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

doi:10.3233/SHTI230126

Enhancing Data Protection via Auditable Informational Separation of Powers Between Workflow Engine Based Agents: Conceptualization, Implementation, and First Cross-Institutional Experiences

Felix ERDFELDER[a,b,c], Henning BEGERAU[a,b], David MEYERS[a,b],
Klaus-Jürgen QUAST[a,b], Daniel SCHUMACHER[a,*], Tobias BRIEDEN[d],
Roland IHLE[d], Danny AMMON[e], Henner M. KRUSE[e], and Sven ZENKER[a,b,c,1]
[a] *Staff Unit for Medical & Scientific Technology Development & Coordination (MWTek), Commercial Directorate; University Hospital Bonn, Germany*
[b] *Institute for Medical Biometry, Informatics, and Epidemiology, University of Bonn, Germany*
[c] *Department of Anesthesiology and Intensive Care Medicine, University Hospital Bonn, Germany*
[d] *Data Integration Center, Central IT Department, University Hospital Essen, Germany*
[e] *Data Integration Center, IT Department, Jena University Hospital, Germany*
**Present address: Skillbyte GmbH, Cologne, Germany*
ORCiD ID: Felix Erdfelder 0000-0002-4797-1889, Sven Zenker 0000-0003-0774-0725, Henning Begerau 0000-0001-5503-138X, Tobias Brieden 0000-0001-5512-7015, Roland Ihle 0000-0002-8502-7303, Danny Ammon 0000-0001-8960-7316, Henner M. Kruse 0000-0003-0286-8682

Abstract. German best practice standards for secondary use of patient data require pseudonymization and informational separation of powers assuring that identifying data (IDAT), pseudonyms (PSN), and medical data (MDAT) are never simultaneously knowable by any party involved in data provisioning and use. We describe a solution meeting these requirements based on the dynamic interaction of three software agents: the clinical domain agent (CDA), which processes IDAT and MDAT, the trusted third party agent (TTA), which processes IDAT and PSN, and the research domain agent (RDA), which processes PSN and MDAT and delivers pseudonymized datasets. CDA and RDA implement a distributed workflow by employing an off-the-shelf workflow engine. TTA wraps the gPAS framework for pseudonym generation and persistence. All agent interactions are implemented via secured REST-APIs. Rollout to three university hospitals was seamless. The workflow engine allowed meeting various overarching requirements, including auditability of data transfer and pseudonymization, with minimal additional implementation effort. Using a distributed agent architecture based on workflow engine technology thus proved to be an efficient way to meet technical and organizational requirements for provisioning patient data for research purposes in a data protection compliant way.

[1] Corresponding Author: Sven Zenker, Venusberg-Campus 1 53127 Bonn, E-mail: zenker@uni-bonn.de.

Keywords: FHIR, pseudonymization, interoperability, data protection

1. Introduction

The Medical Informatics Initiative (MII) in Germany [1], initiated and funded by the Federal Ministry of Education and Research, has, since 2016, pursued facilitating secondary use of patient data in German University Hospitals by developing a generic, interoperable infrastructure to enable local and cross-site data sharing and usage. In the initial implementation phase of the MII, the SMITH consortium, currently comprising ten German university hospitals, collaborated in realizing and evaluating one of four competitive implementation concepts to achieve this goal [2]. The initial plans to realize these goals by employing an IHE-based infrastructure [2] was found to require adjustment to achieve these goals across the SMITH sites since the complex requirements of secondary use of highly confidential, heterogeneous patient data for research were not mappable to IHE standard compliant processes and interfaces with the available technologies provided by the SMITH commercial partners without breaking standard compliance and thus interoperability with the available resources. In particular, the need to technically and organizationally segregate the information technological (IT) infrastructure at the SMITH sites into two domains – the clinical domain (CD) and the research domain (RD) – was identified. The CD allows processing of all patient related data – including their identifying information – for the purpose of supporting and optimizing patient care, with the legal basis provided by the contractual relationship between patient and healthcare provider, whereas the RD only processes data of patients who have given broad consent [3,4] to secondary use of their pseudonymized patient data for research purposes (Fig. 1). Current German national best practice standards [5] require an informational separation of powers, where no party involved in the pseudonymization process required when performing selective data transfer from CD to RD may at any time simultaneously have access to directly identifying information (IDAT, e.g. a patient's name or date of birth), medical data (MDAT, the “payload”), and definitive pseudonyms (PSN), with the intention to organizationally and technically minimize the risk of malicious or unintentional de-pseudonymization (Fig. 1). To achieve this, the relations between IDAT and PSN are typically entrusted to a trusted third party service (TTA), effectively a third independent domain, which alone can de-pseudonymize by retrieving IDAT for a given PSN, but never sees any MDAT.

