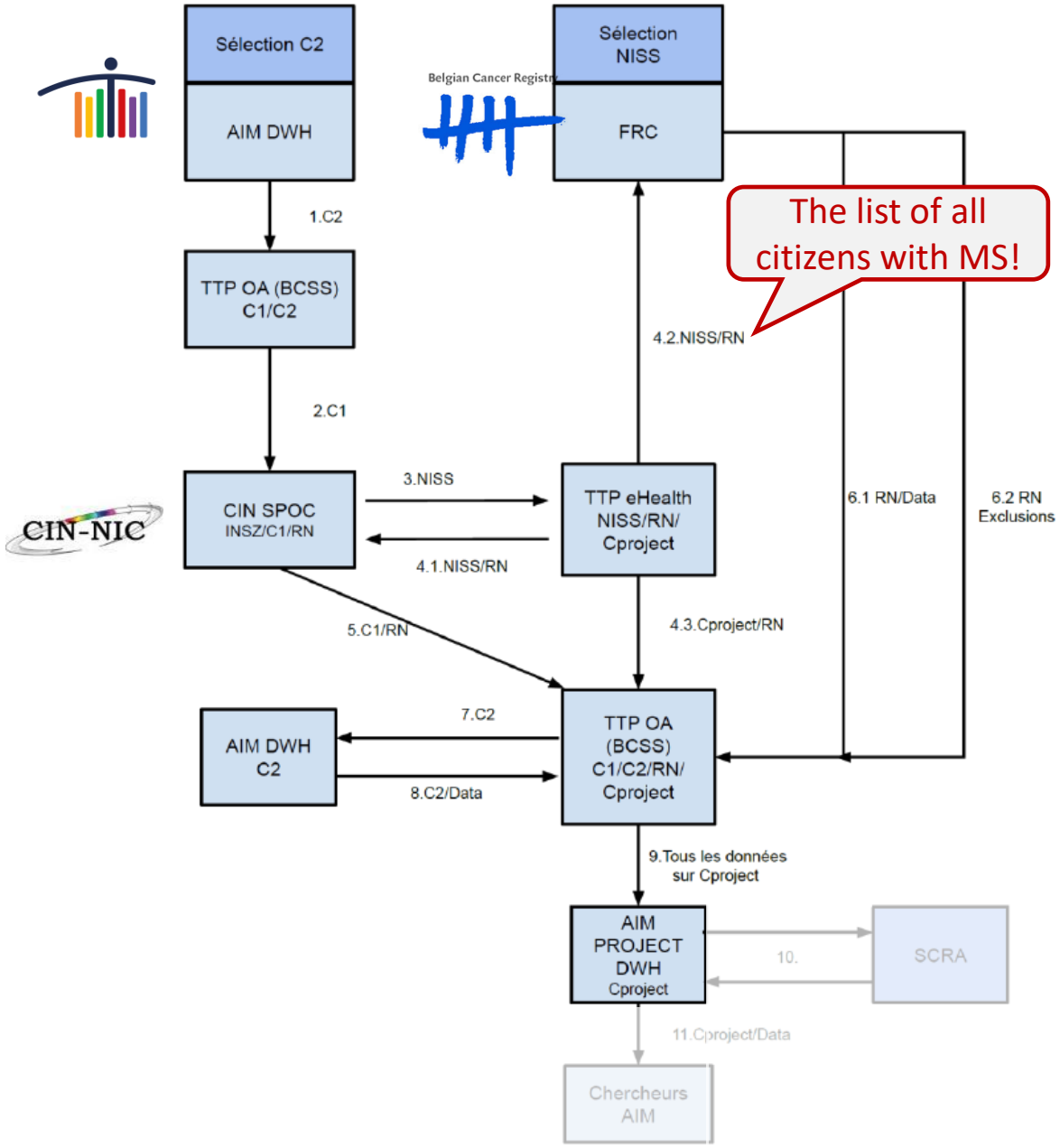
After permit has been granted by permitting HDAB

Observations

- ✗ Complex flow
- ✗ Expensive
- ✗ Bespoke
- ✗ Doesn't scale well
- ✗ Slow
- ✗ Security risk (data leakage)

Other countries

Heavy reliance on combination of trusted parties and strong legal regulations



Selective disclosure

Data Minimisation Challenge

How can BCR deliver only records about MS patients without learning who has MS?

Approach

- BCR de-identifies and encrypts requested data about ***ALL*** cancer patients and sends the result to the anon/pseudon. TTP.
- Cryptography guarantees that the anon/pseudon. TTP can link and decrypt ***ONLY*** records about citizens with MS

Other requirements

- **Access to micro data**
  Researcher needs access to individual pseudonymized data records. No data aggregation.
- **Uniform flow**
  Avoid bespoke flows, despite diversity of research questions. We no longer want to waste resources on designing flows.
- **Fast technical execution**
  Minimal human intervention

Smals ICT for society

Oblivious Join

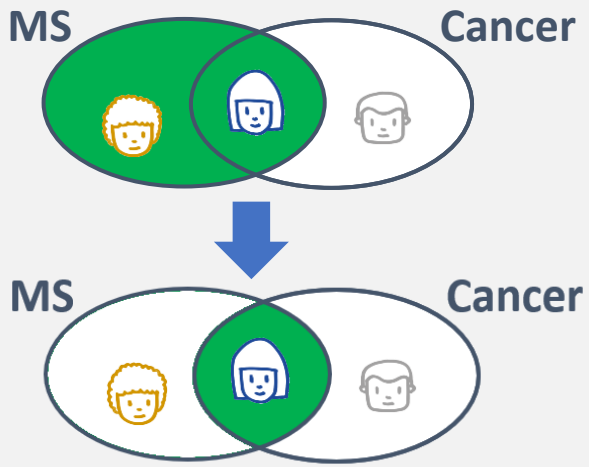
- Problem statement
- **Concept**
- Proof-of-Concept
- Required extensions
- Conclusion

