



TEHDAS

**Towards
European
Health
Data
Space**

Milestone M7.4 Validation report on the proposed services and architecture and infrastructure solutions

21 April 2023

TEHDAS - Towards the European Health Data Space Joint Action

Start date of the project: 1.2.2021

Duration: 30 months

This project has been co-funded by the European Union's 3rd Health Programme (2014-2020) under Grant Agreement no 101035467.

Document information

Document authors	
Author	Partner
Jaakko Lähteenmäki	VTT Technical Research Centre of Finland
Juha Pajula	VTT Technical Research Centre of Finland
Juan Gonzalez-Garcia	IACS Aragon Health Sciences Institute
Helena Lodenius	CSC – IT Center for Science

Accepted in Project Steering Group on 28 March 2023.

Disclaimer

The content of this deliverable represents the views of the author(s) only and is his/her/their sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Consumers, Health, Agriculture and Food Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use of its contents.

Copyright Notice

Copyright © 2023 TEHDAS Consortium Partners. All rights reserved. For more information on the project, please see www.tehdas.eu.



Outline

- Objective
- Method
- Workshop plan
- Workshops
- Workshops practices
- Workshop 7: presentations, background, discussion
- Workshop 8: presentations, background, discussion
- Conclusions



Objective

- Collect comments of the WP7 permanent advisory group (WPAG 7) on the service, architecture and infrastructure options identified as a result of TEHDAS WP7 activity
- Provide input to the deliverable 7.2, which presents guidelines for EHDS in the area of metadata publication systems, data permit application management systems and secure processing environments

