Proceedings of the eHealth 2010 conference (paper 10579 - electronic proceedings).

Integrating consumer-oriented vocabularies with selected professional ones from the UMLS using Semantic Web Technologies

Elena Cardillo[1,3], Genaro Hernandez[2,3], and Olivier Bodenreider[3]

[1]Fondazione Bruno Kessler, Povo, Trento, Italy
cardillo@fbk.eu

[2]University of Maryland Baltimore County, Baltimore, Maryland, USA
genaroh1@umbc.edu

[3]National Library of Medicine, Bethesda, Maryland, USA
olivier@nlm.nih.gov

Abstract. During the past few years, many consumer-oriented vocabularies have been developed to reflect the multitude of ways consumers express health topics, in order to help lay users access health information and manage their personal healthcare data. To address problems such as interoperability, ambiguity, and heterogeneity of medical information, the lay expressions from consumer vocabularies need to be mapped to the specialized vocabulary used by professionals (i.e., mapped to terms from existing medical vocabularies). This paper presents an approach for creating an Integration Framework for the General Practice domain. Our work leverages the Italian consumer-oriented medical vocabulary and its mapping to the ICPC2 terminology. We exploit mappings to four other professional vocabularies available through the UMLS Metathesaurus. Semantic Web technologies provide a platform for the representation of medical terms and their inter-relations. This framework could facilitate the exchange of information between consumers and healthcare professionals in health information systems.

Keywords: Medical-Terminology Integration, UMLS, Semantic Web.

1. Introduction

During the past few years healthcare researchers have spent considerable effort on the development of applications devoted to healthcare consumers, such as Personal Health Records (PHR). Such applications make it possible for patients and their families to organize health information, gather medical records from doctors, and receive personalized reminders and alerts. The terminology used in these applications is generally different from that in electronic health records. Lay terminology reflects the different ways consumers and patients express health topics, and is embodied by consumer-oriented vocabularies, “collections of forms used in health-oriented communication for a particular task or need by a substantial percentage of consumers from a specific discourse group and the relationship of the forms to professional con-

cepts" [1]. For example, terminologies such as MeSH do not provide the lay term "Hearth attack" as a synonym for "Myocardial Infarction". Therefore existing medical terminologies and classification systems are generally poorly suited for consumer-oriented applications, because they mostly represent the physicians' perspective. Furthermore lay expression and topics need to be mapped to standard medical terms in order to enable effective health information integration and ex-change.

The objective of this work is to create an integration framework for the General Practice domain in order to bridge the gap between the terminology used in consumer-oriented health information systems and the one used by systems for healthcare professionals. Toward this end, our approach focuses on the integration of a consumer-oriented vocabulary, namely the Italian Consumer-oriented Medical Vocabulary (ICMV) [2], with several professional vocabularies available through the Unified Medical Language System (UMLS), using the International Classification of Primary Care 2Ed (ICPC2) as a pivot for the integration. We also take advantage of Semantic Web technologies [3] to represent the medical terms and their inter-relations.

The paper is organized as follows: Section 2 presents the background; Section 3 describes the resources under investigation; Section 4 presents our approach to integrating biomedical vocabularies through our framework; Section 5 reports some preliminary results; and finally Section 6 offers some conclusions.

2. Background

2.1 Consumer-oriented vocabularies

To help healthcare consumers fill the communication gap existing when accessing healthcare information on the web, many initiatives have been created to try to sort out the different ways they communicate within distinct discourse groups. The Consumer Health Vocabulary Initiative, by Q. Zeng *et al*. [1], aimed to develop the Open Access Collaborative Consumer Health Vocabulary for English. Soergel *et al*. [4] collected expressions used by lay people and health mediators, associated a Mediator Medical Vocabulary with the consumer-oriented one, and mapped them to a Professional Medical Vocabulary. Few applications have resulted from these academic efforts. Regarding the multilingual aspect, the Multilingual Glossary of Popular and Technical Medical Terms[1], in nine European languages, is a limited medical vocabulary (1,400 terms) for medicinal product package inserts accessible to consumers. The specific contribution of our work is the integration of language-dependent consumer-oriented vocabularies with language-independent professional medical terminologies.

