Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210112

Metadata of Registries: Results from an Initiative in Health Services Research

Jürgen STAUSBERG[a,1] and Sonja HARKENER[a]
[a] *University Duisburg-Essen, Faculty of Medicine, Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), Germany*

Abstract. Metadata management is an essential condition to follow the FAIR principles. Therefore, metadata management was one asset of an accompanying project within a funding scheme for registries in health services research. The metadata of the funded projects were acquired, combined in a database compatible with the metamodel of ISO/IEC 11179 "Information technology - Metadata registries" third edition (ISO/IEC 11179-3), and analyzed in order to support the development and the operation of the registries. In the second phase of the funding scheme, six registries delivered a complete update of their metadata. The mean number of data elements increased from 245.7 to 473.5 and the mean number of values from 569.5 to 1,306.0. The conceptual core of the database had to be extended by one third to cover the new elements. The reason for this increase remained unclear. Constraints from the grant might be causal, a deviation from an evidence-based development process as well. It is questionable, whether the revealed quality of the metadata is sufficient to fulfill the FAIR principles. The extension of the metamodel of ISO/IEC 11179-3 is in agreement with the literature. However, further research is needed to find workable solutions for metadata management.

Keywords. Documentation, health services research, metadata, registries.

1. Introduction

Metadata availability and metadata management are essential conditions to assure the FAIR principles [1]. Data one is interested in could be located by using information on data elements like a specific sex category ("**F**indability"). Information about the data structure enable third parties to access data automatically ("**A**ccessibility"). A mapping of metadata supports the combination of different data sources to answer a question at hand ("**I**nteroperability"). The description of a data source allows their use even if the context of the data acquisition has changed ("**R**eusability"). Consequently, metadata management was one focus of an accompanying project within a funding scheme from the German Federal Ministry of Education and Research about registries in health services research. Sixteen projects were supported in the development of a registry protocol in the first phase of the funding scheme. The registries' catalogs of data elements were collected from 15 out of the 16 projects, mapped to the metamodel of ISO/IEC 11179 "Information technology - Metadata registries (MDR)" third edition (ISO/IEC 11179-3) [2], and analyzed [3]. The projects received feedback concerning overlaps and

[1] Corresponding Author, Institute for Medical Informatics, Biometry and Epidemiology, Faculty of Medicine, University Duisburg-Essen, Hufelandstrasse 55, 45122 Essen, Germany; E-mail: stausberg@ekmed.de.

discrepancies in order to initiate a quality improvement of metadata a) through their harmonization, e.g. by using the same value set for sex, and b) by adopting existing standards, e.g. for the social status of a person. Fourteen projects applied for funding to realize their registry. Six projects were accepted and started the implementation of their registry in spring 2019 (cf. Table 1) [4].

The funding covered the implementation followed by an operating phase for a total of three (one project) to five years (five projects). One year after the start of the follow-up funding, the accompanying project again asked the projects to provide the catalogs of the data elements, which may have been revised in the meantime. Aim of this work is to describe the metadata, to analyze differences between the catalogs of data elements from the development to the realization phase, and to report on necessary extensions of the ISO/IEC 11179-3 metamodel.

Table 1. Medical fields of the registries (D = development phase, R = realization phase).

Area	Medical field	D	R
Acute conditions	Acute respiratory distress syndrome	x	
	Fever in childhood	x	x
	Heart attacks in Brandenburg	x	
	Pulmonary embolism	x	
	Recurrent urolithiasis of the upper urinary tract	x	x
Chronic diseases	German celiac registry	x	
	Lifelong monitoring of paraplegic patients	x	x
	Treatment exit options for uveitis	x	x
Oncology	Breast cancer care for patients with metastases	x	
	Hereditary breast and ovarian cancer	x	x
Rare diseases	Prader-Willi-Syndrome	x	
	Systemic lupus erythematosus in Germany	x	
Interventions	Appropriateness of total knee replacement for osteoarthritis	x	
	Vaccination information system Saarland	x	
Other conditions	National mortality registry	x	
	Safety of living kidney donors in Germany	x	x

2. Methods

2.1. Metadata model

The accompanying project setup a database with Microsoft Access to maintain the registries' metadata. The implementation of a web-based metadata registry (MDR) did not get the consent of the ministry. Therefore, a decentral access to joined metadata was not budgeted. As an alternative to the proprietary Microsoft Access database, the use of Samply.MDR [5] was considered. This option was discarded due to some constraints in Samply.MDR for our setting.