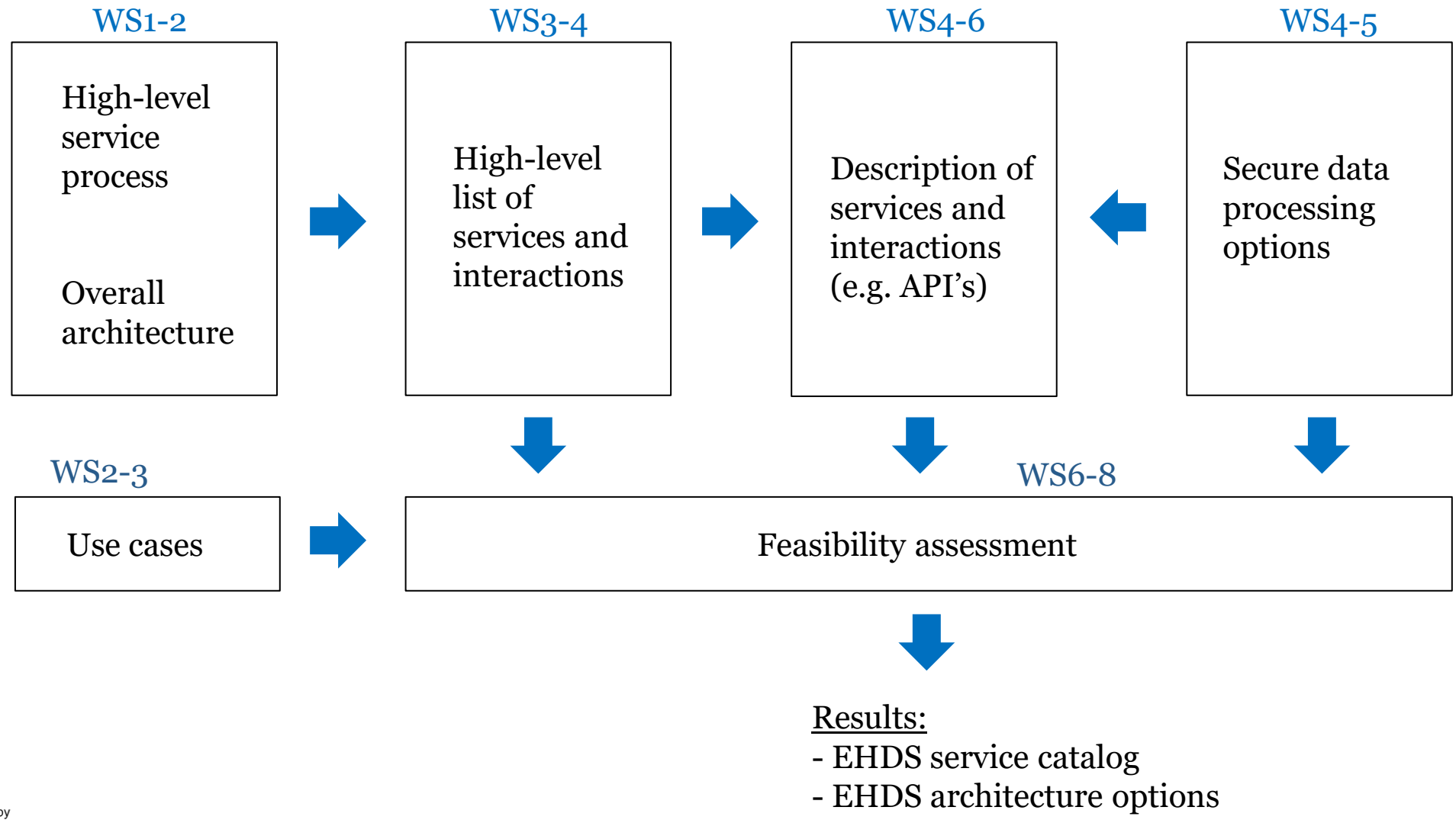


Method

- WPAG7 has been gathered for a series of workshops to collect views and opinions on EHDS architecture and services (see workshop topics next slide)



Workshop plan



Workshops

Reported in milestone 7.2 report:

- Workshop 1 (online), 18.05.2021, 40 participants
- Workshop 2 (online), 22.06.2021, 46 participants
- Workshop 3 (online), 14.09.2021, 39 participants

Reported in milestone 7.3 report:

- Workshop 4 (online), 7.12.2021, 38 participants
- Workshop 5 (online), 15.02.2022, 36 participants
- Workshop 6 (online), 10.05.2022, 39 participants

Workshops reported in this milestone 7.4 report:

- Workshop 7 (online), 29.11.2022, 52 participants
- Workshop 8 (online), 28.2.2023, 46 participants

Workshop practices

- In all workshops, options and approaches for EHDS architecture and services have been presented to the meeting participants
- The materials have been provided to the participants before the workshop to ensure time to prepare for the workshop
- The participants have been invited to comment the presented architectural approaches:
 - orally or in chat during the workshop (Teams meeting)
 - on whiteboard (Jamboard) during or before the workshop
- Inputs have been recorded in minutes shared after the workshop to all participants

Workshop 7

Presentations

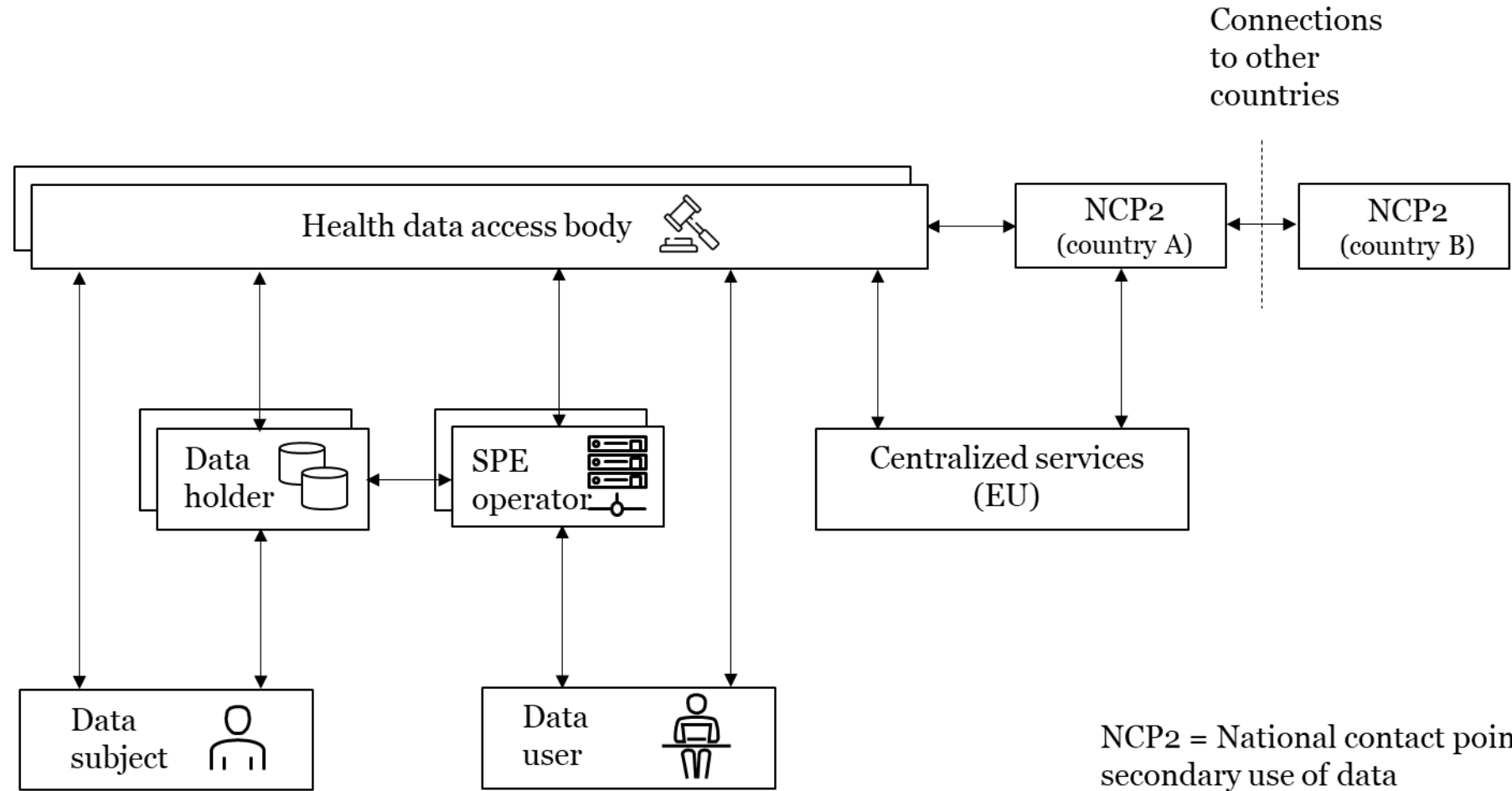
- Secure processing environments – questionnaire/interview results, Juan González, IACS
- Findata's secure processing environment services, Heikki Lanu, Findata
- SPE functionalities/guidelines for EHDS, groupwork introduction Jaakko Lähteenmäki, VTT

