A Trusted Data Sharing Environment based on FAIR Principles and Distributed Process Execution

Marcel Klötgen[1,2][a], Florian Lauf[2][b], Sebastian Stäubert[1,3][c], Sven Meister[5][d]
and Danny Ammon[1,4][e]

[1]SMITH consortium of the German Medical Informatics Initiative, Germany
[2]Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany
[3]Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig, Germany
[4]Data Integration Center, Jena University Hospital, Jena, Germany
[5]Faculty of Health/School of Medicine, Witten/Herdecke University, Witten, Germany

Keywords: Health Information Interoperability, Workflow, Data Science, Modeling Information System Architecture.

Abstract: Provision and usage of distributed secondary-use data for medical research requires the implementation of a distributed data use & access process and several sub-processes. The SMITH Service Platform (SSP) manages process-based interactions with several Data Integration Centers (DIC), each being responsible for the management and provision of suitable data sets in the context of a data use project. A trusted data sharing environment is specified based on the implementation of Trusted Connectors as specified by the International Data Spaces (IDS) Reference Architecture Model. Thus, a distributed data delivery sub-process enforces data sovereignty aspects between all involved parties. In the future, a solidification of the concepts and a further implementation of the trusted data sharing environment should be addressed.

1 INTRODUCTION

1.1 Background

In Germany, the need and demand for a rapid development of new and appropriate treatment methods, pharmaceuticals and technologies for healthcare is increasingly important. Reusing, linking and analyzing healthcare-related data from different sources leverages the potential for the implementation of secondary-use purposes and medical research (Celi et al. 2013). Especially, processing extensive real world data sets enables a well-aimed validation of scientific hypotheses, optimization of artificial intelligence applications, or the derivation and recognition of influences and correlations. Yet, several requirements regarding data security and protection, regulatory and legal aspects, and interoperability must be met for the provision of data in the context of different secondary-use purposes.

Smart Medical Information Technology for Healthcare (SMITH) is one of four funded consortia of the Medical Informatics Initiative (MII) (Löffler et al.), (Semler et al. 2018), specifically striving to optimize and interlock primary patient care and secondary-use of data for medical research based on the joint collaboration of universities, university hospitals and IT-companies in Germany. SMITH specifies a reference architecture for its distributed Data Integration Centers (DIC) to be established at all seven participating sites, thus ensuring a modular architecture for processing and providing medical data from consenting patients (Winter et al. 2018) in an interoperable core data set format (Ammon et al. 2019). Cross-site use cases and operations involving

[a] https://orcid.org/0000-0003-4109-8641
[b] https://orcid.org/0000-0003-0844-3722
[c] https://orcid.org/0000-0002-7221-7415
[d] https://orcid.org/0000-0003-0522-986X
[e] https://orcid.org/0000-0001-8960-7316

Klötgen, M., Lauf, F., Stäubert, S., Meister, S. and Ammon, D.
A Trusted Data Sharing Environment based on FAIR Principles and Distributed Process Execution.
DOI: 10.5220/0010869800003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 5: HEALTHINF*, pages 583-589
ISBN: 978-989-758-552-4; ISSN: 2184-4305

several DICs, such as feasibility queries, data use project management and data delivery, are supported through interactions with the SMITH Service Platform (SSP) and its standards-based interfaces, allowing for a cross-site and centralized process management and identical application of quality principles. Yet, the concepts incorporate a low degree of data sovereignty aspects, especially regarding delivered data sets.

The International Data Spaces (IDS) (Otto et al. 2018) represent a rising concept for data spaces especially addressed in Europe and steadily developed by the International Data Spaces Association (IDSA), striving to anchor its concepts and architecture in European digitization strategies such as Gaia-X (Cinzia Capiello et al. 2020). IDS combines standards and technologies to create a data-driven data economy while ensuring data sovereignty (IDSA 2019). IDS connectors are gateway tools for accessing such data spaces and are available in different implementations (e.g., Trusted Connector or Dataspace Connector). Different domains constitute different IDS verticalizations (e.g., industrial data space, agriculture data space, or medical data space). To foster the medical domain, the project PanDa@IDS realizes the concept of data donations using IDS connectors (Fraunhofer Institute for Software und Systems Engineering ISST 2021). By embedding the Trusted Connector in SMITH, scientists can retrieve data sets via IDS technology, thus incorporating domain-specific negotiation processes missing from the IDS-based technology.

