*Caring is Sharing – Exploiting the Value in Data for Health and Innovation*
*M. Hägglund et al. (Eds.)*

*doi:10.3233/SHTI230066*

# No Transfer Without Validation: A Data Sharing Framework Use Case

Hauke HUND[a,1], Reto WETTSTEIN[b], Christopher HAMPF[c], Martin BIALKE[c], Maximilian KURSCHEIDT[a], Simon T SCHWEIZER[a], Christoph ZILSKE[a], Simon MÖDINGER[a] and Christian FEGELER[a]
[a] *GECKO Institute, Heilbronn University of Applied Sciences, Heilbronn, Germany*
[b] *Institute for Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany*
[c] *Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany*

**Abstract.** Availability and accessibility are important preconditions for using real-world patient data across organizations. To facilitate and enable the analysis of data collected at a large number of independent healthcare providers, syntactic- and semantic uniformity need to be achieved and verified. With this paper, we present a data transfer process implemented using the Data Sharing Framework to ensure only valid and pseudonymized data is transferred to a central research repository and feedback on success or failure is provided. Our implementation is used within the CODEX project of the German Network University Medicine to validate COVID-19 datasets at patient enrolling organizations and securely transfer them as FHIR resources to a central repository.

**Keywords.** Validation, Data Sharing, Framework, Open Source, FHIR, BPMN

## 1. Introduction

To facilitate medical research at German university hospitals, the Network University Medicine (NUM) was created at the onset of the COVID-19 pandemic. One of the projects funded by the Federal Ministry of Education and Research (BMBF) within NUM is the COVID-19 Data Exchange Platform (CODEX) with the goal of harmonizing, collecting, distributing and analyzing real-world patient data.

While the usage of real-world data promises new insides for clinical research, it also brings its own challenges and limitations [1,2], with availability and accessibility of the data being important preconditions. Within the German Medical Informatics Initiative (MII)[2], infrastructure components at university hospitals have been established to provide access to data for researchers, including harmonized data definitions, accessible ontologies and data integration centers, including use and access committees.

The conceptual approach of the CODEX project, building on the infrastructure of the MII, was described by Prokosch, et al. in [3]. With this paper we want to report on

[1] Corresponding Author: Hauke Hund, GECKO Institute, Heilbronn University of Applied Sciences, Max-Planck-Straße 39, 74081 Heilbronn, Germany; E-mail: hauke.hund@hs-heilbronn.de.
[2] https://www.medizininformatik-initiative.de/en/start (accessed 2023 March 4).

the data transfer process deployed, with a focus on data validation as well as feedback on success or failure of data transfers.

## 2. Methods

The architecture and data transfer process of the CODEX project were developed based on existing tools from the MII as well as requirements defined by the data protection guideline created within the CODEX project.

A process plugin for the Data Sharing Framework (DSF) was implemented to enable automatic patient data validation and transfer. The DSF allows distributed business processes to be executed across organizations, with processes modeled using BPMN 2.0 and data exchange using HL7 FHIR R4. The DSF consists of a FHIR server accessible from other organizations and a private Business Process Engine (BPE) to integrate local and remote systems [4].

The CODEX project utilizes the German Corona Consensus Dataset (GECCO) designed by Sass, et al. [5] with data transfers to a central research repository after informed consent [6].

Components from the HAPI FHIR library[3] were used to implement a client for the federated Trusted Third Party service at Greifswald University Hospital [7], the FHIR terminology server at Köln University Hospital [8], the central research repository's FHIR Bridge [9] and to generate *StructureDefinition* snapshots and validate FHIR resources at patient enrolling organizations.

## 3. Results

### *3.1. Requirements and Architecture*

Several requirements were defined for the CODEX project's data transfer architecture and process: GECCO data (MDAT) should be stored in FHIR servers only accessible from local networks. The transfer process needs to be able to send complete datasets or only modified resources. Only valid datasets should be transported, with resources considered valid, if they follow the FHIR implementation guide and terminologies defined by the GECCO dataset. Directly identifying information (IDAT) needs to be removed from FHIR resources before transport to the central repository.

Privacy-preserving record linkage based on one-way hashed IDAT must be performed across organizations with unique pseudonyms ($PSN_s$) for all enrolling organizations and the central repository ($PSN_t$). The enrolling organizations (Source) need to be hidden from the central repository (Target) to reduce the risk of patient re-identification. The datasets (MDAT) and error messages (Error) need to be encrypted during transport so that they can only be read by the central repository and the enrolling organization.

Figure 1 shows a generalized version of the data transfer architecture used within the CODEX project. Patient enrolling organizations are depicted as *Source* and the central repository as *Target*. Record linkage and generation of organization-specific pseudonyms are performed by the *federated Trusted Third Party* (fTTP), with the *Data*

[3] https://github.com/hapifhir/hapi-fhir (accessed 2023 March 4).

*Transfer Hub* (DTH) acting as a middle man to hide the sending organization from the receiving central repository and to enforce organization-specific pseudonyms.