Workshop materials and minutes:

https://drive.google.com/drive/folders/1MsyK4jBTJDR37UVdXmK02wKkvCqqLNBm?usp=share_link

Jamboard with comments ([link](#))

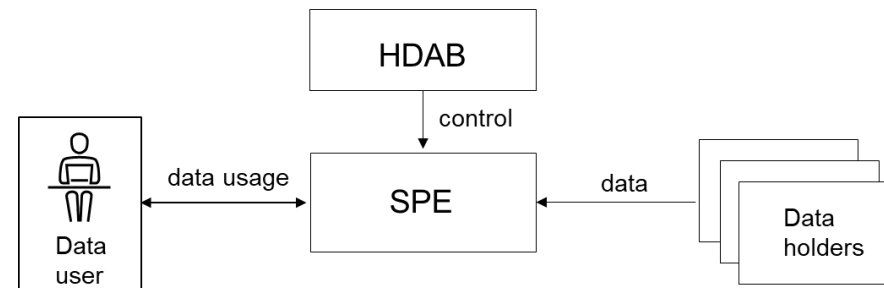
WS 7 – background: Architecture aligned with EHDS legal proposal



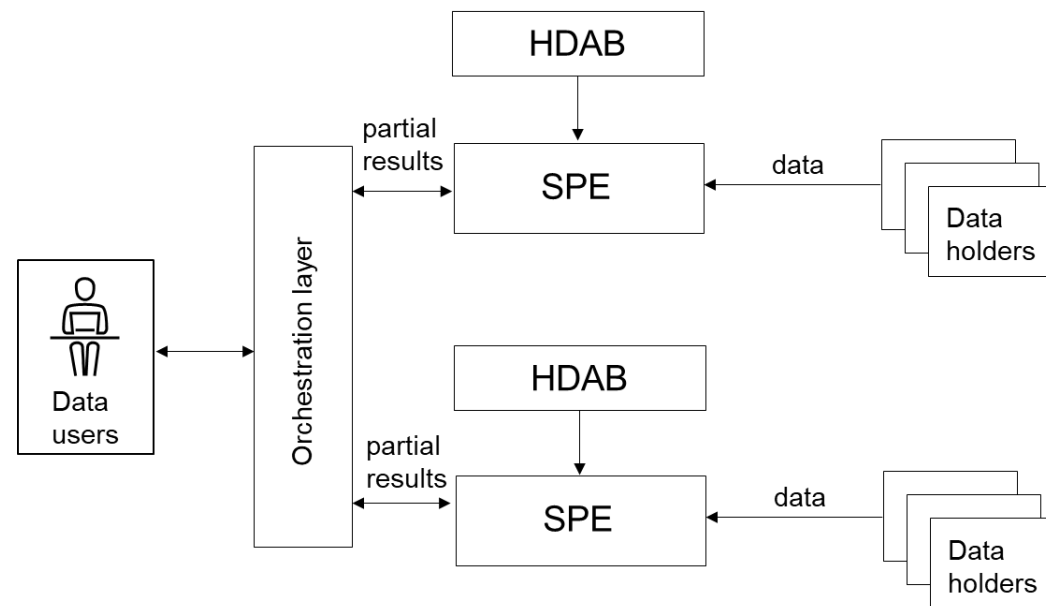
NCP2 = National contact point for secondary use of data
HDAB = Health data access body