1.2 Requirements

Sharing healthcare-related data for secondary-use purposes raises great potential, yet requires fundamental constraints at the same time. Scientists require the ability to retrieve and process large amounts of distributed primary care data from different DICs. This implies the specification of cohorts and data items to be retrieved, such as data on the prevalence of cardiac diseases among a specific age-class, the submission and evaluation of a common data usage proposal form, and the conclusion of contracts between all parties. In order to enforce data sovereignty aspects, such as a limited utilization period of the retrieved data sets, scientists must employ trusted applications or environments for accessing or analyzing the data sets, e.g. a web interface implementing structured views and traversing operations for the data. When the contract becomes invalid, the trusted application should prevent access to the data sets.

Several requirements are realized by SMITH and PanDa@IDS, addressing FAIR data principles (Wilkinson et al. 2016) with the MII Core Data Set based on HL7 FHIR (Ammon et al. 2019) and legal and regulatory aspects of data sharing including anonymization or pseudonymization. Findability (the capability to easily find data and metadata), interoperability (the capability of data sets to integrate with one another and to be used for analysis, processing and storage) and reusability (the capability of data sets to be replicated or combined in different environments) of data sets and metadata according to the FAIR principles are ensured by the use of the MII Core Data Set for data management tasks at each DIC, the use of FHIR Search as a compatible query language incorporating the core data set and the Data Use & Access macro-process described in chapter 3.1. Accessibility (the capability of data sets to be accessed by users) is ensured by the setup of the data sharing environment, including the use of open interfaces based on international standards, and the provision of user interfaces through the SSP.

Therefore, the data sharing environment must incorporate the following additional requirements:

- **Scalability:** The data sharing environment must be able to integrate different actors and systems and therefore support different requirements and actions of its users, such as different legal frameworks or ethical assumptions.
- **Process Interoperability:** The data sharing environment must support the integration of different systems for the visualization and execution of distributed tasks based on a common process and representation.
- **Information Interoperability:** The data sharing environment must provide its usage and project management data in a common interoperable representation.
- **Trusted Data Sharing:** The involved systems and components must provide a trusted environment for data sharing based on technical measures, such as certification or authentication of components and secure data transfer.
- **Trusted Data Usage:** The involved systems and components must provide a trusted environment for data utilization in the context of a data use project, technically managing access and utilization according to the contract.

This work focuses on the definition of an architecture, which is able to address the aforementioned requirements for a trusted data sharing environment, thus leveraging data sovereignty concepts in the context of medical research. Additionally, a concept for distributed standards-based process execution is presented.

2 STATE OF THE ART

Several recent projects and organizations have produced results and inspiration for the idea of a trusted data sharing environment. The LIFE Research Center for Civilization Diseases acts as a data provider for secondary-use data gained from patient care of the university hospital Leipzig. The LIFE data sharing approach incorporates the OAIS reference model (Kirsten et al. 2017) and covers the needs for organizational and IT infrastructure for data sharing. LIFE is able to provide data sets in the context of a data use project, yet lacks the capabilities to coordinate cross-organizational data provision with different data providers and especially enforce the contracted usage policies for delivered data.

The IDS Reference Architecture Model defines a common infrastructure for data sharing scenarios based on standardized interfaces and self-descriptive services, but has not yet achieved the desired maturity level realizing the requirements of secondary-use data sharing of the healthcare domain, such as the integration of existing processes (IDSA 2019).

Usage control enforces constraints on the usage of data beyond its provision, incorporating definitions for authorization, obligations, and conditions (Park and Sandhu 2004). An implementation for distributed systems is defined in (Pretschner et al. 2006). These principles are relevant for subsequent enforcement of usage control based on negotiation and distribution methods described in the following sections.

