FAIRness of openEHR Archetypes and Templates

Caroline Bönisch[*, 1[0000-0001-7169-6090]], Anneka Sargeant[*, 1[0000-0003-3289-948X]], Antje Wulff[2[0000-0002-2550-2627]], Marcel Parciak[1[0000-0002-6950-929X]], Christian R Bauer[1[0000-0003-2613-419X]], Ulrich Sax[1[0000-0002-8188-3495]]

[1]Department of Medical Informatics, University Medical Center Goettingen, Robert-Koch-Str. 40, 37075 Goettingen, Germany
[2]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Carl-Neuberg-Str.1, 30625 Hannover, Germany

Abstract. *Background:* The FAIR Data Publishing Group designed 15 principles to quantify the FAIRness of scientific data. By using the FAIR Principles it is possible to make scientific data findable, accessible, interoperable and reusable. This paper checks the FAIRness of openEHR archetypes and templates as formalisms to preserve semantic interoperability in electronic health records. *Objectives:* Within the semantic framework of the HiGHmed project, the aim is to exchange harmonized data between various institutions and make them available for research, by modelling archetypes and templates within openEHR. To ensure interoperability across various locations, archetypes and templates have been examined in this paper with regard to the FAIR principles (Findable, Accessible, Interoperable and Re-Useable). *Methods:* Analysis of the archetypes developed in HiGHmed and stored in the HiGHmed Clinical Knowledge Manager to determine the degree of fulfillment of FAIRness. *Results:* All fifteen FAIR Principles are met, respectively partially fulfilled. The openEHR approach and the Clinical Knowledge Manager as a collaborative library are compatible with the FAIR Principles and are well suited for the exchange of research data.

Keywords: openEHR, FAIR, Principles, HiGHmed, Archetypes, Clinical Knowledge Manager, Modelling

1 Introduction

The collection and processing of health data is a prerequisite for medical care and patient management. Health data is often recorded electronically in a variety of application systems (e.g. radiological information system or laboratory information system) in different formats (Bauer, et al., 2016). In clinics, heterogeneity is intensified not only by divergent departmental and personal documentation approaches, but also by the fact that technical and clinical parameters of similar examination devices are not described in the same way by different manufacturers (Krefting, et al., 2010). Often the exchange and comparison of otherwise equivalent data is hindered by e.g. not using a standardized format, which leads to redundancies and inconsistencies and thus has a significant impact on data quality.

*Both authors contributed equally to this manuscript

The HiGHmed project funded by the Federal Ministry of Education and Research (BMBF) aims to establish a shared information governance framework connecting local medical data integration centers. The primary goals are the shared use of heterogeneous data from different clinical departments and the reuse of collected data for research purposes and clinical care. Exemplary use cases established in the HiGHmed project are demonstrating the feasibility of the planned governance framework. These medical driven use cases are located in the areas cardiology, oncology and infection control and pursuing different objectives. For example, the infection use case is developing an early detection system trying to encounter outbreaks of multidrug-resistant germs.

To achieve the different use case objectives, the data has to be introduced and managed in a semantic framework, which allows to separate "*the knowledge and information levels in information systems*." (Beale, 2002) Information means specific characteristics of an entity. This includes information that can be assigned to a dedicated patient (for example John Doe with a blood pressure of 120 to 80). Interpretation of information (knowledge) denotes statements, which apply to all entities of a class and are not dependent on a patient. Archetypes, as described in openEHR, form the distinction of the knowledge: "The term archetype is used to denote knowledge level models which define valid information structures." (Beale, 2002) . Templates contain a context-specific set of archetypes, where the used archetypes can be further constrained to meet the specific requirements. Templates are often used to represent medical reports or findings.

HiGHmed follows the openEHR approach to reach semantic interoperability (Haarbrandt, 2018). Clinical concepts, terminologies and services are combined and converted into a machine-readable form that is still easily understandable by humans.

2 Objectives

We aim to evaluate the FAIRness of our openEHR approach (archetypes and templates) to create semantic interoperable data models within HiGHmed. The evaluation of medical data is not part of this work.