WS 7 – background: Identified approaches for data processing

- **Traditional:** all data is transferred into one SPE for processing



- **Federated:** data is processed in several SPE's and final results are integrated from the partial results



WS 7 – discussion / main points (1)

Traditional vs. federated

- Both approaches (traditional/centralized and federated/distributed) need to be supported
- Several SPE's based on the traditional/centralized approach already exist and can be used as a model
- Federated approach maybe used to increase data protection, in the context of very large data sets (e.g. genomic) or to balance computing load between SPE's

Authentication/authorization

- Common approach needed to authenticate and authorize data users needed
- Also common approach to identify data sources is needed

Who can offer SPE services

- Any capable organisation (incl. private companies) should be allowed in principle, but they need to be audited against common requirements/criteria
- Several SPE's per country should be available
- Centralized SPE will be helpful for countries without own SPE's

WS 7 – discussion / main points (2)

Data transfer to SPE from data holder

- Manual process in current implementations, but API-based approach should be aimed at
- API-based automatic process requires semantic and syntactic interoperability
- Verify integrity and origin of data by electronic signatures
- ”Ready to use datasets” and metadata to make data usage easier

Data preparation for use

- Definition of standard practices and related responsibilities (e.g. pseudonymization and data quality assessment) are needed
- Data quality parameters to be provided along with data to the user

Upload of user contents

- For example the following contents are needed depending on the case: users own software, pre-trained models, users own data (to be linked with registers), code tables and other support materials)

WS 7 – discussion / main points (3)

Trustworthiness of user-originated content

- A common approach is needed
- Maintain a whitelist of scripts and tools which can be safely uploaded and used
- Malware scanning in the environment may be sufficient
- Quarantine environment
- There should be trust towards the data user
- Automatic (AI-based) audit of content intended to be uploaded

Tools

- The environment should have default set of tools with the possibility to ask for other tools needed in a particular project
- Common requirements for the default set are needed
- A "European consolidation center" could maintain a list of recommended tools

WS 7 – discussion / main points (4)

Computing power and disk space

- Different types of SPE's are needed to meet variable requirements of data users
- Cloud-based SPE's should be enabled, but special measures for safety needed

Download (export) of contents

- Control of anonymity should be retained for HDAB or SPE provider, but data user is also responsible for ensuring anonymity
- Guidelines and rules are needed for anonymity verification (also to be noted that it is not always even possible – e.g. in case of rare diseases)
- AI models involve specific privacy risks → specific methods needed
- Various privacy enhancing methods are available and should be used
- Whitelisted algorithms pre-confirmed to produce anonymous results is one solution although not relevant in all cases

WS 7 – discussion / main points (5)

Reaccess

- A mechanism is needed to archive data sets and results so that they can be reaccessed, for example to verify results afterwards when needed

Specific needs for federated processing

- A common data model (e.g. OMOP) is needed so that same algorithms can be run in parallel in different SPE's (semantic and technical interoperability)
- An orchestration layer (e.g. "master SPE") is needed to execute the processing
- The orchestration layer should "trusted" (e.g. maintained by an HDAB or EU) is needed if anonymity of partial results can not be ensured
- In case anonymity of partial results can be ensured the orchestration layer can be hosted by the data user or a third party (simpler approach)

