German Medical Data Sciences 2021: Digital Medicine: Recognize - Understand - Heal
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210548

Data Sharing in Distributed Architectures – Concept and Implementation in HiGHmed

Reto WETTSTEIN[a,1], Hauke HUND[b], Christian FEGELER[b], and Oliver HEINZE[a]
[a] *Department Medical Information Systems, Heidelberg University Hospital, Germany*
[b] *GECKO Institute, Heilbronn University of Applied Sciences, Germany*

Abstract. Medical routine data has the potential to benefit research. However, transferring this data into a research context is difficult. For this reason Medical Data Integration Centers are being established in German university hospitals to consolidate data from primary information systems in a single location. But, small data-sets from one organization can be insufficient to answer a research question adequately. In order to obtain larger data-sets, attempts to merge and provide data-sets across institutional boundaries are made. Therefore, this paper proposes a possible process that can extract, merge, pseudonymize and provide distributed data-sets from several organizations conforming to privacy regulations. This process is executed according to the open standard BPMN 2.0, the underlying process data model is based on HL7 FHIR R4. The proposed solution is currently being deployed at eight university hospitals and one Trusted Third Party in the HiGHmed consortium.

Keywords. Data sharing, distributed processes, secondary use, BPMN, FHIR

1. Introduction

1.1. Background

The large volume of every day produced medical routine data can provide great potential for medical research [1]. However, this potential can only be harnessed if routine data can be transferred into a research context with reasonable effort (i.e., not manually) and in adequate time [2]. To achieve this objective, the HiGHmed consortium [3] has been established as part of the German Medical Informatics Initiative (MII) [4].

Each university hospital involved in HiGHmed is currently in the process of establishing a Medical Data Integration Center (MeDIC) based on open standards. The goal of each MeDIC is to create an infrastructure that consolidates data from primary medical information systems in a single repository, to facilitate the transfer of routine data into a research context and to improve data delivery to research projects [5, 6]. However, the data stored in a single MeDIC can be insufficient to adequately address a research question. Thus, a researcher must have the possibility to request data across several organizations. To provide this functionality, a concept of a distributed data

[1] Corresponding Author, Reto Wettstein, Department Medical Information Systems, Heidelberg University Hospital, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany; E-Mail: reto.wettstein@med.uni-heidelberg.de.

sharing process, aiming to extract, merge, pseudonymize and provide data stored in multiple distributed MeDICs has been developed, implemented and tested.

1.2. Requirements

In order to achieve a high degree of harmonization among the four consortia funded by the MII, the initiative's National Steering Group (NSG) has developed an umbrella process at the organizational level of cross-organizational data sharing [7]. This process is divided into three sub-processes. The first sub-process deals with feasibility queries and is already implemented at the executable level for HiGHmed [8]. The second sub-process is concerned with contract management. An executable implementation must still be realized. The third sub-process manages data provisioning to researchers and was used as the starting point for the executable implementation in HiGHmed. To allow for cross-consortia interoperability, a high degree of conformity between the executable and the umbrella process is required.

In addition to the requirements of the NSG, the data sharing process has to comply with the requirements of HiGHmed: A high degree of automation without any centralized components storing medical data long-term, a process implementation based on open standards that is executable on the HiGHmed Data Sharing Framework (DSF) and compliance with the European General Data Protection Regulation.

2. State of the Art

Based on the open standards Business Process Model and Notation (BPMN 2.0)[2] and Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR R4)[3], HiGHmed is developing the DSF [9]. It is designed to execute various cross-organizational processes within HiGHmed. The DSF consists of two components, a publicly available HL7 FHIR Endpoint and an internal Business Process Engine (BPE). BPMN 2.0 is used to model executable distributed processes which are orchestrated by the BPE. HL7 FHIR R4 is used to define the necessary communication messages as well as process input and output variables.

A distributed feasibility query process for calculating cross-organizational cohort sizes in HiGHmed has already been published for the DSF [8]. The findings from this process development, deployment and testing were used as a starting point to develop the process presented in this paper.

