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


*doi:10.3233/SHTI190194*

# Compatible Data Models at Design Stage of Medical Information Systems: Leveraging Related Data Elements from the MDM Portal

**Martin Dugas[a], Stefan Hegselmann[a], Sarah Riepenhausen[a], Philipp Neuhaus[a], Leonard Greulich[a], Alexandra Meidt[a], Julian Varghese[a]**

[a]*Institute of Medical Informatics, University of Münster, Germany*

**Abstract**

*Compatible data models are key for data integration. Data transformation after data collection has many limitations. Therefore compatible data structures should be addressed already during the design of information systems.*

*The portal of Medical Data Models (MDM), which contains 20.000+ models and 495.000+ data items, was enhanced with a web service to identify data elements, which are frequently collected together in real information systems. Using Apache Solr, a fast search functionality to identify those elements with semantic annotations was implemented.*

*This service was integrated into the metadata registry (MDR) component of MDM to make it available to the scientific community. It can be used to build intelligent data model editors, which suggest and import frequent data element definitions according to the current medical context.*

***Keywords:***
Health Information Interoperability, Data collection, Metadata

## Introduction

Integration of medical data from different sources is a common but complicated task. There are many use cases for data integration, for instance in medical research: comparison of data from different sites, comparison of new study results with published data sources or merging of routine data with study data. In routine care, data integration is needed for quality assurance, for example to transfer a site-specific data set into a central database, or clinical decision support, for instance to apply standardized decision rules on local data sets. This list is not complete. However, there are many reports about data integration issues [1].

In many cases, collected data from different sources is similar in several aspects (e.g. regarding disease domain, patient group, assessment, therapy, outcome), but from a data analysis point of view, similar is not good enough. Instead, data shall be compatible to enable merging of data sources and joint analysis. A common approach to address this problem is data transformation after data collection. Data warehousing is a typical example for this approach: It applies extract-transform-load (ETL) to generate an integrated data set from different sources. However, the transformation step has important limitations, such as information loss during aggregation or bias due to semantic differences of data elements.

Why are there so many similar but incompatible data sources in medicine? This is caused by the semantic richness of medical terminology. SNOMED CT [2] contains more than 300.000 non-synonymous terms. Even for a small data model like a case report form with only one sheet (e.g. 40 data elements), there is an *astronomical* number of different models. As a consequence, if two medical experts design a data model for a given medical topic independently, the probability for compatible models is extremely close to zero.

Table 1 presents an example for data elements in two information systems, which are similar but incompatible: Myocardial infarction and atrial fibrillation are only a subset of heart diseases, i.e. a patient with myocardial infarction or atrial fibrillation has heart disease, but a patient with heart disease has not necessarily myocardial infarction or atrial fibrillation. The value sets of pain level are not transformable between system 1 (5 levels) and 2 (4 levels). This similarity has major drawbacks: joint data analysis of those systems is very limited and a joint algorithm for decision support is not applicable.

*Table 1– Similar data are problematic for data integration: Data from systems 1 and 2 are similar, but incompatible.*

| System 1 | System 2 |
|---|---|
| heart disease yes/no | myocardial infarction yes/no<br>atrial fibrillation yes/no |
| pain level 0-1-2-3-4 | pain level 0-1-2-3 |

The similarity of medical data models is a big challenge, but also an opportunity for standardization and thereby better compatibility. Data integration would be a lot easier if data models from different sources were more compatible, i.e. had a larger proportion of compatible data elements. Overall, it would be highly desirable to foster compatible data models at the design stage, because post-hoc transformations have limitations.

The Portal of Medical Data Models (MDM-Portal) [3] addresses this problem with an open access approach to foster best practice sharing and thereby promotes re-use of existing data models. Users can search, comment, upload and download data models. An integrated metadata registry (MDR) component provides search functionality at the data element level to identify most frequent data element definitions for a medical concept such as weight or potassium.

The objective of this work is to go one step further and answer the following question: For a given data element (or a set of data elements) - what other data elements are collected most frequently together with this data element in real information systems?

This list of related data elements can contribute to the design of new information systems. It supports completeness of data models by identification of data elements, which are frequently covered in many related data sources; and, importantly, it fosters compatibility of data models by re-use of data element definitions from existing data sources wherever possible.

A concrete example: What lab values (or any other medical characteristics) are frequently collected together with "bilirubin"? And what are the precise data element definitions of these related data elements?

This type of relatedness goes beyond terminological relationship (like atrial fibrillation is a subtype of heart disease) - it is about actual data elements used in real medical data sources.