Workshop 8

Presentations

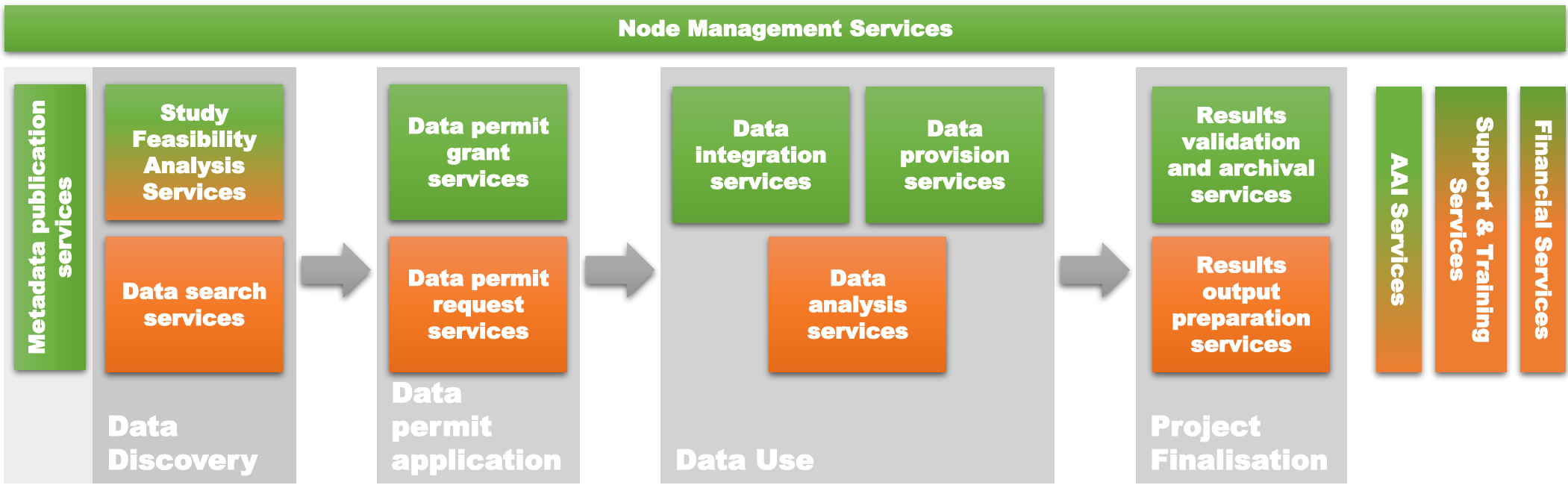
- TEHDAS project and EHDS status update, Tapani Piha, Sitra
- TEHDAS WP7 update, Milestone 7.6 overview and groupwork introductions, Juan González, IACS

Workshop materials and minutes:

https://drive.google.com/drive/folders/1zO2e0UnwYfiW53i8z4myHbreVeiWtJFb?usp=share_link

Jamboard with comments ([link](#))

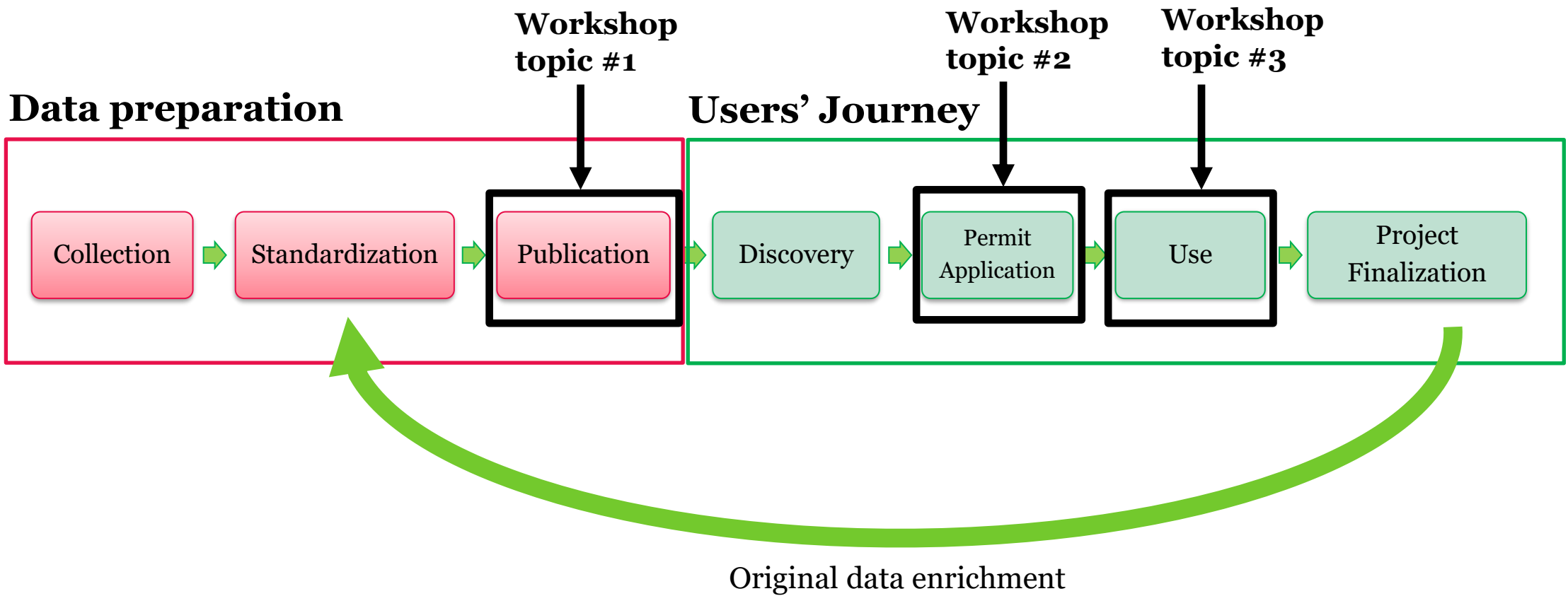
WS 8 – background: Current version of the data user’s journey



