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

doi:10.3233/SHTI230952

Case-Reported Data Management Methodology Using an RDF Data Model for Building a Multicenter Clinical Registry

Masamichi ISHII [a,1], Hiroyuki HOSHIMOTO[a] and Kengo MIYO[a]
[a] *Center for Medical Informatics Intelligence, National Center for Global Health and Medicine, Tokyo, Japan*
ORCiD ID: Masamichi Ishii https://orcid.org/0000-0002-4228-1723

Abstract. In multicenter clinical research, case-reported clinical data are managed for each research project. Participating institutions manage the mapping between standardized codes and in-house codes. To use the data extracted from electronic medical records in case report forms, it is necessary to pay attention to the gap in the semantic hierarchy. Managing mapping information between in-house and standardized codes is centralized in Resource Description Framework (RDF) stores. The relationship between standardized and in-house codes is described in RDF and stored in RDF stores. RESTful APIs for accessing RDF stores in SPARQL was developed and verified. The relationship between standardized codes and in-house codes of pharmaceuticals was expressed in RDF triples. As a +result of the operational verification of the implemented APIs, it was confirmed that data management with knowledge bases expressed in RDF graphs is possible. The ability to dynamically modify mapping definitions enables flexible data management and ease of operational restrictions.

Keywords. RDF, semantic web, clinical registry

1. Introduction

Direct Data Capture directly collects real-world data stored in electronic medical record systems for research purposes and has attracted considerable attention [1]. However, many interoperability issues persist with electronic health-record-derived research data [2]. One such error is the mapping of standardized codes to in-house codes. Many Japanese medical institutions use in-house drug codes in their electronic medical record systems. Consequently, many research projects that collect clinical research data from electronic medical record systems require in-house codes to be converted to standardized codes at the time of data submission. In most cases, the standardized codes for in-house codes are preset in the electronic medical record system's drug master and used during data conversion. Because this is a manual operation, human errors, such as mistaken mapping and registration operation errors, are inevitable. Suppose the standard codes are

[1] Corresponding Author: Masamichi Ishii, NCGM (National Center for Global Health and Medicine) Address: 1-21-1 Toyama, Shinjuku-ku, Tokyo 162-8655, Japan Phone:+81(3)3202-7181, email: masaishii@hosp.ncgm.go.jp.

set incorrectly after the research data are submitted. In that case, the medical institution must go back in time and recreate the data for submission, and the database center must repair the database. Such problems often occur in laboratory test result codes, disease codes, etc., in addition to drug codes. However, medical record data freely written in electronic medical record systems are not structured. Thus, the Japan Diabetes Comprehensive database project based on an advanced electronic medical record system (J-DREAMS) [3], a joint project of the Japan Diabetes Society and NCGM, provides clinical templates for diabetes and recommends that medical records be created as structured data using the templates. Each input data point in the medical templates has its own Object Identifier (OID), which can be used for data identification across facilities. However, this OID is a J-DREAMS proprietary definition, and its interoperability with other research projects is not ensured.

In a study on the mapping between standardized codes and in-house codes, Tanaka [4] proposed a method to prepare pair information of the Concept Unique Identifier (CUI) of the Unified Medical Language System (UMLS) and in-house codes in advance on the server and to provide the pair information at all times. Hayashi et al. [5] proposed a data model for survey questionnaires using the Resource Description Framework (RDF) and the RDF schema of the Semantic Web technology of the World Wide Web Consortium (W3C) recommendations [6,7]. They showed that new relations between resources could be added flexibly. On the other hand, as an attempt to apply RDF technology to research databases, Kawazoe et al. [8] presented a method of converting HL7 V2.5 messages to RDF to enhance semantic search. Gaudet-Blavignac et al. [2] promoted national-level interoperability by mapping existing research data to multiple Common Data Models (CDMs) in RDF triples. Based on these previous studies, this study verifies that data center operations can be flexibly implemented by converting and managing the mapping information between standardized and in-house codes, which were previously managed in-house, into a knowledge base expressed in RDF triples.

2. Methods

This study verified that mapping information management can be flexibly implemented by constructing a knowledge base expressed in RDF triples from mapping information between in-house codes and multiple standardized drug codes used in Japan [2].

2.1. Materials

The following will be used as resource data.

- In-house drug master of the electronic medical record system in operation at NCGM
- The standardized master for pharmaceutical products in Japan

YJ code: A drug code commonly used in Japan to identify a drug as a product.

Receipt computerization code: A dedicated code used for electronic insurance claims

HOT code:13-digit standardized codes for integrated management of other existing Japanese drug codes. The 13-digit HOT code (HOT13) corresponds to the Japan Article Number (JAN) / European Article Number (EAN) code on a one-to-one basis. The upper nine digits (HOT9) correspond to the YJ or Receipt computerization code for identifying drug products. The upper seven digits (HOT7) of the HOT code correspond to the insurance-listed prescription drug price codes.

