MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220026

AHD2FHIR: A Tool for Mapping of Natural Language Annotations to Fast Healthcare Interoperability Resources - A Technical Case Report

Raphael Scheible[a,b], Deniz Caliskan[c], Patrick Fischer[d], Fabian Thomczyk[b], Susanne Zabka[b], Henning Schneider[d], Martin Boeker[a,b], Stefan Schulz[e], Hans-Ulrich Prokosch[f], Christian Gulden[f]

[a] Institute of Medical Informatics, Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Munich, Germany,
[b] Institute of Medical Biometry and Statistics, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany,
[c] Medical Center for Information and Communication Technology, University Hospital Erlangen, Erlangen, Germany,
[d] Institute of Medical Informatics, Faculty of Medicine, Justus-Liebig-University Giessen, Giessen, Germany,
[e] Averbis GmbH, Freiburg, Germany,
[f] Chair of Medical Informatics, Department of Medical Informatics, Biometrics and Epidemiology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

Abstract

A significant portion of data in Electronic Health Records is only available as unstructured text, such as surgical or finding reports, clinical notes and discharge summaries. To use this data for secondary purposes, natural language processing (NLP) tools are required to extract structured information. Furthermore, for interoperable use, harmonization of the data is necessary. HL7 Fast Healthcare Interoperability Resources (FHIR), an emerging standard for exchanging healthcare data, defines such a structured format. For German-language medical NLP, the tool Averbis Health Discovery (AHD) represents a comprehensive solution. AHD offers a proprietary REST interface for text analysis pipelines. To build a bridge between FHIR and this interface, we created a service that translates the communication around AHD from and to FHIR. The application is available under an open source license.

Keywords:

Health Information Interoperability, FHIR, NLP, Electronic Health Records

Introduction

Major parts of the Electronic Health Record (EHR) are only available as unstructured narratives, e.g. surgical or finding reports, clinical notes and discharge summaries. In order to extract structured data for secondary use, it is therefore necessary to leverage natural language processing (NLP) tools. For text analysis and information extraction a broad range of tools and systems have been developed, tools [7,13,16,24,27–29], among which there are also specialized ones for medical NLP [10,11,14] which find its application in many areas [17,26]. Many of them make use of medical terminologies. Unfortunately, many of these tools embed their results into different formats. In order to homogenize the heterogeneous landscape of data formats in the clinical area, HL7 created Fast Healthcare Interoperability Resources (FHIR), an information modelling standard for health care data exchange. First published in 2014 as a draft for trial use, FHIR is a standard for exchanging healthcare data [30]. FHIR describes data formats and elements as so-called resources, e.g. Patient, Medication and DocumentResource. In addition, interfaces for the exchange of data are defined, i.e. the web-based API technology REST, whereby the data are represented in XML or JSON.

The NLP2FHIR project has already taken advantage of FHIR as a basic data structure. Hong et al. map in their work formats from various NLP tools to FHIR [4–6]. Unfortunately, this work along with most approaches related to medical NLP have been done for English documents only [15].

However, with recent investments in German medical informatics under the Medical Informatics Initiative (MII) [2,22], the application of NLP technology to clinical documents in German language have increased significantly. The MII, funded by the German Federal Ministry of Education and Research, has set itself the goal of making data from patient care and research more usable. Four consortia, composed by large hospitals in Germany, have committed themselves to this task. One of these consortia is MIRACUM [18] which is composed of 10 university hospitals and one industrial partner, Averbis, who provides the natural language processing (NLP) toolkit. Their tool Averbis Health Discovery (AHD) is an NLP solution specifically developed for German medical text analysis. For cross-consortium and cross-site communication FHIR was chosen as a format for data homogenization, due to its increasing popularity, despite alternatives such as openEHR, which were already used in the context of German NLP [25]. An important advantage for FHIR in MIRACUM is that it was chosen as the target format to which structured data obtained from clinical information systems are already mapped. In recent years, the popularity of FHIR has steadily increased, especially in the context of data models for mobile and web applications as well as medical devices [12]. Leveraging the standard simplifies data integration into existing FHIR infrastructures. With Averbis Health Discovery (AHD) becoming more and more established in German-speaking countries [3,21,23], but which does not yet provide a FHIR output in its current version (planned for future versions), we see a strong need to address this issue.