Other architectures and processes for cross-organizational data sharing already exist. With the Clinical Communication Platform, the German Cancer Consortium (DKTK) [10] is pursuing an approach based on centralized and decentralized components. Medical data-sets are stored in bridgeheads. Specific parameters from these data-sets are regularly sent to a central platform to enable simple feasibility queries. A decentralized search tool, where queries are evaluated directly in the bridgeheads and results returned to the central platform, allows for precise patient location and for collaboration requests in specific research projects.

Another cross-institutional infrastructure providing communication, security and terminology services to eleven hospitals and ten pharmaceutical companies across

[2] https://www.omg.org/spec/BPMN/2.0
[3] https://www.hl7.org/fhir/R4

Europe was built by the trans-European Electronic Health Records for Clinical Research project [11]. The infrastructure enables protocol feasibility and patient recruitment scenarios.

The Clinical Research Platform [12] of the German Center for Cardiovascular Research maintains a central data management system that is populated with data-sets of multiple organizations at regular intervals. This enables cross-organizational data sharing for research using a centralized approach.

To the best of our knowledge, no solution exists that meets all requirements of HiGHmed. Therefore, this paper presents a novel process approach for distributed extraction, merging, pseudonymization and provisioning of disease independent data-sets for research.

3. Concept

The proposed data sharing process has been developed using BMPN 2.0 models shown in Figures 1 and 2. These process models consist of four distinct user/ organization types, each shown in a pool. Each pool represents a sub-process. The pool at the very top represents the researcher. The pool underneath shows the coordinating organization where the researcher submits his request. This organization type orchestrates the data sharing process. The pool at the bottom shows all request-receiving organizations. They process the data sharing request locally and provide the extracted data-sets to a Trusted Third Party (TTP), represented by the second-lowest pool. The TTP does not hold any data permanently, but rather merges and pseudonymizes the data-sets, which are eventually transmitted to the coordinating organization and made available to the researcher.

The data sharing process consists of the following steps:

First, a researcher authenticates himself against a request management tool, provided locally at his organization. This tool enables the researcher to define the data usage request. It is regarded as an external service that is not part of the process implementation and conceptually also supports the contract management process. The proposed data sharing process assumes that a data usage contract exists and thus all organizational and legal issues regarding data sharing have been resolved. If this is the case, the request management tool sends a message trigger containing the released data usage request to the coordinating organization's BPE to start the data sharing process.

For automated execution the request has to contain inclusion and exclusion criteria represented as data extraction queries, for each cohort that should be analyzed in the research project. Furthermore, all participating organizations and the participating TTP should be part of the message trigger. Finally, the initial message trigger needs to be configured using the two parameters *consent checks* (Figure 1) and *record linkage* (Figure 2).

After receiving the initial message trigger, the coordinating organization generates a request specific AES encryption key and organization specific correlations keys. Following, the request is registered at the TTP and forwarded to all participating organizations. The TTP receives all correlation keys (i.e., one unique correlation key for each participating organization) but not the actual request or the encryption key. Only after a data-set is available for each correlation key or if the maximum execution time timer expires, the process can continue at the TTP with merging and pseudonymizing the

data-sets. Simultaneously, the participating organizations receive a message containing the whole data sharing request, their individual correlation key and the encryption key. They execute the queries supplied by the request for each cohort definition to retrieve the Medical Data (MDAT). Each row in the retrieved data-set is then supplemented by the local Patient Identifier (PID). In the case that the *consent check* parameter is activated, a further process step verifies that a rule exists in the Policy Decision Point (PDP) for each PID, that allows access to the MDAT. If the *record linkage* parameter is enabled, the PID is used to retrieve each patients' demographic data from the Master Patient Index (MPI) in order to generate a privacy-preserving Record Bloom Filter (RBF), which is then also attached to the corresponding data-set row. After that, the PID is replaced by a local first order pseudonym and the MDAT's are encrypted with the provided key. Last, the data-set is sent together with the organizations' correlation key to the TTP.