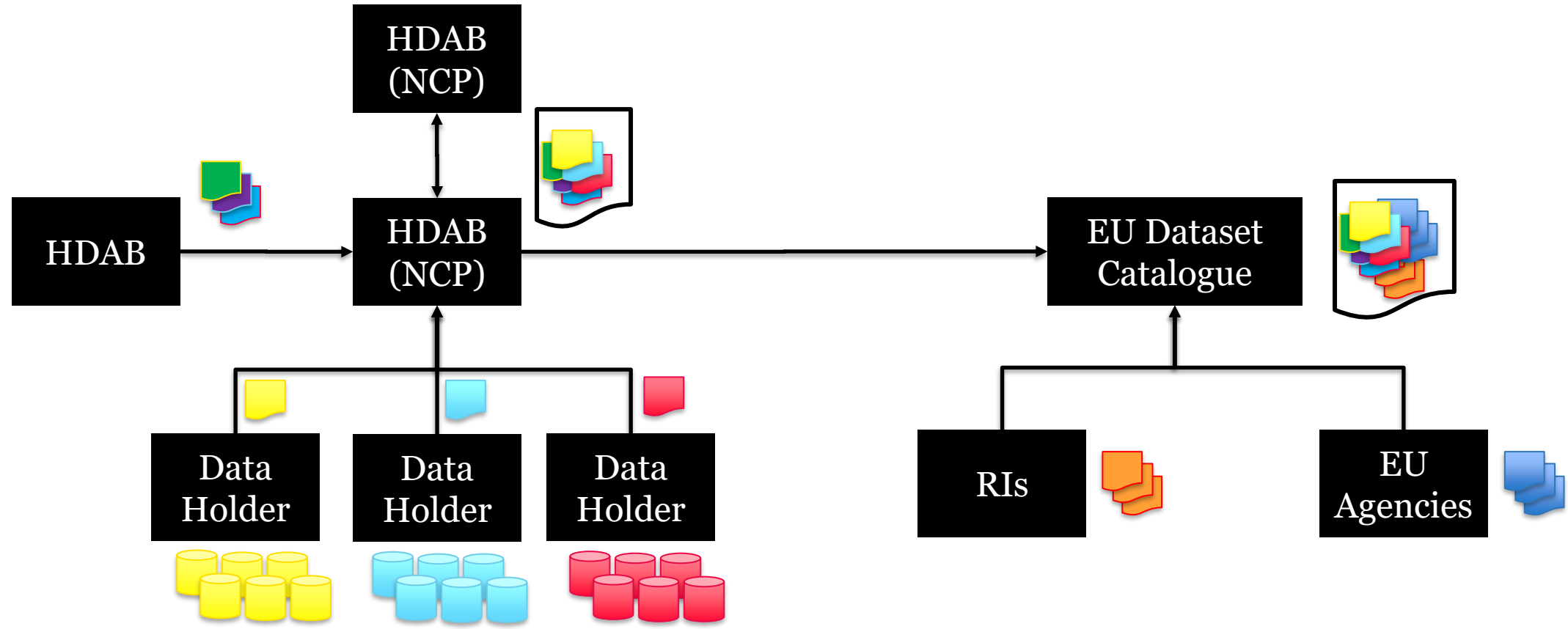
 **Infrastructure services**



 **Services with data user interaction**

WS 8 – background: Data lifecycle