In this paper, we report on the implementation of a service that maps extracted codes and context information from the proprietary annotation format to FHIR. supporting fhir both as an input and output format. We name the tool AHD2FHIR.

Methods

Averbis Health Discovery

An essential feature of Averbis Health Discovery (AHD) [31] is its optimisation towards the analysis of German clinical language. It is a commercial product which is based on UIMA, similar to other established tools such as e.g. cTAKES [20]. Therefore, annotators can efficiently be developed using the extensive rule-based scripting language RUTA [9]. Furthermore, AHD annotations use terminologies such as ICD10, LOINC and ATC. Averbis has suggested FHIR additions for NLP [1] and is working on FHIR integration. This functionality is not available in the current version 5.39. Documents need to be sent as plain text or html to AHD's REST interface. The result is a proprietary JSON format, which contains one or multiple annotations, each one defined by the Averbis type system [32], which can be extended by custom annotators.

Averbis annotation types describe the output of the annotator with information like the text position in the document, a terminology code (e.g. ICD-10, SNOMED CT, RxNorm, Abdamed), a normalized representation of the concept, together with context information like negation, certainty etc. The structured layout of this type system allows to infer mappings to FHIR resources which has already been formally discussed in [1].

Design

The main requirement for AHD2FHIR is the support of receiving resources and rendering the results as FHIR output. Clinical text is sent from the client to the application in the form of DocumentReferences. AHD2FHIR takes over further communication with AHD in proprietary formats. The result of the text analysis is subsequently returned to the client in FHIR (see Figure 1). To simplify deployment and integration with established FHIR patterns, the design of a web service defines another requirement. The tool is built upon FHIR R4.

Results

Process

AHD2FHIR is a server-based software, which can be used in two forms: via a REST API for synchronous and additionally via Apache Kafka [33] for asynchronous processing. In the case of the REST API the client sends the document as DocumentReference to the API. The mapper forwards the unpacked document text to the AHD API and maps the returned annotation to FHIR. Finally, the client receives these FHIR resources as a result. Besides the REST interface, the service is also capable of integrating with Apache Kafka, allowing it to asynchronously process FHIR DocumentReferences. Therefore, the document is sent as DocumentReference with a HTTP client. Thereafter, the mapping is processed as described before in the Kafka workflow. The only difference is, that the obtained FHIR resources are directly sent back via HTTP, instead of being transformed to a topic by the Kafka producer. Figure 2 depicts the API workflows.

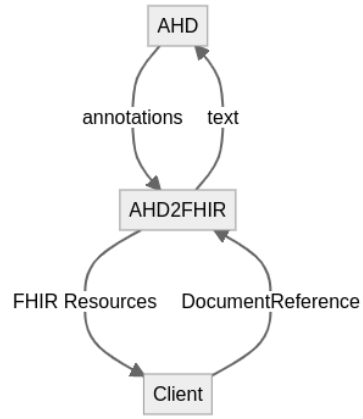

Figure 1 – AHD2FHIR takes as input a Document-Reference, handles the communication with AHD and returns the results as FHIR Resources.

Implementation

The mapping service was implemented using Python 3.9, the web framework FastAPI [19], and the fhir.resources library [8]. Both processing methods, Kafka and REST API, expect a JSON-encoded FHIR Bundle of DocumentReferences or a single Document-Reference as input. The textual content is extracted from each DocumentReference resource and sent to the AHD REST API using a library provided by Averbis [34]. The implementation currently supports mapping the returned result to FHIR Condition, Medication, and MedicationStatement resources. An example mapping is shown in Figure 3.

AHD supports detecting negated conditions, i.e. the absence of a specific diagnosis. We apply post-processing to filter out such results.

Deployment

The service is packaged as a container and can be run using tools such as Docker and Docker Compose. Further, deployment on Kubernetes is fully supported, providing additional benefits such as self-healing, horizontal scalability, and simplified observability of the service.

Availability

The source code of AHD2FHIR is available at https://github.com/miracum/ahd2fhir.