## Methods

### Semantic annotation of data elements

Semantic annotation is needed to identify meaningful related data elements because item names can be ambiguous [4]. Figure 1 presents an example: an attribute named "size" could refer to body height, tumor size or shoe size. By assigning a semantic code – such as a UMLS [5] or SNOMED CT code – the medical concept of an item name can be specified and a reference to a more detailed description is available.

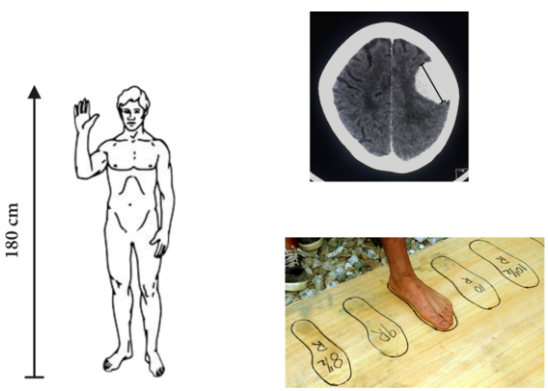


*Figure 1– The need for semantic annotation: "size" is ambiguous, it can refer to body height, tumor size or shoe size.*

### Open access to medical data models

To identify related data elements, the MDM portal is used as a source of data models. Figure 2 presents the overall architecture of this system, which is described in more detail in [3,6]. As of March 2019, MDM contains more than 20,000 data models with approximately 495,000 semantically annotated data elements. To our knowledge, it constitutes the largest collection of medical data models in Europe. Each data element consists of name, textual description (which can be multilingual), data type, and semantic code. Optionally, a code list, value range or unit can be provided. These data elements are defined according to ISO 11179 [7], i.e. for each element, a concept domain and a value domain are specified. An MDR component provides search functionality and frequency counts based on data element name, data type, range, unit, values (code list) and semantic coding.

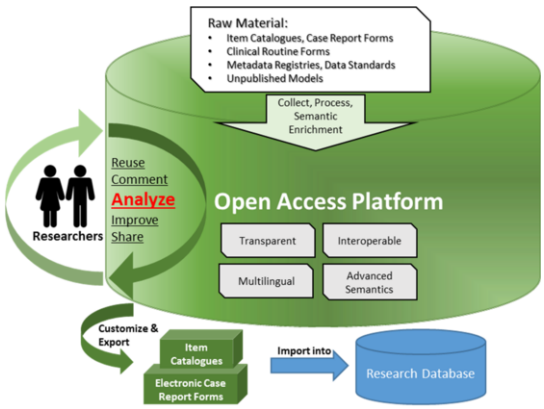


*Figure 2– Overall approach of the MDM portal: an open access platform for medical data models*

### Relationship of data elements

Each data model consists of a list of context-specific data elements; therefore any two data elements can be defined as related, if these elements belong to the same data model. This means that these two data elements are collected together in the same documentation setting.

## Results

### Web service to identify related data elements

Apache Solr is used to index data elements in the MDM portal [8]. This open-source search platform is selected due to its fast full-text search and query language. A web service is implemented with the Java web framework Spring Boot. Java libraries SolrJ and JAXB are applied for the communication with Solr and to process the XML-format of ODM files within the MDM portal.

The overall architecture of the web service "related items" is depicted in figure 3. There are two variants of this service: Variant 1 identifies all data elements, which are jointly collected with a given data element (indicated by data element ID). Variant 2 identifies all data elements, which are jointly collected with a set of data elements with the same name (e.g. "age"). In both variants, results are ordered by frequency of co-occurrence.

The MDR component of MDM is enhanced with this web service using JavaScript to demonstrate the search functionality for related data elements. It is available at https://medical-data-models.org/ under actions / metadata registry (after free registration).

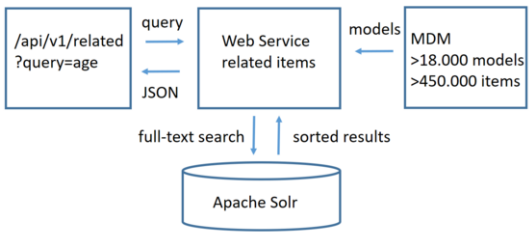


*Figure 3– Architecture of the web service "related items". MDM items are indexed with Apache Solr. For a given query, jointly collected data elements are reported in JSON format, ordered by frequency of co-occurence.*

Figure 4 presents the output of the MDR for search term "Bilirubin". The most frequent definition (n=30) for items named "Bilirubin" is data type float, unit mg/dl and UMLS code C0005437. By clicking on the green button "Search for related items" a list of related data elements, ordered by frequency, is displayed (figure 5). For instance, AST is on this list. This is plausible from a medical perspective, because AST and bilirubin are laboratory tests, which are frequently elevated in certain liver diseases.

