MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230950

Large-Scale Standardized Image Integration for Secondary Use Research Projects

Hannes ULRICH[a,1], Michael ANYWAR[a], Benjamin KINAST[a] and Björn SCHREIWEIS[a]
[a] *Institute for Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Campus Kiel, Kiel and Lübeck, Schleswig-Holstein, Germany*
ORCiD ID: Hannes Ulrich https://orcid.org/0000-0002-8349-6798, Michael Anywar https://orcid.org/0000-0001-8028-803X, Benjamin Kinast https://orcid.org/0000-0003-2554-4381, Björn Schreiweis https://orcid.org/0000-0002-1748-1563

Abstract. Imaging techniques are a cornerstone of today's medicine and can be crucial for a successful therapy. But in addition, the generated imaging series are an important resource for new informatics' methods, especially in the field of artificial intelligence. This paper describes the success of integrating clinical routine imaging data into a standardized format for research purposes. Thus, we designed an integration flow and successfully implemented it in the local data integration center of University Hospital Schleswig-Holstein. The flow integrates imaging series and radiological reports from the primary system into an openEHR repository with enrichment by semantic codes for better findability and retrieval using HL7 FHIR. As a result, 6.6 million radiological studies with 29 million image series are now available for further medical (informatics) research.

Keywords. Clinical data, radiology, data integration, big data

1. Introduction

In modern healthcare, imaging procedures are of crucial importance in patient treatment and research. However, central Picture Archiving and Communication Systems (PACS) only store data from specific modalities (e.g., MRI, CT, ultrasound) consequently. In addition, physicians only report these data in the radiological information system (RIS). But the majority of data are either not stored in well-established IT systems or are just available in special departmental systems in heterogeneous, non-standardized formats [1]. The situation is further compounded by the fact that the medically relevant information like radiological reports is usually only available in a free-text form which needs further processing. Thus, sustainable and reproducible reuse of these data is elusive.

The junior research group IMPETUS is dedicated to precisely these challenges. The main goal of the IMPETUS project is to extend the Medical Data Integration Center

[1] Corresponding Author: Dr. Hannes Ulrich, Institute for Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Campus Kiel, Kiel and Lübeck, Schleswig-Holstein, Germany; email: hannes.ulrich@uksh.de.

(MeDIC) [2] established at UKSH as part of the HiGHmed consortium [3] to enable the integration and further reuse of all multimedia objects and reports, regardless of the format, storage or presentation in a standardized manner. The clinical routine IT systems mainly transmit data to the MeDIC via HL7 messages, which ultimately stores them in the openEHR repository for further research projects. The MeDIC uses big-data technologies to handle the vast amount of incoming data [4]. Technically speaking, processing frameworks like Apache Nifi and Kafka, and an on-premise Ceph S3 object storage solution are already in productive use. This study focuses on integrating routine imaging data and the corresponding diagnostic reports into the MeDIC environment to enable later research reuse under the consistent use of two interoperable standards, HL7 FHIR and openEHR.

2. Methods

In order to identify and categorize the challenges facing large-scale image integration, we used an established requirements catalog [5] and determined the major requirements for our integration scenario:

R1. The integration must be based on healthcare interoperability standards.
R2. The integration must use international terminologies to support interoperability.
R3. The integration must handle unstructured data by using natural language processing (NLP).
R4. The data must be linked from the clinical routine IT systems.
R5. The integration must preserve the raw data.
R6. The integration must create, transfer and process logs.
R7. The integration must provide metadata processing.
R8. The integration must follow European, national and local data privacy and legal regulations.
R9. The integration must provide pseudonymization.

According to the stated requirements for integrating image data from clinical routine, in addition the following infrastructural requirement should be considered: the existing MeDIC [4] infrastructure should be reused as extensively as possible.

3. Results

Due to the intended reuse of the MeDIC components, the technical foundation was already set, so the focus during implementation could be placed on the other requirements, such as standards usage and process traceability.

We base the established integration process, shown in Figure 1, on two daily CSV reports transferred into the MeDIC object storage from different systems. One report contains the metadata for the images originating from the PACS, the other the corresponding radiological reports received from the RIS. The metadata CSV report contains crucial information for further processing and findability, like patient and image series identifiers. In addition, contextual information, like a series description, the modality, and laterality are included. The integration flow does not require retrieval of

the actual images, but the image reference according to the metadata (R4, R9). These can be accessed from the PACS at any time via the metadata. The second CSV export contains the reports and the physicians' assessments of the imaging series, including the encounter information. Thus, the image series and reports can be linked with other medical data in the MeDIC, such as laboratory or medication data.