Smals ICT for society

Simplifications

- For didactical reasons
- Aligns better with current state of PoC

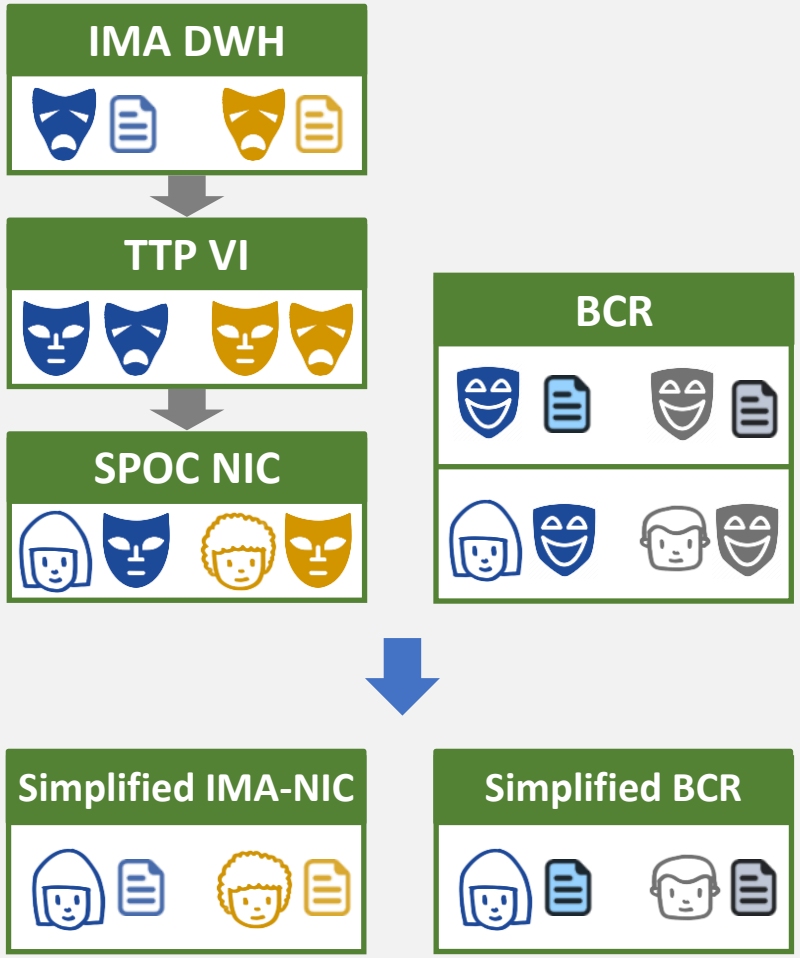
Focus on intersection



Personal data hidden from SPE

Personal data disclosed to SPE

Data holders know data under ids

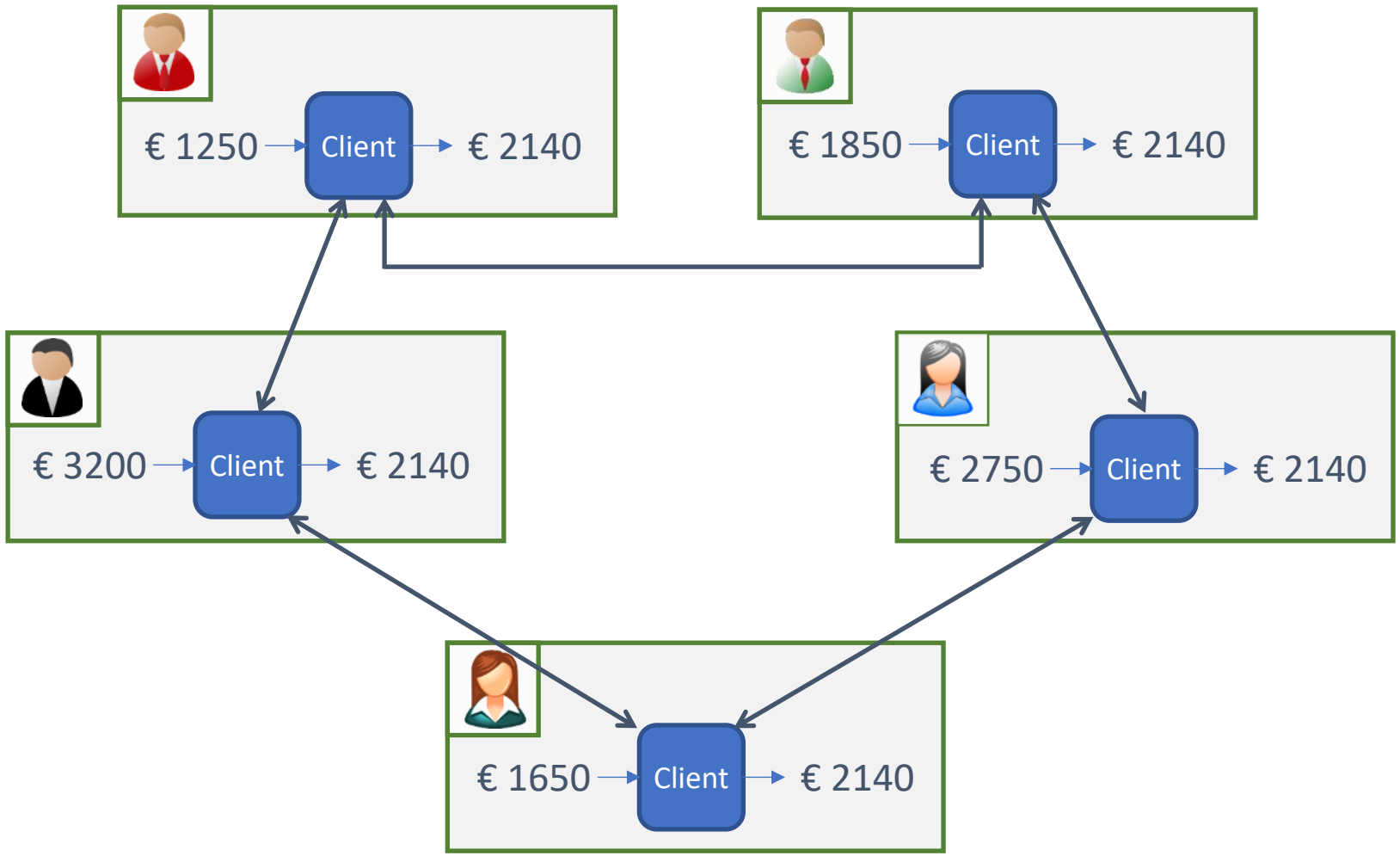


From there, we can in the future extend towards realistic cases!

