pHealth 2022
B. Blobel et al. (Eds.)

doi:10.3233/SHTI220975

Decentralized EHRs in the Semantic Web for Better Health Data Management

Dagmar CELUCHOVA BOSANSKA[a,1], Michal HUPTYCH[b] and Lenka LHOTSKÁ[a,b]
[a] *Faculty of Biomedical Engineering, Czech Technical University in Prague, Czech Republic*
[b] *Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, Czech Republic*

Abstract. Electronic Health Record (EHR) systems currently in use are not designed for widely interoperable longitudinal health data. Therefore, EHR data cannot be properly shared, managed and analyzed. In this article, we propose two approaches to making EHR data more comprehensive and FAIR (Findable, Accessible, Interoperable, and Reusable) and thus more useful for diagnosis and clinical research. Firstly, the data modeling based on the LinkML framework makes the data interoperability more realistic in diverse environments with various experts involved. We show the first results of how diverse health data can be integrated based on an easy-to-understand data model and without loss of available clinical knowledge. Secondly, decentralizing EHRs contributes to the higher availability of comprehensive and consistent EHR data. We propose a technology stack for decentralized EHRs and the reasons behind this proposal. Moreover, the two proposed approaches empower patients because their EHR data can become more available, understandable, and usable for them, and they can share their data according to their needs and preferences. Finally, we explore how the users of the proposed solution could be involved in the process of its validation and adoption.

Keywords. Distributed electronic health records, FAIR principles, HL7 FHIR, bio-data management, ontology

1. Introduction

Burnout in healthcare takes a toll on physicians and their impact on patient health they could have without it. We believe that overly complicated healthcare institutions are part of the problem. According to Clayton Christensen's work, "The Innovator's Prescription," these institutions mix three separate business models. Our goal is to research and validate a decentralized electronic health record (EHR) that would be optimized for the "solution shop" business model. Solution shops are designed to diagnose patients, meaning to solve unstructured problems with the help of experts who have the appropriate tools and resources.

Centralized electronic health records (EHRs) are established digital solutions that are part of hospital systems and outpatient clinic software. It may not be realistic to replace them. Therefore, decentralized EHRs must operate in synergy with established systems, focus on relevant patient problems that can be documented more efficiently in

[1] Corresponding Author: Dagmar Celuchova Bosanska, Sítná 3105, 272 01 Kladno, Czech Republic; Email: celucdag@fbmi.cvut.cz.

a decentralized cooperative way, and add significant value to diagnosing. In other words, decentralized EHRs must provide more FAIR (Findable, Accessible, Interoperable, and Reusable) bio-data that can be created efficiently and analyzed in the complex healthcare ecosystem and contribute to faster and better diagnosis in solution shops.

2. Methods

Because the availability of high-quality data is the key to better diagnostic processes, we consider a decentralized EHR a data-driven application. It can operate on a diverse set of harmonized data collected from various environments. However, this feature poses significant challenges to the underlying data model. Firstly, it must evolve more flexibly than a schema for a relational model. Secondly, it must accommodate different levels of sophistication in data modeling - from flat tables up to formal ontologies expressed in RDFS / OWL [1]. Thirdly, it must be easy and practical to compile this data model to other frameworks used by different experts (e.g., clinicians, data scientists, terminology curators, developers). The promising approach that addresses these challenges is the LinkML framework. It already has traction in successful applications such as cancer and environmental microbiome data harmonization and biological knowledge graph integration [2]. LinkML (linkml.io) is a data modeling language for describing the structure of an instance collection. Each instance instantiates a class from the LinkML metamodel depicted in Figure 1.