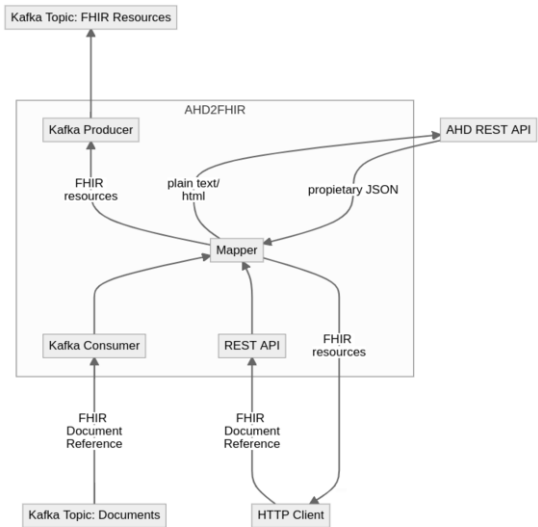

Figure 2 – Implementation Diagram of the AHD2FHIR service

Discussion

AHD2FHIR deploys FHIR as both an input and output format so that it can be used together with different NLP tools independently of their implementation details. While this requires an initial mapping step from existing documents to FHIR DocumentReferences, the advantage is that this structure encodes additional meta-data about the text; both the encounter and the patient the document relates to can be captured and used as part of the mapping (see also Figure 3). To provide provenance information for the mapped FHIR resources, Daumke et al. [1] and Hong et al. [4] propose NLP-specific extensions containing, amongst others, a confidence measure and the offset inside the document where the concept was detected. To provide stable identifiers when creating multiple FHIR resources from text is challenging. Unambiguous identifiers prevent the creation of duplicate resources in case of multiple processing of the same document. We are currently using a concatenation of the document identifier, the offset into the text, and the concept value to construct each mapped resource's identity. However, this approach is only stable as long as the document contents, used annotators and pipelines don't change, or the document is only appended to. Currently, only a limited subset of AHD's annotations are mapped to FHIR resources by our implementation. Mapping additional types such as laboratory values and scores and classifications is planned.

Conclusions

In this work, we presented AHD2FHIR, a tool acting as a FHIR-based abstraction layer over the proprietary NLP tool Averbis Health Discovery. Both input and output have to be structured with FHIR resources. This step is a prerequisite to design cross-site queries with the tools developed in MIRACUM and beyond. The presented tool successfully maps annotations, extracted by AHD, to FHIR Condition, Medication, and MedicationStatement resources. Moreover, pre- and post-processing of data before mapping to FHIR is possible, which can improve the quality of the mapped resources. Future work will investigate the compatibility with different NLP tooling and open questions related to resource provenance. By publishing the source code, we hope to foster cooperation between the sites in the field of German medical NLP.

Acknowledgements

This work was supported by the German Ministry for Education and Research (BMBF FKZ 01ZZ1801B, 01ZZ1804A,

