A terminology in General Practice / Family Medicine to represent non-clinical aspects for various usages: the Q-Codes

Marc JAMOULLE [a,1] , Julien GROSJEAN [b], Melissa RESNICK [c], Ashwin ITTOO [d], Arthur TREUHERZ [e], Robert VANDER STICHELE [f], Elena CARDILLO [g], Stéfan J. DARMONI [b], Frank S. SHAMENEK [h] and Marc VANMEERBEEK [a]

[a] *Department of General Practice, University of Liège, Belgium*
[b] *Département d'Information et d'Informatique Médicale, University of Rouen, France*
[c] *University of Texas, Health Science Center at Houston, TX USA*
[d] *HEC Management School, University of Liège, Belgium*
[e] *Department of Health Sciences Terminology, BIREME/PAHO/WHO, Sao Paulo, Brazil*
[f] *Heymans Institute of Pharmacology, University of Ghent, Belgium*
[g] *Institute of Informatics and Telematics, Rende, Italy*
[h] *Consultant, New York, NY USA*

Abstract. The hereby proposed terminology called "Q-Codes" can be defined as an extension of the International Classification of Primary Care (ICPC-2). It deals with non-clinical concepts that are relevant in General Practice/Family Medicine (GP/FM). This terminology is a good way to put an emphasis on underestimated topics such as Teaching, Patient issues or Ethics. It aims at indexing GP/FM documents such as congress abstracts and theses to get a more comprehensive view about the GP/FM domain. The 182 identified Q-Codes have been very precisely defined by a college of experts (physicians and terminologists) from twelve countries. The result is available on the Health Terminology/Ontology Portal (http://www.hetop.org/Q) and formatted in OWL-2 for further semantic considerations and will be used to index the 2016 WONCA World congress communications.

Keywords. Q-Codes, general practice, family medicine, ICPC, categorization, controlled vocabulary, qualitative analysis, abstracts

1. Introduction

The field of medicine is blessed with a rich array of terminologies that support structured documentation of clinical information, and storage and retrieval of research publications. Overarching resources are the International Classification of Diseases (ICD), the Medical Subject Headings (MeSH) and the SNOMED CT. For most of the specialized medical domains have been built proper nomenclatures and classifications (e.g., the Systematized Nomenclature of Pathology, SNOP for pathologists) [1].

General Practice / Family Medicine (GP/FM) is a peculiar domain characterized by a very broad scope and a large array of research methods, and encompassing both

[1] Corresponding author: Dr Marc Jamoulle, MD GP, Charleroi, Belgium. E-mail: marc.jamoulle@doct.ulg.ac.be

clinical and non-clinical issues. Clinical issues are those aspects pertaining to signs and symptoms, reasons for encounter, and processes and diagnoses covered by the International Classification of Primary Care (ICPC) [2].

The non-clinical content of GP/FM is not fully represented either in indexes of textbooks in GP/FM, or in specific classifications or terminologies. The existing representation suffers from a top-down approach which is not always reflecting the complexity of the discipline.

To tackle this lack of non-clinical content in vocabularies, we hereby propose a new terminology for pragmatic use in real-life situations which can be used to index specific content of GP/FM (communications, literature or event managerial content of the contact with patients). Thus this resource is complementary to the ICPC-2.

In this paper, we aimed to a) describe the methodology applied for the creation of a taxonomy, b) on this basis, to present a domain-specialized terminological resource called the "Q-Codes" to facilitate the indexing of GP/FM non-clinical content and c) show how this resource could support the information retrieval of specific bibliographic information.

By doing this, we aim to contribute defining the limits of GP/FM.

This is also a preliminary work to address the future necessity of the application of semantic web technology on the GP/FM domain [3]. This approach refers to an ontology building process [4].

2. Methods

The first step of the methodology had the aim to design the taxonomy by identifying relevant concepts in a compiled corpus that includes GP/FM texts. We have studied the concepts identified in hundreds of communications of GPs during congresses from a bottom-up approach. The relevant concepts belong to the fields that are focusing on GP/FM activities (e.g. teaching, ethics, or environmental hazard issues).

The second step was the development of a terminological resource for each category of the resulting taxonomy. This has been formalized by defining concepts, hierarchy and mappings relationships. Several methods and tools were used to perform this step: i) Cimino's standard set of desiderata was applied to build the terminology content and structure [5], ii) we highlighted each concept by relevant bibliographic citations as well as by linking them to BabelNet (http://babelnet.org), DBpedia (http://wiki.dbpedia.org) and to other reference terminologies, iii) we relied on the HeTOP multi-lingual and multi-terminology portal to fulfill each conceptual content and to manage the translations of terms and their definitions.