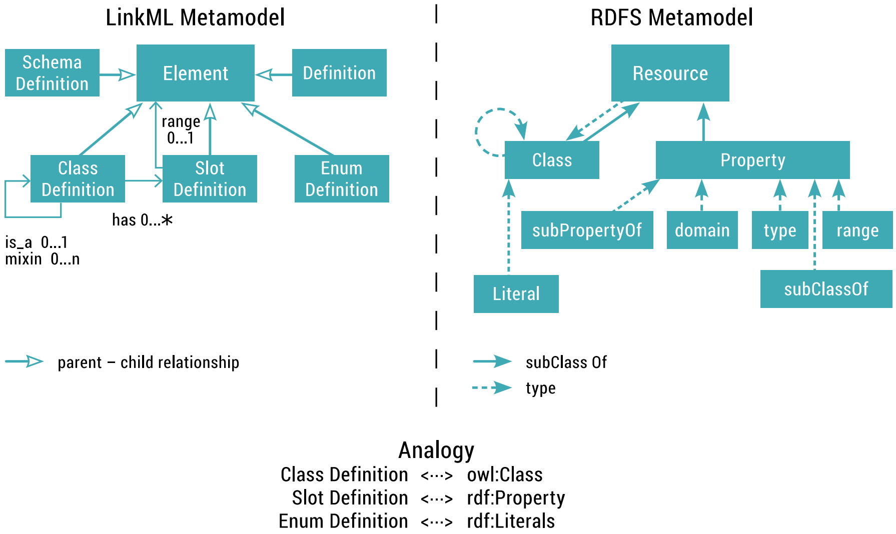


Figure 1. The LinkML Metamodel for data-driven applications and its analogy to the RDFS Metamodel.

Furthermore, the LinkML framework provides developer and curator-friendly tools for working with data, such as data validators, data converters, compatibility tools, and schema inference. Data models are created in simple YAML files, optionally annotated using ontologies. Furthermore, this data model can be compiled to other frameworks such as JSON-LD contexts, JSON schema, and Python data classes. One could argue that the LinkML metamodel is too simple to accommodate any clinical model that is the foundation of a well-designed EHR. However, we will show in the Results section that even with this simplified modeling, we can store and process clinically relevant data. The added value is the availability of more FAIR data because this approach can easily integrate data from various sources that differ in the sophistication of their data model.

Two promising technology stacks appropriate for the decentralized EHRs have been researched. These stacks represent two competing advancements of the World Wide Web (WWW), namely blockchain within Web3 [3] and Solid - a decentralized platform for social Web applications within the Semantic Web or Web 3.0 [4]. Table 1 summarizes the main features and limitations of both.

Table 1. Main features of two promising technologies for decentralized EHR

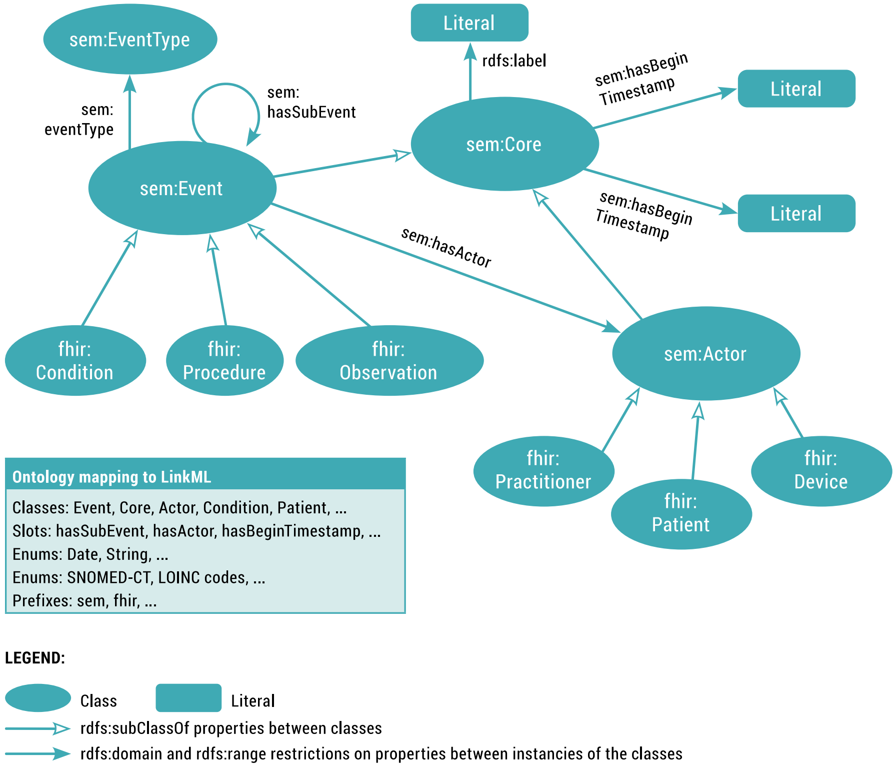
Feature comparison		
Feature	**Blockchain**	**Solid (CSS)**
Advancement of WWW	Web3	Semantic Web or Web3.0
Data privacy and security	Data widely accessible but patient identity is a secret via cryptography	Data not available without an access grant to a pod
Data sharing	Smart Contracts	Consents (Verifiable Credentials)
Patient identification	Unique patient identifier	WebID
Data safety	Many decentralized copies	Centralized locations of providers
Interoperability in general	Lack of open standards	W3C standards, conformance tests
Data interoperability	HL7 FHIR and other standards only in off-chain data	HL7 FHIR and other standards
Information findability	None defined	Link traversal based queries [5]
Data integrity	Immutable ledger	End-to-end encryption
Auditability	Systematic examination of blocks and network	None defined
Semantic data modelling	RDFS / OWL or LinkML only in off-chain data	RDFS / OWL or LinkM
Data representation	Transaction blocks with references. Graph data - JSON-LD off-chain	Graph data - JSON-LD
Data storage challenges	Blockchain size is a limit but pointers to off-chain data possible	Graph size influences the performance of queries and analysis
Ways of data monetization	Bitcoin and other cryptocurrencies	None defined

