MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190384

Bridging Documentation and Metadata Standards: Experiences from a Funding Initiative for Registries

Jürgen Stausberg[a], Sonja Harkener[a]

[a] University of Duisburg-Essen, Faculty of Medicine,
Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), Essen, Germany

Abstract

Within a funding initiative for patient registries in Germany, data specifications from 15 projects were collected in a structured format. Subsequently, the data specifications were transferred into a structure following the ISO/IEC 11179-3 standard for metadata registries. The data specifications included a median of 16 documentation objects and 196 data elements per project. Only Common designations were used. The bridging to ISO/IEC 11179-3 revealed several overlaps between the specifications. For example, bridging can be used to align the multifold designations of sex and sex categories as well as to harmonize the respective value lists. Further work intends to create a metadata repository based on a community-driven approach as part of the IT-infrastructure of this funding initiative. Without this infrastructure, comparability of the data specifications is unlikely.

Keywords:

Documentation, metadata, registries

Introduction

From the perspective of medical informatics, registries are a type of medical documentation system aiming at the analysis of research questions regarding a group of individuals, typically patients. From this perspective, registries could be differentiated from health records supporting health care for individuals. In a broader view, registries are projects that maintain staff, rooms, and IT infrastructure and sometimes have a specific legal status [1]. However, even in the broader view, data recorded on the observational units are the assets of registries. The definition of those data in terms of variables and value lists is therefore a main task in the development of registries [1, 2, 3]. Ideally, this definition is supported by documentation standards. Documentation standards guarantee that the data on the one hand cover all information needed to answer the research questions. On the other hand the data must support the analysis methods applied in statistical evaluations in reports and in scientific and public publications. Proposals for those documentation standards are available [4].

On a level going beyond a single registry, the interaction between projects and between different types of documentation systems gains importance. This interaction requires a description of the data independent from individual research questions and individual statistical analyses. Use cases on that level can include the identification of projects sharing the same approach to quality of life or addressing the same population, for example patients suffering from dementia [5]. ISO/IEC 11179 "Information technology - Metadata registries (MDR)" [6] is a standard that relates to such kinds of use cases. ISO/IEC offers an approach to maintain data about data, i.e. metadata, in a systematic and unambiguous approach. Although ISO/IEC 11179 is noticed in health care [7, 8], the relationship between documentation standards and the metadata approach is complex [9].

With the goal of supporting both the development of individual registries and the interaction between different registries, a German funding initiative defined a bridging between documentation standards and metadata standards that will be introduced below. The bridging will be used to establish a metadata repository for the funding initiative in the future.

Methods

Registries

The German Ministry for Research and Education launched a funding initiative for registries in health services research. Sixteen projects were accepted for a concept development phase lasting nine months. At the end of this phase, the projects were invited to submit a proposal for an implementation phase lasting up to five years. In parallel to the funding of individual registries, an accompanying project (abbreviated REGISVF-AP) was accepted that supports the registries in development and implementation. REGISVF-AP started one month later than the registries which began in September 2017 and will last two years. Thus, it covered eight of the nine months of the concept development phase. The decision about funding of the registry implementation will be announced at the end of 2018.

The 16 projects of the development phase covered different areas of health care: rare diseases, oncology, acute conditions, chronic diseases, interventions, and other conditions. Further information about the projects is provided in table 1.

Table 1 – Projects of the development phase

Acronym	Institution	Area
Rare diseases		
sLEGER	University clinics Düsseldorf	Systemic lupus erythematosus in Germany
PAREMIS	University clinics Leipzig	Prader-Willi-Syndrome
Oncology		
HerediCaRe	University clinics Cologne	Hereditary breast and ovarian cancer
BRE-4-MED	University clinics Würzburg	Breast cancer care for patients with metastases
Acute conditions		
RADaR	University clinics Regensburg	Acute respiratory distress syndrome
FieberApp-Register	Private university Witten/Herdecke	Fever in childhood
HIRB	Brandenburg Medical School	Heart attacks in Brandenburg
EMBO-Lung	Ludwig-Maximilians-Universität Munich	Pulmonary embolism
RECUR	University clinics Freiburg	Recurrent urolithiasis of the upper urinary tract
Chronic diseases		
TOFU	University clinics Bonn	Treatment exit options for uveitis
GeCeR	Competence network bowel diseases	German celiac registry
ParaReg	University clinics Heidelberg	Lifelong monitoring of paraplegic patients
Interventions		
INDICATE-TKR	Technical University Dresden	Appropriateness of total knee replacement for osteoarthritis
IISAAR	Saarland University	Vaccination information system Saarland
Other conditions		
SoLKID-GNR	University clinics Münster	Safety of living kidney donors in Germany
CDDD_ENMR	Center for Population and Health Wiesbaden	National mortality registry

