Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220432

Metadata Definition in Registries: What Is a Data Element?

Jürgen STAUSBERG[a,1], Sonja HARKENER[a], Markus BURGMER[b], Christoph ENGEL[c], Robert FINGER[d], Carsten HEINZ[e], Ekkehart JENETZKY[f,g], David MARTIN[f,h], Rüdiger RUPP[i], Martin SCHOENTHALER[j], Christian SCHULD[i], Barbara SUWELACK[k] and Jeannine WEGNER[k]

[a] *University Duisburg-Essen, Faculty of Medicine, IMIBE, Essen, Germany*
[b] *Department of Psychosomatics and Psychotherapy, LWL-Hospital and University Hospital Münster, Münster, Germany*
[c] *Leipzig University, IMISE, Leipzig, Germany*
[d] *Department of Ophthalmology, University Hospital Bonn, Bonn, Germany*
[e] *Department of Ophthalmology, St. Franziskus-Hospital Münster, Münster, Germany*
[f] *Faculty of Health/School of Medicine, Witten/Herdecke University, Witten, Germany*
[g] *Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Center of the Johannes-Gutenberg-University, Mainz, Germany*
[h] *Department of Pediatrics, Eberhard-Karls University Tübingen, Tübingen, Germany*
[i] *Spinal Cord Injury Center, Heidelberg University Hospital, Heidelberg, Germany*
[j] *Department of Urology, University of Freiburg, Freiburg, Germany*
[k] *University Hospital Münster, Transplant Nephrology, Münster, Germany*

Abstract. Observational research benefits from a rich methodological foundation of registry development and operation published in international and national guidelines. Metadata management is an essential part of registry implementation based on concepts of data elements and value sets. The metadata from six German registries revealed vastly divergent interpretations of the concept of data elements. The different perspectives of research questions, data acquisition and data storage were all represented in the registries' catalogs of data elements. Consequently, the whole life cycle of a registry needs to be accompanied by a catalog of data elements, which has to be continuously adapted to the changing perspectives. A standard for the representation of those metadata is still missing. The FAIR Guiding Principles introduce important methodological requirements, but the tools for their fulfillment in respect to the management of metadata are still in its infancy.

Keywords. Data element, documentation, metadata, registries.

1. Introduction

Registries are an established method in observational studies in health services research, in epidemiology and public health, and in clinical research. Dating back to the National Leprosy Registry of Norway [1] established in 1856, registry research is nowadays in a

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

consolidated phase with a rich set of standards and recommendations for best practices [2]. The management of metadata, i.e. mainly the definition of data elements and value sets, represents an essential task within development and operation of a registry. However, it is a real challenge to reduce strongly or to remove the mess of concepts and terms related to metadata [3].

There is consent in international and national guidelines about research questions as starting points for the definition of the registries' metadata. The US-American Agency for Healthcare Research and Quality (AHRQ) describes the translation of these questions into measurable exposures and outcomes as an essential step of implementation [4]. These measures are represented by data elements that are grouped to subcategories and assigned to one of the three domains characteristics, treatments, and outcomes. The AHRQ further points out that data elements can also be defined from the perspective of data analyses, e.g. by calculating new data elements from others as the body mass index from body weight and height. Furthermore, a data element can be associated to a source, for example a case report form (CRF), documenting the type of the source or the time of recording [5]. Unfortunately, a standardized organization structure of registries' metadata with the levels measures, analysis and source is neither explicitly consented in registry research [6] nor reflected by the third edition of the international standard ISO/IEC 11179-3 "Information technology - Metadata registries (MDR)" [7].

A consensus approach to the definition of registry data elements was the aim of this work, because of the lack of consensus in this definition.

2. Methods

2.1. Material

Six registries were included in a survey within a funding initiative for health services research (cf. [8] for details). The registries covered different fields of health care, lifelong monitoring of people with spinal cord injury (A), treatment exit options for uveitis (B), hereditary breast and ovarian cancer (C), safety of living kidney donors (D), recurrent urolithiasis of the upper urinary tract (E), and fever in childhood (F). In order to perform a cross-project comparison, a template based on Microsoft Access for a common representation of the individual metadata was forwarded to the registries [9].

Three projects reported the metadata of their registries based on this template (A, B, C). One delivered a codebook exported from an electronic data capture system (D); another provided a set of text-files based on Microsoft Word with metadata organized in tables (E). The sixth project used a proprietary tabular format (F). Only one project delivered their metadata fully compliant with the template's format (A). Consequently, 1) all material was transferred by the independent group mentioned above into the template's structure without changing the project's interpretation of the concept "data element" and 2) information was added where necessary, e.g. a documentation object as an umbrella structure for data elements. The complete list of metadata was finally embedded into the meta model of ISO/IEC 11179-3 for further analyses [10].