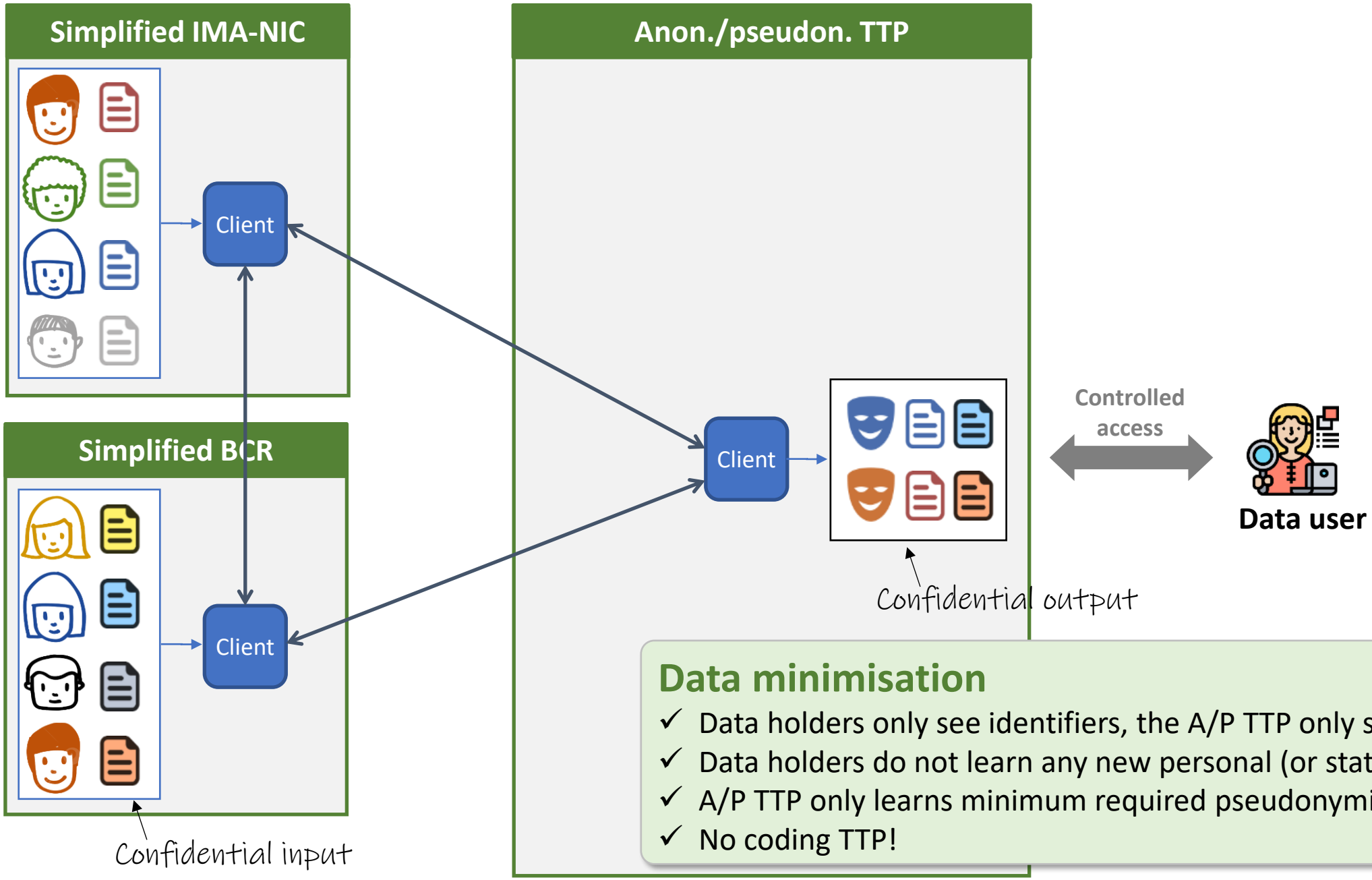
Smals ICT for society

Secure Multi-Party Computation

- Multiple parties jointly compute a function over their confidential inputs
- Example: **Calculate average wage without disclosing individual inputs**
  No one learns anything new aside from the average wage



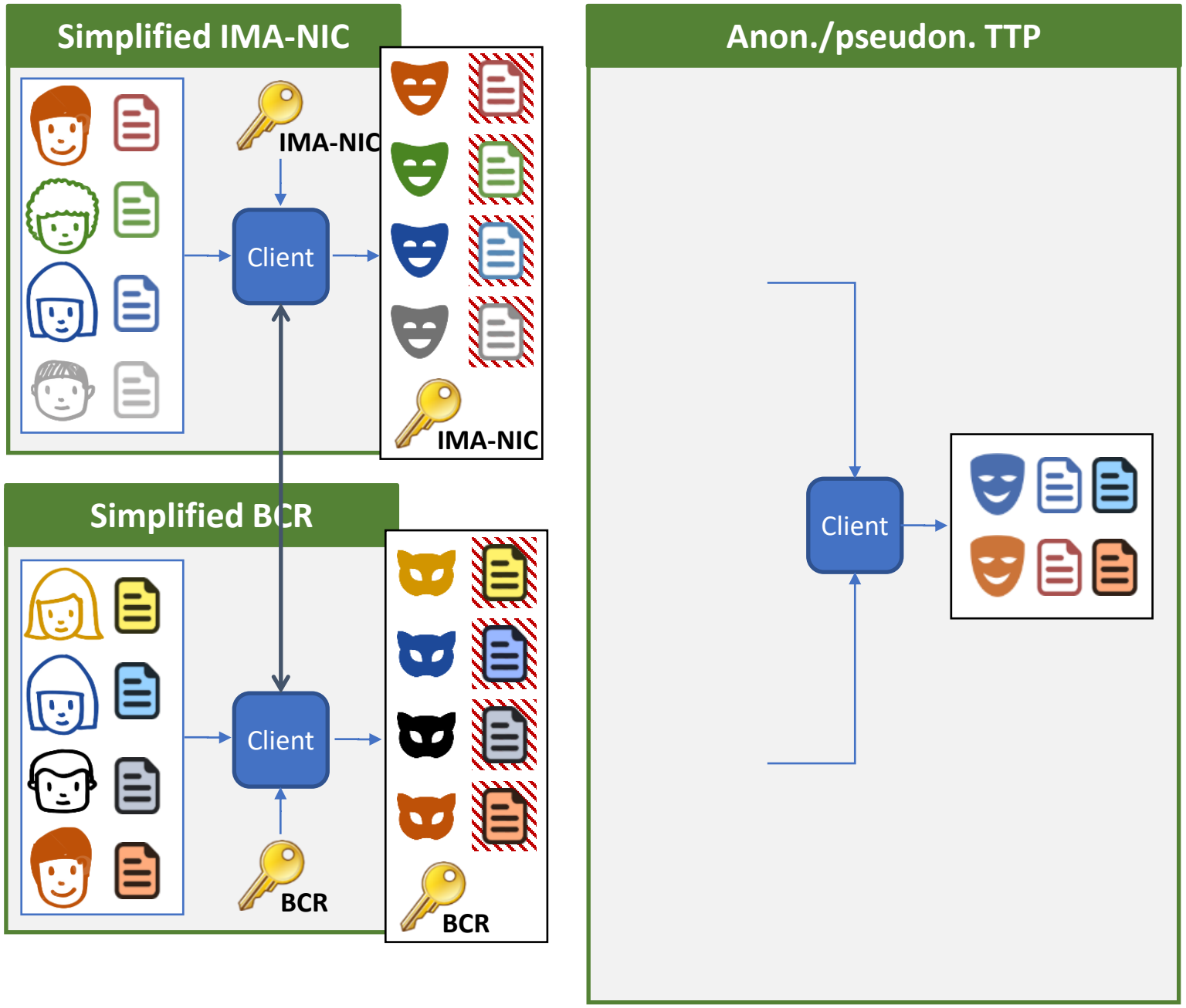
Concept



Data minimisation

- ✓ Data holders only see identifiers, the A/P TTP only sees pseudonyms
- ✓ Data holders do not learn any new personal (or statistical) data
- ✓ A/P TTP only learns minimum required pseudonymised personal data
- ✓ No coding TTP!