2.2. Semantic Web technologies and medical information integration

The Semantic Web provides a common framework for the integration, sharing and reuse of data from multiple sources on the web. Its technologies include Uniform

[1] http://users.ugent.be/~rvdstich/eugloss/information.html

Resource Identifiers (URIs), the Extensible Markup Language (XML), the Resource Description Framework (RDF), the Web Ontology Language (OWL), and the SPARQL query language for RDF and OWL repositories. Here we focus in particular on RDF, the prescribed framework for representing resources in a common format. It describes information in the form of subject-predicate-object triples, enabling information to be represented in the form of a graph. RDF graphs can be stored in specialized databases called triple stores and can be queried using the SPARQL query language. RDF has several serialization formats including RDF/XML and N-Triples. This latter format (used in this study) is a line-based, plain text serialization format for RDF. The use of Semantic Web technologies in the biomedical domain has delivered promising results for the issue of information integration across heterogeneous resources [5]. Semantic Web technologies have been used not only for creating mashups of biomedical data [6], but also for terminology integration purposes. For example, [7] and [8] exploits RDF for comparing formal definitions in SNOMED CT and the NCI Thesaurus, and between LOINC and SNOMED CT.

3. Materials

In this section we briefly review the main characteristics of the terminologies we took under investigation in our work, namely ICVM, ICPC-2, ICD-10, UMLS, SNOMED CT, MeSH and LOINC.

ICMV is the Italian Consumer-oriented Medical Vocabulary, developed by Cardillo *et al*. [2], composed of ~3000 lay terms and expressions used by Italian speaking people to indicate "symptoms", "diseases", and "anatomical concepts", including synonyms, and clinical mapping to ICPC2. ICMV was mapped to ICPC2, considering 4 types of relations between pairs of consumer-oriented terms and ICPC2 rubrics (exact match - *Fever/Fever*; synonym - *Nosebleed/Epistaxis*; broader term - *Bronchitis*/*Acute Bronchitis*; and narrower term - *Absence of Voice*/*Voice Symptom*).

ICPC2 is the current revision (2000) of the international classification for primary care published by the World Organization of National Colleges, Academies (WONCA) to allow health care providers of the Family Practice domain to classify patient data and clinical activity [9]. In this study, we use the OWL version of ICPC2 (English), where the original biaxial structure is preserved through multiple inheritance. Disjoint statements have also been added (e.g., between siblings), as well as the Italian translation and lay mapping to lay terms from the ICMV [10].

ICD10 is the tenth revision of the International Classification of Diseases published by the World Health Organization (WHO). The goal of the system is to allow the systematic collection and statistical analysis of morbidity and mortality data from different countries around the world. In this work, we use the OWL version of ICD-10 developed by Cardillo *et al*. [10], in which ICD-10 chapters, sections and categories are represented as classes. Subsumption and disjunction relations are defined, and the concepts representing each ICD category are labeled by the ICD codes.

The **Unified Medical Language System** (**UMLS)** is compendium of a large number of biomedical national and international vocabularies developed at the National Library of Medicine [11], providing clinical and semantic mapping among them. Here knowledge is organized by concept, synonymous terms are clustered together to form

a concept and concepts are linked by means of various types of relationships. The UMLS integrates over 9 million names for some 2 million concepts from more than 100 families of biomedical vocabularies. All the terminologies under investigation in this study are integrated in version 2009AB of the UMLS Metathesaurus.

SNOMED Clinical Terms (SNOMED CT) is the reference terminology for clinical concepts developed by the College of American Pathologists and managed by the IHTSDO[2]. It provides clinical content and expressivity for clinical documentation and reporting. It also includes concepts, terms and relationships with the objective of precisely representing clinical information across the scope of health care. In this work, we use the July 2009 version of the international release of SNOMED CT.

Medical Subject Headings (MeSH) is a controlled vocabulary developed by the NLM for the indexing and retrieval of the biomedical literature, including MEDLINE. The main MeSH entities are the descriptors, organized in a hierarchical structure. Descriptors contain concepts, and each concept consists of a set of terms. In this work, we use version of MeSH 2009.

Logical Observation Identifier Names and Codes (LOINC) is a terminology system for laboratory tests and clinical observations, developed by the Regenstrief Institute and the LOINC Committee [12]. The purpose is to facilitate the exchange and pooling of clinical or laboratory results for clinical care. LOINC is composed of two main types of entities: lab test and observation "concepts" on the one hand, and "part" concepts used to support the description of the tests and observations on the other. In this work, we used the version of LOINC (June 2009) extracted from the UMLS Metathesaurus.

