German Medical Data Sciences: A Learning Healthcare System
U. Hübner et al. (Eds.)

doi:10.3233/978-1-61499-896-9-50

Samply.MDR – A Metadata Repository and Its Application in Various Research Networks

Dennis KADIOGLU[a,1], Bernhard BREIL[c], Christian KNELL[b], Martin LABLANS[d], Sebastian MATE[b], Danijela SCHLUE[h], Hubert SERVE[f], Holger STORF[a], Frank ÜCKERT[d], Thomas WAGNER[g], Paul WEINGARDT[e] and Hans-Ulrich PROKOSCH[b]

[a] *Medical Informatics Group, University Hospital Frankfurt, Frankfurt am Main, Germany*

[b] *Chair of Medical Informatics, Friedrich-Alexander-University Erlangen-Nuernberg, Erlangen, Germany*

[c] *Niederrhein University of Applied Sciences, Krefeld, Germany*

[d] *Medical Informatics in Translational Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany*

[e] *Cancer Registry Rhineland-Palatinate, Mainz, Germany*

[f] *Department of Hematology and Oncology, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany*

[g] *Frankfurt Reference Center for Rare Diseases, University Hospital Frankfurt, Frankfurt am Main, Germany*

[h] *Center of Clinical Epidemiology, University Hospital Essen, Essen, Germany*

Abstract. Collaboration in medical research is becoming common, especially for collecting relevant cases across institutional boundaries. If the data, which is usually very heterogeneously formalized and structured, can be integrated, such a collaboration can facilitate research. An absolute prerequisite for this is an extensive description about the formalization and exact meaning of every data element contained in a dataset. This information is commonly known as metadata. Various research networking projects tackle this challenge with the development of concepts and IT tools. The Samply Metadata Repository (Samply.MDR) is a solution for managing and publishing such metadata in a standardized and reusable way. In this article we present the structure and features of the Samply.MDR as well as its flexible usability by giving an overview about its application in various projects.

Keywords. Metadata, data integration, interoperability, research networks

1. Introduction

Collaboration in the analysis of medical data is becoming common, as collecting data through the recruitment of patients at a single institution may prove insufficient for, inter alia, large cohort studies or research projects on rare diseases. Collaborators must

[1] Corresponding Author, Dennis Kadioglu, Medical Informatics Group, University Hospital Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany; E-mail: kadioglu@med.uni-frankfurt.de.

therefore share their data with others. This, however, requires the integration of data from all participating sites and their various heterogeneous sources. Generally, this is achieved by consenting on a common dataset. Based on its specification, every participating site has to map all of their relevant data elements. Traditional approaches such as defining and exchanging data element definitions with Excel or Word documents have many limitations, especially in larger research networks. Furthermore, this information would have to be manually translated into mapping or transformation rules. One way to address this is to formally describe all data elements in a metadata repository (MDR) according to internationally accepted standards that provide clear guidelines as to the content and structure of the metadata. Finally, open and standardized access to metadata for data collected in healthcare and medical research would ease data sharing and integration [1]. In order to tackle this challenge, the Samply.MDR was designed and stepwise enhanced in functionality throughout its application in various research networking projects.

2. Methods

One such common metadata standard is ISO 11179, whose six parts generically describe how metadata can be used to describe data elements to preferably cover all possible use cases and domains, e.g. within the manufacturing industry or the healthcare system. Our work focused on ISO 11179 part 3 edition 3 [2], which specifies a metamodel divided into two layers of metadata. A representational layer comprises information about data storage in a database, e.g. the datatype and restrictions such as a range or a list of permitted values of a specific data element. The conceptual layer outlines how to formalize the meaning of data elements, e.g. the relation between a value and its corresponding code item from a classification. To ensure that the metadata content can be adapted to different use cases, any further necessary metadata attributes can be specified as slots (key value pairs), e.g. additional implementation specific information which is needed by the respective software solution relying on the metadata.

Samply [3], a collection of IT tools and components to support collaboration and data sharing in research networks implements, inter alia, parts of ISO 11179. Its development began in 2011 with the idea of supporting smaller biobanks in sharing their biomaterial collections including their annotation data [4]. The intent in using metadata as a natural requirement was (1) to foster the reuse of existing data elements and (2) to support the formulation of inquiries to networked biobanks based on their respective data elements. Samply is published as open source [https://bitbucket.org/medicalinformatics/] to facilitate vendor and implementation independent data and system integration.