2.2. Reference system

Figure 1 shows the reference system applied in this work. For answering research questions, information needs to be collected to allow analyses. The information is formally represented in a catalog of data elements [11], independently from technical aspects related to data storage or data acquisition. The catalog of data elements defines the structure of the data storage as well as the content of CRFs or other means of data acquisition. Reusing data from existing sources will possibly set constraints for the latter.

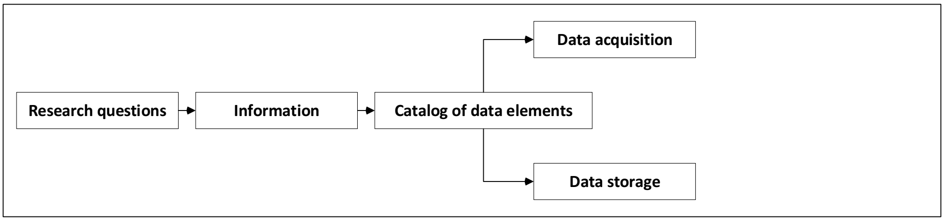

Figure 1. Reference system for metadata of registries.

3. Results

Registry A delivered the catalog of data elements as requested. Data elements represented information independently from technical constraints through the storage or the acquisition of data. Data elements were designated with generally understandable terms, e.g. "gender" or "date of birth". In contrary, registry C filled the designations of data elements with shortcuts like "SEX" or "BIRDAT", called labels in the following. The relationship from labels to information was then provided through a description in the template, however with a non-uniform format. Registry B, also delivering the metadata with the template, partially used questions as data elements, like "What was the date of birth?" Furthermore, registry B added context to a data element in its denomination, for example the role of the person that is asked for his or her date of birth.

The codebook from registry D contained names of variables and fields. The entries were mainly short forms of designations like "dateofbirth", complemented through a counter as "drug_1" where needed. The counters were related to multiple answers as well as to multiple time points. Registry E used the full capacity of free text, e.g. by defining more than one data element in a table cell. Such data element collections will be divided for data storage and data acquisition. Some registries used data element collections specifically in case of established questionnaires. Some entries from registry E were tagged as calculated, representing entries that are not relevant for data acquisition but relevant for data storage. Registry F delivered an elaborated legacy format of its metadata offering consequently a short format of a data element like "date_birth" and a description like "date of birth". In case of questionnaire data elements, the description included a direct question to the recording person. Furthermore, the description established links between two data elements by repeating the counterpart in square brackets.

Technical information was found in the registries' deliveries too. Registries A, C, and F reported primary or foreign keys. Denominations from registry E included guiding notes related to data acquisition, e.g. a request to record more information in case of a certain selection. Overall, data elements were filled with the following types of entries:

- Generally understandable full or abbreviated terms
- Terms with context information, terms with counters, terms with guiding notes for data acquisition, terms tagged as calculated
- Data element collections with terms combining different data elements

- Shortcuts or labels
- Questionnaires or instruments, questions or prompts
- Primary or foreign keys

4. Discussion

The registries provided different interpretations of the concept “data element” independently from the use of the recommended template. Comparing the feedbacks of the registries with the reference system in figure 1, it becomes clear that different perspectives were used all having an own value in the management of metadata. The feedbacks comprised the perspectives of data storage and data acquisition additionally to the perspective of the information needed to answer research questions. Consequently, the catalog of data elements was also used for the listing of technical labels as well as for the listing of questions presented on CRFs. Therefore, this work extends the understanding of the challenge. Not only the “variety of ways in which similar data are represented” [12] has to be handled, but also the different perspectives indentified here.

The example of quality of life (QoL) should highlight the dependencies between the different perspectives. Patient reported outcome could be in the focus of a research question. Health-related QoL is defined as information needed for answering this research question. The EQ-5D-5L is selected as instrument to measure QoL and listed in the catalog of data elements. Implementing the EQ-5D-5L in a patient app requires the presentation of five questions with five categories each, along with a vertical line with a length of 20 centimeters. This lead to five further data elements - mobility, self-care, usual activities, pain/discomfort, anxiety/depression - to store the results of data acquisition. Furthermore, an element labelled VAS (visual analogue scale) with a numerical datatype is added for the visual analogue scale. The numerical value of QoL is calculated from the five dimensions of the EQ-5D-5L using a value set available for different nations [13]. This calculation requires three further data elements, two for the master data of the value set and one for the result that will be used in statistical analyses. The catalog of data elements is finally filled from all three perspectives regarding QoL.

