Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

© 2022 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220486

Large-File Raw Data Synchronization for openBIS Research Repositories

Aleksei DUDCHENKO[a,1], Friedemann RINGWALD[a], Felix CZERNILOFSKY[b], Sascha DIETRICH[b], Petra KNAUP[a] and Matthias GANZINGER[a]
[a] *Institute of Medical Informatics, Heidelberg University Hospital, Germany*
[b] *Department of Hematology, Heidelberg University Hospital, Germany*

Abstract. In a systems medicine research consortium, openBIS is used as a research data repository. To facilitate efficient upload of large files, openBIS is complemented by a Nextcloud data cloud system. Using a Nextcloud client, raw mass spectrometry data is automatically imported into the repository in the background, enabling comprehensive data provenance.

Keywords. systems medicine, research platform, mass spectrometry

1. Introduction

SMART-CARE (Systems Medicine Approach to Stratification of Cancer Recurrence) is a research project of several clinical, biomedical, and medical informatics partners [1]. Its goal is to build a robust mass spectrometry (MS)-based pipeline for predicting relapse of cancer patients. For all steps of the pipeline, standard operating procedures are developed to ensure reproducible proteome and metabolome analyses. The requirements of individual lab partners for a complex solution for MS data management led to the SMART-CARE Linked Data Repository (Smart LDR) based on the open-source openBIS software [2] to assure correct sample and experiment identification. In this solution, security and access management are provided for large data sets. To ensure complete reproducibility of the analyses and allow for the application of new analytical methods in the future, the consortium decided to add raw MS data files to the Smart LDR. Since the regular web-based interface of openBIS proved to be inconvenient for sharing up to 90 files per experiment with sizes of three to eight gigabytes each, an extension for automatically synchronizing these files had to be newly developed.

2. Methods

For the Smart LDR, openBIS was adopted for collecting, managing, and exchanging complex data and metadata provided by clinics, laboratories, and analytical partners. The raw files generated by mass-spectrometry equipment have to be uploaded to a data server via a web browser interface and are attached to corresponding objects in the Smart LDR.

[1] Corresponding author, Aleksei DUDCHENKO; E-mail: aleksei.dudchenko@med.uni-heidelberg.de

Users suggested the upload process should be done in the background and should handle possible internet disruptions. The uploaded files should be linked to the corresponding objects in Smart LDR containing annotations without manual interaction.

We researched existing approaches allowing for smooth upload of large files. A class of software that is well-established, are systems for setting up private data clouds [3], such as Seafile or Nextcloud. Further requirements were the existence of a local client program in addition to the web browser interface. Since local firewalls only allow for network access via HTTP proxies, only systems with a client suitable for such an environment were considered.

3. Results

As a result of our research, we decided to use the open-source software Nextcloud. The software has a mature state of development and is complemented by client software that works well with HTTP proxies. Nextcloud was set up on a dedicated virtual server and solves the task of transferring data from laboratory machines to the Smart LDR data store server. This was achieved by defining an external storage device in Nextcloud for each lab. The labs were instructed to install the client software and configure it to automatically synchronize a local drop directory with the respective folder on the server.

The server-side storage devices are linked into the openBIS data storage server via Network File System (NFS). As soon as completely transferred files are detected by openBIS, files are moved to the final storage location and deleted from the local directory. Customized openBIS scripts and rules assure correct assignment between files and samples respectively experiments based on a naming convention of the files.

4. Discussion and Conclusions

The solutions we implemented for MS raw data exchange allow seamless data upload for labs and fully automated data binding to responsible samples and objects in the Smart LDR. The major outcome of openBIS extension is the decrease of workload and reduction of burden for lab staff while at the same time ensuring a complete and error-free raw data repository. Since the concept is generic, our approach can be adopted by other projects using openBIS.

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

References

[1] 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 27;281:1104-1105. doi: 10.3233/SHTI210365. PMID: 34042858.

[2] 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 8;12:468.

[3] Kanbar LJ, Shalish W, Robles-Rubio CA, Precup D, Brown K, Sant'Anna GM, Kearney RE. Organizational principles of cloud storage to support collaborative biomedical research. Annu Int Conf IEEE Eng Med Biol Soc. 2015 Aug;2015:1231-4. doi: 10.1109/EMBC.2015.7318589. PMID: 26736489.