WS 8 – Groupwork #1: National datasets catalogue systems - architecture

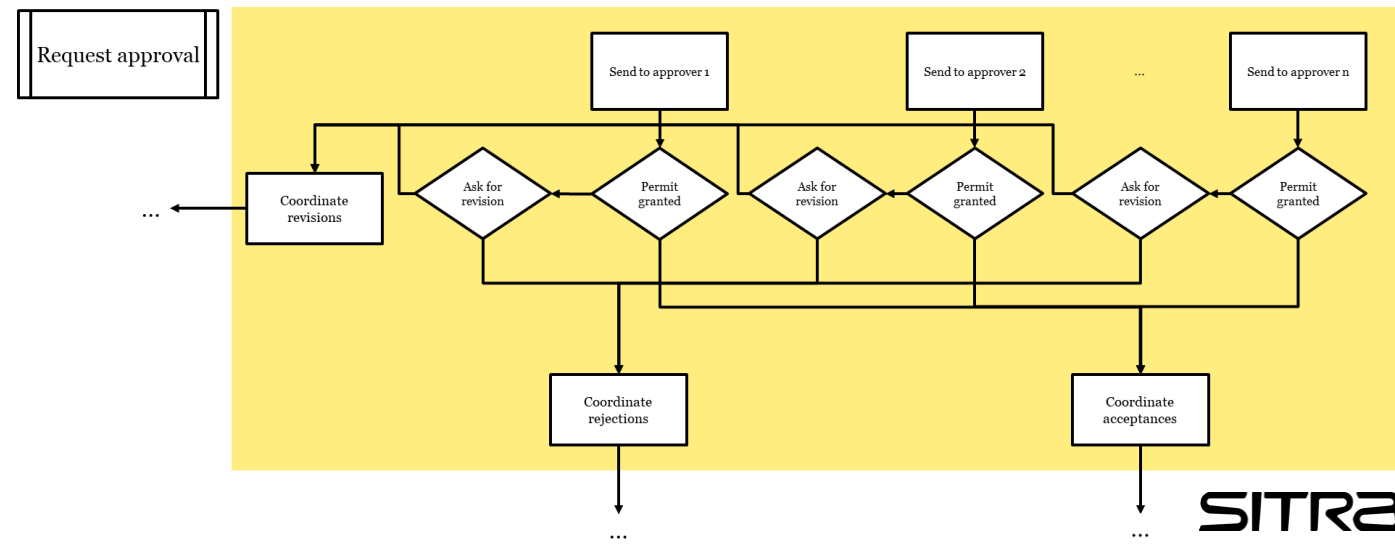
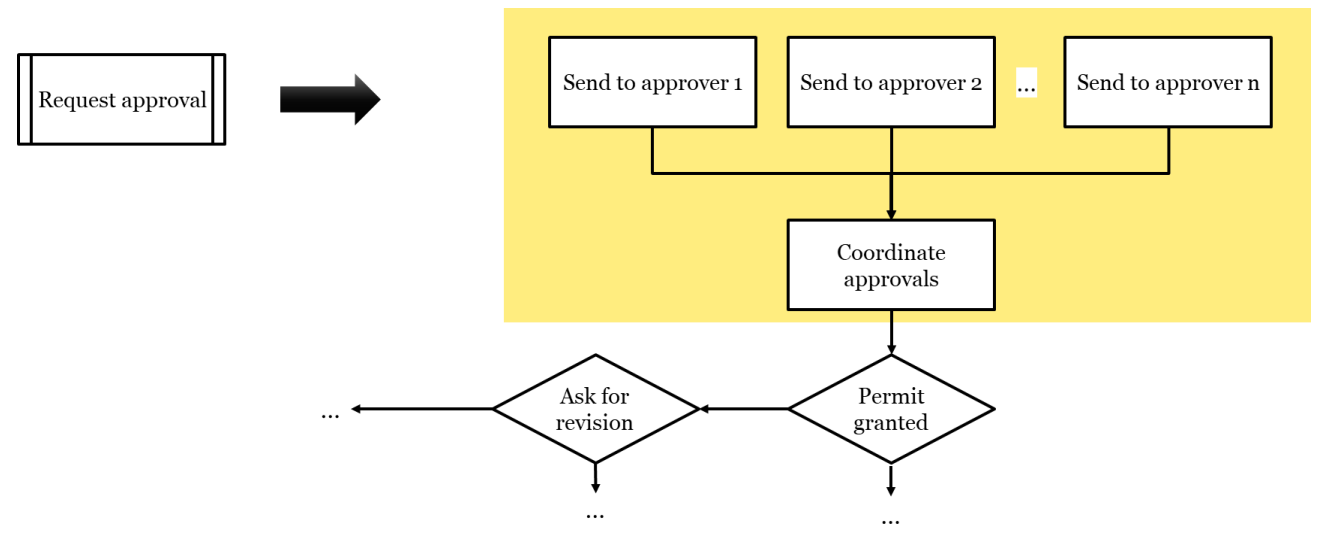
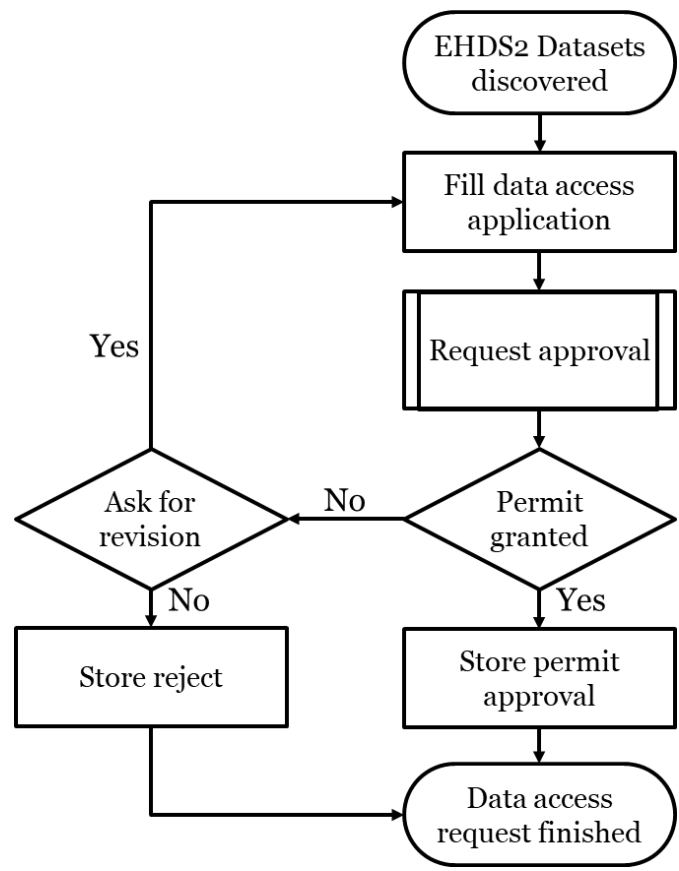