- Metadata to classify drugs: Anatomical Therapeutic Chemical Classification System (ATC code): A drug classification code managed by WHO. Drug Classification of Japan Standard Commodity Classification Number: Drug Classification code used in the upper four digits of the YJ code.
- Drug Knowledge Base: Kyoto Encyclopedia of Genes and Genomes (KEGG) MEDICUS [9]: The database includes mapping information from YJ Code to the ATC code.

2.2. Experimental environment

We use VIRTUOSO [10] for RDF stores, Ubuntu for the Operating system, Python for the development language, FastAPI [11] for the web framework, and Swagger [12] for the description framework (Figure 1).

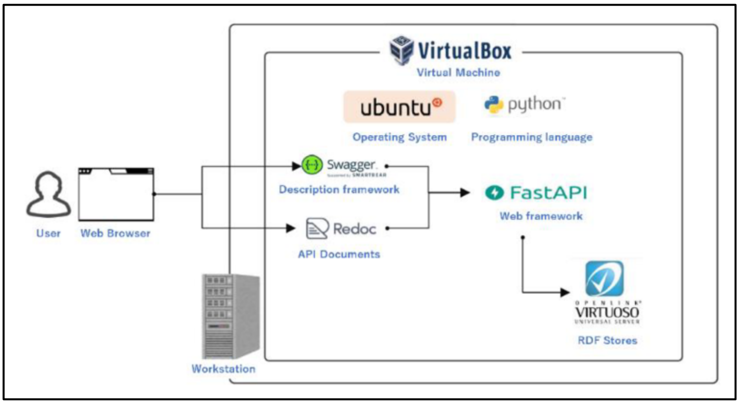

Figure 1. Experimental environment of RESTful API applications.

2.3. Implementation procedure

Using the standardized master for pharmaceutical products, the following procedure was used to express mapping:

- Define the "standardized master" and "in-house master" as nodes.
- Define the relationship between the masters as predicates.
- The mapping information between the masters is defined as triples using the above nodes and predicates.
- Register the defined triples in the RDF stores for centralized management.
- Develop RESTful APIs that can access RDF stores using SPARQL (SPARQL protocol and RDF query language) [13].
- Run the developed APIs on Swagger to verify its operation.
- Develop a mapping management user interface application using the developed APIs.

3. Results

The relationship between the standardized master code and in-house codes for drugs can be represented as a graph model using RDF (Figure 2). Among the standardized masters, metadata-derived resources were identified during API implementation by defining a "Classification Class Layer attribute. Standardization master-derived resources other than metadata were identified by defining the "Standardization Class Layer" group attributes.

In the J-DREAMS project, medical institutions were asked to recreate the previously collected data when an error in the mapping setting of standardized codes occurred. The data center had to wait 8 to 12 weeks until the data were resubmitted. This method confirmed that the mapping destination of in-house codes could be immediately corrected at the database center side, eliminating the need for data re-creation and re-submission.

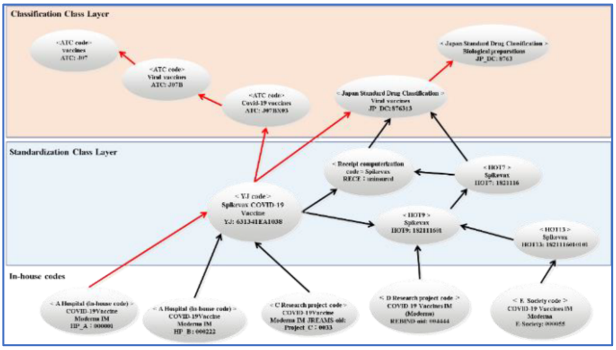
Classification Class Layer
<ATC code> vaccines ATC: J07
<ATC code> Viral vaccines ATC: J07B
<ATC code> Covid-19 vaccines ATC: J07BX03
< Japan Standard Drug Classification > Biological preparations JP_DC: 8763
< Japan Standard Drug Classification > Viral vaccines JP_DC: 876313
Standardization Class Layer
< Receipt computerization code > Spikevax RECE : uninsured
< HOT7 > Spikevax HOT7: 1821116
< YJ code> Spikevax COVID-19 Vaccine YJ: 631341EA1038
< HOT9 > Spikevax HOT9: 182111601
< HOT13 > Spikevax HOT13: 1821116010101
In-house codes
< A Hospital (in-house code) > COVID-19Vaccine Moderna IM HP_A : 000001
< A Hospital (in-house code) > COVID-19Vaccine Moderna IM HP_B : 000222
< C Research project code > COVID-19Vaccine Moderna IM JREAMS oid: Project_C : 0033
< D Research project code > COVID 19 Vaccines IM (Moderna) REBIND oid: 004444
< E Society code > COVID-19 Vaccines IM Moderna E Society: 000055

Figure 2. Graphical model for pharmaceutical standardized codes and in-house codes. An example is the Spikevax intramuscular injection (IM) of COVID-19 Vaccine. The in-house code for Spikevax at Hospital A is "000001" as the value of node “HP_A” in this knowledge graph. The Spikevax have a code of "631341EA1038" for YJ code. and ATC code "J07BX03" (Covid-19 vaccines) and Japan Standard Drug Classification code "876313 (Viral vaccines)" as metadata.

