MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-858

The Portal of Medical Data Models: Where Have We Been and Where Are We Going?

Sophia Geßner[a], Philipp Neuhaus[a], Julian Varghese[a], Philipp Bruland[a], Alexandra Meidt[a], Iñaki Soto-Rey[a], Michael Storck[a], Justin Doods[a], Martin Dugas[a]

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

Abstract

To address current key problems of medical documentation: lack of transparency, overwhelming amount of medical contents to be documented and missing interoperability, the Portal of Medical Data Models (http://medical-data-models.org/) was established in 2012. Constantly evolving, four years later, the portal displays more than 8900 medical data models with more than 250000 items, of which 84 % have been semantically annotated with UMLS codes to support interoperability. Giving an update on new functions and contents of the portal, two additional export formats have been implemented, allowing the reuse of forms such as HL7's framework Fast Health Interoperability Resources (FHIR) Questionnaires, as well as the OpenDataKit format. Future projects include the implementation of an ODMtoOpenClinica converter, as well as supporting the reuse of forms with Apple's ResearchKit and Android's ResearchStack.

Keywords:
Surveys and Questionnaires, Semantics, Clinical Trial

Introduction

The Portal of Medical Data Models (MDM), established in 2012, is a constantly evolving and fast-growing German and European information infrastructure for medical research and healthcare [1,2]. The multilingual platform allows the upload, download, discussion, ranking and reuse of medical questionnaires or documentation forms. These "medical data models" are created in Operational Data Model (ODM) format, developed by the Clinical Data Interchange Standards Consortium (CDISC) [3]. ODM is XML-based and represents the standard exchange format for research metadata in order to facilitate interoperability amongst various software systems [4]. Apart from ODM, the portal offers various download formats, enabling the import of metadata into different medical information systems. To improve interoperability and data integration, medical concepts are semantically annotated with Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS), developed by the U.S. National Library of Medicine [5]. The identification and coding of medical concepts is manually performed by medical experts respecting the published coding principles [6]. The creation of medical data models has been standardized, allowing uniform semantic annotation, using ODMedit [7]. The integrated editor ODMedit proposes possibly matching data elements, which have been defined and semantically annotated before and may be reused.
The value of interoperability in the United States was evaluated in 2005, showing that the fully standardized exchange and interoperability of health care information between various providers has the potential to save 77.8 bil- lion dollars per year [8].
The German Federal Ministry of Education and Research has launched a funding initiative to establish data integration centers, for which the portal provides an infrastructure for a standardized exchange of medical data models between participating consortia in order to analyze their medical (meta-) data landscapes [9].
The overwhelming amount of data documented in medicine, as well as the number of distinct medical concepts in a clinical terminology, such as Systematized Nomenclature of Medicine (SNOMED CT) [10], indicate the "astronomical" number of potential medical data models used by health professionals [11]. Up to now, most of the medical forms and data from clinical trials remain unpublished. This leads to a cost and time intensive process of re-developing and re-implementing case report forms (CRFs) for clinical studies and documentation forms in EHR systems. Researchers, practicing physicians and their patients are calling for "open (meta-) data", starting to consider clinical trial data as a public good [12]. Unpublished data hampers the systematic review and reproduction of published results leading to redundancies in clinical research as well as uncertainties in patients and treating physicians. MDM addresses the lack of transparency by granting open access to all of its contents. Furthermore its contents may be downloaded under different Creative Commons Licenses, allowing sharing and adapting the material for different purposes.
In this paper we will analyze the current contents and describe new functions of the portal. Further research objectives give an overview of the user's activities, planned additional download formats, functions and further research based on MDM. The research questions of this paper can be summarized as followed:

1. What kind of medical contents and functionalities are available in the system and how did it evolve over time?
2. Who is using the system and what export formats are selected most frequently?

Methods

Architecture of the Portal and its contents

The technical background of the portal and the editor have been described before [7,13]. To summarize, the portal has been implemented in Ruby on Rails. The data models are stored on a web server. After uploading an ODM file, its structure is additionally stored in a MySQL database on the server. [7,13]

The medical data models are sustainably archived by the University Library of Muenster [14].

