January 2021

openEHR is FAIR-Enabling by Design

Francesca FREXIA [a,1], Cecilia MASCIA [a], Luca LIANAS [a], Giovanni DELUSSU [a], Alessandro SULIS [a], Vittorio MELONI [a], Mauro DEL RIO [a] and Gianluigi ZANETTI [a]
[a] *CRS4: Center for Advanced Studies, Research and Development in Sardinia (Italy)*

Abstract. The FAIR Principles are a set of recommendations that aim to underpin knowledge discovery and integration by making the research outcomes *Findable*, *Accessible*, *Interoperable* and *Reusable*. These guidelines encourage the accurate recording and exchange of structured data, coupled with contextual information about their creation, expressed in domain-specific standards and machine readable formats. This paper analyses the potential support to FAIRness of the openEHR e-health standard, by theoretically assessing the compliance with each of the 15 FAIR principles of a hypothetical Clinical Data Repository (CDR) developed according to the openEHR specifications. Our study highlights how the openEHR approach, thanks to its computable semantics-oriented design, is inherently FAIR-enabling and is a promising implementation strategy for creating FAIR-compliant CDRs.

Keywords. openEHR, FAIR Principles, Semantics, Semantic Interoperability, Data Integration, Information Models, Representation Standards, Archetypes

1. Introduction

Data semantics plays a central role in the study, design and implementation of methods and tools for extracting meaningful information from the explosion of data we have been experiencing over the past decades [1]. In particular, increasing attention has recently been dedicated to the semantic interoperability between systems and to the importance of an extensive data enrichment to improve their analysis and the preservation of their meaning when they are reused in different contexts. The crucial value of an accurate expression of semantics, both for humans and machines, is underlined also in the FAIR Guiding Principles, which strongly encourages the association of large sets of context and content metadata to the research results, to make them *Findable*, *Accessible*, *Interoperable* and *Reusable* [2]. The FAIR Principles are general guidelines towards data-driven research and it is then up to each specific community to identify relevant domain standards and ontologies to achieve this goal. At present, the first practical implementations are mainly in the biomedical research domain [3], but the healthcare sector is also exploring the feasibility of applying the FAIR Principles [4, 5]. This paper analyses the potential FAIRness of openEHR [6], which is recognised as one the main stacks, i.e. collection of tools or standards that work together, in health informatics [7]. Starting from the openEHR specifications, we assessed the feasibility of creating a FAIR compliant resource containing Electronic Health Record (EHR) data, and showed how the openEHR design principles can handle the FAIR requirements.

[1]Corresponding Author: Francesca Frexia, CRS4: Center for Advanced Studies, Research and Development in Sardinia, Loc. Piscina Manna - Edificio 1, 09050 Pula (CA), Italy; E-mail:francesca.frexia@crs4.it.

2. Methods

2.1. openEHR

openEHR [6] is an open standard for e-health data, including a set of open specifications, clinical models and open-source software components, finalised to the construction of an open, vendor neutral platform for EHRs and interoperable clinical and research data.

The foundation of openEHR paradigm is the principle of separation between data representation and domain content, achieved through a multilevel information model stack. Datatypes (e.g., Text, URI), structures (e.g., single, list, tree) and containers (e.g., EHR, Composition, Demographics) are part of the *Reference Model* (RM), while the domain-related content is expressed through archetypes and templates in the *Archetype Model* (AM). Only the RM structures are implemented in software, thus decoupling the deployment of systems from the heterogeneity and variability of data. Archetypes are used to model specific domain concepts with all the possible data points that would be universally needed in any context. Technically, archetypes are extensible formal constraint definitions of object structures expressed in the Archetype Definition Language (ADL), where each data point corresponds to a specific path and can be bound to external terminologies and ontologies, to better specify its meaning. Built for specific use cases, templates define a tree of one or more archetypes, possibly hiding non-relevant data points and tightening existing constraints while maintaining all the terminological binds. Templates are used at runtime to create forms and data instances as well as to validate data input, ensuring that all data conform to the constraints defined in the included archetypes or in the template when tighter. Notably, templates preserve the path structure of each archetype's node, ensuring in this way the inclusion in the data instances of the notion of which archetypes were used at data creation time. Paths are expressed in an XPath-compatible syntax and contain archetype attribute names and node identifiers, therefore supporting semantic querying. Queries are defined in AQL (Archetype Query Language), combining SQL and paths drawn from the archetypes.

openEHR clinical models are computable artifacts to model healthcare contents, consisting of archetypes which can be created from scratch or reusing previously published models. The openEHR Clinical Modelling Program gathers a wide international community of clinicians, researchers and developers working on the design and maintenance of the clinical models, collected in the international Clinical Knowledge Manager (CKM). At present, it includes a library of 535 governed archetypes (i.e., over 8000 data points) and other local CKMs exist too, to address specific projects' or national needs.