3 Methods

In the HiGHmed project, archetypes and templates are created to enable sharing of harmonized data across various institutions for three different use cases (cardiology, oncology, infection control). As described in Wulff et al. (Wulff, 2019), 79 archetypes had been identified in the context of HiGHmed. Most of these archetypes were already available at the public instance of the Clinical Knowledge Manager (CKM) (https://www.openehr.org/ckm/) and could be used of-the-shelf with minor changes to fit the HiGHmed requirements. The CKM in general serves as a collaborative tool to support the modelling process and acts as a repository of archetypes and templates. The identified archetypes are constrained and serve as building blocks for templates. Cur-

rently, 12 templates, are part of the HiGHmed modelling process. The of-the-shelf archetypes were already FAIR compliant and have not been modified within the HiGHmed Project. Some archetypes such as “ethnic background” could not be adopted from the international CKM and were created according to the requirements of the clinicians involved.

All templates were newly created in the course of the project, taking care that the FAIR principles were taken into account. The role of the HiGHmed project is not only to re-use archetypes and templates but also to create FAIR compliant archetypes and templates, specific to the HiGHmed use cases, where needed.

The FAIR Data Publishing Group designed fifteen principles to quantify levels of FAIRness, such as the principle F2. “data are described with rich metadata”, in 2016. To check the developed HiGHmed archetypes and templates for their FAIRness, the FAIR Principles of Wilkinson et al. (Wilkinson, et al., 2016) are taken into account. The FAIR Principles define further characteristics for the terms Findable, Accessible, Interoperable and Re-usable. To make data findable, it must be equipped with a globally unique, persistent identifier and enriched with extensive metadata. Data has to be registered or indexed in a searchable resource and the metadata needs to be specified by the data identifier.

In order to be accessible, data sets shall be retrievable by their unique identifier using a standardized communication protocol that is open, free and universally implementable. Furthermore, the protocol has to allow an authentication and authorization procedure, where it is necessary. Metadata should be accessible, even when the data is no longer available. For data to be interoperable, data should use a formal, accessible, shared and broadly applicable language for knowledge representation and use FAIR compliant vocabularies. The reference between data and other (meta)data needs to be included (Wilkinson, et al., 2016).

To be re-usable, data must have a plurality of accurate and relevant attributes and need to be released with a clear and accessible usage licence. Moreover, data have to be associated with their provenance and meet the domain-relevant community standards.

In scope of this examination, we assessed the archetypes and templates in their compliance to the fifteen FAIR Principles in the HiGHmed CKM. They have been analysed, regarding to the generally characteristic of an archetype or template such as header information, concept name or reference information.

4 Results

The following sections are divided into the four categories of the FAIR principles (Findable, Accessible, Interoperable and Re-Useable). In the category “Findable”, four out of four principles could be met in the HiGHmed CKM. However, in the category “Accessible” there were only two of the four principles that are fulfilled and two principles that are partially fulfilled in the HiGHmed project. Furthermore, four out of four

principles could be met in the category “Re-usable” and three out of three principles in the category “Interoperable”.

The subsequent paragraphs describe which criteria have contributed to fulfil the particular FAIR principles.

Findable

The principle F1 is achieved by the Attribution Build Uid and Major Version ID. The Build Uid is unique to the corresponding instance of the archetype and is initially set during the creation of the archetype. It changes whenever the archetype is uploaded, checked out or committed. Based on these two IDs, archetypes and templates can be kept persistent at any time.

Principle F2 is closely connected with principle R1. To fulfil the principle F2 archetypes and templates provide multiple mandatory attributes. In addition to the Build Uids and Major Version ID, attributes contain information on the archetype ID, the assigned licence, the original author, the current custodian, and other contributors or translators. All these aspects are also available in the XML representation of the archetype. Within the HiGHmed project, the tool Archetype Editor from Ocean Informatics was used to create archetypes. In the process of creating an archetype, a unique ID is assigned to every archetype created. This ID is checked every time an archetype is uploaded to any instances of the CKM, so that the ID of an archetype is ensured to be the same across all CKMs. It is therefore possible to describe which object is referenced by means of an ID.