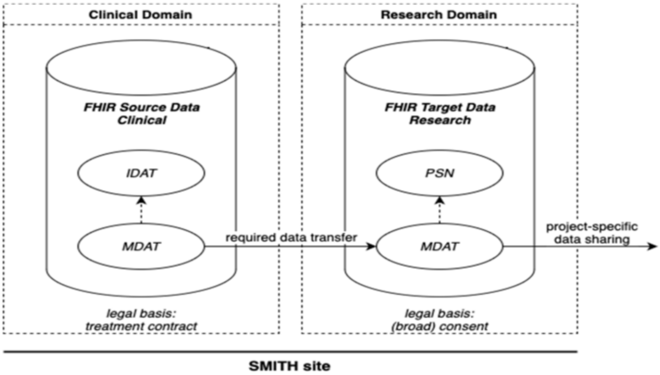
Clinical Domain
Research Domain
FHIR Source Data Clinical
FHIR Target Data Research
IDAT
PSN
MDAT
MDAT
required data transfer
project-specific data sharing
legal basis: treatment contract
legal basis: (broad) consent
SMITH site

Figure 1. Domain separation and data storage at SMITH sites.

2. Methods

To achieve this goal, the target overall data selection, pseudonymization, and physical data transfer workflow was mapped to an agent-based architecture where a Clinical Domain Agent (CDA) and a Research Domain Agent (RDA) interact with a trusted third party agent (TTA) to jointly execute a distributed workflow. Informational separation of powers is realized via generation of a temporary transport PSN (TPSN) for each potentially identifying data item. This is used to tag MDAT for transferral from CDA to RDA, resolved via TTA by RDA, and deleted irreversibly after successful completion of the entire transferral workflow (Fig. 2). CDA consumes FHIR data from heterogeneous sources while RDA synchronizes to a standard compliant FHIR server. All agents were implemented in Java™ SE 11 using the Spring™ framework [6], allowing lightweight, low-maintenance operation on independent virtual machines in separated network domains. CDA and RDA use the open source version of the Camunda™ workflow engine [7] to auditably execute BPMN workflows modeled using the Camunda™ modeler, while TTA uses gPAS from the MOSAIC toolbox [8] for pseudonym generation and persistence. IDAT and MDAT were represented in MII core dataset FHIR profile [9] compliant JSON. All agent interactions were realized via certificate authentication based, TLS secured REST APIs using Spring Boot and Spring Security mechanisms, including the workflow handover between separated Camunda™ instances via the Camunda™ REST API. IDAT->PSN substitution was extended beyond classical IDAT items to FHIR identifiers that could facilitate identification attacks using a generically parameterizable substitution/deidentification mechanism that made use of the inherent structure of FHIR based data representations.

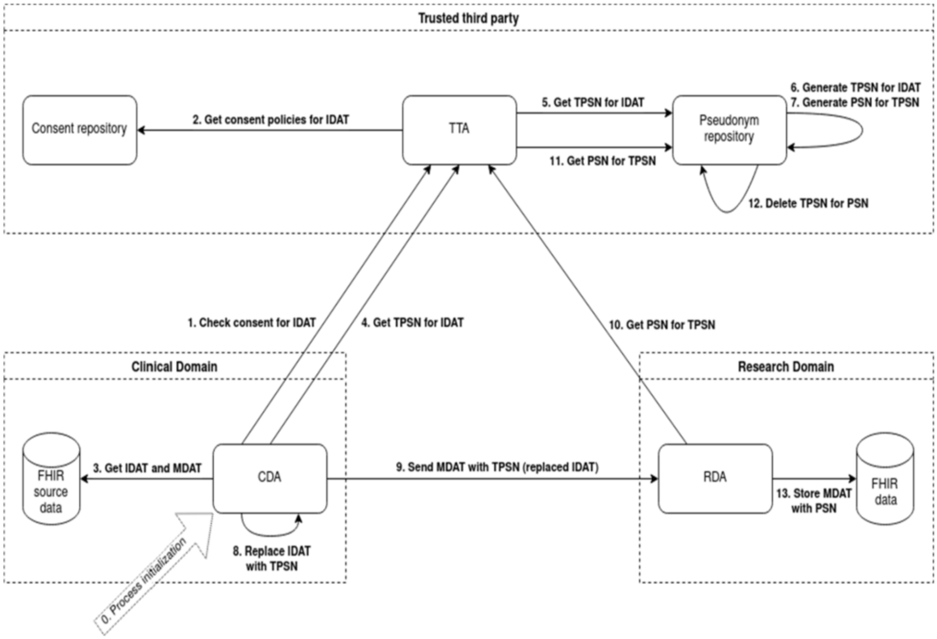

Figure 2. Distributed workflow of CDA and RDA with TTA.