Concept



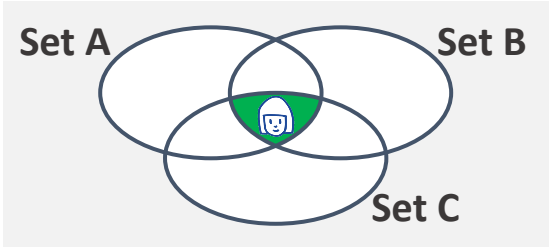
Smals
ICT for society

Scenario A: Every data holder delivers data about a specific citizen

Step 1
Key generation

Step 2
Key split

Step 3
Distribution key shares



Step 5:
Transfer ciphertext & key shares

Step 6:
Recompose keys

Stap 7:
Record decryption

Data holder A

encrypt

Data holder B

encrypt

Data holder C

encrypt

Anon./pseudon. TTP

decrypt

decrypt

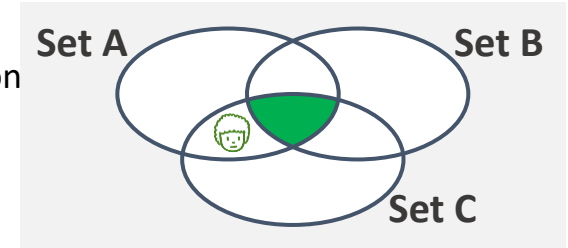
decrypt

Scenario B: NOT every data holder delivers data about a specific citizen

Step 1
Key generation

Step 2
Key split

Step 3
Distribution key shares



Step 5:
Transfer ciphertext & key shares

Step 6:
Recompose keys
FAILS

Stap 7:
Record decryption
IMPOSSIBLE

Data holder A

encrypt

Data holder B

Data holder C

encrypt

Anon./pseudon. TTP

?

?

?

?

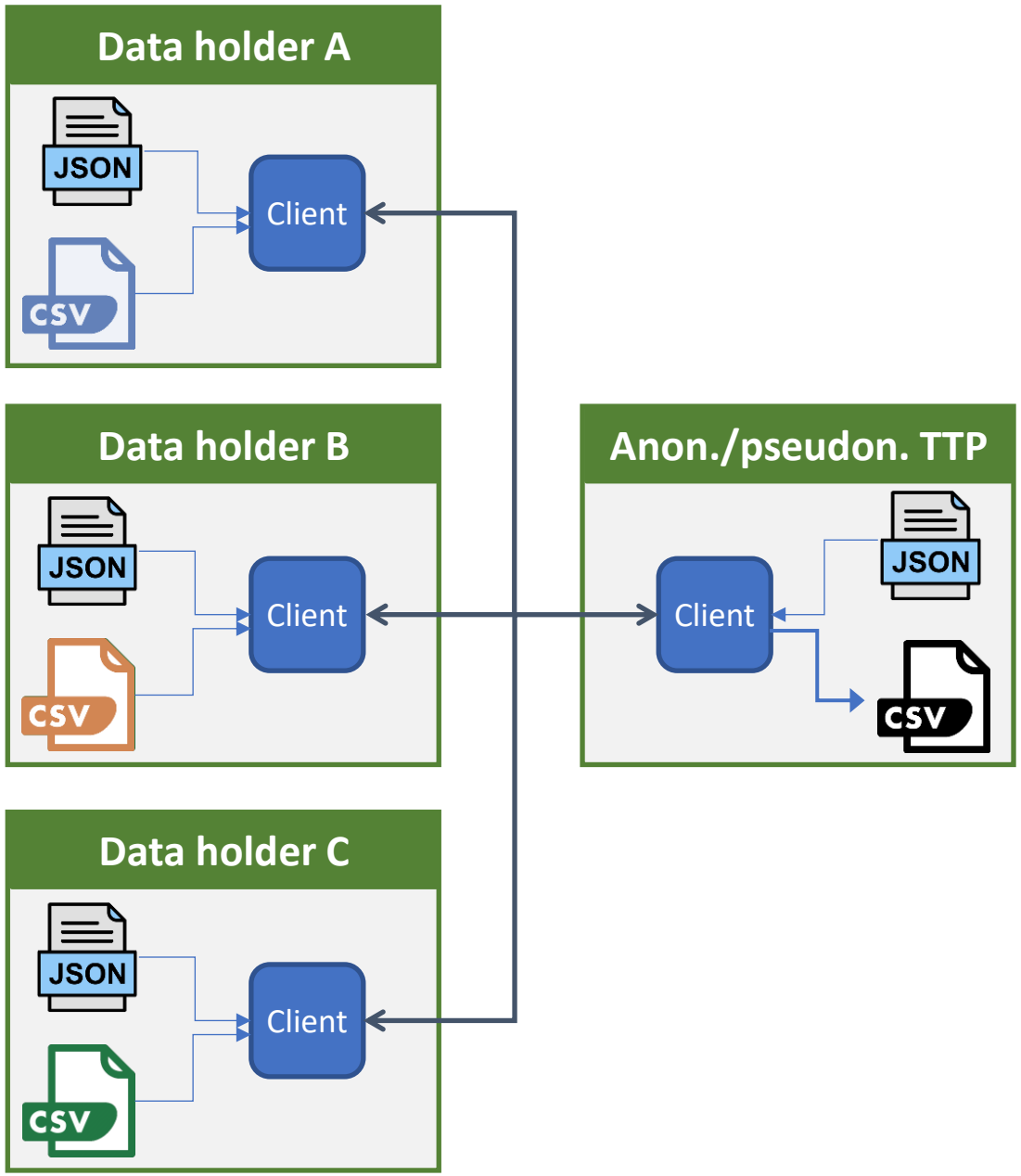
?

Oblivious Join

- Problem statement
- Concept
- **Proof-of-Concept**
- Required extensions
- Conclusion

Smals
ICT for society

Proof-of-Concept



client

Client

- Stand-alone → Usable without integration
- Command-line interface

JSON

Parameters

- Machine-readable (JSON)
- Contains all info required to execute protocol
- Created by coordinating party (**HDAB**)
- All parties use same parameters

CSV

Input files

- CSV file (’simple excel’)
- Contains all, potentially relevant, identified personal data
- Created by individual data holders (out of scope)

CSV

Output file

- CSV file
- Output by SPE client after protocol execution
- Contains minimal required de-identified and linked personal data

Test with fictional data

Extract input CSV

Data holder A (IMA-CIN)

60.01.03-231.73	Teriflunomide
60.01.03-562.33	Alemtuzumab
60.01.03-697.92	Glatiramer acetate
60.01.04-606.56	Interferon beta
60.01.04-681.78	Dimethyl fumarate
60.01.05-045.05	Teriflunomide
60.01.05-186.58	Tysabri
60.01.05-617.15	Ocrelizumab
60.01.05-715.14	Alemtuzumab

200 000 records

E.g. Citizens with MS

Extract input CSV

Data holder B (BCR)