4. Discussion

The following effects can be expected from the proposed method in the future.

When a medical institution participates in a multicenter clinical registry project, it is necessary to map in-house codes to standardized codes before data submission. If the mapping information is managed by the database center using this method, data submission can be started without waiting for the completion of the mapping work by medical institutions. This is expected to reduce costs and speed up research, even for small registry research projects.

This method allows for adding a mapping to a controlled term master, such as Snomed-CT or LOINC, after data collection for clinical data collected with a project-specific data identifier code. This makes it possible to increase interoperability among arbitrary research projects while preserving the original data [14].

We want to consider using the proposed method in several national multicenter clinical research registry database projects that the NCGM will undertake.

5. Conclusions

In this study, we developed a prototype knowledge base described in the RDF that manages interrelationship information between standardized masters, mapping information between standardized masters and in-house codes, and mapping information between observation items without standardized codes and in-house codes on an RDF store. This enables flexible interoperability in using research data among institutions while maintaining the originality of clinical data. It is expected to promote clinical research by reducing the burden of establishing and operating a multicenter research registry.

In Japan, the Ministry of Health, Labor, and Welfare (MHLW) has established a standard master applicable to medical information but does not enforce its use. As a result, many Japanese medical institutions continue to use in-house codes in their electronic medical record systems. In the future, we hope that an incentive program to promote the introduction of standardized masters will be developed, following the example of the Meaningful Use Initiative in the United States. By expanding the proposed methodology, we hope to develop a clinical research infrastructure that promotes the utilization of clinical data derived from electronic medical record systems.

Acknowledgments

The National Center for Global Health and Medicine supported this study.

References

[1] Electronic Source Data in Clinical Investigations. US Food and Drug Administration; 2013. Available from: https://www.fda.gov/media/85183/download. (cited on 2023 Mar 9).

[2] Kawazoe Y, Imai T, Ohe K. A querying method over RDF-ized health level seven v2.5 messages using life science knowledge resources. JMIR Med Inform. 2016 Apr;4(2):e12, doi: 10.2196/medinform.5275.

[3] Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the swiss personalized health network: methodological study. JMIR Med Inform. 2021 Jun;9(6):e27591, doi: 10.2196/27591.

[4] Sugiyama T, Miyo K, Tsujimoto T, Kominami R, Ohtsu H, Ohsugi M, Waki K, Noguchi T, Ohe K, Kadowaki T, Kasuga M, Ueki K, Kajio H. Design of and rationale for the Japan Diabetes compREhensive database project based on an Advanced electronic Medical record System (J-DREAMS). Diabetol Int. 2017 Jun;8(4):375-82, doi: 10.1007/s13340-017-0326-y.

[5] Tanaka M. Proposal of standardization of medical terminology for exchange and sharing of medical information. Japan Journal of Medical Informatics. 2001;21(1):3-11, doi: 10.14948/jami.21.3.

[6] Hayashi M, Horii H, Kweon I, Yoshida T. Construction of questionnaire survey supporting system using semantic web technology. Japan Journal of Medical Informatics. 2007;27(1):109-16, doi: 10.14948/jami.27.109.

[7] W3C Semantic Web. 2013. Available from: https://www.w3.org/2001/sw/ (cited on 2022 Sep 1).

[8] World Wide Web Consortium (W3C) RDF Working Group. Resource Description Framework (RDF). 2014. Available from: https://www.w3.org/RDF/ (cited on 2022 Sep 1).

[9] MEDIS Standardized master. MEDIS-DC. Available from: https://www.medis.or.jp/4_hyojyun/medis-master/index.html (cited 2023 Mar. 9).

[10] KEGG MEDICUS. Available from: https://www.genome.jp/kegg/medicus.html (cited on 2023 Mar 9).

[11] VIRTUOSO. Available from: https://virtuoso.openlinksw.com/ (cited on 2023 Mar 9).

[12] FastAPI. Available from: https://fastapi.tiangolo.com/ (cited on 2023 Mar 9).

[13] Swagger. Available from: https://swagger.io/ (cited on 2023 Mar 9).

[14] SPARQL. Available from: https://www.w3.org/TR/sparql11-query/ (cited 2023 Mar 9).