Archetypes and templates include identifier of the data they describe (Archetype ID and Template ID). Therefore, it is possible for machines to identify an archetype or a template without appropriate support (principle F3).

Archetypes and templates are indexed with several keywords to make them searchable across the HiGHmed CKM (principle F4.) We consider F4 therefore as fulfilled.

Accessible

The HiGHmed clinical knowledge governance framework, as described in (Wulff, et al., 2018), defines requirements to make archetypes and templates accessible within the HiGHmed project. Hence, it is used as starting basis to investigate the accessibility of archetypes and templates. The archetypes and templates are retrievable via the CKM REST API, but the information of the server, which has to be requested, is not included in the XML representation of archetypes nor templates. Principle A1 can therefore be considered only as partially fulfilled.

The openly documented CKM REST API (https://ckm.openehr.org/ckm/rest-doc/) defines mechanisms to list selection of archetypes or update an archetype. It is also possible to get a file set for all archetypes used in a template (principle A1.1).

In addition, the CKM uses a role and rights management that supports authorization and authentication (principle A1.2). To create and review archetypes and templates, researchers and data stewards need appropriate roles so that special user-dependent

rights can be assigned. As part of the role and rights management of the CKM, a differentiation can be made between system-wide roles, subdomain roles and project roles. System-wide roles can create, change and delete ontology classes as well as release sets. In addition, complete classification schemes can be created or deleted. The translation of classification schemes is also part of the system-wide roles. At subdomain level, the user can only take over the above rights for one project (e.g. HiGHmed-specific). On project level (e.g. use case-specific), there are roles such as Editor or Reviewer. These have their rights only for the corresponding project and cannot create or change archetypes or templates in any other projects.

The long-term archiving of data and documents faces several challenges. It must be ensured that data can be retrieved over a long period of time and that data cannot be changed unnoticed. When an archetype or template is completely and irrevocably deleted in the CKM, all dependencies such as review rounds, comments, discussions, change requests and history are removed. However, the deletion is only anticipated if the archetype or template is no longer needed. It is preferred to set the status of an archetype to “rejected” or “deprecated”. The status “deprecated” is used, when an archetype has already been published in advance, however, the “rejected” status is applied for archetypes still in development. Both status allow that archetypes to be accessible under the tab "checked-out resources” within the CKM. It should be noted, that the metadata is not kept separate from the actual data and thus not kept independent of each other. Therefore, we consider principle A2 as partially fulfilled.

Interoperable

OpenEHR makes use of the Archetype Definition Language (ADL) to represent knowledge formally, accessible, shared and broadly applicable (principle I1). ADL is established and documented as an open, domain-relevant standard in the openEHR environment.[1] In addition to the representation in ADL, archetypes and templates can also be represented in XML format.
In addition, terminologies such as LOINC can also be integrated for each data item of an archetype to be able to exchange data more easily and be interoperable according to the FAIR Principles (principle I2).

Within archetypes, cluster slots are defined, in which further archetypes can be referenced to nest other archetypes. Furthermore, templates mainly contain references to various archetypes. Templates map context-specific documents in which archetypes are referenced as representations of clinical information. The relationships between individual archetypes within a template can be queried using the CKM REST API. The service queries the ID of the template together with the archetypes that are required for the template (principle I3).

Complying with all three interoperability principles, openEHR archetypes and templates provide a FAIR basis for information exchange.

[1] https://specifications.openehr.org/releases/AM/latest/ADL2.html

Re-Usable

Every archetype or template in every CKM instance is licensed under the Creative Commons Attribution-ShareAlike 3.0 License or higher, so that the (meta)data are released with a clear and accessible data usage license (R1.1).

To keep data reusable, a detailed history of editing (creation, modification or deletion) is of high importance. The time, initiator, content and logged processes are recorded. This audit trail is stored in the CKM as the history of archetypes and templates and can be viewed without user authentication. Information such as current status, date of last changes and modifier can be pictured. The history also includes all previous versions of archetypes and templates. Additionally, the earlier versions can be compared to avoid inconsistencies. Data and workflow provenance within openEHR is stored in elements from the openEHR reference model. The *feeder audit class* and *feeder audit details class* describe the semantic content of an audit trail, for example, the source system, feeder system, and other audit information transferred. Statements about when, by who and where which information was provided, can be referenced to an archetype (principle R1.2).

