J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231118

Architecture of the Mass Spectrometry Data Management Pipeline in the SMART-CARE Project

Friedemann G. RINGWALD[a,1], Aleksei DUDCHENKO[a,b], Petra KNAUP [a], Felix Czernilofsky[b], Sascha Dietrich[b,c] and Matthias GANZINGER[a]

[a] *Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany*

[b] *Department of Medicine V, Hematology, Oncology and Rheumatology, University of Heidelberg, Heidelberg, Germany*

[c] *Department of Hematology and Oncology, University Hospital Düsseldorf, Düsseldorf, Germany*

ORCiD ID: Friedemann G. Ringwald https://orcid.org/0000-0002-0572-4820

Abstract. In the SMART-CARE project- a systems medicine approach to stratification of cancer recurrence in Heidelberg, Germany - a streamlined mass-spectrometry (MS) workflow for identification of cancer relapse was developed. This project has multiple partners from clinics, laboratories and computational teams. For optimal collaboration, consistent documentation and centralized storage, the linked data repository was designed. Clinical, laboratory and computational group members interact with this platform and store meta- and raw-data. The specific architectural choices, such as pseudonymization service, uploading process and other technical specifications as well as lessons learned are presented in this work. Altogether, relevant information in order to provide other research groups with a head-start for tackling MS data management in the context of systems medicine research projects is described.

Keywords. Systems medicine, mass spectrometry, data management

1. Introduction

Relapse of cancer remains the most common reason for death related to cancer. By the means of personalized medicine, various new treatment approaches could be developed and are making their way into clinical practice. Especially proteome and metabolome analyses by means of mass-spectrometry (MS) are coming up more [1]. In order to successfully conduct proteome and metabolome analyses, carefully planned, harmonized and streamlined processes need to be established. This is important when translating the results into clinical practice. In order to discover metabolomic changes connected to cancer relapse, and to build up a standardized - MS pipeline, the Systems

[1] Corresponding Author: Friedemann G. Ringwald, email: friedemann.ringwald@med.uni-heidelberg.de.

Medicine Approach to Stratification of Cancer Recurrence (SMART-CARE) project was launched. The main objective of SMART-CARE is to predict cancer relapse by establishing standard operating procedures to ensure reproducible proteome and metabolome analyses. Multiple research facilities with distinct research goals take part in this project:

- Clinical partners are responsible for sample extraction and preparation.
- Laboratory partners develop harmonized and reproducible wet lab workflows.
- Computational partners focus on downstream analysis and model development of raw and peak-processed MS data.

To achieve optimal data exchange, documentation, and interaction between the partners, the SMART-CARE Linked Data Repository (SMART-LDR) was developed [2]. Different architectural choices and the software landscape are described and discussed in this article in order to provide other research groups with a head-start for tackling MS data management in the context of systems medicine research projects.

2. Methods

For a safe and reliable data exchange platform many requirements have to be met. First, it has to be fast and highly performant, stable and accessible for all project members at all times. Users should be able to have accounts with different configurations for access rights to different areas of the system. Additionally, an identity management solution should support existing identity management procedures such as lightweight directory access protocol (LDAP). Due to the sensitive nature of patient data, which is uploaded to the system, pseudonymization procedures are required to prevent re-identification of patients. The solution has to be in line with the general data protection regulation (GDPR) for countries of the European Union. Pseudonymization has to be reversible, so clinical partners are able to identify their patients for adding follow-up data.

Data, which is provided by clinical partners, should be harmonized where applicable to make joint analyses possible. Modifying specifications of the data model as well as the definitions of entry forms and descriptions should be possible at all times, so additional metadata fields can be added if needed. Moreover, data entry should be possible manually, via the graphical user interface (GUI), or semi-automatically via file upload for example in spreadsheet format.