3 CONCEPT

3.1 Distributed Process Management

The delivery of data is part of a centrally managed, scalable and distributed macro-process, addressing several regulatory and administrative aspects of the management of cross-site data use projects. This macro-process is specified by the MII as Data Use & Access (DUA) and consists of different tasks, grouped into several phases:

1. **Data Usage Proposal Provision Phase:** A scientist describes the planned data use project and provides all necessary information for the retrieval of the desired data based on a common data use proposal form to selected DICs.
2. **Data Usage Proposal Evaluation Phase:** The Use & Access Committees (UAC) of the selected DICs receive and evaluate the provided data usage proposal individually, based on individually relevant formal, regulatory and content-related aspects.
3. **Contracting Phase:** The scientist and project management actors of all participating DICs agree on the rules and regulations determining the usage of the data sets and sign a common contract.
4. **Data Preparation and Delivery Phase:** Once the contracting phase has been successful, the DICs process and provide anonymized or pseudonymized data sets meeting the data use project criteria individually, thus implementing the required security requirements. The involved scientists retrieve the provided data sets for utilization.
5. **Exploitation and Closing Phase:** At the end of the data use project, a closing phase begins, allowing scientists and DICs to exchange relevant project results.

These phases of the distributed DUA macro-process involve several tasks to be executed by different roles, organizations or systems. The macro-process is managed centrally in order to maintain control of the overall process state and execution, while the assignment of tasks to different organizations, systems or roles realizes task distribution. Data delivery is part of the data preparation and delivery phase and implemented as a distributed sub-process.

3.2 Data Space Transactions

Trust between two interacting parties can be achieved through the involvement of a Trusted Third Party (TTP), which is trusted by the two parties each. The TTP is responsible for performing sensitive tasks, such as the management and merge of distributed data sets or the authorization of data delivery and utilization. The provision and retrieval of distributed data sets involve the following actors according to the IDS Reference Architecture Model (IDSA 2019):

- **Data Owner:** The DICs managing and processing data for primary and secondary use purposes according to DUA.
- **Data User:** The scientists utilizing data sets in the context of data use projects based on previously agreed contracts according to DUA.
- **Data Provider:** A component of the DIC, responsible for and implementing the actual delivery of processed data sets in the context of a data use project.
- **Data Consumer:** A component operated by the Data User actor, implementing a secure environment and reception point for data sets from multiple Data Provider actors in the context of a data use project.

Aiming at a secure and reliable cross-site retrieval and usage of distributed data sets, the TTP system is partially implemented by the Data Consumer component. Relocation of TTP functionality of the data delivery sub-process allows for the central execution of several services and operations without the need to involve another substitutional system. The relocated tasks of the data delivery sub-process to the Data Consumer include the merge of previously distributed data sets and authorization of data processing operations. The Data Consumer thus represents a fully integrated system extension and strengthens data sovereignty and usage control by providing a secure environment for data utilization according to the contracts, thus addressing the requirements *trusted data sharing* and *trusted data usage* by creating a secure and closed data space.

As represented in figure 1, the DUA macro-process and its process model and instances are managed centrally by a Business Process Engine. Process instances are mapped to a representation that ensures interoperability with distributed actors and service providers for distributed process execution, thus ensuring the requirements *scalability* and *process interoperability*. Distributed sub-processes are addressed as tasks within the macro-process, allowing for individual implementations of task execution workflows by each Data Provider based on individually relevant regulations or demands. The Data Providers now act as Task Performers with their own workflow management. Therefore, all required input and output parameters must be provided with the tasks, along with assignment and orchestration data for the notification and activation of responsible actors and systems. The input parameters are used for displaying or processing of information during the execution of individual sub-processes. Once the distributed execution of a task is complete, the results are registered as output parameters, along with a resulting status of the task for process management.

Data delivery is represented as a service task, which is added on demand (e.g. functionality of a user interface) for a distributed implementation of sub-processes. All input parameters of the data delivery task must include references to the data use project or the prepared data sets, identification or recipient information of the Data Consumer, and metadata of the data use project, including relevant contracting parameters.