Firstly, all developed HiGHmed archetypes and templates are fully compliant to the openEHR approach. The openEHR approach serves as a relevant standard for the medical community. The semantic framework defines specifications regarding the structuring of the data to store them in a standardized way. Specifications and APIs of the openEHR community can be viewed via the openEHR specifications[2] and form the basis of every development in openEHR. In addition, the HiGHmed project has its own community guidelines regarding the translation of international archetypes. This ensures compliance with community standards within the project and within the openEHR community (principle R1.3).

Table 1. Overview of the FAIR Principles and their equivalent in openEHR archetypes and templates.

FAIR Principles	Equivalent in openEHR	Degree of compliance
F1. (meta)data are assigned a globally unique and eternally persistent identifier.	The Build UID and Major Version ID acts as a globally unique and persistent identifier.	Fulfilled
F2. data are described with rich metadata (defined by R1 below)	The Attribution tab of archetypes and templates provides information about additional metadata such as authors, contributors and licenses.	Fulfilled

[2] https://specifications.openehr.org/

F3. Metadata clearly and explicitly include the identifier of the data	The archetype UID and template ID are contained as an attribution in archetypes and templates.	Fulfilled
F4. (meta)data are registered or indexed in a searchable resource	Archetypes and templates are indexed via keywords in the Clinical Knowledge Manager.	Fulfilled
A1. (meta)data are retrievable by their identifier using a standardized communications protocol	The archetypes and templates are retrievable via the CKM REST API, but the information of the server, which has to be requested, is not included.	Partially fulfilled
A1.1. the protocol is open, free and universally implementable	Archetypes and templates can be derived via the openly available CKM REST API.	Fulfilled
A1.2. the protocol allows for an authentication and authorization procedure, where necessary	The HiGHmed CKM is access controlled through a username and password.	Fulfilled
A2. metadata are accessible, even when the data are no longer available	By setting the status of an archetype to DEPRECATED or REJECTED metadata is still retrievable, but not if an archetype is deleted.	Partially fulfilled
I1. (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation	ADL and XML are used as formal syntax languages.	Fulfilled
I2. (meta)data use vocabularies that follow FAIR principles	The terminologies that are used include SNOMED-CT and LOINC.	Fulfilled
I3. (meta)data include qualified references to other (meta)data	Nesting of archetypes due to Slots and within templates. Reference to FHIR resources and IHE concepts can be set.	Fulfilled
R1. meta(data) are richly described with a plurality of accurate and relevant attributes	All attributes are displayed under the attribution tab of the archetype in the HiGHmed Clinical Knowledge Manager.	Fulfilled
R1.1. (meta)data are released with a clear and accessible data usage license	The license used is stored under Attribution tab of the Archetype.	Fulfilled

R1.2. (meta)data are associated with their provenance	Data Provenance is supported by Audit-Trailing and Use/Misuse information. Workflow Provenance is managed via feeder audit classes, which contain information about the workflow process.	Fulfilled
R1.3. (meta)data meet domain-relevant community standards	Clinical and technical reviewers check whether archetypes and templates comply with the Community Standards.	Fulfilled

5 Discussion

The FAIR Principles are designed to make research data findable, accessible, interoperable and reusable for the public. As a part of the HiGHmed project, future research data will be collected and stored by using openEHR archetypes and templates. Therefore, the CKM is used as a library of clinical knowledge artefacts (archetypes and templates).

In the context of this work the archetypes and templates have been explored with regard to the FAIR Principles. As a result, all 15 principles can be regarded as fulfilled or partially fulfilled. Archetypes and templates have extraordinary strengths, especially in the area of findability and interoperability. By using a clear ID management and meaningful keywords, the required archetypes and templates can easily be found. Data models are subject to a formal syntax and are enriched with terminologies and metadata to make them exchangeable across various institutions.