In addition to handling metadata, controlled terminologies, and pre-processed numerical data, the system should be able to store raw data reaching up to 50 GB in individual file size. This kind of data will be provided by the different laboratory partners. Thus, upload and storage should be managed independently for each partner. Raw data should be linked to existing objects which should be linked to a patient object in a hierarchical relationship. Data upload should recover automatically from internet interruptions and work as background process on workstations. Due to local firewall and software restrictions, only solutions capable of HTTP proxies were considered. Finally, the combined software solution has to be modifiable and by adding self-developed plugins. This is important for the computational members of SMART-CARE, who conduct down-stream analysis using deep learning and other computational methods. All requirements were derived in consultation with users and clinicians. The architecture described below aims to fulfil the criteria mentioned above as far as possible.

3. Results

State-of-the-art server hardware was acquired to host all software used in the SMART-CARE project. All sub-systems mentioned below are installed on dedicated virtual machines. The open-source digital laboratory notebook management software openBis [3], developed by the ETH Zurich was identified as a comprehensive solution for many of the above-mentioned requirements. With the built-in functionality, the system was connected to the universities existing active directory system, allowing for a fine-grained access control concept addressing the access control requirements of the different research groups. Inside openBis, data entry is possible manually via a web-based user interface providing input forms called *masks*. In addition, upload of Excel-files covering multiple data objects in a structured way is supported.

Each SMART-CARE collaboration partner is assigned a custom area within openBis, called *space*, where only persons with respective permissions granted can read and write data. Inside each space, data objects owned by the corresponding partner are stored. For the clinics, this is mostly data regarding patients, samples, and aliquots. Each clinic has custom masks, since not all cancer types investigated within SMART-CARE have the same kind of metadata available. Between clinical partners, all information on the masks is harmonized where possible to facilitate automated analysis. Standardized terminologies are used where possible.

For keeping patient identifying information outside the system, the pseudonymization service *Mainzelliste* [4] was chosen and setup on dedicated a virtual machine. It is integrated as a plugin into the openBis web interface by using the representational state transfer (REST) interface provided. A fast and safe conversion of the sensitive patient identifiers to SMART-CARE Identifiers (SCIDs) which can be used for identification in the system is possible. Only the SCID is stored for each patient within openBis. The workflow for creating a pseudonym is displayed in Figure 1.

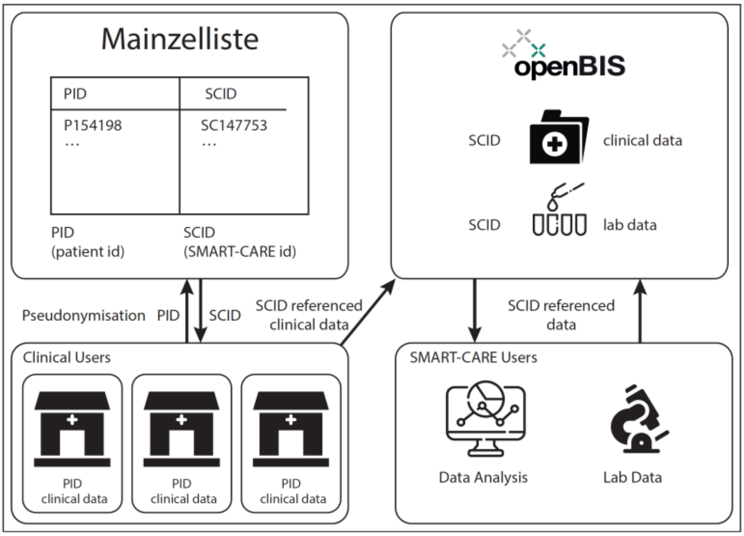

Figure 1. Workflow of the pseudonymization process with Mainzelliste in the SMART-CARE project: The clinical partners create a SCID which is mapped to their sensitive patient identifier. Only the SCID is stored in the SMART-LDR and openBis.