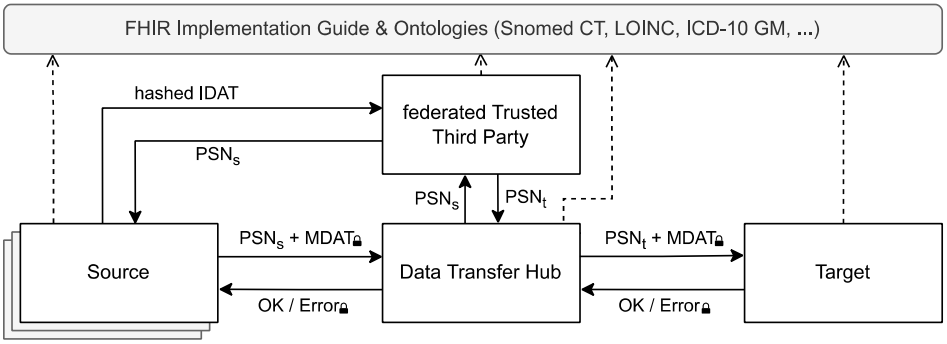


**Figure 1.** Generalized data sharing architecture.

All connections are based on the FHIR API with transport encrypted using TLS 1.3. Datasets are encrypted asymmetrically at the *Source*- using a public-key ($p$) from the *Target* organization to hide the content from the *DTH*. A hybrid cryptosystem (RSA+ AES) is used with two AES keys ($a$, $b$) generated at the *Source* organization for every data transfer, resulting in the data format $rsa_p(a) + aes_a(b + MDAT)$ for transferring encrypted MDAT and $aes_b(Error)$ for the return of encrypted validation errors.

### *3.2. Data Transfer Process*

At the *Source* organizations datasets are read, pseudonymized, directly identifying information removed, validated and encrypted. Encrypted datasets and pseudonyms ($PSN_s$) are send to the *DTH*. The *DTH* replaces the pseudonym ($PSN_s$) with a pseudonym for the *Target* organization ($PSN_t$) using the *fTTP*. At the *Target* organization, datasets are decrypted, validated and stored. Validation errors that may contain patient information are encrypted for the return path or an Ok message is send back.

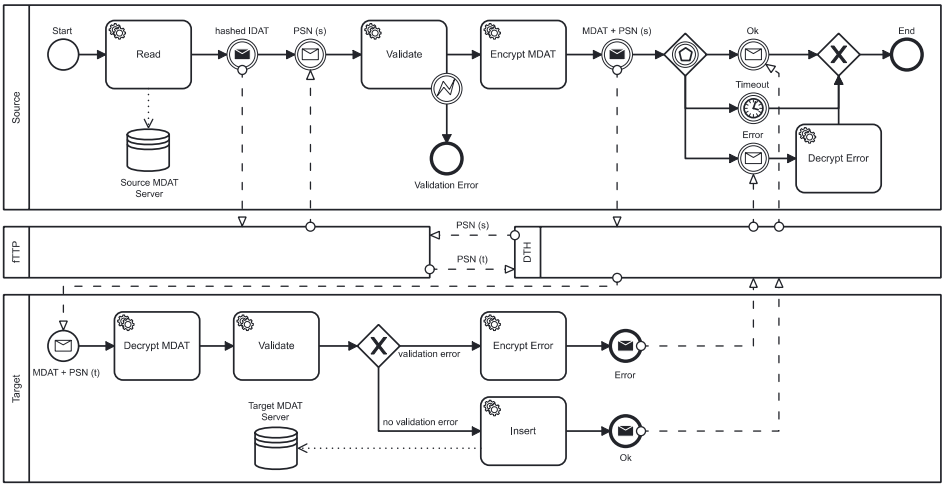


**Figure 2.** Data sharing process, simplified to improve readability.

Other technical errors at the *fTTP*, *DTH* or *Target* organizations that do not contain patient data are transported back without additional encryption, including high-level codes that enable monitoring. A simplified version of the data transfer process is depicted in Figure 2. The complete processes for the *Source*, *DTH* and *Target* organizations are implemented as a process plugin for the DSF, including executable BPMN models. The process plugin is available as open source under the Apache 2.0 license on GitHub[4].

### *3.3. Data Validation*

Datasets (MDAT) are transferred as FHIR *Bundle* resources of type *transaction*, with individual resources included using *conditional updates*. Before the *Bundle* is encrypted for transport at the *Source* organization, the content is validated against the GECCO FHIR implementation guide[5], with the process failing if validation errors occur.

Preparatory steps need to be performed during the startup of the DSF BPE to complete the resource validation offline without sending patient data to an external terminology- or validation server: First, the implementation guide package and dependent packages are downloaded[6]. Second, required metadata resources are extracted from the downloaded packages. Third, all *ValueSet* resources are *expanded* locally or, if necessary, using the central terminology server. Fourth, profile *snapshots* are calculated for all required *StructureDefinition* resources. And finally, all downloaded and generated resources are stored locally in a file system cache to improve subsequent BPE startups.