60.01.03-782.07	Melanoma	3	G1
60.01.04-124.53	Colorectal	1	G3
60.01.04-345.26	Prostate	2	G2
60.01.04-562.03	Breast	2	G1
60.01.05-045.05	Lung	1	G3
60.01.05-893.30	Pancreas	4	G2
60.01.06-401.07	Breast	3	G1
60.01.06-696.03	Stomach	2	G1
60.01.07-203.78	Thyroid	1	G3

500 000 records

E.g. Citizens with cancer

Extract input CSV

Data holder C (FPS Health)

60.01.03-542.53	C
60.01.03-559.36	G
60.01.03-606.86	D
60.01.03-697.92	A
60.01.04-697.62	G
60.01.04-816.40	B
60.01.05-045.05	D
60.01.06-701.95	B
60.01.06-886.07	F

1 000 000 records

E.g. Citizens with high-risk profile

Extract output CSV

SPE

50 000 records

99338454821…	Teriflunomide	Lung	3	G1	F
12056965607…	Alemtuzumab	Cervix uteri	2	G2	B
15380767762…	Daclizumab	Pancreas	1	G2	A
15380767762…	Teriflunomide	Lung	1	G3	D
31309444464…	Ocrelizumab	Stomach	3	G1	C
99921347021…	Dimethyl fumarate	Breast	2	G2	H
69025938558…	Ofatumumab	Prostate	3	G3	A
38469942453…	Alemtuzumab	Melanoma	4	G1	E
18048091119…	Aubagio	Prostate	3	G3	D

Smals ICT for society

Test with fictional data

Extract input CSV

Data holder A (IMA-CIN)

60.01.03-231.73	Teriflunomide
60.01.03-562.33	Alemtuzumab
60.01.03-697.92	Glatiramer acetate
60.01.04-606.56	Interferon beta
60.01.04-681.78	Dimethyl fumarate
60.01.05-045.05	Teriflunomide
60.01.05-186.58	Tysabri
60.01.05-617.15	Ocrelizumab
60.01.05-715.14	Alemtuzumab

200 000 records
E.g. Citizens with MS

Extract input CSV

Data holder B (BCR)

60.01.03-782.07	Melanoma	3	G1
60.01.04-124.53	Colorectal	1	G3
60.01.04-345.26	Prostate	2	G2
60.01.04-562.03	Breast	2	G1
60.01.05-045.05	Lung	1	G3
60.01.05-893.30	Pancreas	4	G2
60.01.06-401.07	Breast	3	G1
60.01.06-696.03	Stomach	2	G1
60.01.07-203.78	Thyroid	1	G3

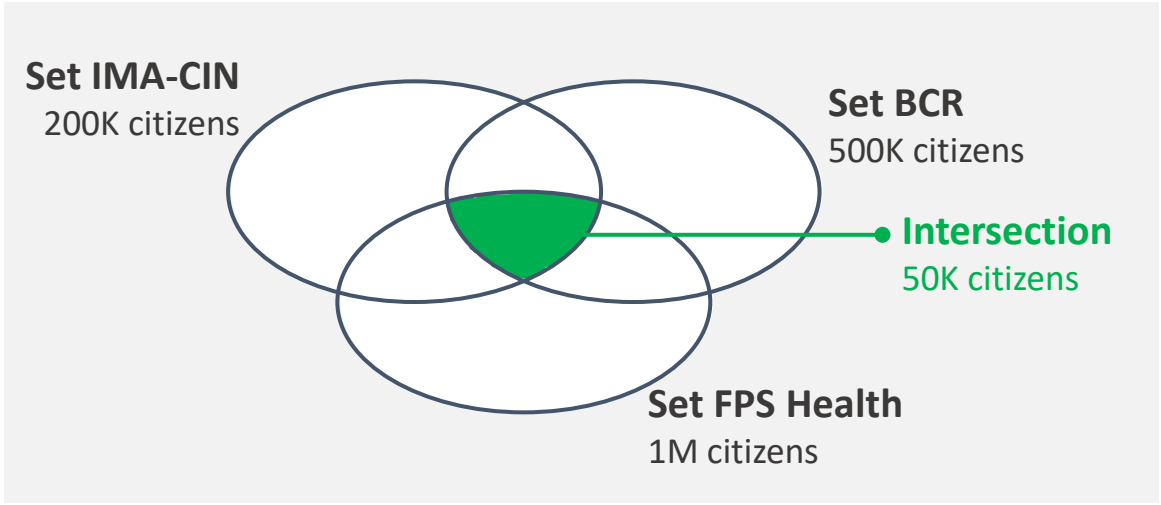
500 000 records
E.g. Citizens with cancer

Extract input CSV

Data holder C (FPS Health)

60.01.03-542.53	C
60.01.03-559.36	G
60.01.03-606.86	D
60.01.03-697.92	A
60.01.04-697.62	G
60.01.04-816.40	B
60.01.05-045.05	D
60.01.06-701.95	B
60.01.06-886.07	F

1 000 000 records
E.g. Citizens with high-risk profile



Optional lock

- If intersection too small (i.e. < 1000), no statistically relevant sample & privacy risks
- In that case, nothing can be linked or decrypted by the anon./pseudon. TTP.

Smals ICT for society

Experimental performance results

Distributed pseudonym calculation

Data holders provide citizen identifiers to their client.
Client SPE outputs the pseudonyms of citizens in intersections
(without MinNbRecords parameter)

	LAN						WAN					
	$\kappa = 128, \lambda = 40$			$\kappa = 256, \lambda = 80$			$\kappa = 128, \lambda = 40$			$\kappa = 256, \lambda = 80$		
m	$n = 3$	$n = 5$	$n = 7$	$n = 3$	$n = 5$	$n = 7$	$n = 3$	$n = 5$	$n = 7$	$n = 3$	$n = 5$	$n = 7$
2^{16}	3	6	10	3	6	10	4	7	12	5	8	13
2^{18}	11	21	37	11	21	35	12	23	38	15	26	45
2^{20}	45	86	156	47	88	156	50	91	170	58	108	176
2^{22}	205	390	667	212	387	667	215	406	688	244	439	757
2^{24}	983	1828	2926	1068	1825	/	1024	1896	2979	1141	2055	/

Table 2: Total runtime in seconds of the SIKA protocol, starting with loading CSV files by the data providers and finishing after the collector has stored the output CSV file. The intersection size is $m/2^4$.

3 data holders, each delivering 1 million records, 128 bit security → 50 seconds
7 data holders, each delivering 16,8 million records, 128 bit security → 50 minutes

LAN	Speed 1Gbs, latency 1ms
WAN	Speed: 150 Mbs, latency 30ms
m:	Number of records per data holder
κ, λ	Security parameters
n	Number of data holders

AWS EC2 r7i.8xlarge VMs, with 32 vCPUs powered by Intel Xeon Platinum 8588C processors at a frequency of 3.2Ghz and 256 GB RAM.
Ubuntu 24.04 LTS as operating system.

Better performance results expected soon!