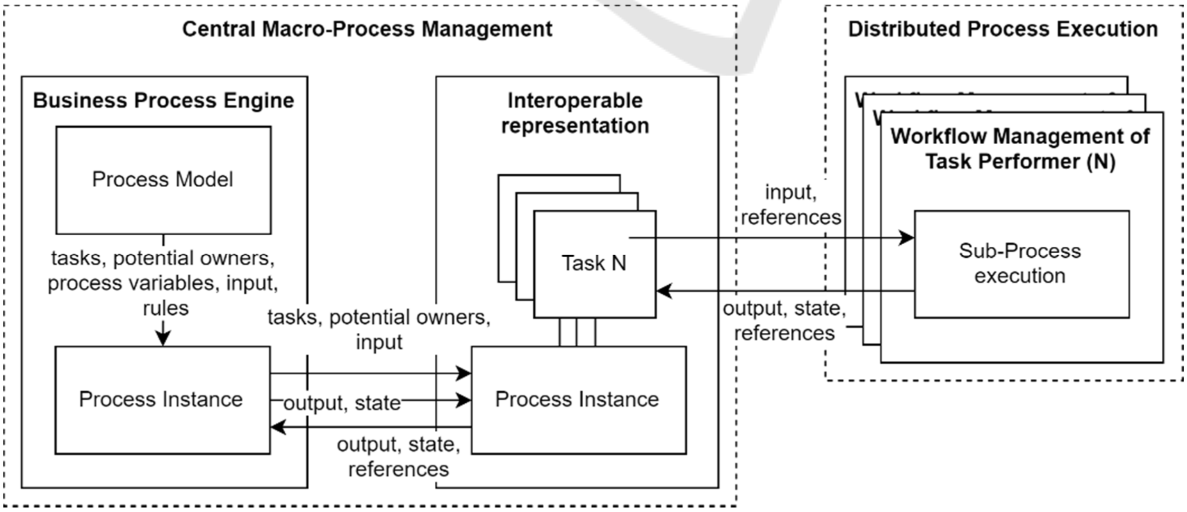
Central Macro-Process Management
Business Process Engine
Process Model
tasks, potential owners, process variables, input, rules
Process Instance
tasks, potential owners, input
output, state
output, state, references
Interoperable representation
Task N
Process Instance
input, references
output, state, references
Distributed Process Execution
Workflow Management of Task Performer (N)
Sub-Process execution

Figure 1: Distributed process management.

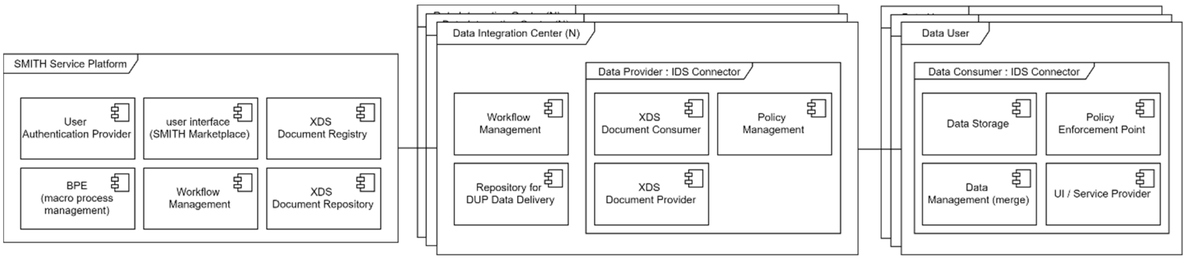

Figure 2: Components of the data sharing environment.

3.3 The Data Delivery Sub-Process

The data delivery sub-process is comprised of several service tasks to be performed by the Data Provider and the Data Consumer, respectively. This allows for an integration of the Data Consumer in the distributed project management through the implementation of a secure data space environment, making it a central actor of the DUA process. The tasks of the Data Providers include the retrieval of the data delivery parameters provided by the task of the macro-process, the retrieval of prepared data sets for data delivery, the creation of policies representing the contract parameters, and the sending these data sets and policies. The Data Consumer receives, stores and merges the data sets of a data use project and activates the provided policies. All operations performed on the retrieved data sets are authorized through policy enforcement, representing and enacting the contract parameters. Policies are created by the Data Provider based on the task input parameters such as the data use project metadata, and delivered to the Data Consumer with the provided data sets.

4 IMPLEMENTATION

As depicted in figure 2, the SSP supports the execution and documentation of all DUA macro-process tasks of the involved actors through a dedicated web interface, the SMITH Marketplace. When the requested data sets are available, the involved scientist receives a task and uses the SSP web interface to authenticate and start the delivery of data sets from all involved DICs based on the distributed process management concept. The Data Consumer application of the scientist receives and processes the data sets, allowing supported interactions with the data by the scientist, such as viewing or accessing (e.g., the scientist can use the Data Consumer's web interface to view the received data sets).