The structure of the database mixed elements of a catalog of attributes [6] with the metamodel of ISO/IEC 11179-3 (cf. figure 1). We fictively represented the core metamodel of ISO/IEC 11179-3 between DATA_ELEMENT_CONCEPT (DEC), DATA_ELEMENT (DE), VALUE_DOMAIN (VD) and CONCEPTUAL_DOMAIN (CD). A DEC joins an OBJECT like patient with its PROPERTY like sex. The CD points to the conceptual background like a karyotype. The VD lists the possible expressions of the PROPERTY. Finally, the DE establishes a recordable item through a combination of a DEC with a VD. Differently to the Cancer Data Standards Registry and Repository (caDSR) and Australia's Metadata Online Repository (METeOR), we retained CD in

order to map DEs between projects and to link DEs among each other mirroring the ISO/IEC Concept metamodel region. Furthermore, we added the element CONCEPTUAL_DOMAIN_GROUP (CDG) to be able to group thematically related CDs. In particular, this allowed a better overview to be fed back to the projects. Until now, we did not implement multilingualism by using the ISO/IEC Designation and Definition metamodel. DEs in different languages are represented twice in our database yet. However, to completely represent a recommendation on the value set of sex we added DESIGNATABLE_ITEM to oppose German and English denominations.

2.2. Metadata acquisition

As mentioned before, the accompanying project received complete new versions of the registries' catalogs of data elements. Four projects submitted the catalog as Excel file, one as Access database, and one as a set of Word documents. Two Excel files and the Access database followed a structure agreed upon for the exchange of metadata. The six catalogs were semi-automatically imported in the already existing database. Finally, the database included metadata of 15 projects from the development and six projects from the realization phase. Metadata were available from both phases for those six projects. Additionally, the recommendations of the accompanying project concerning metadata were integrated as a separate project. Data elements, documentation objects and values were unambiguously identified in the database through a combination of an ID and the language. Documentation objects represent a useful combination of data elements, e.g. all demographic items of a patient. The new metadata were manually assigned to existing CDs if possible. Otherwise, a new CD was created and if necessary a new CDG as well.

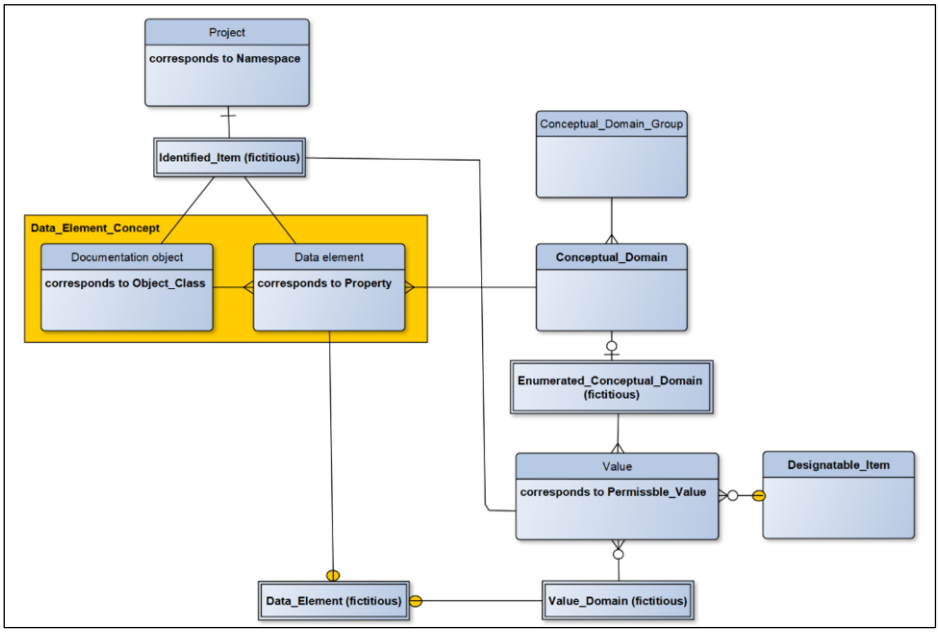

Figure 1. Data model for metadata management. Each entity not marked as fictitious is realized as table in the database. Classes from ISO/IEC 11179-3 in bold.

3. Results

Five projects delivered their catalogs in German, one in English. Table 2 shows the number of elements included in the six catalogs in comparison with the key figures of the development phase. Comparing the mean numbers of the 15 projects of the concept development phase with those of the six projects of the realization phase, the number of data elements has increased by around 80%. The mean number of values has even increased by 180%. On the contrary, the number of documentation objects decreased slightly. The comparison of the projects funded in both phases revealed an increase of the number of values in all six projects. Five projects increased the number of data elements resulting in a median of 482 data elements per project. Two projects increased the number of documentation objects. The number of data elements belonging to a CD with a defined value list (called ENUMERATED_CONCEPTUAL_DOMAIN) decreased slightly for the six projects present in both phases, from 59.6% (879 out of 1,474 data elements) to 54.9% (1,560 out of 2,841).

2,528 out of 2,841 data elements from the revised catalogs (89%) could be assigned to 274 already existing CDs. Further 121 new CDs captured the remaining 313 new data elements. Thirty-six CDs from the development phase were no longer used by the six projects. Four CDGs were introduced to capture 14 new CDs: clinical trial, coronavirus pandemic, mobile application, and quality-assured program for early detection of cancer.