The reference system must be extended to represent domain specific supplements as well as technical elements found in the registries’ documents (cf. figure 2). Unfortunately, existing work does not support this complex system. For example, the Common Healthcare Data Interoperability Project focusses on the simplified model in figure 1 with a data acquisition and a data storage perspective only [6]. ISO/IEC 11179 defines a data element as “a unit of data that is considered in context to be indivisible” [7]. Taking the example of QoL, this definition supports both, the recording of EQ-5D-5L as data element with the score as value and the recording of its five dimensions as data element with the value set of five categories each. However, the term “data element” has not been adopted by the Clinical Data Interchange Standards Consortium (CDISC) standards (cf. http://www.cdisc.org/). We found items, variables and labels depending on the specific CDISC module. From the point of view of metadata science, the registries’ interpretation of the term “data element” comprises descriptive, structural and technical aspects [14].

The FAIR guiding principle that “(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation” [15] is theoretically relevant, but nowadays not fully satisfiable. The concept of a data element must be interpreted at least from the perspectives of research questions, data acquisition and data storage to fully capture the life cycle of a registry. The variety of different interpretations of this concept in the metadata from the six registries in our study highlight this notion. We plan to make the metadata of the six registries publicly available to further stimulate the transformation of the reference system into metadata methodology.

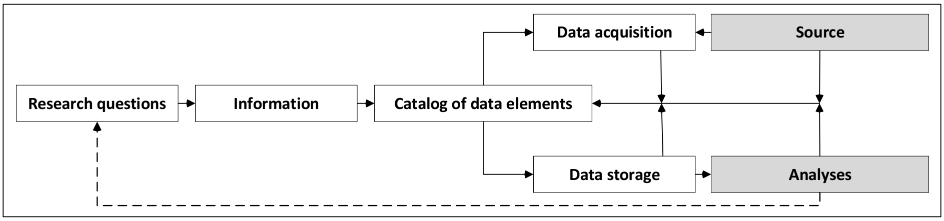

Figure 2. Extended reference system for metadata of registries.

Acknowledgements

Funded by the German Federal Ministry of Education and Research under contracts 01GY1901, 01GY1902, 01GY1903, 01GY1904, 01GY1905, 01GY1906, 01GY1917B.

References

[1] Irgens LM. The origin of registry-based medical research and care. Acta Neurol Scand. 2012;126 (Suppl 195):4-6.

[2] Baig, MA, Bazarbashi MS, AlDakhil H, Baig SM. The journey of clinical registries through various phases of the digital age: a technical perspective. Stud Health Technol Inform. 2022;289:345-8.

[3] Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, Gött R, Kern J, Lablans M, Majeed RW, Stöhr MR, Stausberg J, Varghese J, Dugas M, Ingenerf J. Understanding the nature of metadata: systematic review. J Med Internet Res. 2022;24:e25440.

[4] Glicklich RE, Leavy MB, Dreyer NA, eds. Registries for evaluating patient outcomes: a user's guide. 4th ed. AHRQ Publication No. 19(20)-EHC020. Rockville, MD: Agency for Healthcare Research and Quality; September 2020.

[5] Dugas M. Design of case report forms based on a public metadata registry: re-use of data elements to improve compatibility of data. Trials. 2016;17:566.

[6] Tcheng JE, Drozda JP, Gabriel D, Heath A, Wilgus RW, Williams M, Windle TA, Windle JR. Achieving Data Liquidity: Lessons Learned from Analysis of 38 Clinical Registries (The Duke-Pew Data Interoperability Project). AMIA Annu Symp Proc. 2020;2019:864-73.

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

[8] Stausberg J, Harkener S, Semler S. Recent trends in patient registries for health services research. Methods of Information in Medicine. 2021; 60 (S 01):e1-e8.

[9] Stausberg J, Harkener S. Metadata of registries: results from an initiative in health services research. Stud Health Technol Inform. 2021; 281:18-22.

[10] Stausberg J, Harkener S. Bridging documentation and metadata standards: Experiences from a funding initiative for registries. Stud Health Technol Inform. 2019;264:1046-50.

[11] Leiner F, Haux R. Systematic planning of clinical documentation. Methods Inf Med. 1996; 35: 25-34.

[12] German E, Leibowitz A, Shahar Y. An architecture for linking medical decision-support applications to clinical databases and its evaluation. J Biomed Inform. 2009;42:203-18.

[13] Ludwig K, Graf von der Schulenburg JM, Greiner W. German Value Set for the EQ-5D-5L. Pharmacoeconomics. 2018;36:663-74.

[14] Riley J. Understanding metadata. What is metadata, and what is it for? Baltimore, MD: NISO, 2017.

[15] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.