The DUA macro-process is managed by the SSP and its integrated Business Process Engine (BPE). As shown in figure 2 and in order to ensure interoperability with distributed actors and systems for task execution, each instance of a DUA macro-process is synchronously provided according to the IHE Cross-Enterprise Workflow Document (XDW) profile based on a central Cross-Enterprise Document Sharing (XDS) Affinity Domain, thus implementing the requirements *scalability* and *process interoperability*. The XDW tasks are routed to their designated performers using the name and notificationRecipient attributes. The Data Provider and Data Consumer components are implemented as IDS Connectors based on the Trusted Connector reference implementation (Schütte et al.) and extended to support the distributed data delivery sub-process. The data sets to be delivered are retrieved from the XDS-based data source of a DIC, and additional XACML policies are created and generated, representing the contract parameters such as duration of the data use project. In order to fulfil the *requirements trusted data sharing* and *trusted data usage*, the Data Consumer additionally implements the merge of received data sets of a data use project and the enforcement of the attached XACML policies through a Policy Enforcement Point (PEP). The User Interface (UI) component implements possible usage scenarios of the data sets representing the data use project.

5 DISCUSSION

A basic range of functions of the envisioned SMITH infrastructure, including a first iteration of the DIC architecture and the distributed DUA macro-process, has been developed and is envisioned to be productive at the end of 2021. The IDS-based data delivery sub-process is not part of the current setup, yet has been developed and integrated as a demonstrator. Using a set of test data, the demonstrator proves the ability to perform all phases of the DUA process with all actors, including the transfer and provision of data sets and usage policies representing crucial contracting parameters. The demonstrator also contains a basic web interface of the Data Consumer, which is capable of displaying the retrieved data and exemplifies the enforcement of policy rules. Thus, the requirements addressing *scalability*, *process interoperability* and *information interoperability* could be met. The requirements of *trusted data sharing* and *trusted data usage* could only be met partially with the need to further implement certain aspects, such as a certification of involved components, proper authorization of data retrieval by the Data Provider, and the implementation of a user interface or API of the Data Consumer, allowing the required flexible and authorized use of received data sets.

Limitations of the current system affect the specification and selection of desired data based on the core data set definition. Thus, only data items previously defined in the MII Core Data Set can be retrieved, such as observation resources encoded as ICD. Another major limitation is attributed to the implementation of data interaction mechanisms of the Data Consumer. Utilization and processing of retrieved data sets will be limited to the Data Consumer's API or user interface, such as viewing the data sets in the case of the demonstrator. It is therefore relevant to find a balanced solution based on future research and requirements analyses incorporating the needs of researchers and citizens equally, which addresses data sovereignty as well as utilization requirements, e.g. a limited FHIR-based API.

The context of data usage, organizational dependencies and regular constraints according to EU's General Data Protection Regulation (GDPR) or even regional laws determines the application of data protection measures. Thus, each DIC is responsible for the appropriate application of data protection measures before providing data sets. Data Use Projects involving several DICs usually provide fully anonymized data sets based on the application of Privacy Preserving Record Linkage (PPRL) methods implemented by additional components (the FHIR Transfer Services) of a DIC in compliance with GDPR, while some regional laws allow the provision of data sets containing personal information within the same organization.

Future steps will address user validation activities as well as the implementation of additional requirements and possible operation of Data Provider and Data Consumer components by the SMITH DICs until 2022.

6 CONCLUSION

The union of a centrally managed, distributed DUA process through the central SSP with distributed IDS components in the DIC infrastructure and relocation of TTP functionality to trusted Data Consumers as a part of a data delivery sub-process addresses several current requirements of data sharing. The concept applies and realizes several data sovereignty aspects to distributed secondary-use data for medical research, addressing the technical enforcement of a data use contract on joined data sets, and thus creates a trusted data sharing and usage environment. Future steps should address the solidification of the concepts within SMITH and the implementation of further aspects of the trusted data sharing and usage environment.

ACKNOWLEDGEMENTS

This publication of the SMITH consortium was supported by the German Federal Ministry of Education and Research, grant number 01ZZ1803J, and the Fraunhofer Internal Programs under Grant No. Anti-Corona 840239. The research is based on the results of several working groups of the Medical Informatics Initiative (MII), the SMITH consortium and the PanDa@IDS consortium.