Table 2. Distribution of elements. For the development phase, figures are reported separately for the complete set of projects and the six projects that received a funding for the realization phase.

Phase/Element	Number	Mean	Stddev.	Minimum	Maximum
Realization phase (6 projects)					
Documentation object	103	17.2	6.74	8	28
Data element	2,841	473.5	301.63	88	864
Values	7,833	1,305.5	1,146.63	355	3,356
Development phase (6 projects also funded in the realization phase)					
Documentation object	200	33.3	45.75	8	126
Data element	1,474	245.7	257.27	48	756
Value	3,417	569.5	505.55	114	1,514
Development phase (15 projects)					
Documentation object	352	23.5	28.80	8	126
Data element	3,905	260.3	194.90	48	756
Value	7,016	467.7	438.53	4	1,514

Values counted only from ENUMERATED_CONCEPTUAL_DOMAINS; Stddev. = standard deviation.

4. Discussion

Surprisingly, the number of elements per registry increased from the concepts derived in a competitively organized development phase to the realization phase. The increase in the mean number of values can be explained by filling fields like “country of birth” with respective entries from standard lists. However, the increase in the number of data elements appears inadequate, even if reference data are rare. Compared with the median of 482 data elements for our six registries, about 80% of registries reported lower values for their basic data set in a former survey (11 out of 14 registries [7]). The reason for the increase remained unclear. There could be constraints from the funder to extend the field of interest leading to an increase in research questions and consequently in data elements on the one hand. On the other hand, the evidence-based approach in registry planning and design might be lost after receiving the grant. From the authors’ experiences, the

number of data elements is negatively correlated with the acceptance by the study sites and should be carefully balanced. It is questionable, whether the revealed quality of the metadata is sufficient to fulfill the FAIR principles. As a limitation of our work, some changes might be due to a different notion of the authors regarding an appropriate transformation of the catalogs to the model in figure 1.

Not all of the registries handled inclusion and exclusion criteria as part of the data, because they were separately defined in the registry protocol. Then, even if the data are completely available, a disease or a procedure responsible for the recruiting of a patient remained unknown without further information. Therefore, to be compliant with the FAIR principles, it is advisable to represent those criteria explicitly as data elements. However, from the perspective of metadata methodology, this might intermingle descriptive with administrative metadata [8].

Several concerns are reported in regards to the metamodel of ISO/IEC 11179-3. Frequently, the CD element is skipped [9,10]. Furthermore, features that support the daily use of an MDR are missed. For example, Milward claims for the ability to group data elements [10]. Exactly for this reason, we add a CDG to our metamodel. However, according to Park and Kim [11], it might be possible to achieve the same functionality with a self-referencing relationship. Much more efforts are needed to realize a hands-on management of registries' metadata.

Acknowledgements

The work was funded by the German Federal Ministry of Education and Research under contract 01GY1917B. The authors acknowledge the cooperation with the six registries.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
[2] ISO/IEC 11179-3:2013(E), Information technology – Metadata registries (MDR). Part 3: registry metamodel and basic attributes, Third edition, 2013-02-15.
[3] Stausberg J, Harkener S. Bridging documentation and metadata standards: Experiences from a funding initiative for registries. Stud Health Technol Inform. 2019;264:1046-50.
[4] Stausberg J, Harkener S, Semler S. Recent trends in patient registries for health services research. Methods Inf Med. 2021.
[5] Kadioglu D, Breil B, Knell C, Lablans M, Mate S, Schlue D, Serve H, Storf H, Ückert F, Wagner T, Weingardt P, Prokosch HU. Samply.MDR - A metadata repository and its application in various research networks. Stud Health Technol Inform. 2018;253:50-4.
[6] Leiner F, Haux R. Systematic planning of clinical documentation. Methods Inf Med. 1996;35:25-34.
[7] Stausberg J, Altmann U, Antony G, Drepper J, Sax U, Schütt A. Registers for networked medical research in Germany. Situation and prospects. Appl Clin Inf. 2010;1:408-18.
[8] Riley J. Understanding metadata: What is metadata, and what is it for? Baltimore: NISO; 2017.
[9] Löpprich M, Jones J, Meinecke MC, Goldschmidt H, Knaup P. A reference data model of a metadata registry preserving semantics and representations of data elements. Stud Health Technol Inform. 2014;205:368-72.
[10] Milward D. Model driven data management in healthcare. In: MODELSWARD 2019 - Proceedings of the 7th International Conference on Model-Driven Engineering and Software Development. SCITEPRESS - Science and Technology Publications; 2019:107-18.
[11] Park YR, Kim JH. Achieving interoperability for metadata registries using comparative object modeling. Stud Health Technol Inform. 2010;160:1136-9.