Laboratory partners are using customized open BIS masks within their space, too. In general, all laboratories share the same masks with slight differences between those focusing on proteomics or metabolomics, respectively. While clinics only provide metadata, laboratories create raw mass-spectrometry files and peak processed files. Due to constraints in the open BIS upload process, a custom upload was configured based on the open-source software *Next cloud* [5]. It was setup on a dedicated virtual machine and configured to allow laboratory partners to either install the local Next cloud client or the web browser for uploading data to the SMART-LDR. After the upload is complete, the data is automatically assigned to the corresponding objects inside open BIS by using a filename convention. This enables the upload of files up to 50 GB in size to a network file system (NFS) device. Integration NFS was chosen to gain more flexibility when the demand for storage sizes increased over time. For example, it is possible to migrate data to a large-scale data service of our university's data center.

For the computational members of SMART-CARE, the Python application programming interface (API) of open BIS was enabled, in order to allow retrieval of raw and tabular data from the SMART-LDR using scripts. Additional extensions, which are currently under development can be added into the existing architecture. The final architecture with all components is displayed in Figure 2.

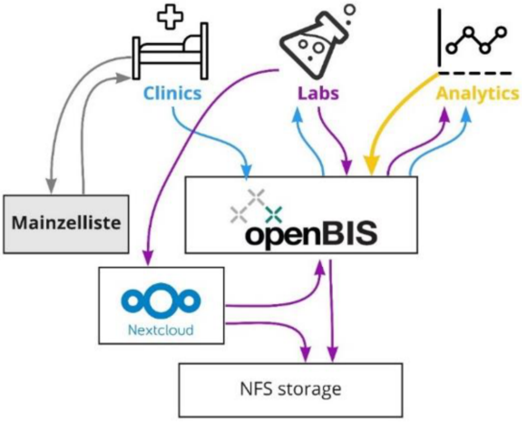

Figure 2. Architecture of the data management of the SMART-CARE project with all components mentioned in this work.

4. Discussion

We have presented architecture for the safe and efficient management of MS data in a medical research context based on open BIS. At the beginning, the system had to be adapted to the needs of the project. For this task, the relevant data structures had to be defined and harmonized across research partners. For this task, the tight collaboration of medical informatics professionals, physicians, laboratory technicians, and data analysts proved to be a big success factor. Due to the nature of the MS data, SMART-LDR is a complex system, which is not very intuitive for first-time users. However, after a short period of training session, all types of users were able to successfully use the system for their respective tasks. This training also helped to establish a culture of systematic data sharing throughout the project.

5. Conclusions

The implemented and configured solutions make up a complete data management solution suitable for optimal documentation and data exchange in SMART-CARE. The centralized approach is useful for protection of patient data, avoiding redundancy and increase collaboration between partners. Additional extensions to improve data analysis and enable downstream computational processes are under development.

Acknowledgment

SMART-CARE is funded by the German Ministry of Education and Research (Grant id: 031L0212A).

References

[1] Srivastava A, Creek DJ. Discovery and Validation of Clinical Biomarkers of Cancer: A Review Combining Metabolomics and Proteomics. Proteomics 2019 may;19 (10):1700448, doi: 10.1002/pmic.201700448.

[2] Ringwald F, Czernilofsky F, Dudchenko A, Ganzinger M, Dietrich S, Knaup P. Data Management for Systems Medicine: The SMART-CARE Joint Environment. Stud Health Technol Inform. 2021 May;281:1104-05.

[3] Bauch A, Adamczyk I, Buczek P, Elmer FJ, Enimanev K, Glyzewski P, Kohler M, Pylak T, Quandt A, Ramakrishnan C, Beisel C, Malmström L, Aebersold R, Rinn B. openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics. 2011 Dec;12:468, doi: 10.1186/1471-2105-12-468.

[4] Lablans M, Borg A, Ückert F. A REST ful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak. 2015 Feb;15:2

[5] Dudchenko A, Ringwald F, Czernilofsky F, Dietrich S, Knaup P, Ganzinger M. Large-File Raw Data Synchronization for openBIS Research Repositories. Stud Health Technol Inform. 2022 May;294:409-10.