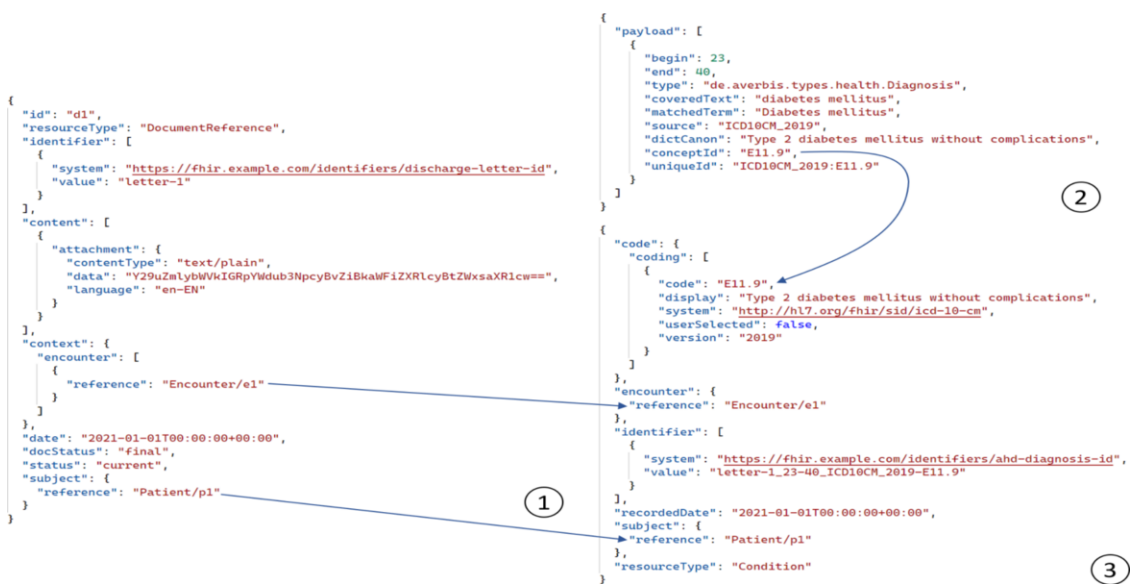

Figure 3 – An example for mapping a FHIR DocumentReference (1) to a FHIR Condition resource (3) using extracted information from the AHD API (2). The analyzed content string is "confirmed diagnosis of diabetes mellitus" encoded as base-64. The arrows indicate the mapping of the structured diagnosis and show preservation of the care context via the DocumentReference's subject and encounter information.

01ZZ1801D, 01ZZ1801J, 01ZZ1801A).

This study was performed for author CG to (partially) fulfill the requirements for obtaining the academic degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

Conflicts of Interest

Stefan Schulz is employed by Averbis GmbH, the company that develops the tool used in this work.

References

[1] P. Daumke, K.U. Heitmann, S. Heckmann, C. Martínez-Costa, and S. Schulz, Clinical Text Mining on FHIR, *Stud. Health Technol. Inform.* **264** (2019) 83–87. doi:10.3233/SHTI190188.

[2] S. Gehring, and R. Eulenfeld, German Medical Informatics Initiative: Unlocking Data for Research and Health Care, *Methods Inf. Med.* **57** (2018) e46–e49. doi:10.3414/ME18-13-0001.

[3] B. Grundel, M.-A. Bernardeau, H. Langner, C. Schmidt, D. Böhringer, M. Ritter, P. Rosenthal, A. Grandjean, S. Schulz, P. Daumke, and A. Stahl, Merkmalsextraktion aus klinischen Routinedaten mittels Text-Mining, *Ophthalmol.* **118** (2021) 264–272. doi:10.1007/s00347-020-01177-4.

[4] N. Hong, A. Wen, M.R. Mojarad, S. Sohn, H. Liu, and G. Jiang, Standardizing Heterogeneous Annotation Corpora Using HL7 FHIR for Facilitating their Reuse and Integration in Clinical NLP, *AMIA. Annu. Symp. Proc.* **2018** (2018) 574–583.

[5] N. Hong, A. Wen, F. Shen, S. Sohn, S. Liu, H. Liu, and G. Jiang, Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data, *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.* **2017** (2018) 74–83.

[6] N. Hong, A. Wen, F. Shen, S. Sohn, C. Wang, H. Liu, and G. Jiang, Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data, *JAMIA Open*. **2** (2019) 570–579. doi:10.1093/jamiaopen/ooz056.

[7] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, Zenodo, 2020. doi:10.5281/zenodo.1212303.

[8] M.N. Islam, nazrulworld/fhir.resources, 2021. https://github.com/nazrulworld/fhir.resources (accessed April 21, 2021).

[9] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, UIMA Ruta: Rapid development of rule-based information extraction applications, *Nat. Lang. Eng.* **22** (2016) 1–40. doi:10.1017/S1351324914000114.

[10] T.A. Koleck, C. Dreisbach, P.E. Bourne, and S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, *J. Am. Med. Inform. Assoc. JAMIA*. **26** (2019) 364–379. doi:10.1093/jamia/ocy173.

[11] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S.F. Jones, R. Forshee, M. Walderhaug, and T. Botsis, Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review, *J. Biomed. Inform.* **73** (2017) 14–29. doi:10.1016/j.jbi.2017.07.012.

[12] M. Lehne, S. Luijten, P. Vom Felde Genannt Imbusch, and S. Thun, The Use of FHIR in Digital Health - A Review of the Scientific Literature, *Stud. Health Technol. Inform.* **267** (2019) 52–58. doi:10.3233/SHTI190805.

[13] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, and D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: Assoc. Comput. Linguist. ACL Syst. Demonstr., 2014: pp. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

[14] M. Neumann, D. King, I. Beltagy, and W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proc. 18th BioNLP Workshop Shar. Task, Association for Computational Linguistics, Florence, Italy, 2019: pp. 319–327. doi:10.18653/v1/W19-5034.

[15] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: opportunities and challenges, *J. Biomed. Semant.* **9** (2018) 12. doi:10.1186/s13326-018-0179-8.