REFERENCES

Ammon, D., Bietenbeck, A., Boeker, M., Ganslandt, T., Heckmann, S., Heitmann, K., Sax, U., Schepers, J., Semler, S., Thun, S. & Zautke, A. (2019). Der Kerndatensatz der Medizininformatik-Initiative. Interoperable Spezifikation am Beispiel der Laborbefunde mittels LOINC und FHIR. *mdi – Forum der Medizin, Dokumentation und Medizin-Informatik* (21), 113–117.

Celi, L. A., Mark, R. G., Stone, D. J. & Montgomery, R. A. (2013). "Big Data" in the Intensive Care Unit. Closing the Data Loop. *Am J Respir Crit Care Med* (187(11)), 1157–1160. doi:10.1164/rccm.201212-2311ED

Cinzia Capiello, Avigdor Gal, Matthias Jarke, Jakob Rehof & Cinzia Cappiello. (2020). Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391). *Dagstuhl Reports 9* (9), 66–134. https://drops.dagstuhl.de/opus/volltexte/2020/11845.

Fraunhofer Institute for Software & Systems Engineering ISST. (2021). *PanDa@IDS: Pandemic-related data donation based on data sover-eignty principles of the Medical Data Space*. Dortmund. Zugegriffen: 10. Mai

2021. https://www.isst.fraunhofer.de/en/business-units/healthcare/projects/PanDaIDS.html.

IDSA. (2019). *Reference Architecture Model: Version 3.0 | April 2019*. Berlin. Zugegriffen: 10. Mai 2021. https://internationaldataspaces.org/use/reference-architecture/.

Kirsten, T., Wagner, J., Kiel, A., Rühle, M. & Löffler, M. (2017). Selecting, Packaging, and Granting Access for Sharing Study Data. Experiences and Recent Software Developments in the LIFE Study. *INFORMATIK 2017 2017*.

Löffler, M. Univ.-Prof. Dr. med. habil., Scherag, A. Univ.-Prof. Dr. rer. physiol. & Marx, G. Univ.-Prof. Dr. med. (Hrsg.). Smart Medical Information Technology for Healthcare. https://www.smith.care/konsortium/. Zugegriffen: 25. März 2020.

Otto, B., Hompel, M. ten & Wrobel, S. (2018). Industrial Data Space. In R. Neugebauer (Hrsg.), *Digitalisierung* (Bd. 51, S. 113–133). Berlin, Heidelberg: Springer Berlin Heidelberg.

Park, J. & Sandhu, R. (2004). The UCON ABC usage control model. *ACM Transactions on Information and System Security (TISSEC) 7,* 128–174. doi:10.1145/984334.984339

Pretschner, A., Hilty, M. & Basin, D. (2006). Distributed usage control. *Communications of the ACM 49. doi:*10.1145/1151030.1151053

Schütte, J., Brost, G. & Wessel, S. *Der Trusted Connector im Industrial Data Space*. Garching. https://www.aisec.fraunhofer.de/content/dam/aisec/Dokumente/Publikationen/Studien_TechReports/deutsch/IDS-Paper_Datensouveraenitaet.pdf.

Semler, S. C., Wissing, F. & Heyder, R. (2018). German Medical Informatics Initiative. A National Approach to Integrating Health Data from Patient Care and Medical Research. *Methods of Information in Medicine* (57). doi:10.3414/ME18-03-0003

Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, Joost, Lusher, Scott J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*.

Winter, A., Stäubert, S., Ammon, D., Aiche, S., Beyan, O., Bischoff, V., Daumke, P., Decker, S., Funkat, G., Gewehr, J. E., Greiff, A. de, Haferkamp, S., Hahn, U., Henkel, A., Kirsten, T., Klöss, T., Lippert, J., Löbe, M., Lowitsch, V., Maassen, O., Maschmann, J., Meister, S., Mikolajczyk, R., Nüchter, M., Pletz, M. W., Rahm, E., Riedel, M., Saleh, K., Schuppert, A., Smers, S., Stollenwerk, A., Uhlig, S., Wendt, T., Zenker, S., Fleig, W., Marx, G., Scherag, A. & Löffler, M. (2018). Smart Medical Information Technology for Healthcare (SMITH). Data Integration based on Interoperability Standards. *Methods of Information in Medicine* (57). doi:10.3414/ME18-02-0004