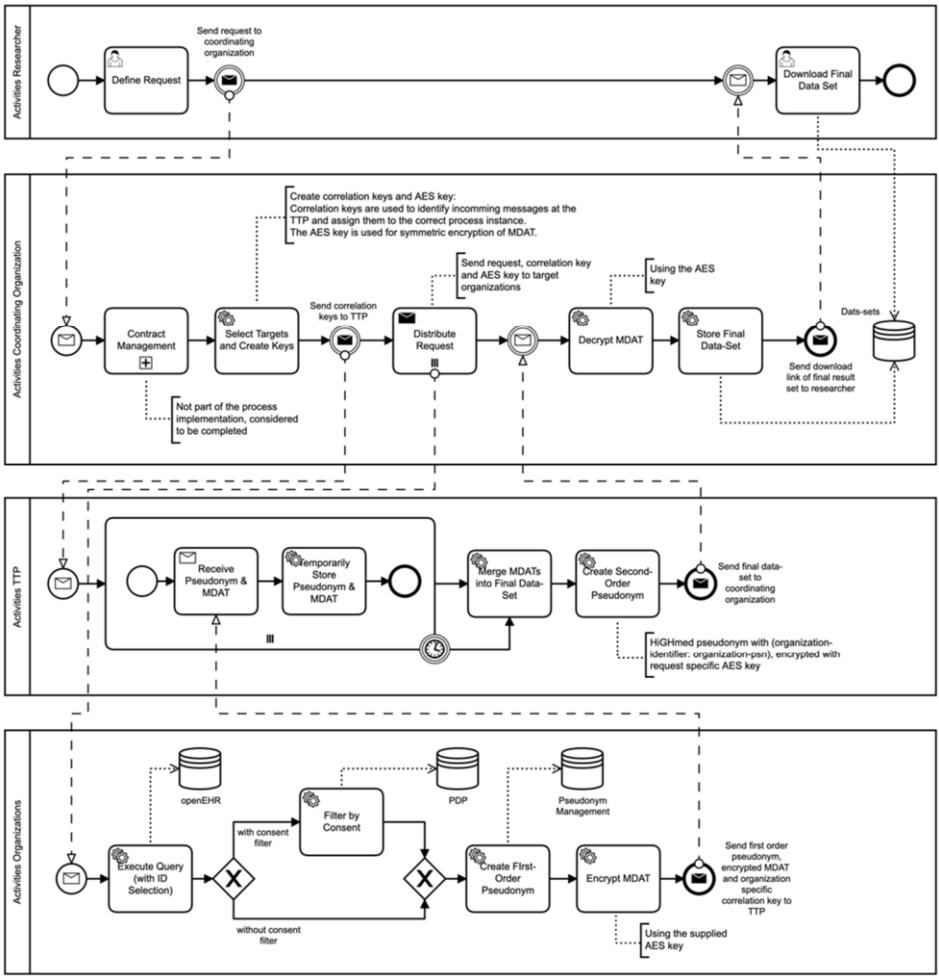

Figure 1. BPMN 2.0 model of the data sharing process using the *consent check* parameter.