 Dataset
 Data catalogue

WS 8 – Groupwork #1: national datasets catalogue systems – main discussion points

- Existing metadata standards, catalogs and data discovery solutions (e.g. Beacon) should be used
- For publishing metadata, most support was given to updating initiation order: data holder → HDAB (→ HDAB/NCP) → EU.
- The role of research infrastructures seems to need clarification. They might be better to interact with HDABs instead of EU-level catalogue
- HDAB-HDAB synchronization there are different opinions. Some do not consider it necessary, because synchronization at EU-level is anyway available. Some consider it still to be important e.g. to maximize resilience.
- Data search services. Most comments favour the approach where the data user interacts with the EU-level data discovery service (instead of national services)
- API-based catalog search and reproducibility of searches should be supported
- The potential leak of personal information in the context of data discovery should be taken into account

WS 8 – Groupwork #2: Data access requests management systems



WS 8 – Groupwork #2: Data access requests management systems – main discussion points

- Active dialogue between data user and authority during the application process should be supported
- Several approvers may be involved → they need to be regularly updated with the acceptance process status
- Separate ethical committee approval maybe needed in some cases although not explicitly visible in the process diagrams
- Mutual recognition should be applied to enhance efficiency
- More elaboration of the process needed for: multi-center study support and handling of conflicting approvals
- Most supported combination of responsibilities: (1) centralized approach for data permit request services, (2) distributed approach for the approval process

WS 8 – Groupwork #3: Secure processing environments – main discussion points (1)

- The discussion in WS8 mostly confirmed the conclusions of WS7. Some new points were raised and are listed below
- Who integrates data to SPE? Data integration takes place before data user accesses data. In the typical case several data holders are involved so that HDAB is a natural integrator.
- Can data users process their own data if they are also a data holder? Yes, but they need an SPE.
- Two types of “data preparation” should be taken into account: (1) overall data curation activities at data holder level, (2) preparation of data to be used in a specific project
- It was commented that version control tools are needed in the SPE to support the process of developing and executing data analysis scripts by the data user (it should also be possible to export version control information from SPE)
- Some federated processing architectures would require SPE-SPE connections. More analysis (e.g. from security perspective) is needed concerning such connections.

WS 8 – Groupwork #3: Secure processing environments – main discussion points (2)

- Container-based analytics tools were considered important to be supported by the SPE's.
- Automation of anonymization process and results is important. Guidelines are needed. Criteria for anonymity are needed.
- The importance of training was highlighted. There should be mechanisms in place ensuring that all data user representatives accessing SPE pass a training program before getting access to data.
- The need for more accurate regulation and guidelines for SPE security requirements was raised.

Conclusions

- WPAG7 workshops 7 and 8 raised active discussion among the external experts and TEHDAS WP7 representatives
- The workshops were focused in three phases of the data users' journey based on earlier feedback by the EC:
 - National datasets catalogues management systems
 - Data access requests management systems
 - Secure Processing Environments
- The discussion confirmed a number of high-level requirements and options for the architecture and services related to the three process phases listed above
- The results of the workshops will be used as input to the final deliverable of TEHDAS WP7: "Options for the services and services' architecture and infrastructure for secondary use of data in the EHDS" (D7.2)
- Workshops 7 and 8 were the last ones in the series of WPAG7 workshops. The WP7 task leader team wishes to thank all external experts and TEHDAS WP7 representatives for active participation and contributions in the workshops