2.2. Analysis Approach

The hypothesis of the innate FAIR compliance of the openEHR approach is primarily justified by the possibility of establishing a direct mapping between the openEHR foundational principles and the FAIR dimensions. In the archetype-driven openEHR architecture, in fact, archetypes and templates perform two key functions: (i) *semantics preservation*, long term and implementation independent, by defining structured and detailed open models for data validation at data capture or import time, therefore supporting the "I" and "R" dimensions; (ii) *semantic querying*, reachability of every data item through "semantic paths" deriving from the composition of archetypes within a template, mechanism relevant for the "F" and "A" dimensions.

In order to check this hypothesis, we have analysed if and how it would be possible to create an openEHR-based Clinical Data Repository (CDR) fulfilling the requirements expressed by each specific FAIR Principle. In particular, in this analysis we considered as "data" all the instances of clinical information in the repository EHR, and we defined two main categories of "metadata": (i) *instance metadata*, that are all the relevant attributes describing the data, included at various levels in the EHR data instances, further sortable into *general* (such as authorship, copyright, licences, languages and translations) and *contextual* (attributes describing information on data acquisition, such as protocol and subject's state); (ii) *knowledge metadata*, the archetypes and templates representing the domain-content models associated with the data. The inclusion of the clinical models in the metadata is motivated by the fact that they are detailed and structured sets of attributes describing the real world contents which generated the data, since the creation and validation of openEHR data instances is controlled by the knowledge artefacts derived from these specific models.

3. Results

Table 1 illustrates the results of our theoretical FAIRness analysis for a generic EHR system, implemented following the openEHR specifications and evaluated from the perspective of the compliance with each single FAIR Principle. As shown in detail in the table, openEHR envisages by design all of the structural components necessary to create a FAIR resource, since every FAIR Principle can be fulfilled by developing a Clinical Data Repository according to the openEHR design fundamentals and specifications.

Table 1.: Compliance of an openEHR-based Clinical Data Repository with the FAIR Principles

FINDABILITY
F1. (Meta)data are assigned a globally unique and persistent identifier - By assigning a Uniform Resource Locator to an openEHR CDR it is possible to refer to an interior (meta)data node from anywhere, as each (meta)data item is locatable in the `EHR` space from the outside via an ADL path, expressed in a W3C XPath compatible syntax. *Ref. Architecture Overview [8]*
F2. Data are described with rich metadata (defined by R1 below) - openEHR data instances includes wide sets of metadata, which we classified into: *instance* (*general* and *contextual*) and *knowledge* metadata. *Ref. Architecture Overview [8]*
F3. Metadata clearly and explicitly include the identifier of the data they describe - Metadata are bounded to an openEHR data instance in different ways: *knowledge metadata* are stored in a separate repository and linked by the `Archetype_ID` embedded in the data instance; *instance metadata*, automatically captured or describing the acquisition context, are part of the instance. *Ref. Architecture Overview [8]*
F4. (Meta)data are registered or indexed in a searchable resource - Within an EHR system, the `EHR` object is a searchable resource: the system ID can be used as a key to access to the underlying (meta)data layers that, in turns, are indexed via the archetype's path mechanism. *Ref. Architecture Overview [8]*
ACCESSIBILITY
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol - (Meta)data can be retrieved by the URI assigned to the `EHR` object or using AQL, by filtering with the specific `EHR_ID`. Both this ways are implemented as a part of openEHR REST API, which is based on standard HTTP protocols for requests. *Ref. Implementation Technologies [8]*
A1.1 The protocol is open, free, and universally implementable - openEHR REST API is a free open protocol, based on standard HTTP requests, that can be implemented in any programming language. *Ref. Implementation Technologies [8]*

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary - Policies and rules for data access are enclosed in the `EHR_ACCESS` section of the `EHR` object. Technically, the preferred end-point for data access is openEHR REST API, so it is up to the specific infrastructure the implementation of any kind of standard HTTP access rule to protect these APIs and, therefore, data. *Ref. Implementation Technologies [8]*
A2. Metadata are accessible, even when the data are no longer available - In fulfilment of medico-legal requirements, both data and metadata are never deleted. *Ref. Reference Model [8]*
INTEROPERABILITY
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation - ADL is the formal language used to express content models, designed as a human-readable and processable syntax. Model serializations in other widely used formats (i.e., XML, JSON) are available too and no restrictions are given for data instance format as long as their structures respect the declared model. *Ref. Archetype Model [8]*
I2. (Meta)data use vocabularies that follow FAIR principles - The Terminology package features fulfill the requirement, in particular nodes can be bound to one or more external terminology codes to improve automatic processing. *Ref. Terminology [8]*
I3. (Meta)data include qualified references to other (meta)data - Different types of cross-references are identified, mostly expressed by the classes `LINK` and `DV_URI`. In the former, the connection meaning is embedded, while in the latter the relation with the target of the link is not "qualified", so it has to be clarified by the knowledge metadata. *Ref. Reference Model [8]*
REUSABILITY
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes - openEHR data instances can be described by two categories of structured metadata: *instance* and *knowledge* metadata. Part of them derives directly from some attributes of the openEHR RM, such as the `protocol` and `guideline_id` attributes of the `CARE_ENTRY` class, which allow to record metadata about methods, instruments used and followed guidelines. *Ref. Archetype Model and Reference Model [8]*
R1.1. (Meta)data are released with a clear and accessible data usage license - Each resource is an instance of an `AUTHORED_RESOURCE` class, that carries metadata relating to: authorship, copyright, licences, languages, translations and other related metadata. *Ref. Architecture Overview [8]*
R1.2. (Meta)data are associated with detailed provenance - The versioning mechanism ensures that all the changes of EHR data and demographic services are audit-trailed with user identity, time-stamp, digital signature, etc. *Ref. Archetype Model and Reference Model [8]*
R1.3. (Meta)data meet domain-relevant community standards - openEHR requires the adherence to some well-known standards for datatypes and formats (e.g., ISO 8601 and UCUM) and enable the binding of data nodes to standard terminologies like SNOMED-CT and LOINC. *Ref. Architecture Overview [8]*