3. Results

Samply.MDR [5] is developed as an extensible and integrable MDR implementation to support the specification of data elements in a structured, formalized and standardized way. Its development began in 2012 as part of the IT for the Clinical Communication Platform (CCP-IT) of the German Cancer Consortium (DKTK) [6]. As another project, the development of an Open Source Registry System for Rare Diseases (OSSE) [7], posed additional requirements, the decision was made to reimplement the PHP-based prototype in JSF - in accordance with the other components of the Samply toolset – and to publish it as open source software in November 2015. It comprises a relational database and tailored implementations for the data access layer, data transfer objects and data access objects in the backend. A browser-based graphical user interface as well as a REST-API can be used to manage and access the metadata. All data elements are organized in namespaces and can be identified by a unique Uniform Resource Name (URN). A user can edit data elements according to his or her access rights for the respective namespace. Reading access to data elements can also be restricted to authorized users by hiding namespaces.

Reusing existing (not hidden) data elements is explicitly desired. A user can issue a full text search for matching data elements and simply import a match into his or her own namespace to reuse it. If necessary, parts of the metadata such as the designation and definition can be modified, e.g. by adding a translation of both in another language. To preserve comparability, any other change such as adding a new value to the value domain is not possible, as this would break comparability of the data element in its old and new revision. In this case the user can use the existing data element as a template for a new one. Table 1 provides a simplified overview of possible metadata attributes.

Table 1. Simplified overview about metadata attributes in the Samply.MDR to describe a data element.

Metadata Attribute	Description	Example Value
Identifier	The identifier as URN [namespace: element type:identifier:revision]	urn:miracum:dataelement:17:1
Designation	The name or label (1-n languages)	Heart Rate
Definition	The extensive human-readable description (1-n languages)	Heart Rate of a patient, measured by palpation on the wrist (radial pulse)
Validation Type	The data type, the list of permissible values or the reference to a catalog	Integer
Validation Rules	Restrictions on the validation type	Range between 0 and 250
Unit of Measure	The unit of measure	bpm
Slot	Key value pairs to describe further aspects (use case specific)	'sharing_allowed':'true'

As a first part of an implementation of the conceptual layer, classifications and terminologies can be imported as coding systems or catalogs. All or any subset of codes contained in a catalog can then be defined as permitted values of a data element.

In the following the authors representing the projects give a brief overview about how these are using and developing Samply.MDR.

CCP-IT: Samply.MDR holds the harmonized common dataset upon which all DKTK partner sites have consented. The dataset and its metadata are used for the formulation of inquiries at the central search broker. An inquiry is then downloaded to the respective bridgehead, the local part of the CCP-IT at every partner site, and is executed against the data stored in the local data warehouse, which involves a locally defined and stored tailored mapping to its used data schema.

OSSE [8]: In addition to CCP-IT, OSSE relies on the metadata stored in the same instance of Samply.MDR to design, render and validate electronic case report forms. The user can drag and drop the desired data elements into a form which can then be published and enabled in the registry. Based on the metadata of the enabled forms, the OSSE registry software creates the database at runtime and validates the user's input.

OnkoWiki [9]: The project aims to establish a resource for medical documentation staff to look up data elements and usage guidelines for the documentation of tumor cases. The user interface of Samply.MDR has been adapted to include additionally required metadata attributes (i.a. Pubmed links). Slots have been used in the backend to minimize necessary modifications. To support the process of maintaining the Common ADT/GEKID record format for tumor cases [10], collaboration features, such as commenting on data elements, have been added.

BBMRI-ERIC CS-IT/ADOPT [11,12]: In the research network Biobanking and BioMolecular resources Research Infrastructure - European Research Infrastructure Consortium (BBMRI-ERIC) and its subprojects all the required IT tools are developed to interconnect biobanks across Europe, enabling them to share their resources and facilities. Currently, every biobank stores its information about samples in different ways and data integration is therefore necessary to reach the expected level of interoperability. The BBMRI search tool Sample Locator relies on the metadata stored in Samply.MDR.

GBA: The German Biobank Alliance (GBA) [http://bbmri.de/ueber-gbn/german-biobank-alliance/] networks eleven German biobanks with a harmonized IT platform for the efficient exchange and usage of biomaterial and its associated clinical data as well as for participation in BBMRI-ERIC. As one of the first steps, an installation of Samply.MDR has been set up and the biobanks have started to specify their data elements within their individual namespaces. The collected metadata is being analyzed and the results will be used to identify common data elements which all or at least most of the biobanks could take into account in inquiries.