We argue that a new approach to decentralization that could have a potential to transform the diagnostic processes within the healthcare system must make use of both technology stacks to overcome all limitations mentioned in Table~1. Our proposed technology stack in the Results section is therefore mainly based on Solid due to its high potential to apply FAIR principles of data sharing and to support data-driven applications.

3. Results

We used the LinkML framework to show how clinically relevant data can be modeled, represented, and stored in a decentralized EHR. Firstly, we show how to easily integrate data from various sources. We model data as events with their time frames and we use the Simple Event Ontology [6] as the basis for the LinkML framework (Figure 2). As described in the Methods section, the LinkML framework makes this data model more accessible to various experts and scenarios. We show its application on an example of

EHR data that can be in the HL7 FHIR RDF format combined with data from wearables in CSV in the following Python implementation[2]. Solid pods have several containers where data can be stored according to different data models. Alternatively, one user can have many Solid pods. This possibility allows the original HL7 FHIR RDF dataset to be stored in a separate container or pod. It can be queried anytime for additional information or unclear clinical context resulting from the simplified LinkML model.



Ontology mapping to LinkML

Classes: Event, Core, Actor, Condition, Patient, ...
Slots: hasSubEvent, hasActor, hasBeginTimestamp, ...
Enums: Date, String, ...
Enums: SNOMED-CT, LOINC codes, ...
Prefixes: sem, fhir, ...

Figure 2. The proposed ontology for the EHR data integration and its mapping to the LinkML Metamodel

Secondly, we explored how the phenotype information can be stored in decentralized EHRs. Currently, EHR systems record virtually no phenotype information. We show how an event in the form of a condition from the HL7 FHIR RDF data can be easily transformed into the Phenopacket Schema (phenopacket-schema.io) with the help of the created LinkML model. A phenopacket links detailed phenotype descriptions with a disease, patient, and genetic information. How to correctly convert all relevant EHR data in the HL7 FHIR RDF format to a phenopacket and enrich it with more data in the LinkML framework is a topic of ongoing research. With the help of our proposed decentralized EHRs, sharing these phenopackets as FAIR data is effortless. Increasing the sharing of FAIR phenopackets will support large-scale computational disease analysis using the combined genotype and phenotype data.

Furthermore, we propose the infrastructure for the decentralized EHRs based on the Community Solid Server (CSS)[2]. To overcome its limitations mentioned in Table 1, this

[2] https://github.com/CommunitySolidServer/CommunitySolidServer

technology stack needs to be extended by Web3 features, especially for data integrity and auditability. Nevertheless, based on the features mentioned in Table 1 and our experimentation, it can significantly contribute to the following FAIR principles:

1. It can make use of the proposed data modelling in the linkML framework to support interoperability.
2. It implements many W3C standards that support data linking and findable data such as W3C Linked Data, SPARQL for query and W3C Link Data Platform.
3. It can be easily connected to terminology services for concept alignments – the same information can be presented to different audiences.
4. It is based on mydata.org principles, therefore users can grant consents for specific personal data usage and data sharing such as data donation to research. This approach to data sharing is in line with the General Data Protection Regulation.
5. It can be the underlying infrastructure for data intermediaries participating in the European Health Data Spaces (EHDS).

As mentioned in Table 1, data stored in Solid pods are without an access grant private. Privacy hinders findability and accessibility of the data, therefore safe mechanisms on how to share data in an anonymized or pseudonymized way must be researched. One way of ensuring the user's privacy while making data accessible is by allowing for selective disclosure of only those attributes that are necessary for authentication and authorization like in [7]. Another approach are methods for answering SPARQL queries through zero-knowledge link traversals that are being researched, for example in [8].