Designing medical data models with ODMedit

As data models represent an "interoperable image" of actual medical forms, they are created in the CDISC ODM format using original templates from clinical trials, registries, routine

documentation forms, common data elements (CDEs) , data standards or patient-reported outcomes. Medical experts, such as physicians, medical documentalists and medical students, create and semantically annotate forms with UMLS codes using ODMedit [7]. Currently 14 medical experts with more than 120 hours per week are creating and annotating medical data models manually. The UMLS CUI of a medical concept is stored as an alias property of the items or codelist items. If a medical concept can be described by a single CUI, it is called a pre-coordinated concept. If a medical concept needs more than one CUI to represent its full meaning, the process is called post-coordination. Uploaded medical data models are reviewed and verified for accuracy before being released for public use.

Analysis of the contents

Database queries were performed using the MySQL Workbench 6.3. To determine the time course of items and medical data models offered by the portal, the cumulative number of items and data models from November 2011 until November 2016 were analyzed. Furthermore, the UMLS codes mapped to the items and the Medical Subject Headings (MeSH) descriptors of the category C for diseases mapped as keywords to the medical data models were analyzed to display the medical contents of the portal.

To display the portal's user distribution, a world map was configured using Leaflet, OpenStreetMap and Mapbox. The retrieval of geolocations of the users' most recently used IP addresses was conducted using freegeoip.net. To analyze the geolocation of potential users who didn't register but visited the portal and accessed forms, Apache Log-Files of the previous eight months were assessed, ignoring Internet bots.

An online user survey was conducted in February 2016 using LimeSurvey. The online questionnaire, available in English and German, contained questions on missing functions and further export formats requested by users.

The new functions of the portal were reviewed: the table of contents was expanded and modified after reviewing and analyzing the contents of the portal. By choosing a subentry of the table of contents, the search function is called with predefined terms, similar to a manual search. Two additional export formats were implemented.

Results

Contents of the Portal of Medical Data Models

As of November 2016 the Portal of Medical Data Models contains 8948 active forms, 16794 forms in total, with a total of 256751 items. The analysis of the time course of medical data models shows that the number of medical data models as well as the number of items are continuously growing (see Figure 1). Since December 2015, about 400 new medical data models per month were uploaded to the portal. As for the number of items, approximately 10000 are added per month. 84 % of the items are semantically annotated with UMLS codes, representing their medical concepts. The amount of semantically annotated items totals to 216586 items. The three most common UMLS codes tagged to items, representing medical contents, are "C0022885" for "Laboratory Procedures", "C0031809" for "Physical Examination" and "C0006826" for "Malignant Neoplasms".

Other UMLS codes that were assigned most frequently are mainly administrative concepts, such as "C0011008" – "Date in time" or "C2348585" – "Clinical Trial Subject Identifier".

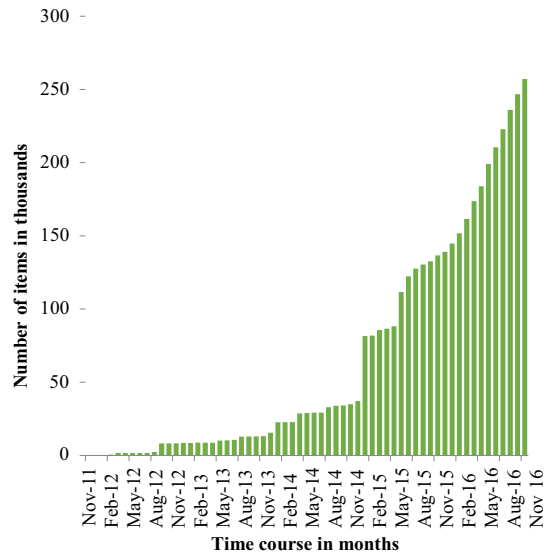

Figure 1 - Time course of items in active medical data models represented in the Portal of Medical Data Models

The results of the analysis of MeSH descriptors is presented in a logarithmic scale in figure 2. A total of 13747 assigned descriptors were identified. The prevailing majority of medical data models is related to oncology. More than 4300 medical data models have been tagged with a MeSH descriptor of the category C04 for neoplasms. The second most used MeSH terms are branches of the category C20 for Immune System Diseases, including diseases such as multiple sclerosis and asthma. The third most represented category is C17 for Skin and Connective Tissue Diseases.