Smals ICT for society

Oblivious Join

- Problem statement
- Concept
- Proof-of-Concept
- **Required extensions**
- Conclusion

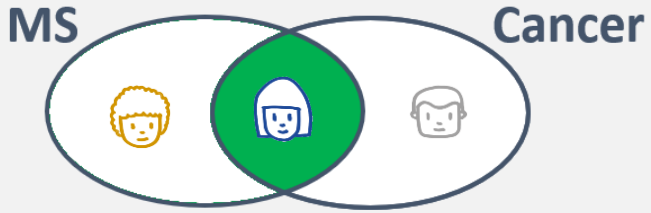
Smals
ICT for society

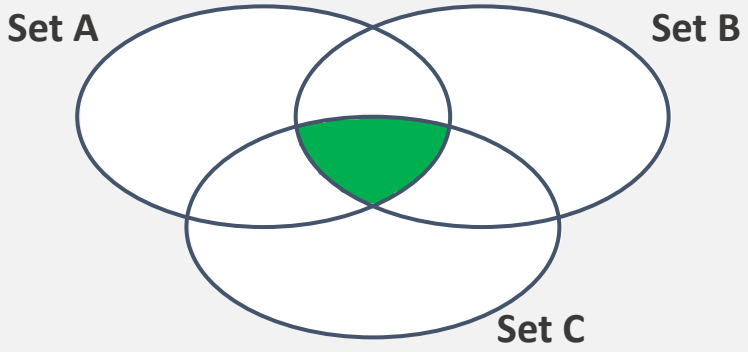
Extension 1: Selective disclosure

Personal data hidden from SPE

Personal data disclosed to SPE

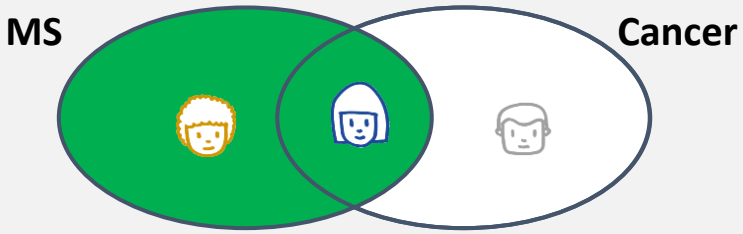
Currently, focus on intersection

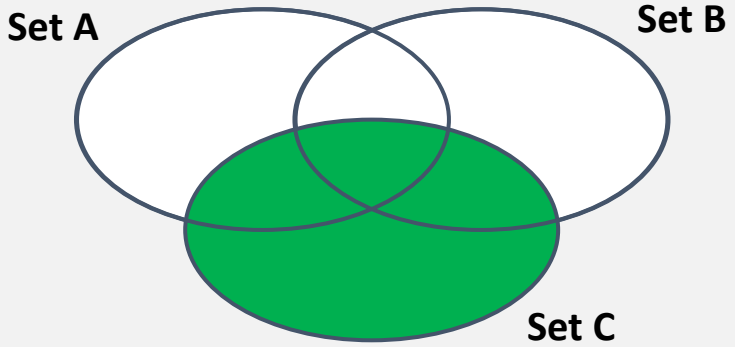


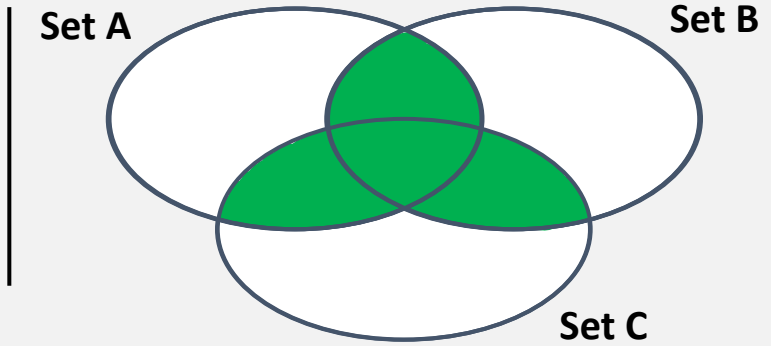


Multi-region support

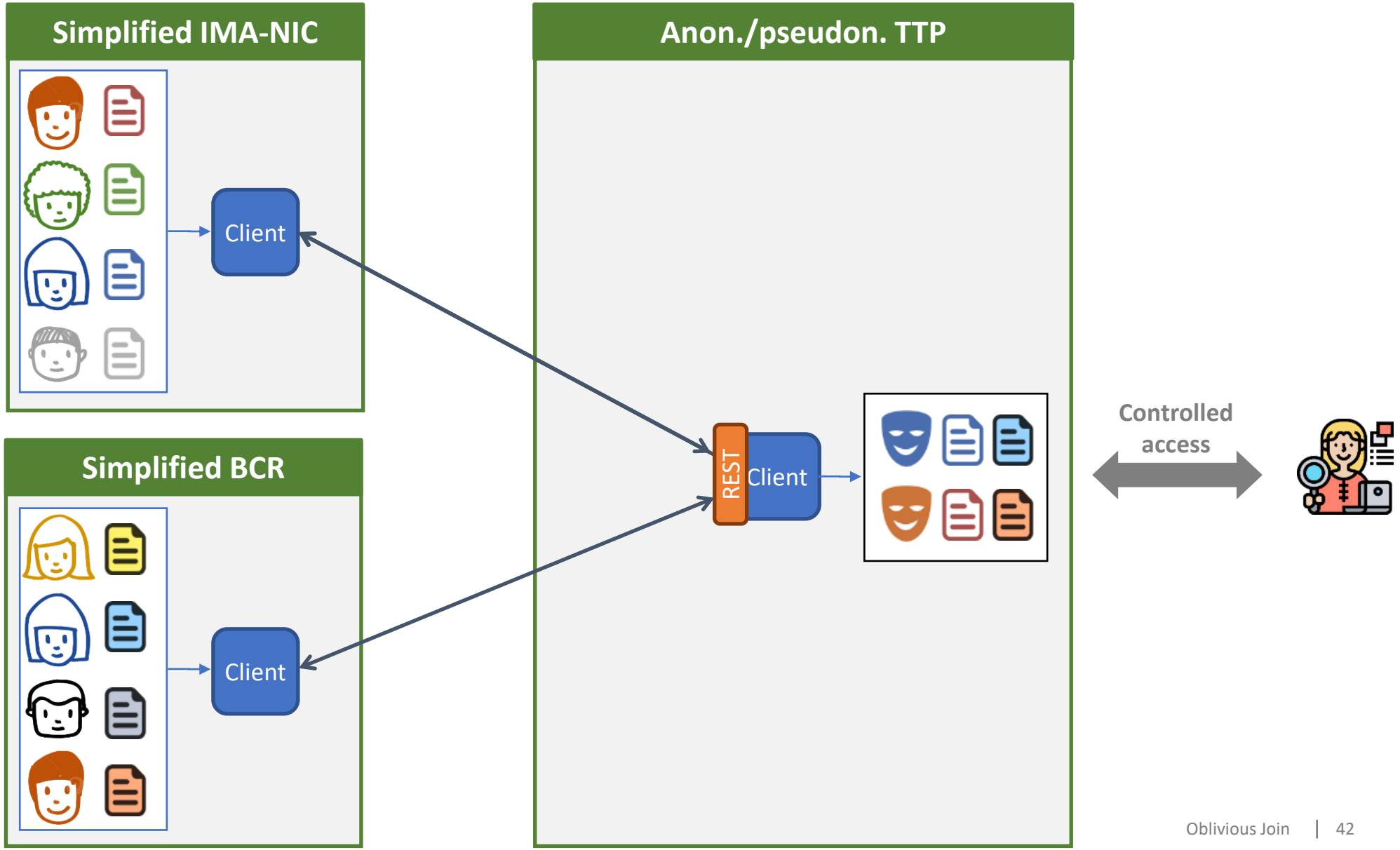
Conceptually possible, not implemented





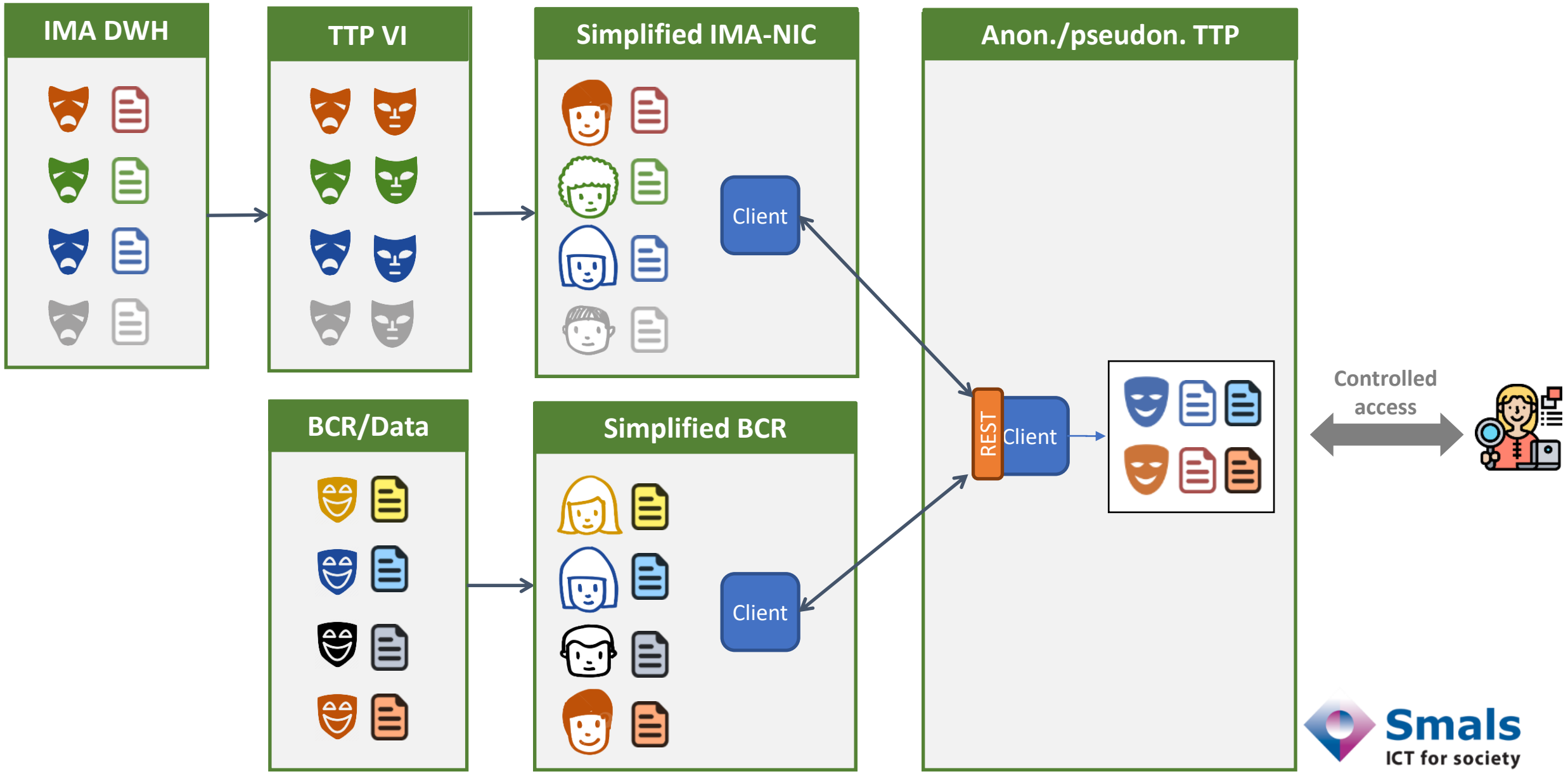


Extension 2: Simplified communication



Smals ICT for society

Extension 3: Pre-pseudonymized data



Smals
ICT for society

Extensions

Selective Disclosure

A/P TTP only able to link and decrypt data in pre-specified regions of the Venn diagram

Communication

- Data holders don't communicate with each other
- REST interface to communicate with A/P TTP.

Pre-pseudonymized data

- Data holders knows personal data under local pseudonyms
- Other entity/entities required to link data to citizen ids

Re-identification

- In case of a *significant finding* (e.g. high risk to develop disease)
- EHDS article 61(5)

Multi-batch transfer

- Data holders deliver multiple times data to the SPE
- A/P TTP should be able to link records about the same citizen

Conceptually possible, but not yet implemented

Other requirements?

Let me know!

Smals ICT for society

Oblivious Join

- Problem statement
- Concept
- Proof-of-Concept
- Required extensions
- **Conclusion**

Smals
ICT for society

Conclusion

Data minimization (1)

Distributed protocol

- No coding TTP that could learn anything
- Surprising that this is conceptually possible/efficient

Data minimization (2)

Data holders learn nothing

Data holders don't learn any statistical or personal data by participating in the code-and-link process, even if they collude

Additional data

- Distributed over multiple entities
- Both A/P TTP and data holder need to be involved in order to re-identify data.

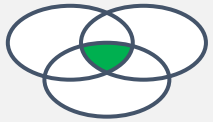
Data minimization (3)

Selective disclosure

Anon./pseudon. TTP only learns minimal required de-identified personal data.