3. Results

Combined usage of the Spring framework and the Camunda™ engine allowed for rapid prototype implementation of the overall system at one SMITH site (Bonn), while a second SMITH site (Essen) implemented the generic deidentification mechanisms. The Camunda™ user interface and audit features, which facilitated straightforward monitoring of transfer processes and rapid post hoc analysis of system behavior, combined with the broad set of convenience features provided by the Spring framework, enabled rapid prototype implementation and verification, with less than three months from initial conceptualization to roll-out to the first non-development SMITH site. Site-specific customization requirements driven by the inter-site infrastructural heterogeneity resulting from very different clinical IT setups at SMITH sites were subsequently identified, extracted, and exposed via Spring declarative configuration mechanisms, allowing SMITH sites with limited software development resources to adapt the system to their specific requirements and thus achieve successful integration. Within 5 months, 3 SMITH sites were thus enabled to demonstrate successful, data protection compliant and fully auditable CD to RD data transfers involving a TTA, contributing to all SMITH sites successfully passing the obligatory MII project audit in spring 2021.

In practice, system stability and performance, even without any investment in specific tuning, was found to be quite satisfactory, with a typical dataset of roughly 5 million FHIR-resources transferred in 5:40 hours.

4. Discussion

The clean decomposition of functionalities and processing steps enforced by the workflow-centric architecture facilitated distribution of development tasks between the Bonn and Essen teams, enabling an agile distributed development workflow without exacerbating integration efforts when merging contributions from the participating development sites. The capabilities of the workflow engine proved crucial to successfully verifying and validating the system within the available timeframe and resource constraints and remain essential to comply with auditability and monitoring requirements imposed by local data protection and IT security regulations. Declarative configuration enabled by the Spring mechanisms proved essential to enable partner sites without sufficient software development capabilities to rapidly adapt the developed system to their local requirements.

Since completion of the primary prototype implementation, the solution has continuously evolved, improving robustness and adding features such as a configurable cohort selection for scheduled data transmission tasks. During this agile and iterative improvement process, the workflow engine based approach proved valuable in reducing the complexity of adaptation tasks and simplifying debugging of system behavior.

Future work will include adaptation of the developed system to the evolving requirements of MII use cases, including managing data types such as biosignals and images not amenable to efficient representation in FHIR, and integration of the developed system with the evolving and increasingly convergent national MII infrastructure of the current funding phase recently initiated, which is currently being constructed from components derived from the results of the four consortia from the previous funding phase.

5. Conclusions

We have shown that an agent-based decomposition of a de-identification and data transfer workflow between different domains of a medical IT architecture involving a trusted third party service to achieve informational separation of powers can successfully, rapidly, and efficiently be implemented and rolled out using an off-the-shelf workflow engine combined with modern Java™ frameworks. Modern workflow engines additionally resolve various overarching concerns of critical relevance when processing sensitive patient data, including full auditability of all data processing steps.

Acknowledgements: The authors wish to thank Dr. Jan Gewehr & team at the SMITH partner site Hamburg Eppendorf for their invaluable contributions to devising the described state of the SMITH architecture and Dr. Thomas Wendt, Sebastian Stäubert & the entire SMITH project management team at the SMITH coordinating site Leipzig for organizational and documentation support. This work was supported, in part, by German Federal Ministry of Education and Research (BMBF) grants no. 01ZZ1803Q, 01ZZ1803P, and 01ZZ1803C.

6. References

[1] Semler SC, Wissing F, Heyder R, German Medical Informatics Initiative. Methods Inf. Med. 2018;57:e50-6. doi:10.3414/ME18-03-0003.

[2] Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, Daumke P, Decker S, Funkat G, Gewehr JE, de Greiff A. Smart medical information technology for healthcare (SMITH). Methods of information in medicine. 2018 May;57(S 01):e92-105. doi:10.3414/ME18-02-0004.

[3] Bild R, Bialke M, Buckow K, Ganslandt T, Ihrig K, Jahns R, Merzweiler A, Roschka S, Schreiweis B, Stäubert S, Zenker S. Towards a comprehensive and interoperable representation of consent-based data usage permissions in the German medical informatics initiative. BMC Medical Informatics and Decision Making. 2020 Dec;20:1-9. doi:10.1186/s12911-020-01138-6.

[4] Zenker S, Strech D, Ihrig K, Jahns R, Müller G, Schickhardt C, Schmidt G, Speer R, Winkler E, von Kielmansegg SG, Drepper J. Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio) medical research: towards a new German national standard. Journal of Biomedical Informatics. 2022 Jul 1;131:104096. doi:10.1016/j.jbi.2022.104096.

[5] Pommerening K, Drepper J, Helbing K, Ganslandt T. Leitfaden zum Datenschutz in medizinischen Forschungsprojekten: Generische Lösungen der TMF 2.0. MWV Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. doi:10.32745/9783954662951.

[6] Johnson R. Expert One-on-One J2EE Design and Development, John Wiley & Sons, 2004.

[7] The Universal Process Orchestrator, *Camunda*. (n.d.). https://camunda.com/ (accessed January 13, 2023).

[8] Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, Hoffmann W. MOSAIC–A modular approach to data management in epidemiological studies. Methods of information in medicine. 2015 Jul;54(04):364-71. doi:10.3414/ME14-01-0133.

[9] Basic modules of the MII core data set | Medical Informatics Initiative, (n.d.). https://www.medizininformatik-initiative.de/en/basic-modules-mii-core-data-set (accessed January 13, 2023).