The portal offers contents in multiple languages. Though most 7731 forms (86 %) are monolingual, either in English or German, 1199 forms (13 %) are available in two languages, mainly in English and German. A total of 14 forms (0,16 %) are available in more than two languages, ranging from 3 languages up to 30 languages. The most common language is English with 8414 forms, followed by German with 1858 forms, French with 13 forms and Spanish, Italian and Polish with 5 forms.

Online Survey and User's distribution

The analysis of the users requests for further export formats revealed the demand for HL7's framework Fast Health Interoperability Resources (FHIR) Questionnaires, combining the most advantageous features of HL7 v2, v3 and CDA for exchanging health data [15]. Other export requests were OpenClinica, an electronic data capture system, using Excel templates to import metadata and OpenDataKit (ODK), a toolset for mobile data collection.

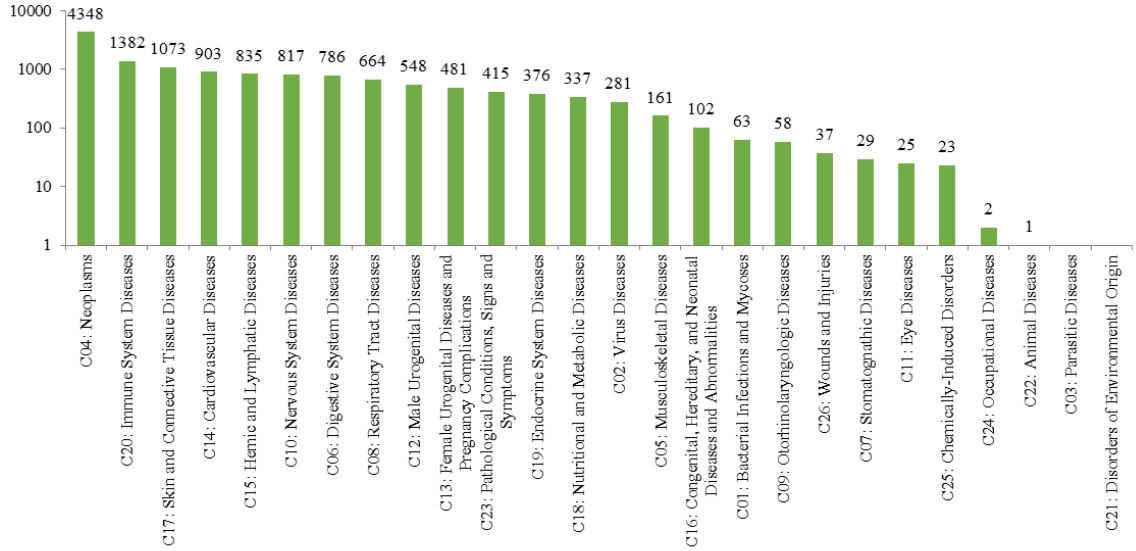

Figure 2 - MeSH descriptors of the category C for diseases assigned to active forms in a logarithmic scale

Currently the Portal has 611 registered and active users, distributed worldwide (Figure 3, green pins). Analyses of users visiting the portal without registration showed the distribution as presented in Figure 3 by the blue pins. The majority of users are located in Central Europe and the United States.

Figure 3 - Geographical distribution of MDM Portal users worldwide. The users' locations are focused on central Europe and North America. The green pins represent registered users (n=611), the blue pins are unregistered users (n=5307).

The total numbers of users, registered and potential are displayed in Table 1, grouped by their geographical origin.

Table 1 - Total number potential users' page views, grouped by their geographical origin

Region	Number of registered users	Number of form visits by unregistered users
Europe	454	3186
Northern/Central America	102	1514
South America	4	60
Asia	34	424
Africa	6	65
Oceania	11	58

Further functions and export formats

In response to the users' requests, two new export formats were implemented. A converter, transforming forms from ODM to FHIR Questionnaire Resources was implemented in Java and integrated into the portal. This included identifying equivalent elements and mapping these ODM elements to the elements of the FHIR Questionnaire Resources [16]. A similar approach was conducted implementing the requested converter ODM to ODK.