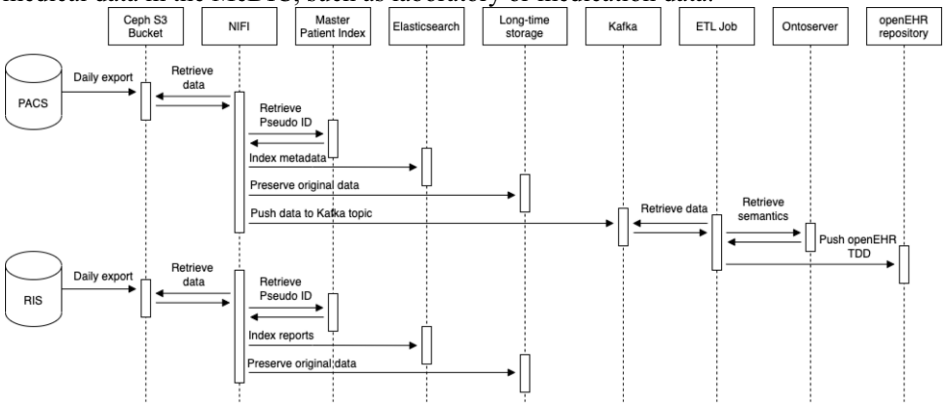

Figure 1. The raw data are exported from the clinical routine IT system into an on-premise Ceph S3 bucket and then continuously progresses through the shown integration flow. Nifi monitors the import bucket and processes each incoming CSV file in order to request the pseudonymization identifier from the central master patient index, register the image metadata in Elasticsearch, store the raw data into a long-time archive and share the data via a Kafka topic with the ETL cluster to generate openEHR template data documents (TDDs).

The CSV reports are thenceforward processed using the MeDIC ingestion procedures: a Nifi flow digests the raw data, extracts the valuable information, links the patient information to the central master patient index (R9), stores the raw data into a Ceph S3 object storage for long-time archiving (R5) and populates the raw data into Kafka topics for distribution to the downstream ETL cluster. The imaging data is converted to a standardized openEHR data model (R1) and stored in the central openEHR repository for further use. For this purpose, we use the archetype "openEHR-EHR-CLUSTER.imaging_series.v0" [6], which is modified locally to represent additional information such as laterality or body part. To store these additional fields in a sustainable and reusable way, we map them to FHIR R4b CodeSystems used by the ImagingStudy resource and provided as a semantic-coded value (R2). The mapping of the laterality contains three entries, while the body part mapping contains over 150 entries.

Quantitatively, about 29 million imaging series, e.g., 6.9 million magnetic resonance series, are now available for further scientific use.

4. Discussion

This work has accelerated the technical integration and processing within MeDIC in terms of clinical image series findability and accessibility. Seven out of nine of the requirements previously identified are addressed and greatly influenced the development of the integration flow. To address more closely the first requirement of using healthcare interoperability standards, the initial integration concept was based on HL7 messages expected between HIS and PACS. Unfortunately, there was no possibility to acquire the needed information as in our setting the communication between HIS and PACS/RIS is

only internally as a monolithic system. Therefore, the data has to be extracted via the CSV reports. The MeDIC architecture, primarily designed to handle message-based communications, is also able to prove its effectiveness in dealing with large files (ca. 500 megabyte), which tend to plug the integration flows and create bottlenecks. NIFI, in particular, is able to handle the high volume of data effectively and without performance degradation. In addition, Nifi includes an advanced auditing system that meticulously documents each processing step as well as changes and stores the audit trail in distinct provenance repositories [7].

The final representation in openEHR fits into the overarching MeDIC architecture allowing us to offer a feasibility portal to researchers which also allows them to query for imaging data. The use of openEHR allows data to be presented according to FAIR principles [8,9], so our integration has taken routine data a big step towards FAIRification. Only the archetype [6] had to be extended locally to store relevant information: *body part* and *laterality* and we are looking at how we can feedback our adaptation to the international community. Adding and semantically enhancing those two attributes enables effective prefiltering in the downstream openEHR repository. However, the mapping of the *body part* to SNOMED CT is not fully complete. About 1% of the data could not be reliably identified due to typos and abbreviations in the source data. However, it should be mentioned that the original entries are preserved next to the semantically annotated entries.