Moreover, data-driven applications can be developed on top of this Solid technology stack, because:

1. There is no crucial data storage challenge of potentially large personal health knowledge graphs [9].
2. Personal data stored in pods can be linked to external knowledge graphs and organized according to various data models such as the mentioned phenopackets.
3. It supports cooperation of several users on data stored in a particular pod.
4. It is an enabler for analytic services based on querying, reasoning, graph analysis and graph neural networks. These services could be used for more advanced clinical decision support systems and for smart guidance for patients.
5. Linked Data Shapes, Forms and Footprints[3] are defined to further support data entry, retrieval, and user interface in data-driven applications.

4. Discussion and Conclusions

The technology stack based on the Community Solid Server, as discussed in the Results section, can be extended by the components of the Blockchain stack:

- An immutable ledger for enhanced data integrity: It may not be used for all data stored in pods but only for a subset for which a high level of integrity is required

[3] https://www.w3.org/DesignIssues/Footprints.html

by the clinical processes (for example, the blood type, current allergies, selected prescribed medications such as narcotics).

- A way of data monetization based on Bitcoin or other cryptocurrencies for data sharing: This feature can reward physicians for sharing part of their records to the decentralized platform or creating unique entries in the decentralized EHRs, thus motivating them to contribute. It can also be used to pay patients for donating their data, for example, to research.

A decentralized EHR could become an authoritative source in areas where some form of a decentralized record exists and where patients, pregnant women, or parents usually proactively collect data. The examples include a pregnancy health record, a newborn and toddler progress note, immunization records, and medical history or anamnesis from the patient narrative.

Moreover, the data integration at the patient's side supports the mentioned "solution shop" business model for better diagnosis. This business model requires multidisciplinary cooperation. The proper bio-data modeling and management based on the LinkML framework and the resulting data representation that can be easily shared via the proposed decentralized technology will significantly contribute to this multidisciplinary cooperation and FAIR Health Data Sets for further research and analysis. The initial data modeling can be simplified to the event-based model, as presented in Figure 2. The model of the actual events of various types can be further elaborated to match the best approaches to clinical modeling. Finally, the potential of proposed approaches to save time, improve diagnosis and increase sharing of FAIR data must be validated by a properly designed study in the future.

References

[1] Uschold M, Gruninger M. Ontologies and Semantics for Seamless Connectivity. ACM SIGMOD Record, 2004:33(4):58–64, 2004, https://doi.org/10.1145/1041410.1041420.

[2] Unni DR, Moxon SAT, Bada M, et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clin Transl Sci. 2022 Aug; 15(8): 1848–1855, doi: 10.1111/cts.13302

[3] Mayer AH, da Costa CA, da Rosa Righi R. Electronic health records in a Blockchain: A systematic review. Health Informatics J. 2020 Jun;26(2):1273-1288. doi: 10.1177/1460458219866350.

[4] Mansour E, Sambra AV, Hawke S, et. al. A Demonstration of the Solid Platform for Social Web Applications. In Proceedings of the 25th International Conference Companion on World Wide Web, pages 223–226, 2016, doi: 10.1145/2872518.2890529

[5] Hartig O, Freytag JC. Foundations of traversal based query execution over linked data. HT'12 - Proceedings of 23rd ACM Conference on Hypertext and Social Media, pages 43–52, 2012, doi: 10.1145/2309996.2310005

[6] Van Hage WR, Segers VMR, Hollink L, Schreiber G. Design and use of the Simple Event Model (SEM). Journal of Web Semantics, 2011;9(2):128–136, doi: 10.1016/j.websem.2011.03.003

[7] Braun CH, Käfer T. Attribute-based Access Control on Solid Pods using Privacy-friendly Credentials. In SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, pages 1–5, 2022.

[8] Fafalios P, Tzitzikas Y. Answering SPARQL queries on the web of data through zero-knowledge link traversal. ACM SIGAPP Applied Computing Review, 2019;19:18–32, , doi: 10.1145/3372001.3372003

[9] Ammar N, Bailey JE, Davis RL, Shaban-Nejad A. The personal health library: A single point of secure access to patient digital health information. Stud Health Technol Inform. 2020; 270: 448–452, 2020, doi: 10.3233/SHTI200200