Table 2 - Number of downloads for each export format since February 2016

Format	Number of downloads	%
ODM	254	35.2%
PDF	99	13.7%
REDCAP	86	11.9%
CSV	50	6.9%
PDF WITH COMMENTS	48	6.7%
FHIR-XML	37	5.1%
FHIR-JSON	32	4.4%
SQL	26	3.6%
CDA	23	3.2%
XLSX	23	3.2%
MACRO	13	1.8%
SPSS	13	1.8%
ADL	9	1.2%
R	8	1.1%

Table 2 gives an overview of the most frequently downloaded formats. Since the implementation of the ODMtoFHIR converter in February 2016, the FHIR questionnaire format is at 5% for FHIR-XML and 4% for FHIR-JSON as one of the six most common download formats. Most forms were downloaded in ODM (35%), PDF (14%) and REDCap (12%). The least downloaded formats were ADL [17] and R with 1%.

The table of contents (see Figure 4) was modified and adapted to the medical contents of the portal. It contains 7 main categories, arranging the contents by type of documentation within the medical data model, such as "Clinical Trial", "Routine Documentation" or "Patient-Reported Outcomes". Furthermore the contents are indexed by specialty under

"Specialty-related forms", containing 16 subitems, such as "Internal Medicine", "Neurology" and "Surgery".

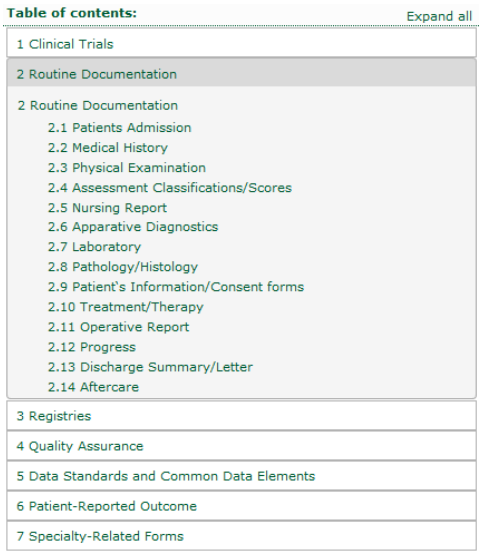

Figure 4 - Table of contents of the Portal of Medical Data Models. The subentry "Routine Documentation" is expanded, showing the contents of the medical documentation in clinical routine. The subitems are chronologically arranged, following the patients path through the clinic from admission to discharge.

Discussion

The analysis of the contents of the Portal of Medical Data Models show major progress over the past year.

In November 2015 the portal contained about 4300 models with about 136500 items. One year later, the number of medical data models more than doubled to the amount of over 8900, increasing by about 400 data models per month. The amount of items also almost doubled to 256751 items, reaching an upload rate of about 10000 items per month. Our goal is to reach an upload rate of 600 medical data models per month in order to ensure the availability of data models in a timely manner. Furthermore the portal shall cover most contents in medical research and clinical practice. Difficulties concerning the open access to metadata in routine patient care and clinical and epidemiological research are still far from being resolved. Small steps in the right direction are being made, as journals like the New England Journal of Medicine have committed to data sharing [18].

To enable semantic interoperability of medical data models, they are semantically annotated. With an annotation coverage of 84%, the portal provides the possibility to access more than 246500 semantically annotated items for reuse. The most frequently assigned UMLS codes with medical content, "laboratory procedures" and "physical examination" represent very well two of the most common medical items documented and examined in clinical trials. It has to be mentioned that semantic annotation is not a trivial task. From our experience, manual review by medical experts is highly required. Even respecting coding principles, mapped UMLS codes still differ between independent coders. When generating CDEs, this still represents an issue, leading to the inevitable, time-consuming process of "code-cleaning". Steps towards uniform semantic annotation are being conducted using the integrated editor ODMEdit [7]. Furthermore, the impact of coding principles and ODMEdit on inter-coder reliability are being evaluated.