Documentation Standard

REGISVF-AP suggested the use of the catalogue of attributes from Leiner and Haux [10] as a format for data definition in the registries' concepts. The registries agreed to use this approach in an adapted structure. The catalogue of data elements offers a view on data definitions that is communicable to domain experts on the one hand and usable for computer scientists to implement the data definitions in the data management infrastructure on the other hand. The catalogue of data elements differentiates between documentation objects like "patient" or "quality of life" and data elements like "sex" or "VAS" (= visual analogue scale). Data elements are assigned to one and only one documentation object. Documentation objects and data elements share several attributes like designation and description. Both could be related to predefined research questions, to project modules, to statistical methods defined in a statistical analysis plan, to health care related categories such as therapies, and to project-related events such as follow-ups. It was furthermore possible to enrich each data element with a) a value set that is either a list or a data type, b) a coding of values within a value list, c) a unit in case of a numerical data type, d) a differentiation between single and multiple occurrences, and e) plausibility checks.

The registries were advised to start the development with the agreement on research questions, to extract target variables, influencing variables and confounders as potential data elements, and to check any proposed data elements according to the following aspects:

- Is the data element needed to answer a research question?
- Does the data element adequately cover the information?
- Is the recording of the data element feasible for the study sites?
- Could the data element be used in the process of analysis?

Templates for the registries' catalogues of data elements were provided as Microsoft Excel and Microsoft Access files. The templates were available in German and English. A short explanation supported the use of the templates. Additionally, a few integrity constraints were implemented in the Access file.

Metadata Standard

The Data Description metamodel and the Concepts metamodel parts of the ISO/IEC 11179 third edition (ISO/IEC 11179-3) were used [6]. In short, a DATA_ELEMENT is defined by a DATA_ELEMENT_CONCEPT (e.g. sex of a patient) which uses a CONCEPTUAL_DOMAIN (sex) combined with a specific VALUE_DOMAIN (e.g. consisting of the PERMISSIBLE_VALUES male, female, diverse, undetermined, and unknown, representing the legislative definition of sex for citizens in Germany). A VALUE_DOMAIN has two possible subclasses, namely the ENUMERATED_VALUE_DOMAIN (represented by a list of distinct values) and the DESCRIBED_VALUE_DOMAIN (represented by a textual definition).

Bridging Approach

The documentation object of the data element catalogue was mapped to the OBJECT_CLASS of ISO/IEC 11179-3. Each data element of this documentation object establishes a PROPERTY leading to a DATA_ELEMENT_CONCEPT combining both. The value set of the data element catalogue was defined as ENUMERATED_VALUE_DOMAIN in case of a list of values and as DESCRIBED_VALUE_DOMAIN otherwise. Values of the value list became PERMISSIBLE_VALUES of ISO/IEC 11179-3. These transformations were automatically performed based on the received specification of the registries. To close the loop between the DATA_ELEMENT_CONCEPT and the VALUE_DOMAIN at the conceptual level of the Data Description metamodel, a CONCEPTUAL_DOMAIN was created in a stepwise-approach. Firstly, all designations of data elements that occur at least twice established automatically a CONCEPTUAL_DOMAIN. This CONCEPTUAL_DOMAIN was assigned to all data elements that share the designation. In a second step, all remaining data elements were assigned to a preexisting CONCEPTUAL_DOMAIN if possible. Otherwise a new CONCEPTUAL_DOMAIN was created. Finally, the CONCEPTUAL_DOMAINS were manually corrected for duplicates and overlaps. To support understandability and retrieval of the results, each CONCEPTUAL_DOMAIN was

related to a single broader CONCEPT via a LINK of the ISO/IEC Concept metamodel region. In the preliminary stage of the presented work, the CONCEPT was arbitrarily defined by the authors.

Table 2 shows an overview of the bridging concept. The bridging was performed semi-automatically using Microsoft Access.

Table 2 – Bridging concept

Catalogue of data elements	ISO/IEC 11179-3	Comment
Documentation object	OBJECT_CLASS	
Data element	PROPERTY	
Value set	ENUMERATED_ VALUE_DOMAIN	in case of a list of values
Value set	DESCRIBED_ VALUE_DOMAIN	in case of data types
	PERMISSBLE_ VALUE	items in the list of values
	CONCEPTUAL_ DOMAIN	semi-automatically determined
	CONCEPT	manually created

Results

Fifteen out of the 16 projects submitted a specification of their data. Two projects submitted a Word file with free text; 13 projects made use of the template of the catalogue of data elements: seven submitted an Excel file, five submitted an Access file, one submitted an Excel and an Access file in parallel. Six projects defined their data in German, six in English, and three in both languages. Nine projects primarily used phrases for the designation of data elements, one used questions, one used phrases and questions, and four used labels.