Because the system architecture is web service based, this search functionality for related items can also be used by external systems.

MetaData Registry (MDR): Data Elements of MDM English

Item Name Bilirubin ↪ Search for related items

(e.g. weight, height, pulse, ASA, temperature, systolic blood pressure, age, potassium)

Showing 1 to 10 of 1,635 entries

| Frequency | Data element | Data type | Range check | Unit | Values | Semantic Code | Models | |
|---|---|---|---|---|---|---|---|---|
| 30 | Bilirubin | FLOAT | | mg/dL | | C0005437 | | ↪ Related |
| 9 | Bilirubin | FLOAT | | | | C1278039 | | ↪ Related |
| 17 | bilirubin measurement | INTEGER | | mg/dl | | C0344395 | | ↪ Related |
| 8 | Bilirubin | INTEGER | | | Grade 0 (0)<br>Grade 1 (1)<br>Grade 2 (2)<br>Grade 3 (3)<br>Grade 4 (4)<br>Grade 5 (5) | | | ↪ Related |

*Figure 4– Search for items "Bilirubin" in the MDR component*

MetaData Registry (MDR): Data Elements of MDM English

Item Name Bilirubin × Close related items

(e.g. weight, height, pulse, ASA, temperature, systolic blood pressure, age, potassium)

Showing 1 to 10 of 2,436 entries

| Count | Data element | Data type | Range check | Unit | Values | Semantic Code | Models |
|---|---|---|---|---|---|---|---|
| 30 | AST | FLOAT | | U/L | | C0201899 | |
| 30 | Leukocytes | INTEGER | | | | C1271681 | |
| 30 | Urea | FLOAT | | mg/dl | | C0523961 | |
| 30 | Basophils percentage | FLOAT | | % | | C2237945 | |
| 30 | Neutrophils | FLOAT | | 1000/l | | C0200633 | |
| 30 | Date and time sample taken | DATETIME | | | | C0011008, C0040223, C0200345 | |
| 30 | Sodium | FLOAT | | mmol/L | | C0337443 | |
| 30 | Erythrocytes | FLOAT | | Mio/l | | C0014772 | |
| 30 | Neutrophils percentage | FLOAT | | % | | C2238207 | |
| 30 | Lymphocytes percentage | FLOAT | | % | | C2200256 | |

*Figure 5– Related items to "Bilirubin": AST is a marker for liver disease like bilirubin*

## Discussion

Data from different sources should be as compatible as possible to facilitate data integration. Syntactic and semantic interoperability of Electronic Health Records (EHR) is a well-known challenge [9]. Definition of data elements is a key step in the design of medical information systems. A huge number of similar, but incompatible data models can be designed due to the semantic richness of medical terminology. For this reason, re-use of data element definitions from already existing information systems should be fostered. The MDR component of MDM supports this re-use with lists of data element definitions for a given item name.

This work goes one step beyond: It identifies data elements, which are collected frequently together with a given data element in real information systems.

The aspect of frequency is important to select most suitable elements for data collection: In a typical medical setting, data consumers want to analyze many data elements with high data quality, but data producers can provide only a limited number of data elements due to resource constraints. For example, what laboratory tests should be collected in a certain medical setting? Each additional lab value provides potentially interesting information, but is also associated with costs. The cost of data collection is a limiting factor for the number of patients, which can be documented.

With the new web service, designers of information systems can learn from already available data sources. If certain data elements were collected frequently together with a given data element, then those elements should also be considered for the new system. This facilitates joint data analysis in the future, in particular if the data element definitions from previous sources can be re-used without modifications.

Related data elements can provide a link between medical terminology and real information systems. The MDM portal provides a large collection of real data models from different sources. In contrast to terminology services, MDM contains information, which data elements are collected together. This goes beyond terminological relationships like "atrial fibrillation is a subtype of heart disease". MDM can answer the following questions: What data element definitions are available for "atrial fibrillation" and how frequent are these definitions being used? The new web service extends this functionality: What data elements are frequently collected together with the data element "atrial fibrillation"? The answer to this question provides the opportunity to harmonize and improve information systems already at the design stage.

At present, most IT frameworks for information systems provide its own data model editor. Relatively few of these editors can import external data model definitions. If it was easier to re-use data models from existing data sources, compatible data structures would be fostered and data integration would be easier. In addition, a lot of resources to maintain and use thousands of different data model editors could be saved.

A key idea of the new web service is to make data model editors more intelligent: with this web service, the editor can suggest new data elements, which are related to the current model context. For instance, suitable laboratory values for a given setting could be suggested.