[16] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, fairseq: A Fast, Extensible Toolkit for Sequence Modeling, in: Proc. NAACL-HLT 2019 Demonstr., 2019.

[17] E. Pons, L.M.M. Braun, M.G.M. Hunink, and J.A. Kors, Natural Language Processing in Radiology: A Systematic Review, *Radiology*. **279** (2016) 329–343. doi:10.1148/radiol.16142770.

[18] H.-U. Prokosch, T. Acker, J. Bernarding, H. Binder, M. Boeker, M. Boerries, P. Daumke, T. Ganslandt, J. Hesser, G. Höning, M. Neumaier, K. Marquardt, H. Renz, H.-J. Rothkötter, C. Schade-Brittinger, P. Schmücker, J. Schüttler, M. Sedlmayr, H. Serve, K. Sohrabi, and H. Storf, MIRACUM: Medical Informatics in Research and Care in University Medicine, *Methods Inf. Med.* **57** (2018) e82–e91. doi:10.3414/ME17-02-0025.

[19] S. Ramírez, tiangolo/fastapi, 2021. https://github.com/tiangolo/fastapi (accessed May 13, 2021).

[20] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, and C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, *J. Am. Med. Inform. Assoc.* **17** (2010) 507–513. doi:10.1136/jamia.2009.001560.

[21] S. Schulz, S. Fix, P. Klügl, T. Bachmayer, T. Hartz, M. Richter, N. Herm-Stapelberg, and P. Daumke, Comparative evaluation of automated information extraction from pathology reports in three German cancer registries, *GMS Med. Inform. Biom. Epidemiol.* **17** (2021) Doc01. doi:10.3205/mibe000215.

[22] S.C. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative, *Methods Inf. Med.* **57** (2018) e50–e56. doi:10.3414/ME18-03-0003.

[23] C. Weber, L. Röschke, L. Modersohn, C. Lohr, T. Kolditz, U. Hahn, D. Ammon, B. Betz, and M. Kiehntopf, Optimized Identification of Advanced Chronic Kidney Disease and Absence of Kidney Disease by Combining Different Electronic Health Data Resources and by Applying Machine Learning Strategies, *J. Clin. Med.* **9** (2020) 2955. doi:10.3390/jcm9092955.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, HuggingFace's Transformers: State-of-the-art Natural Language Processing, *ArXiv191003771 Cs*. (2020). http://arxiv.org/abs/1910.03771 (accessed May 3, 2020).

[25] A. Wulff, M. Mast, M. Hassler, S. Montag, M. Marschollek, and T. Jack, Designing an openEHR-Based Pipeline for Extracting and Standardizing Unstructured Clinical Data Using Natural Language Processing, *Methods Inf. Med.* **59** (2020) e64–e78. doi:10.1055/s-0040-1716403.

[26] W.-W. Yim, M. Yetisgen, W.P. Harris, and S.W. Kwan, Natural Language Processing in Oncology: A Review, *JAMA Oncol.* **2** (2016) 797–804. doi:10.1001/jamaoncol.2016.0213.

[27] Apache OpenNLP, (n.d.). https://opennlp.apache.org/ (accessed March 31, 2021).

[28] John Snow Labs - Spark NLP, (n.d.). https://nlp.johnsnowlabs.com/ (accessed March 31, 2021).

[29] deepset-ai/FARM, deepset, 2020. https://github.com/deepset-ai/FARM (accessed May 14, 2020).

[30] Index - FHIR v4.0.1, (n.d.). http://hl7.org/fhir/ (accessed March 30, 2021).

[31] Averbis Health Discovery - Analyse von Patienten Daten, *Averbis GmbH*. (n.d.). https://averbis.com/de/health-discovery/ (accessed March 30, 2021).

[32] averbis/health-typesystems, Averbis GmbH, 2021. https://github.com/averbis/health-typesystems (accessed March 31, 2021).

[33] apache/kafka, The Apache Software Foundation, 2021. https://github.com/apache/kafka (accessed May 13, 2021).

[34] averbis/averbis-python-api, Averbis GmbH, 2021. https://github.com/averbis/averbis-python-api (accessed March 31, 2021).

Address for correspondence

Raphael Scheible, IMedIS, Klinikum rechts der Isar der Technischen Universität München, Ismaninger Str. 22, 81675 München, Germany, raphael.scheible@tum.de