4. Discussion

This paper shows how openEHR design fundamentals can comply with the FAIR Principles and, to the best of our knowledge, this is the first study highlighting this conceptual analogy. Only a previous work, in fact, already explored the FAIRness of a resource containing openEHR archetypes and templates [9], but its focus was on the fulfillment of the Principles for a determined repository rather than on the overall affinity between openEHR and FAIR.

Our objective was to examine the interrelation between the basis of two sets of theoretical guidelines, not the evaluation of a specific implementation. We therefore have highlighted how openEHR specifications natively support the creation of a FAIR data resource, not that any openEHR repository is FAIR. Indeed, the effective FAIRness of a repository also depends on the construction of the persistence solution and it has to be assessed for any repository, whatever kind of data it contains.

A central point of the study is the very specific assumption about the definition we adopted for metadata, specialised to include the openEHR clinical models. However, our wider interpretation is consistent both with the general FAIR suggestions about metadata [10] and with other relevant guidelines for metadata formalisation, like those enounced by the Research Data Alliance. Therefore, we can conclude that even if the assumption is very tailored to the openEHR context, it doesn't affect the reliability of the results. Moreover, this approach enforces Interoperability and Reusability, explicitly supporting the availability of detailed description of domain knowledge.

5. Conclusion

Our analysis reveals the intrinsic potential of openEHR to build a FAIR-compliant Clinical Data Resource, by applying the openEHR specifications in combination with some *ad hoc* deployment configurations. This native support to FAIRness is a direct consequence of the centrality of semantic preservation and querying in the openEHR philosophy since its inception, long before the FAIR Guiding Principles were formulated. Despite its abstract nature, our analysis highlights how the openEHR approach can be considered a viable choice to have data more Findable, Accessible, Interoperable and Reusable in the clinical and biomedical context. Future work will involve exploring how this considerations can be applied to real implementations, by assessing the actual FAIRness of openEHR resources developed for a particular use case.

6. Acknowledgments

We dedicate this work to the memory of Gianluigi Zanetti, whose legacy continues to inspire us. This work has been partially supported by the DIFRA Project (funded by the Sardinian Regional Authority) and by the European Joint Programme on Rare Diseases (grant agreement N. 825575).

References

[1] P. Ceravolo, et al., Big Data Semantics, Journal On Data Semantics 7 (jun 2018). doi:10.1007/s13740-018-0086-2.
[2] M. D. Wilkinson, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (1) (2016) 160018. doi:10.1038/sdata.2016.18.
[3] M. van Reisen, et al., Towards The Tipping Point For FAIR Implementation, Data Intelligence 2 (1-2) (2020) 10.1162/dint_a_00049. doi:10.1162/dint_r_00024.
[4] P. Holub, et al., Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health, Biopreservation and Biobanking 16 (2) (2018) 97–105. doi:10.1089/bio.2017.0110.
[5] A. A. Sinaci, et al., From Raw Data To FAIR Data: The FAIRification Workflow For Health Research, Methods Of Information In Medicine 59 (S 01) (2020) e21–e32. doi:10.1055/s-0040-1713684.
[6] What is openEHR?, `https://www.openehr.org/about/what_is_openehr`.
[7] WHO and ITU, Digital Health Platform Handbook: Building A Digital Information Infrastructure (Infostructure) For Health, 2020.
[8] openEHR International, OpenEHR Specifications, `https://specifications.openehr.org`.
[9] C. Bönisch, et al., FAIRness of openEHR Archetypes and Templates, SWAT4(HC)LS Conference, 2019.
[10] B. Mons, et al., The FAIR Principles: First Generation Implementation Choices And Challenges, Data Intelligence 2 (1-2) (2020) 1–9. doi:10.1162/dint_e_00023.