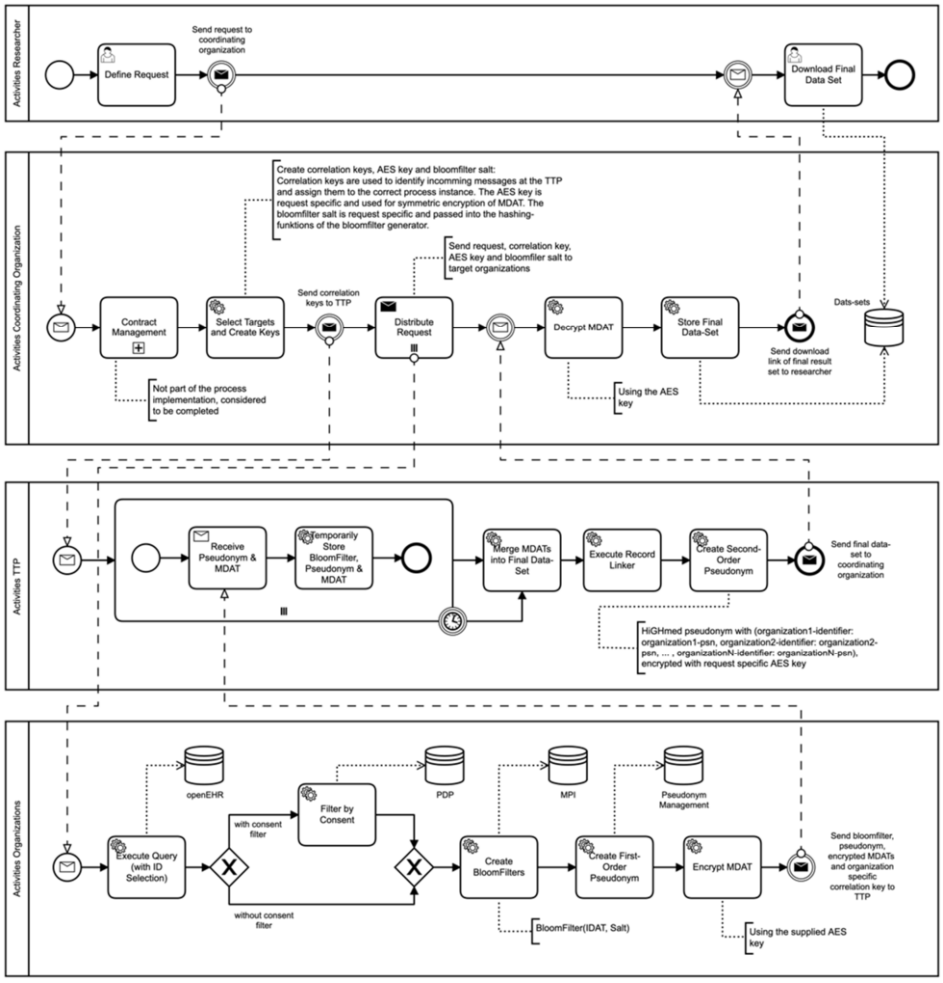

Figure 2. BPMN 2.0 model of the data sharing process using the *record linkage* parameter.

The TTP temporarily stores all data-sets. When a data-set is available for each correlation key or the timer expires, the TTP merges all available data-sets into a single one. In case record linkage was requested, the TTP executes a further process step by running a linkage algorithm, using the supplied RBFs to merge patients that are present in data-sets across multiple organizations. The TTP then replaces the first-order pseudonyms by calculating a second-order pseudonym and sends the final data-set to the coordinating organization.

Finally, the coordinating organization decrypts the MDAT, stores the final data-set in a data-mart and sends a temporary available download link to the researcher so that he or she can download and use the data.

4. Implementation

The data model of the data sharing process is designed using HL7 FHIR R4 resources and is based on the data model of the feasibility query process [8]. Similar to the feasibility query process, cohorts and their inclusion and exclusion criteria are modelled using the *Group* resource. The actual query of each cohort is stored using an extension. The query language depends on the data repository that can be addressed. For example, AQL or CQL queries can be transported and executed. The cohorts are bundled by the *ResearchStudy* resource. As in the feasibility query process, this resource contains extensions representing the participating organizations as well as the TTP. In addition, the *RelatedArtefact* field contains references to the request form and the data usage contract in PDF format as part of *Binary* resources as well as an additional reference to the results of a mandatory preceding feasibility query.

All messages between organizations are transmitted as *Task* resources and stored in the HL7 FHIR Endpoint of the receiving organization in order to be forwarded to the BPE for execution. A *Task* defines a requestor and a recipient organization, the sub-process to be executed, the message name of the corresponding BPMN-Event and the input values required for sub-process execution. A *Task* resources also indicates the current state of a sub-process and stores results or possible execution errors. For the exact handling of *Task* resources during DSF process execution, the reader is referred to [8].