The final step was to evaluate and discuss the conceptual content of each created code of the taxonomy by involving several experts from all over the world (twenty four GPs and terminology experts from ten countries).

Selection of abstracts and corpus compilation

A total of 1,702 abstracts were selected from six sources in English or French: (i) Wonca Europe 2007 (n=998), (ii) Portuguese 18th national conference of family medicine, 2013 (n=128), (iii) Congrès Confédération des Généralistes Enseignants (CNGE) Clermont-Ferrand 2013 (n=205), (iv) Congrès CNGE Lille 2014 (n=289), (v) SwissFamilyDocs Zurich 2014 (n=45), and (vi) Belgian GP/FM research congress

Brussels 2014 (n=37). These six sources were selected due to the ease of accessibility of their abstracts.

Qualitative methods for development of the taxonomy

Data collection stemmed from analyzing abstract proceedings from the mentioned GP/FM conferences. These data were analyzed in a grounded theory approach. It involves construction of a hypothesis or discovery of concepts through data analysis [6].

Tools

HeTOP is the Health Terminology/Ontology Portal (http://www.hetop.org) which provides access to 60 main health terminologies in several different languages. Due to these multilingual terminologies, HeTOP is used for many purposes. It is very useful not only for translators, terminologists and ontologists, but also for physicians coding patient records and using services on demand, e.g. info buttons. Finally, and most important for this research, HeTOP assists in indexing resources on the Internet.

We relied on HeTOP for this study to: a) find similar concepts in other vocabularies; b) create the Q-Codes terminology; c) manage each concept (labels, synonyms, definitions); d) perform manual mappings and e) give access to the final Q-Codes terminology through a web site.

The Q-Codes terminology

The ambition was to create a terminology to represent the non-clinical activities of GP/FM, by extending the 17 chapters of the clinical classification ICPC with an 18th chapter, called Q-Codes ("Q" being a letter in the alphabet not yet used in ICPC).

Each term identified in the first step has been converted into a concept (a "Q-Code"). Each concept has received a definition explaining its conceptual value. The extension of the concept, i.e. its use in several other online databases, has been documented through a careful search in a set of online dictionaries and terminologies. For each Q-Code, a minimum of properties was fulfilled to define the concept: a) the Preferred Label, b) one or more definitions, c) synonyms and linguistic variants, d) a sample of pertinent articles to the understanding of the main subject and e) BabelNet and DBpedia unique Ids and f) relevant MeSH terms related to the Q-Codes were mapped. Those mappings can lead to ease querying bibliographic databases (e.g. PubMed) for each specific Q-Code but they are also an important way to evaluate and assess the quality of Q-Codes definitions and conceptual content.

With regards to implementation, the Q-Codes (concepts, properties and relations) were formalized and implemented in Web Ontology Language (OWL). Based on the Resource Description Framework (RDF) standard, OWL is a knowledge-representation language which is considered the *de facto* language for ontology implementation. This task was a rough OWL-2 export from HeTOP to Web Protégé (http://webprotege.stanford.edu/) without any description logic.

3. Results

The complete analysis of 1,702 French/English abstracts lead to the construction of a taxonomy composed by 182 terms. This taxonomy was enhanced to a terminological

level according to Cimino's desiderata [5] and thanks to a complete support of semantic web technologies (HeTOP, Web Protégé, etc.) The resulting terminology called the "Q-Codes" consists of 182 concepts divided among 8 domains.

Each Q-Code (concept) was contentiously defined and tagged with English terms and definitions. Bibliographic citations and external concept URIs have been added to ensure semantic extension and validity. Table 1 gives an overview of the main Q-Codes domains with included covered topics.