With regards to the representation of disease entities in the portal, we are able to show that a wide range of disease entities is already represented. The majority of contents are related to oncology, explaining the third most frequent semantic code assigned ("Malignant Neoplasms"). This is in line with the contents of ClinicalTrials.gov, a registry for clinical studies, maintained by the U.S. National Institutes of Health [19]. It currently lists a total of more than 231900 studies, 25 % of which are associated with "Neoplasms". The European EU Clinical Trials Register currently displays more than 29300 studies [20]. In accordance with the worldwide registry ClinicalTrials.gov, the amount of clinical trials related to oncology represents, with more than 7000 studies, about 24% of the contents. A great amount of medical data models in the portal have been created based on the trial inclusion and exclusion criteria, indicated in the study record details of studies on ClinicalTrials.gov. With a focus on oncologic, neurologic and cardiovascular diseases, this resembles the numerical distribution of contents. So far, the portal does not contain medical data models, mapped with a MeSH descriptor of the tree branches C21 for disorders of environmental origin and C03 for parasitic diseases. Only one medical data model from the tree branch C22 for animal disease was found.

Another medical research repository is the National Institute of Health's (NIH) Common Data Element Repository, maintained by the U.S. National Library of Medicine. It currently offers 2175 case report forms in four export formats (ODM, two NIH/CDE schemata and SDC). The main research domain focus is on neurological disorders (NINDS) with 1409 elements, as well as on PhenX measures, a consented amount of 627 "standard measures of phenotypes and exposures for use in research" [21]. This is only a small subset of medical forms, compared to the amount of forms offered by our portal. Nevertheless, one must note, that the contents of the NIH CDE Repository consist mainly of previously consented CDEs. Considering the costly and time-consuming task, the creation of CDEs poses, involving the close cooperation of various research communities, the NIH CDE Repository contains already a substantial number of reusable forms. By presenting an image of the current state of documentation in various medical backgrounds, the Portal of Medical Data Models may contribute to a faster, more efficient and effective way to create CDEs, serving as the infrastructure used to identify and generate CDEs. In course of their doctoral thesis, six medical students are currently doing research on various disease entities. To identify CDEs used in myeloid leukemia, the portal already presents a solid and feasible foundation [22].

By adapting the table of contents, clinicians may be able to get a quick overview over the topics of their specialty. To approach the needs of users, we are constantly implementing further export possibilities. As the user survey showed, there is a great need for an export of medical forms to the OpenClinica metadata import format. The converter is currently being developed and will be available in the short term.

Furthermore we soon will offer an export to Apple's ResearchKit, in order to support the reuse of our contents in studies conducted by mobile devices via mobile applications. We are planning the implementation of an export "ODMtoResearchStack" as it represents the equivalent to Apple's ResearchKit for surveys conducted by Android users.

A problem with the current implementation in Ruby on Rails is the lack of scalability and easiness of maintenance. To face these problems the system is currently being re-developed in Java EE, which is planned to go online in the course of 2017.

Once the contents of the portal represent most of the currently used medical data models, the transnational creation of data standards and CDEs may be accelerated and will support the interoperability of clinical data, leading to harmonized documentation, improving cross study comparisons and meta-analyses.

Conclusion

Transparency, interoperability and huge amounts of data and metadata are crucial issues in medical research, approached by the Portal of Medical Data Models. This paper gives an overview of the new contents of the portal, such as the representation of an increasing amount of medical data models and items as well as new functions such as converters from ODM to FHIR or ODK. Additionally, the presentation of content is now structured according to the origin and specialty that the medical data models are related to. The user survey revealed the demand for further export features, which will be developed in the near future.

Acknowledgements

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG grant DU 352/11-1). We would like to thank all students who contributed to this project, especially Jan Kenneweg.

References

[1] MERIL - Mapping of the European Research Infrastucture Landscape, http://portal.meril.eu/converis-esf/publicweb/research_infrastructure/3574, accessed 26 October 2016.