MIRACUM: MIRACUM (Medical Informatics in Research and Care at University Medicine) [http://www.miracum.de/], one of the four consortia funded by the BMBF within the Medical Informatics Funding Scheme [http://www.medizininformatik-initiative.de], is a network of eight German university hospitals, two technical universities and one industrial partner. Three out of four funded consortia consider ISO 11179 for the specification of metadata. MIRACUM has designed hospital-based data integration centers with the Samply.MDR as local installations at every partner site. These hold the metadata of locally-used data elements and a central MDR will be used to publish and communicate all agreed-upon dataset specifications. Local ETL processes will be implemented based on the metadata to ensure that all relevant data can be queried and used as intended by the distributed search concept of MIRACUM.

4. Discussion

Although one of the ideas is to foster harmonization and reuse of data elements, the current implementation of Samply.MDR lacks functionality to actively support the user in avoiding redundancies. As the current implementation does not cover the entire ISO 11179 standard and mainly provides features to register, specify and manage data elements according to the representational layer of the metamodel, the user can only perform a full-text search, which should be replaced, or at least complemented, by a semantics backed feature. This would improve usability, as the result set contains fewer candidate data elements the user can reuse instead of defining a new one. Also, such enhancement would ease the curation of all the metadata. The incorporation of relations and mappings between data elements, which is currently worked on, will further improve the harmonization and deduplication.

Comparing the concept and implementation of Samply.MDR with other existing approaches in detail, i.a. the tools ART-DECOR, caDSR or semanticMDR, as well as comparing ISO 11179 with other standards, i.a. HL7 FHIR or ISO 13606, is out of scope of this article. However, we will further investigate approaches to ensure interoperability between Samply.MDR and other standards and implementations.

At the moment we are in the process of collecting, evaluating and pushing the various project specific modifications to the public repository. Either way, the project welcomes pull requests or other contributions from the community at any time.

5. Conclusion

The broad use of Samply.MDR as well as the existence of many other similar solutions show that a metadata-based approach can help in developing and establishing IT tools to support collaborative research and data integration. Furthermore, the number of research networks relying on Samply.MDR indicates, that using and developing existing open source solutions can be feasible to allow them to focus on other aspects of their goals.

6. Conflict of Interest

The authors declare that there is no conflict of interest.

References

[1] M. Dugas, K.H. Jöckel, T. Friede, et al., Memorandum "Open Metadata". Open Access to Documentation Forms and Item Catalogs in Healthcare, *Methods Inf Med.* **54.4** (2015), 376-378.

[2] ISO/IEC 11179, Information technology – Metadata registries (MDR). Part 3: registry metamodel and basic attributes. 3rd ed. Date: 2013-12-02, *International Organization for Standardization* (2013).

[3] M. Lablans, D. Kadioglu, M. Muscholl, et al., Exploiting Distributed, Heterogeneous and Sensitive Data Stocks While Maintaining the Owner's Data Sovereignty, *Methods Inf Med* **54.4** (2015), 346–352.

[4] M. Lablans, S. Bartholomäus, I. Roderfeld, et al., Vertrauen und Interoperabilität für die Föderation verteilter Biomaterialbanken, *German Medical Science GMS Publishing House* (2011).

[5] D. Kadioglu, P. Weingardt, F. Ückert, et al., Samply.MDR – Ein Open-Source-Metadaten-Repository, *German Medical Science GMS Publishing House* (2016).

[6] M. Lablans, E. Schmidt, F. Ückert, An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium, *JCO Clinical Cancer Informatics* **1** (2017), 1-8.

[7] M. Muscholl, M. Lablans, T.O.F. Wagner, et al., OSSE – open source registry software solution, *Orphanet Journal of Rare Diseases* **9 Suppl 1** (2014).

[8] H. Storf, J. Schaaf, D. Kadioglu, et al., Register für seltene Erkrankungen, *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* **1-9** (2017).

[9] D. Schlue, S. Mate, J. Haier, et al., From a Content Delivery Portal to a Knowledge Management System for Standardized Cancer Documentation, *Stud Health Technol Inform.* **243** (2017), 180-184.

[10] J. Dudeck, W. Wächter, U. Altmann, et al., The definition of a new uniform basic data set for hospital cancer registries in Germany, *MIE Proceedings Freund Publishing House Ltd.* (1993), 489-492.

[11] G.J. van Ommen, O. Törnwall, C. Bréchot, et al., BBMRI-ERIC as a Resource for Pharmaceutical and Life Science Industries: The Development of Biobank-Based Expert Centres, *Eur J Hum Genet* **23(7)** (2015), 893-900.

[12] M. Lablans, D. Kadioglu, S. Mate, et al., Strategies for biobank networks. Classification of different approaches for locating samples and an outlook on the future within BBMRI-ERIC. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* **59** (2016), 373-378.