The calculation of the first-order pseudonyms must be ensured by each participating organization itself and can therefore deviate depending on each organization's preconditions. The second-order pseudonyms are calculated by the TTP. If record linkage has been performed, the second-order pseudonym is defined by a list of *organization-identifier:organization-psn-first-order* entries, with one entry for each organization in which the patient has been treated. If no record linkage was performed, the second-order pseudonym contains only one entry in the list. To prevent a researcher from identifying the first-order pseudonyms, the second-order pseudonym is encrypted using a symmetric AES key only known to the TTP. For a detailed description on how the RBF based linkage algorithm and the calculation of the second-order pseudonyms are implemented, we like to refer the reader to [13].

The data sharing process was implemented using an open source approach in the Java programming language and can be found together with the corresponding specifications in HL7 FHIR R4 and BPMN 2.0 on GitHub[4,5].

5. Discussion

Based on the specifications of the data sharing umbrella process provided by the NSG and the experience gained from developing a decentralized feasibility query process, an automated process for data sharing could be developed. A high level of automation was possible by excluding process steps solving organizational and legal issues regarding data sharing as well as assuming that a data usage contract is in place before the process starts. Most likely, manual verification steps will be added to the process in the future before releasing the final data-set to the researcher.

[4] https://github.com/highmed/highmed-dsf/
[5] https://github.com/highmed/highmed-processes/

The concept of the process follows, compared to other tools providing cross-organizational functionality, a fully decentralized approach without any central components for persistent data storage. This would allow any process role to be taken by any organization, including the one of the TTP. Due to the decentralized approach, medical data only leaves an organization if an actual research project has been approved.

The process is based on the open standards HL7 FHIR R4 and BPMN 2.0, resulting in a reference implementation with great flexibility, where each part is individually interchangeable using other implementations than the provided one. Consequently, the process can be adapted to an organization's local requirements. The selected approach also makes the process independent of the data model in which the medical data is stored and thus different query languages such as AQL and CQL can be supported. The two parameters *consent check* and *record linking* allow tailoring of the process according to the legal foundation and the cohort sizes of the research project. By extending the feasibility query process' data model, a successful feasibility query request can be converted into a data sharing request.

Two-level pseudonymization is applied during data aggregation. Each organization is responsible for the first-order pseudonym. The second-order pseudonym is an encrypted string representing a list of local first-order pseudonyms. This allows the TTP to work economically by only storing the encryption key. Long lists with one-to-one mappings between local first-order pseudonyms of each organization and second-order pseudonyms can be avoided. The actual patient pseudonym for re-identification of the patient is, same as the identifying data used for calculation of the RBF's, only stored at the organization it belongs to. This approach, together with the encryption concept of the transmitted and aggregated medical data, was addressed in the HiGHmed Data Privacy Policy and received a positive vote from the TMF e.V.

So far, the process could be tested with one TTP and three organizations, each containing a small data-set with laboratory values of the same 15 patients. The greatest effort was incurred during the installation of the DSF due to the configuration of externally required systems as well as network and security settings. This depended largely on the organization's preconditions (e.g. experience with containerization, requesting and using of client certificates or security concerns because of required firewall configurations) and took several days to a few weeks. The actual deployment and execution of the data sharing process could be completed within minutes. How execution times change with large data-sets and additional organizations, especially when using record linkage, will be tested after all HiGHmed organizations have deployed the process. In particular, a solution has yet to be found for data-sets that cannot be loaded completely into RAM and need to be transmitted between participating organizations and the TTP.