[2] Deutsche Forschungsgemeinschaft (DFG), RIsources - The Research Infrastructure Portal, http://risources.dfg.de/detail/RI_00396_en.html, accessed 22 November 2016.

[3] Clinical Data Interchange Standards Consortium, Operational Data Model (ODM) - XML, https://www.cdisc.org/standards/foundational/odm, accessed 14 October 2016.

[4] V. Huser, C. Sastry, M. Breymaier, A. Idriss and J.J. Cimino, Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM), *Journal of Biomedical Informatics* **57** (2015) 88–99.

[5] U.S. National Library of Medicine, Unified Medical Language System (UMLS), https://www.nlm.nih.gov/research/umls/, accessed 21 October 2016.

[6] J. Varghese and M. Dugas, Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT, *Methods of Information in Medicine* **54** (2015) 83–92.

[7] M. Dugas, A. Meidt, P. Neuhaus, M. Storck and J. Varghese, ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository, *BMC Medical Research Methodology* **16** (2016) 65.

[8] J. Walker, E. Pan, D. Johnston, J. Adler-Milstein, D.W. Bates and B. Middleton, The value of health care information exchange and interoperability, *Health affairs* (2005) W5-10-W5-18.

[9] Federal Ministry of Education and Research, Medical Informatics Funding Scheme - Networking data – improving health care, http://www.gesundheitsforschung-bmbf.de/_media/Flyer_Medizininformatik_englisch_bar rierefrei.pdf, accessed 21 October 2016.

[10] U.S. National Library of Medicine, SNOMED CT, https://www.nlm.nih.gov/healthit/snomedct/, accessed 25 October 2016.

[11] M. Dugas, P. Neuhaus, A. Meidt, J. Doods, M. Storck, P. Bruland and J. Varghese, Portal of medical data models: information infrastructure for medical research and healthcare, *Database: The Journal of Biological Databases and Curation* 2016 (2016).

[12] Institute of Medicine (IOM), Sharing clinical trial data: Maximizing benefits, minimizing risk. *Washington, D.C: The National Academies Press* (2015).

[13] B. Breil, J. Kenneweg, F. Fritz, P. Bruland, J. Doods, B. Trinczek and M. Dugas, Multilingual Medical Data Models in ODM Format: A Novel Form-based Approach to Semantic Interoperability between Routine Healthcare and Clinical Research, *Applied Clinical Informatics* **3** (2012) 276–289.

[14] Universitäts- und Landesbibliothek Münster, miami ulb Münster: Medical-Data-Models.Org, https://miami.uni-muenster.de/Record/4f9faa6d-cbad-4c96-aade- 2306f76bb642, accessed 27 October 2016.

[15] FHIR v1.0.2, https://www.hl7.org/fhir/summary.html, accessed 25 November 2016.

[16] J. Doods, P. Neuhaus and M. Dugas, Converting ODM Metadata to FHIR Questionnaire Resources, *Studies in Health Technology and Informatics* **228** (2016) 456– 460.

[17] P. Bruland and M. Dugas, Transformations between CDISC ODM and openEHR Archetypes, *Studies in Health Technology and Informatics* (2014) 1225.

[18] J.M. Drazen, Data Sharing and the Journal, *N Engl J Med* **374** (2016) e24.

[19] U.S. National Institutes of Health - ClinicalTrials.gov, https://clinicaltrials.gov/, accessed 12 December 2016.

[20] European Medicines Agency - Clinical Trials Register, https://www.clinicaltrialsregister.eu/ctr-search/search, accessed 12 December 2016.

[21] PhenX Toolkit, https://www.phenxtoolkit.org/, accessed 30 November 2016.

[22] J. Varghese, C. Holz, P. Neuhaus, M. Bernardi, A. Boehm, A. Ganser, S. Gore, M. Heaney, A. Hochhaus, W.-K. Hofmann, U. Krug, C. Muller-Tidow, A. Smith,

[23] Weltermann, T. de Witte, R. Hehlmann and M. Dugas, Key Data Elements in Myeloid Leukemia, *Studies in Health Technology and Informatics* **228** (2016) 282–286.

Address for correspondence

Sophia Geßner
Email: Sophia.Gessner(at)uni-muenster.de