Table 1. Q-Codes domains overview

Q-Code domain	Label	Examples of covered topics
QC	Patient's category	age, gender issues, abuse
QD	Family doctor's issue	communication, clinical prevention, medico legal issues
QE	Medical ethics	bioethics, professional ethics, info ethics
QH	Planetary health	environmental health, biological hazards, nuclear hazards
QP	Patient issue	patient safety, patient centeredness, quality of health care
QR	Research	research methods, research tools, epidemiology of primary care
QS	Structure of practice	primary care setting, primary care provider, practice relationship
QT	Knowledge management	teaching, training, knowledge dissemination

Results of the final step are related to the translation of the labels and definitions of the Q-Codes by different general practitioners in their native languages which are: French, Spanish, Brazilian-Portuguese, Dutch, Turkish, Korean and Vietnamese. Two terminologists reviewed the translations for three of the languages. One terminologist validated the English translation, and the second terminologist validated the Spanish and Portuguese translations. The Q-Codes multi-lingual terminology is available on HeTOP at http://www.hetop.org/Q (authentication required as wicc/wiccdemo). Figure 1 is a screenshot of a Q-Code entry in HeTOP (QD44 quaternary prevention).

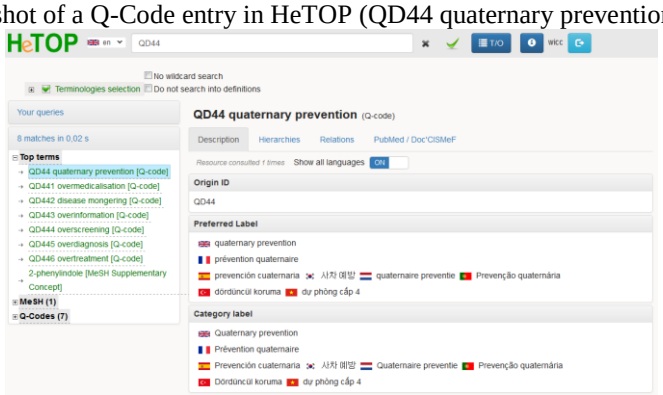

Figure 1. A Q-Code content on HeTOP

4. Discussion

To the best of our knowledge, this is the first attempt to expand the ICPC coding system with an extension for managerial issues, thus covering non-clinical content,

with the intent to improve performance in information storage and retrieval for research purposes in this broad, eclectic, and underserved domain of medicine.

Implications for Practice

We expect that the creation of this terminological resource for indexing abstracts and for facilitating Medline searches for general practitioners, researchers and students in medicine will reduce loss of knowledge in the domain of GP/FM. In addition, through better indexing of the grey literature (congress abstracts, master's and doctoral theses), we hope to enhance the accessibility of research results of general practitioners.

Implications for research

End-users are often not very well-versed in knowledge-representation formalisms, and it remains to be proven that our proposed terminology will help them in dealing with more complex systems, such as MeSH, to support their information storage and retrieval activities. Nevertheless, the Q-Codes base is aimed at several uses: a) online repository of knowledge specific to GP/FM in several languages; b) online PubMed linked bibliographic system easy to use for training in GP/FM; c) resource for online e-learning; d) resource for the analysis of content of congresses in GP/FM in joint usage with ICPC-2 and indexation of gray literature in GP/FM; e) automatic or semi-automatic congresses indexing system; f) linking GP/FM to the web of data and the linked data initiative. Further work could be conducted to enhance the Q-Codes formalism with ontology building as the current OWL-2 version is a rough export.

5. Conclusion

"Q-Codes" are a terminology of non-clinical subjects in GP/FM. This work is the result of a two-year cooperative project between participants from twelve countries and eight languages. The 182 concepts have been elaborated in a bottom-up approach, by retrieving the topics most frequently addressed by GPs when they met in congresses. This work is freely available online at http://www.hetop.org/Q. The data is also available in the OWL-2 language for future use in the semantic web. The product, user guide and e-learning are available at http://3cgp.docpatient.net.

References

[1] R. Cornet & N. de Keizer. Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making* **8** (2008) Suppl 1, S2.

[2] I. Okkes, M. Jamoulle, H. Lamberts & N. Bentzen. ICPC-2-E: the electronic version of ICPC2. Differences from the printed version and the consequences. *Family Practice* **17(2)** (2000), 101–107.

[3] M. Jamoulle, R.H. Vander Stichele, E. Cardillo, J. Roumier, J. Grosjean & S.J. Darmoni. Semantic Web and the Future of Health Care Data in Family Practice. *Merit Research Journal of Medicine and Medical Sciences* **3(12)** (2015) 586–594.

[4] M. Fernández-López & A. Gómez-Pérez. Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review* **17(02)** (2002), 129–156.

[5] J.J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine* **37(4-5)** (1998) 394–403.

[6] C. Faggiolani. Perceived Identity: applying Grounded Theory in Libraries. *JLIS.it* (2011)