While inserting the transported FHIR resources into the central repository, a second validation against the implementation guide is executed, and validation steps across existing FHIR resources of the entire patient collective can be performed.

## 4. Discussion

Availability and accessibility are important preconditions for clinical research using real-world data. As part of the CODEX project, a data transfer and access platform was created with patients being enrolled at 34 university hospitals in Germany. To allow for the analysis of data collected across various new and existing legacy systems, a common data model was created and enforced by validation.

By validating data at the enrolling organizations and the central repository, we can provide fast feedback to data providers, minimize the number of invalid datasets being transported, and perform validation steps across the entire patient collective.

Although FHIR profiles provide a good mechanism for defining and validating data semantics, there is currently no simple mechanism for specifying resource properties, such as the patient's name or address, which may exist locally but should not be transferred to the central repository. Employing two different but related profiles, one allowing additional properties locally and another strictly enforcing data protection rules, could help but would require the maintenance of two sets of FHIR profiles.

With our process plugin for the DSF data for newly enrolled patients or updates for existing patients can be validated and transferred fully automatically. With the included return channel across the distributed system, we can inform data providers about errors

[4] https://github.com/num-codex/codex-processes-ap1 (accessed 2023 March 4).

[5] https://simplifier.net/forschungsnetzcovid-19 (German, accessed 2023 March 4).

[6] FHIR implementation guide packages are downloaded from https://packages.simplifier.net

during transport or data processing. Because the DSF can send emails when errors occur, operators do not need to monitor the system in production manually.

Using a process plugin for the DSF to validate and transport datasets allowed us to reuse existing infrastructure at German university hospitals and should minimize project-specific maintenance requirements in the future.

## 5. Conclusions

The data sharing architecture and processes implemented within the German Network University Medicine CODEX project enable the secure transport of real-world patient data using FHIR resources to conduct prospective studies with centralized data storage.

To analyze real-world patient data collected from a large number of independent healthcare providers, syntactic- and semantic uniformity need to be achieved and verified. To improve feedback to data collectors and to minimize the amount of invalid data being transported, data validation should be conducted at the centralized data storage site and also at every patient enrolling organization.

The described implementation enables validation of FHIR resources based on FHIR implementation guides and improves validation speeds while minimizing data transfers.

## Acknowledgements

The project is funded by the German Federal Ministry of Education and Research (BMBF, grant ids: 01ZZ1802E, 01ZZ1802A, 01ZZ1801M and 01KX2021).

## References

[1] Maissenhaelter BE, Woolmore AL, Schlag PM. Real-world evidence research based on big data. Onkologe (Berl). 2018;24(Suppl 2):91-98. doi: 10.1007/s00761-018-0358-3

[2] Alemayehu D, Ali R, Alvir JMJ, Cappelleri JC, et al. Examination of Data, Analytical Issues and Proposed Methods for Conducting Comparative Effectiveness Research Using “Real-World Data”. Journal of Managed Care Pharmacy. 2011;17(9 Supp A):1–37. doi: h10.18553/jmcp.2011.17.s9-a.1

[3] Prokosch H-U, Bahls T, Bialke M, Eils J, et al. The COVID-19 Data Exchange Platform of the German University Medicine. Stud Health Technol Inform. 2022 May;294:674-678. doi: 10.3233/SHTI220554

[4] Hund H, Wettstein R, Heidt CM, Fegeler C. Executing Distributed Healthcare and Research Processes – The HiGHmed Data Sharing Framework. Stud Health Technol Inform. 2021 May;278:126-133. doi: 10.3233/SHTI210060

[5] Sass J, Bartschke A, Lehne M, Essenwanger A, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Medical Informatics and Decision Making. 2020;20(1). doi: 10.1186/s12911-020-01374-w

[6] MII Consent Working Group. Patient Consent Form Template. 2020 November [cited 2023 March 4] Available from: https://www.medizininformatik-initiative.de/sites/default/files/2020-11/MII_WG-Consent_Patient-Consent-Form_v1.6d_engl-version.pdf

[7] Bahls T, Hampf C, Bialke M, Hoffmann W. Lösungsbaustein fTTP (federated Trusted Third Party) als ein Enabler für vernetzte medizinische Forschung mit dezentraler Datenhaltung. 2022 November [cited 2023 March 4] Available from: https://www.ths-greifswald.de/forscher/num/fttp-fact-sheet

[8] Mateen A, Tielsch-Nebel R, Okereke J. Terminology Server Services Concept. 2021 January [cited 2023 March 4] Available from: https://webstatic.uk-koeln.de/im/dwn/pboxx-pixelboxx-238948/terminology-server-services-concept-uniklinik-koeln.pdf

[9] EHRbase: FHIR Bridge. [cited 2023 March 4] Available from: https://github.com/ehrbase/fhir-bridge