6. Conclusion

This paper presents a possible solution for an automated data sharing process across multiple organizations using the open standards BPMN 2.0 and HL7 FHIR R4. The implementation does not require any central components, ensuring that identifying data does not leave any organization and that medical data is only shared for approved research projects. Therefore, regulations on data privacy and data minimization are observed. The data sharing process will be deployed at all HiGHmed organization in the near future. Finally, due to the generic approach, the independence of the medical data

repository and the compliance to the NSG umbrella process, this process could also be considered for data sharing between HiGHmed and other MII consortia.

Declarations

Conflict of Interest: The authors declare, that there is no conflict of interest.

Author contributions: RW, HH, CF, OH: conception and design of the process; RW, HH: implementation, deployment and testing of the process; RW: writing the manuscript; HH, CF, OH: substantial revising of the manuscript. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.

Acknowledgement: The project is funded by the German Federal Ministry of Education and Research (BMBF, grant id's: 01ZZ1802A and 01ZZ1802E). The authors would like to thank all committers that contributed to the open source implementation and for testing the current release.

References

[1] L.V. Rasmussen, The Electronic Health Record for Translational Research, *J Cardiovasc Transl Res*, 7.6 (2014) 607–614. doi: 10.1007/s12265-014-9579-z.

[2] K. Dentler, A. ten Teije, N. de Keizer, and R. Cornet, Barriers to the Reuse of Routinely Recorded Clinical Data: a Field Report, *Stud Health Technol Inform*, 192 (2013) 313–317. doi: 10.3233/978-1-61499-289-9-313.

[3] B. Haarbrandt et al., HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries, *Methods Inf Med*, 57.S 01 (2018) e66–e81. doi: 10.3414/ME18-02-0002.

[4] P. Knaup, T.M. Deserno, H.-U. Prokosch, and U. Sax, Implementation of a National Framework to Promote Health Data Sharing, *Yearb Med Inform*, 27.01 (2018) 302-304. doi: 10.1055/s-0038-1641210.

[5] S.L. Aguduri, A. Merzweiler, N. Yüksekogul, N. Meyer, A. Brandner, and O. Heinze, Modeling Clinical Data Transformation for a Medical Data Integration Center: An openEHR Approach, *64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)*, DocAbstr.113 (2019). doi: 10.3205/19gmds161.

[6] N. Yüksekogul, N. Meyer, S.L. Aguduri, A. Merzweiler, and O. Heinze, ETL-Processes for a Medical Data Integration Center – First Experiences from the Heidelberg University Hospital, *64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)*, DocAbstr.112 (2019). doi: 10.3205/19gmds169.

[7] T. Wendt et al. (AG Data Sharing der MII), Prozessmodelle des Data Sharing im Rahmen der Medizin-informatik-Initiative, *unpublished*, (2019).

[8] R. Wettstein, H. Hund, I. Kobylinski, C. Fegeler, and O. Heinze. Feasibility Queries in Distributed Architectures – Concept and Implementation in HiGHmed, *Stud Health Technol Inform*, 278 (2021) 134-141. doi: 10.3233/SHTI210061

[9] H. Hund, R. Wettstein, C.M. Heidt, and C. Fegeler. Executing Distributed Healthcare and Research Processes – the HiGHmed Data Sharing Framework, *Stud Health Technol Inform*, 278 (2021) 126-133. doi: 10.3233/SHTI210060

[10] M. Lablans, E. Schmidt, and F. Ückert, An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium, *JCO Clinical Cancer Informatics*, 2 (2018) 1-8. doi: 10.1200/CCI.17.00062.

[11] G. De Moor et al., Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform*, 53 (2015) 162-73. doi: 10.1016/j.jbi.2014.10.006.

[12] German Centre for Cardiovascular Research (DZHK), Clinical Research Platform (CRP), (2020). https://dzhk.de/en/research/clinical-research/clinical-research-platform/ (accessed March 24, 2021).

[13] C.M. Heidt, H. Hund, and C. Fegeler, A Federated Record Linkage Algorithm for Secure Medical Data Sharing, *Stud Health Technol Inform*, 278 (2021) 142-149. doi: 10.3233/SHTI210062