Catalogues of Data Elements

The catalogues of data elements defined a total of 352 documentation objects. Excluding a project that used 126 documentation objects, the number of documentation objects ranged between 8 and 27 (median: 16, variation coefficient: 0.31). Fifty-one documentation objects were additionally described in a second language (German or English) leading to 403 designations. The most frequent designation was “basis data” (four in English, six in German as “Stammdaten”). A unique phrase or question was present in 385 out of the 403 designations.

The documentation objects included 3,935 data elements (range per project: 48 to 756, median: 196, variation coefficient: 0.73), 842 additionally available in a second language. Excluding duplicates for the combination of designation and language, 4,068 different designations remained from 4,777. From 4,068 designations, 3,929 occurred only once, and 53 occurred several times. The most frequent designations were “Geschlecht” (=sex, six times), “Geburtsdatum” (=date of birth, six), “Pseudonym” (German, five), “date of birth” (five), “Vorname” (=first name, four), “gender” (English, four), and “sex” (four). The range of data elements per documentation object was 1 to 191 (median: 16, variation coefficient: 1.62).

Example - The Data Element Sex

All projects defined at least one data element related to the sex of a patient. In total, 25 data elements occurred for sex with eight different designations. The data elements for sex were assigned to 16 different designations of documentation objects, most frequently to “Stammdaten” (=basis data, five times) or “basis data” (four times). The 25 data elements shared 14 different value lists. The languages and acronyms were mapped to the following seven values: männlich/male/m, weiblich/female/w, intersexuell/intersexual, unbekannt/unknown, unbestimmt/not determined/undecided, anderes/other/a, and unclassified. Only one data element used the value list defined on a national level for routine data in secondary care, and not one used the definition defined in Germany for primary care.

Mapping to ISO/IEC 11179-3

1. Identified by using a well formatted structure as the value list, 1,197 ENUMERATED_VALUE_DOMAINS out of 3,935 value sets were identified (30%) in the preliminary stage of the bridging process.
2. Those ENUMERATED_VALUE_DOMAINS included 4,162 different PERMISSIBLE_VALUES.
3. The remaining 70% of the value sets were regarded as DESCRIBED_VALUE_DOMAIN.
4. With a minimum of two occurrences of data element designations independent from languages or project assignments, 670 CONCEPTUAL_DOMAINS were extracted. Those covered 1,709 data elements out of 4,777.
5. By now, another 761 data elements have been manually assigned to an existing or newly created CONCEPTUAL_DOMAIN.
6. After clearing of duplicates, 464 CONCEPTUAL_DOMAINS remained covering 2,470 data elements.
7. Preliminarily, 325 of the 464 CONCEPTUAL_DOMAINS covering 2,143 data elements were aggregated to 21 CONCEPTS with drug (533 data elements) and disease/symptom (302) as the most frequent ones.

Discussion

Fifteen projects submitted their specifications of variables and values as part of the registry development. Thirteen used the recommended format of a catalogue of data elements for the submission. According to the coefficient of variation, the projects agreed concerning a reasonable number of documentation objects. To the contrary, the total number of data elements varied strongly between the projects. Additionally, the number of data elements that were condensed to a documentation object was quite different. In agreement with a survey including 30 already implemented

registries [11], most of the registries intended to record between 100 and 499 data elements (cf. figure 1).

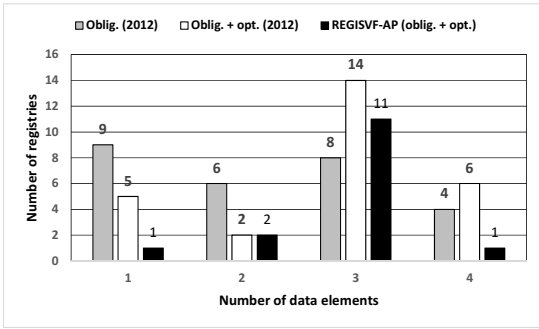

Figure 1 – Number of of data elements per registry (references for 2012 from [11]). Data elements: oblig.=obligatory, opt.=optional.

However, the projects strongly disagreed concerning the designations of documentation objects, data elements and values. They also disagreed concerning the value lists that arise through the aggregation of single values. To some extent this could be explained through the diverse use cases like uveitis, fever, or metastasizing breast cancer. But the example of the values found for the data element sex showed that the projects did not even consider national conventions. An evidence-based recommendation about an appropriate documentation standard was not sufficient to establish homogenous metadata.