There are many format options for data model definitions. In the field of clinical research, CDISC Operational Data Model (ODM) [10] is a growing standard and due to its endorsement by regulatory authorities more and more software solutions adopt it. For this reason MDM provides data models in ODM format, but also in many other formats like HL7 FHIR [11], Archetype Description Language as well as office and statistical formats.

There are many other approaches to foster compatible data models at the design stage of medical information systems. In particular, there are several initiatives to develop and consent common data elements (CDEs), such as the NIH CDE Repository [12]. Archetypes [13] are another major initiative to foster compatible data structures. Several libraries for standardized measures and instruments are being developed, such as PhenxToolkit [14] and ICHOM [15]. To our knowledge, none of these initiatives provide a service to identify related data elements from real data sources.

The current implementation of the web service for related data elements is a proof of concept and therefore has limitations, which should be addressed in future work:

1. It should be validated for a large number of data elements.
2. It relies on the contents of the MDM portal, which covers only a small subset of all medical data models.
3. It needs to be evaluated externally in the design of information systems, specifically regarding its contribution to the compatibility of data models.

## Conclusions

Compatible data models should be implemented at the design stage of information systems. A novel web service to foster re-use of data element definitions is available for the scientific community. This service is based on the MDM portal and identifies data elements which are collected frequently together in real information systems.

## Acknowledgements

Funding from the German Research Foundation (DFG Grant DU 352/11-2, DU 352/11-1) is acknowledged.

## References

[1] R.L. Richesson, J. Krischer, Data standards in clinical research: gaps, overlaps, challenges and future directions, *J Am Med Inform Assoc*. **14** (2007), 687-96
[2] SNOMED CT, Systematized Nomenclature of Medicine Clinical Terms, available from http://www.snomed.org/ [accessed November 15, 2018].
[3] M. Dugas, P. Neuhaus, A. Meidt, J. Doods, M. Storck, P. Bruland, J. Varghese, Portal of medical data models: information infrastructure for medical research and healthcare. Database (2016), pii:bav121.
[4] M. Dugas, Missing semantic annotation in databases. The root cause for data integration and migration problems in information systems, *Methods Inf Med.* **53** (2014), 516-7.
[5] UMLS, available from http://www.nlm.nih.gov/research/umls/ [accessed November 23, 2018].
[6] J. Varghese, M. Fujarski, S. Hegselmann, P. Neuhaus, M. Dugas, CDEGenerator: an online platform to learn from existing data models to build model registries, *Clin Epidemiol*. **10** (2018), 961-970.
[7] International Organization for Standardization/International Electrotechnical Commission: ISO/IEC 11179,

Information Technology - Metadata Registries (MDR) - Part 1: Framework; 2004, page 11, available from http://standards.iso.org/ittf/PubliclyAvailableStandards/c035343_ISO_IEC_11179-1_2004(E).zip [Accessed November 23, 2018]
[8] S. Hegselmann, M. Storck, S. Geßner, P. Neuhaus, J. Varghese, M. Dugas, A Web Service to Suggest Semantic Codes Based on the MDM-Portal, Stud Health Technol Inform. **253** (2018),35-39.
[9] S. Garde, P. Knaup, E. Hovenga, S. Heard, Towards semantic interoperability for electronic health records, Methods Inf Med. **46(3)** (2007),332-43.
[10] Operational Data Model, available from https://www.cdisc.org/standards/data-exchange/odm [accessed November 23, 2018]
[11] HL7 FHIR, available from https://www.hl7.org/fhir/ [accessed November 23, 2018]
[12] NIH CDE Repository, available from https://cde.nlm.nih.gov/ [accessed November 23, 2018]
[13] openEHR, available from https://www.openehr.org/ [accessed November 23, 2018]
[14] C.M. Hamilton, L.C. Strader, J.G. Pratt, D. Maiese, T. Hendershot, R.K. Kwok, J.A. Hammond, W. Huggins, D. Jackman, H. Pan, D.S. Nettles, T.H. Beaty, L.A. Farrer, P. Kraft, M.L. Marazita, J.M. Ordovas, C.N. Pato, M.R. Spitz, D. Wagener, M. Williams, H.A. Junkins, W.R. Harlan, E.M. Ramos, J. Haines, The PhenX Toolkit: get the most from your measures, *Am J Epidemiol.* **174** (2011), 253-60.
[15] International Consortium for Health Outcome Measurement, available from https://www.ichom.org/ [accessed November 23, 2018]

**Address for correspondence**

Prof. Dr. Martin Dugas, dugas@uni-muenster.de