Two missing requirements, NLP usage and privacy regulations, must be addressed separately. The use of NLP methods (R3) had to be stalled, as anonymization is a particular challenge, especially in the free-text findings. The reports for the image series were accordingly imported and registered in the index, but the extraction into the archetypes "Imaging examination result" [10] for the staging in the openEHR repository with the aid of NLP methods [11] is the next integration goal. With respect to data privacy (R8), the UKSH has already a broad consent in use since 2017 allowing re-use of clinical data for research [12]. In addition, procedures will be established to allow anonymization so data collected before 2017 are usable. In addition, for each project requesting imaging data, an analysis is conducted to ensure the scientific purposes are fulfilled in accordance with Articles 9 and 89 of the European General Data Protection Regulation.

5. Conclusions

With our work, 6.6 million radiological studies with 29 million clinical imaging series from one of the largest hospitals in Germany are enabled for further research projects. Due to the enormous amount of data, the technical integration is realized based on established big data technologies.

The next major step will be the flow expansion with automatic image processing. This tagging and indexing should make the image data more effectively accessible and easier to find via content-based image retrieval (CBIR). The follow-up research question emerges: How can well-established CBIR procedures be applied to non-DICOM objects and biosignals?

Acknowledgments

This research was funded by German Federal Ministry of Education and Research, grant number 01ZZ1802T and 01ZZ2011.

References

[1] Benson T, Grieve G. Principles of health interoperability: SNOMED CT, HL7 and FHIR. Cham: Springer; 2016, doi: 10.1007/978-3-319-30370-3.
[2] Kock-Schoppenhauer AK, Schreiweis B, Ulrich H, Reimer N, Wiedekopf J, Kinast B, Busch H, Bergh B, Ingenerf J. Medical Data Engineering–Theory and Practice. In: Bellatreche L, Chernishev G, Corral A, Ouchani S, Vain J, editors. Advances in Model and Data Engineering in the Digitalization Era. Communications in Computer and Information Science; 2021 June 21-23; Cham: Springer; 1481. p. 269-84, doi: 10.1007/978-3-030-87657-9_21.
[3] Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, Knaup-Gregori P, Bavendiek U, Dieterich C, Brors B, Kraus I, Thoms CM, Jäger D, Ellenrieder V, Bergh B, Yahyapour R, Eils R, Consortium H, Marschollek M. HiGHmed–an open platform approach to enhance care and research across institutional boundaries. Methods Inf Med. 2018 May;57(S 01):e66-81, doi: 10.3414/ME18-02-0002.
[4] Cheng KY, Pazmino S, Schreiweis B. ETL processes for integrating Healthcare Data - Tools and Architecture patterns. Oslo: IOS Press; 2022. 151-6 p, doi: 10.3233/SHTI220974.
[5] Kinast B, Ulrich H, Bergh B, Schreiweis B. Functional Requirements for Medical Data Integration into Knowledge Management Environments: Requirements Elicitation Approach Based on Systematic Literature Analysis. J Med Internet Res. 2023 Feb;25:e41344, doi: 10.2196/41344.
[6] openEHR Foundation, Imaging series - openEHR Clinical Knowledge Manager, (2022). https://ckm.openehr.org/ckm/archetypes/1013.1.6012 (accessed October 5, 2022).
[7] Isah H, Zulkernine F. A Scalable and Robust Framework for Data Stream Ingestion. 2018 IEEE International Conference on Big Data (Big Data); 2018 Dec 10; Seattle, WA, USA. p. 2900-5. doi: 10.1109/BigData.2018.8622360.
[8] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar;3(1):160018, doi: 10.1038/sdata.2016.18.
[9] Frexia F, Mascia C, Lianas L, Delussu G, Sulis A, Meloni V, Del Rio M, Zanetti G. openEHR Is FAIR-Enabling by Design. Stud Health Technol Inform. 2021 May;281:113-7, doi: 10.3233/SHTI210131.
[10] openEHR Foundation, Imaging examination result - openEHR Clinical Knowledge Manager, (2022). https://ckm.openehr.org/ckm/archetypes/1013.1.6012 (accessed October 5, 2022).
[11] Maros ME, Cho CG, Junge AG, Kämpgen B, Saase V, Siegel F, Trinkmann F, Ganslandt T, Groden C, Wenz H. Comparative analysis of machine learning algorithms for computer-assisted reporting based on fully automated cross-lingual RadLex mappings. Sci Rep. 2021 Mar;11(1):5529, doi: 10.1038/s41598-021-85016-9.
[12] Richter G, Krawczak M, Lieb W, Wolff L, Schreiber S, Buyx A. Broad consent for health care–embedded biobanking: understanding and reasons to donate in a large patient sample. Genet Med. 2018 Jan;20(1):76-82, doi: 10.1038/gim.2017.82.