Archetypes are closely linked to their metadata. When an archetype is irrevocably deleted, the corresponding metadata will also be removed. Therefore, it is necessary to set archetypes and templates as deprecated or rejected within their lifecycles so that the metadata is still available. This procedure is part of the community guidelines in openEHR. However, it would also be possible to set up a Git repository, in which all archetypes are additionally stored. The HiGHmed CKM would always push the archetypes into the repository during an update. The use of a Git repository could thus become part of the Clinical Knowledge Governance Framework (Wulff, et al., 2018).

Furthermore, there are additional approaches to extend the data and workflow provenance, which can be incorporated into the Clinical Knowledge Governance Framework. As described in (Parciak, et al., 2018), *w3c prov* can be used, to improve the provenance process including further information on provenance for every interaction.

The FAIR Principles are not specified as a mandatory set of rules, but they can be used to provide sustainable data management. However due to the high influence of FAIR Principles in the scientific community, the FAIRmetrics.org working group has developed a framework to objectively test digital objects for FAIRness. Therefore, fourteen FAIR Metrics are created, based on the FAIR Principles. The FAIR Metrics

Framework is currently in a test phase and was therefore not part of this paper. Initiatives such as GO-FAIR[3] evaluate and discuss the use of these metrics to test FAIRness (Wilkinson, et al., 2018). After the successful test phase of FAIR Metrics, all archetypes and templates created within the HiGHmed project will be tested using the FAIR Metrics Framework. It can be assumed that there will be a higher gradation of the degree of compliance after the FAIR Metrics Framework application.

In conclusion, it can be stated that openEHR supports the FAIRification process, as described in the GO-Fair Initiative, with medical data. Using the semantic framework, medical data becomes linkable and semantic-enriched. FAIR archetypes and templates have an effect on the collected medical data. They offer the possibility to store data in a structured way and link information on data provenance. This makes medical data easy to find for future research questions as well as available for subsequent use.

From the beginning, the data management of the HiGHmed project was aimed to establish an infrastructure with findable, accessible, interoperable and reusable data in a distributed environment. The FAIR compliance of the archetypes and templates used, ensures that all clinical models are correct, understandable, available and complete. They serve as the target schema for all data integration pipelines and are therefore the basis for semantic interoperability.

Based on the results of this work, we will investigate whether it is possible to perform an automated examination of medical data for FAIRness on the basis of the featured FAIR archetypes.

Acknowledgment

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the research and funding concepts of the Medical Informatics Initiative (01ZZ1802B/HiGHmed).

References

Bauer, C. R. et al., 2016. Architecture of a Biomedical Informatics Research Data Management Pipeline. In: *Volume 228: Exploring Complexity in Health: An Interdisciplinary Systems Approach.* s.l.:s.n., pp. 262-266.

Beale, T., 2002. Archetypes: Constraint-based Domain Models for Future-proof Informations Systems. *OOPSLA workshop on behavioural semantics*, p. 18.

Brien, E. O. et al., 2001. Blood pressure measuring devices: recommendations of the European Society of Hypertension. *BMJ*, pp. 531-536.

Haarbrandt, B. e. a., 2018. HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. *Methods of Information in Medicine*, July, pp. e66-e81.

[3] https://www.go-fair.org/

Krefting, D., Loose, H. & Penzel, T., 2010. Employment of a Healthgrid for Evaluation and Development of Polysomnographic Biosignal Processing Methods. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*.

Parciak, M. et al., 2018. PROV@TOS. a Java Wrapper To Capture Provenance for Talend Open Studio Jobs.

Wilkinson, M. D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, p. 4.

Wilkinson, M. D. et al., 2018. *Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework*, s.l.: s.n.

Wulff, A., 2019. A Report on Archetype Modelling in a Nationwide Data Infrastructure Project. In: *Studies in health technology and informatics (258).* s.l.:IOS Press, pp. 146-150.

Wulff, A., Haarbrandt, B. & Marschollek, M., 2018. *Clinical Knowledge Governance Framework for Nationwide Data Infrastructure Projects.* s.l.:s.n.