Data minimization (4)

Extra lock



If result set too small, anon./pseudon. TTP cannot link or decrypt anything

Strengths

- Streamlined flows
- Fast technical execution
- Formal academic validation

Challenges

- Only passive interest → Still in research mode
- More complex logic (but no coding TTP)
- Extensions required
- Uncertainties regarding feasibility

Wrapping up

Smals
ICT for society

Comparison

	Blind pseudon. service eHealth	Oblivious Join
Status	**Life**	**Proof of Concept**
Selective disclosure	No Collector learns all received data	Yes A/P TTP only learns minimally required data
Performance	High Thousands of ops / sec. (shared by multiple services)	Extremely high Tens of thousands of ops / sec. (exclusive)
Required infrastructure	Blind pseudonymisation service	None
Integration efforts	Medium Logic integrated into client software relatively easily	Not required Oblivious Join client application runs stand-alone
Number of data holders	Many Practical limit by available capacity pseudon. service	Limited Tests with up to 7 data holders so far

Smals ICT for society

Advanced De-identification & linkage

of Personal Data originating from Multiple Sources for Secondary Use

eHealth Blind Pseudonymisation service

High-security service

Life

Oblivious Join

Distributed protocol

Experimental

Illustrates potential of current state of the art

Smals
ICT for society

REFERENCES

Articles

- **Introductie tot de nieuwe eHealth pseudonimiseringsdienst**
  https://www.smalsresearch.be/basisprincipes-voor-een-moderne-pseudonimiseringsdienst/
- **Introduction au nouveau service de pseudonymisation eHealth**
  https://www.smalsresearch.be/basisprincipes-voor-een-moderne-pseudonimiseringsdienst-2/
- **Kruisen van persoonsgegevens met eHealths blinde pseudonimiseringsdienst**
  https://www.smalsresearch.be/kruisen-van-persoonsgegevens-met-ehealths-blinde-pseudonimiseringsdienst-2/
- **Croisement des données personnelles avec le service de pseudonymisation à l'aveugle d'eHealth**
  https://www.smalsresearch.be/kruisen-van-persoonsgegevens-met-ehealths-blinde-pseudonimiseringsdienst/
- **Privacy-By-Design in the Belgian Public Sector - Pseudonymising and Joining Personal Data Fragmented Over Multiple Organizations**
  https://link.springer.com/chapter/10.1007/978-3-031-84748-6_6

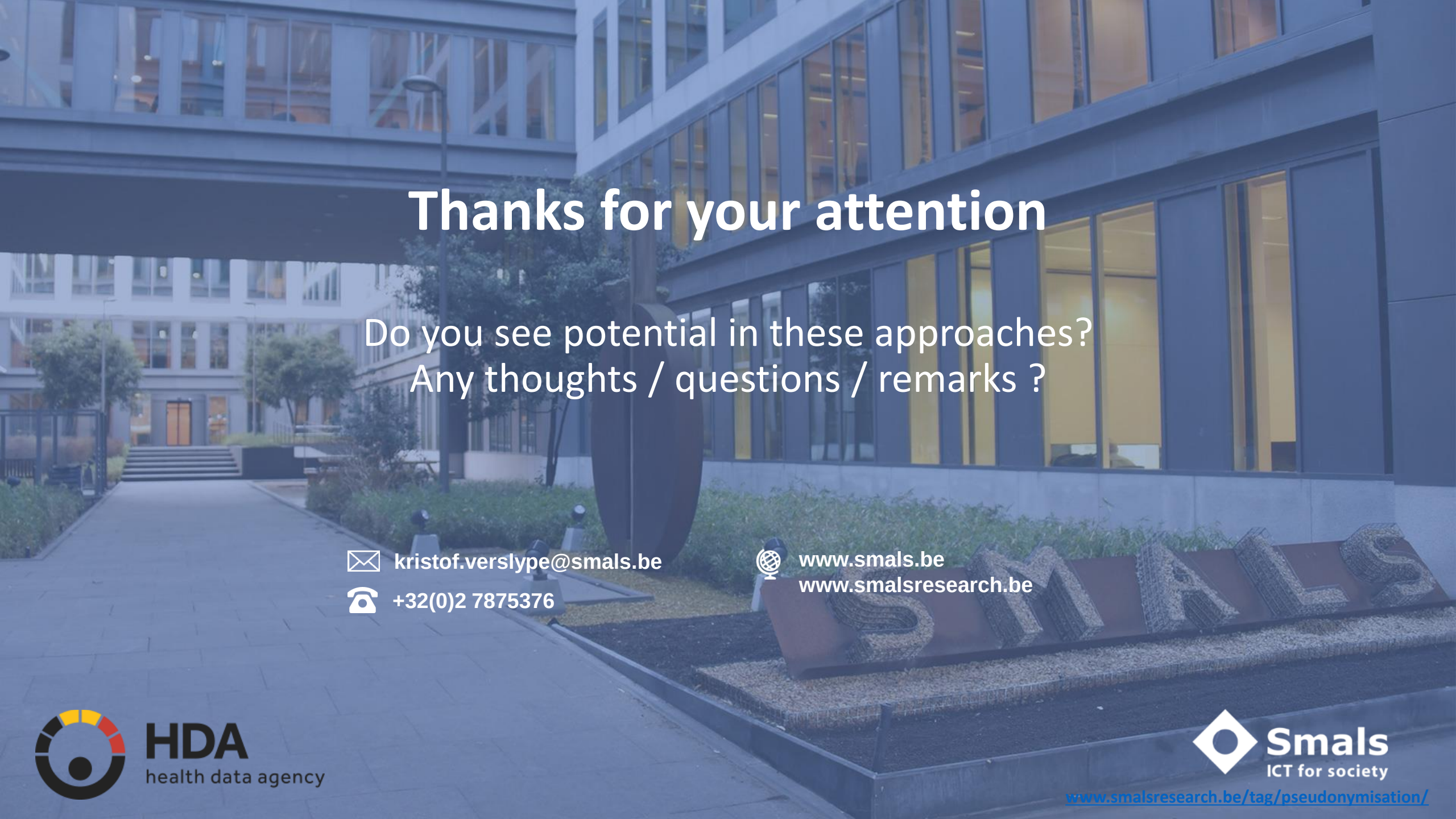
Talks

- **Devoxx 2024: Privacy in Practice - Smart Pseudonymisation**
  Slides: https://www.smalsresearch.be/download/presentations/20240606_webinar_pseudonimisatie_PRINT.pdf
  Recording: https://www.youtube.com/watch?v=-mx9vmdezL4&t=18s

Documentation

- **[DOC] eHealth Pseudonymisation service**
  https://portal.api.ehealth.fgov.be/api-details?apiId=eb8015c0-693b-4c4f-bab9-f671d35ddc15&managerId=1&Itemid=171&catalogModuleId=120
- **[DOC] Smals Pseudonymisation Helper**
  Helps to use the eHealth Pseudonymisation service, and how to integrate it into your projects
  https://github.com/smals-belgium/shared-pseudo-helper-java

Smals ICT for society



Thanks for your attention

Do you see potential in these approaches?
Any thoughts / questions / remarks ?

kristof.verslype@smals.be
+32(0)2 7875376

www.smals.be
www.smalsresearch.be

HDA
health data agency

Smals
ICT for society

www.smalsresearch.be/tag/pseudonymisation/

Images

Judy Dean
Creative Commons
https://flickr.com/photos/peterscherub/53152339550/

Aris Gionis
Creative Commons
Flickr

Smals ICT for society