Therefore, the results clearly emphasize the need for metadata repositories. A metadata repository can be used for several use cases [11]:

- it makes the creation of case report forms and data models easier,
- it improves the quality of case report forms and data models,
- it harmonizes variables and value lists,
- and it supports the mapping of metadata and data.

Nevertheless, the relationship between metadata approaches and documentation standards is ambivalent [4]. For both, the modelling of health-related information could be different. One might consider “disease” as the main entity and “heart attack” as attribute or “heart attack” as main entity and “existence” as attribute. The implementation of models for complex information objects like “blood pressure” or “quality of life” is unclear for both metadata standards and registry definitions.

The interim results of the bridging between the projects’ specifications of data elements and the structure of ISO 11179-3 underline the value of a metadata repository. For many data elements, counterparts in other projects can be identified. The authors expect a significant stimulus for harmonization of variables and value lists through feedback of this information to the projects. However, the time-consuming effort for the mapping demonstrated the necessity of direct involvement of registry staff in maintaining their metadata themselves. Additionally, the high abstraction level of ISO 1179-3 must be concretized to reach acceptance by those users.

Conclusions

A metadata repository is a core component of an IT-infrastructure for registries as well as for health services and health science research in general. The experiences gained from merging data specifications from 15 registries demonstrate impressively that harmonization and comparability will not arise without explicit metadata services. The high resource consumption needed to align data specifications subsequently underpin that a community-driven approach is essential involving the individual projects in the maintenance of the content. REGISVF-AP will further pursue the objective to implement this approach in the funding initiative of registries for health services research in Germany.

Acknowledgements

The work was funded by the German Federal Ministry of Education and Research under contract 01GY1720B. REGISVF-AP is implemented together with the Technology, Methods, and Infrastructure for Networked Medical Research (Berlin, Germany) and the German Network for Health Services Research. We would like to thank all projects funded within REGISVF who submitted their specifications.

References

[1] R.E. Gliklich, N.A. Dreyer, M.B. Leavy, eds., *Registries for Evaluating Patient Outcomes: A User’s Guide. Third edition. Two volumes,* Agency for Healthcare Research and Quality, Rockville, MD, 2014.

[2] D. Muller, M. Augustin, N. Banik, W. Baumann, K. Bestehorn, J. Kieschke, R. Lefering, B. Maier, S. Mathis, S.J. Rustenbach, S. Sauerland, S.C. Semler, J. Stausberg, H. Sturm, C. Unger, and E.A. Neugebauer, [Memorandum registry for health services research], *Gesundheitswesen* **72** (2010), 824-839.

[3] M. Zaletel, M. Kralj, eds., *Methodological guidelines and recommendations for efficient and rational governance of patient registries,* National Institute of Public Health, Ljubljana Slovenia, 2015.

[4] S.M. Ngouongo and J. Stausberg, Terminology-based documentation systems: a systematic comparison of four different approaches, *Stud Health Technol Inform* **190** (2013), 97-99.

[5] J. Stausberg, M. Lobe, P. Verplancke, J. Drepper, H. Herre, and M. Loffler, Foundations of a metadata repository for databases of registers and trials, *Stud Health Technol Inform* **150** (2009), 409-413.

[6] ISO/IEC 11179-3:2013(E), *Information technology – Metadata registries (MDR). Part 3: registry metamodel and basic attributes,* Third edition, 2013-02-15.

[7] D.B. Warzel, C. Andonaydis, B. McCurry, R. Chilukuri, S. Ishmukhamedov, and P. Covitz, Common data element (CDE) management and deployment in clinical trials, *AMIA Annu Symp Proc* (2003), 1048.

[8] Australian Institute of Health and Welfare, *National Health Information Standards and Statistics Committee. Creating nationally-consistent health information: Engaging with the national health information committees,* AIHW, Canberra, 2014.

[9] S.M. Ngouongo, M. Lobe, and J. Stausberg, The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research? *J Biomed Inform* **46** (2013), 318-327.

[10] F. Leiner and R. Haux, Systematic planning of clinical documentation, *Methods Inf Med* **35** (1996), 25-34.

[11] J. Stausberg, Register, Kohorten und Data Repositories, in: *IT-Infrastrukturen in der patientenorientierten Forschung. Aktueller Stand und Handlungsbedarf - 2012/2013,* J. Drepper, S.C. Semler, eds., AKA GmbH, Berlin, 2013, 31-42.

Address for Correspondence

Prof. Dr. med. Jürgen Stausberg, Institute for Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Hufelandstraße 55, D-45147 Essen, Germany
Email: stausberg@ekmed.de.