4. Approach

In this work, ICPC2 serves as a pivot between our consumer-oriented vocabulary, ICMV, and other professional vocabularies integrated in the UMLS. This is possible because ICMV was mapped to ICPC2 and ICPC2 is integrated in the UMLS, along with many other professional vocabularies. For this reason, we use ICPC2 as an entry point into the UMLS in order to find mappings to concepts from SNOMED CT, MeSH, LOINC and ICD10, also integrated in the UMLS. As shown in Figure 1, the major steps of our approach can be described as follows: 1) Convert into RDF all the terminological resources under investigation; 2) Enrich each terminological resource with UMLS attributes; 3) Load these resources into a Triple Store; and 4) Explore relations among terms, through SPARQL queries.

Converting. Two resources, SNOMED CT and MeSH, had already been converted to an RDF representation for other projects at NLM. In the resulting triples, the subject is most often a SNOMED CT concept or a MeSH descriptor. The predicates correspond to concept properties, including type (concept or relationship), preferred name (label), and relations to other concepts (e.g., *subClassOf*). The object of these triples is either a literal corresponding to a property or a node representing another concept. Two other resources, ICPC2 and ICD10, had already been converted to an OWL representation at an earlier stage of this project [10].

2 http://www.ihtsdo.org/

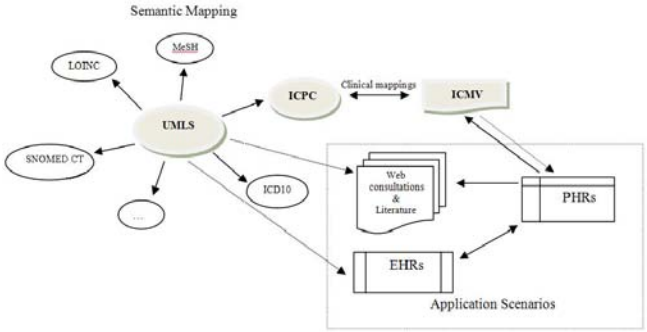

Figure 1. Integrating ICMV-ICPC2 with UMLS Methathesaurus

OWL resources can be serialized in RDF and are therefore directly compatible with other RDF resources. However, unlike for SNOMED CT and MeSH, the subjects of triples from an OWL representation can be blank (anonymous) nodes used for the representation of restrictions of the classes, in particular related to disjunction and quantifiers (*SomeValuesFrom, AllValuesFrom*). Finally, we created RDF triples for LOINC from data in the UMLS Metathesaurus. In all cases, conversion to RDF or OWL has been carried out automatically through simple scripts.

Enriching. In order to facilitate term comparisons among vocabularies, we enriched each terminological resource with UMLS attributes. These attributes include concept unique identifier (CUI), source abbreviation (SAB), lexical unique identifier (LUI) and term type (TTY). Of particular importance are the lexical unique identifiers, which provide a compact representation of normalized terms (e.g., singular and plural forms of the same term share the same LUI). We extracted this information from the corresponding fields of the UMLS table MRCONSO.RRF and automatically created triples as described previously.

Loading. In this phase all the N-triples were loaded into a triple store. We used the Virtuoso RDF store v6.1.1[3], an open source triple store, which can be queried using the SPARQL query language[4]

Querying. We created template queries, which we populated with values of interest before sending them to the Virtuoso SPARQL engine. In practice, we used a Java program to automate the submission of batch queries to Virtuoso and collect the results. We executed the following types of queries for each concept in ICPC2.

- Find concepts in SNOMED CT, MeSH, ICD-10 and LOINC corresponding to a particular ICPC2 concept, using the UMLS CUI as a bridge (Table 1).
- Find the preferred terms and synonyms in SNOMED CT, MeSH, ICD-10 and LOINC corresponding to a particular ICPC2 concept, using the UMLS CUI as a bridge.
- Find the lexical variants (LUI) in SNOMED CT, MeSH, ICD-10 and LOINC corresponding to a particular ICPC2 term (LUI), using the UMLS CUI as a bridge

[3] http://virtuoso.openlinksw.com
[4] http://www.w3.org/TR/rdf-sparql-query/

Table 1. SPARQL query for mapping the ICPC2 concept Headache to the other sources

```
SPARQL
PREFIX ICPC2E: <http://dkm.fbk.eu#ICPC2E:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX UMLS_MT: <http://nlm.nih.gov#UMLS_MT:>
SELECT ?icpc2_id  ?label ?cui  ?code
from <http://nlm.nih.gov/ICPC2E_to_UMLS_Enrichment>
from <http://dkm.fbk.eu/ICPC2E>
from <http://nlm.nih.gov/SNOMED CT_to_UMLS_Enrichment>
from <http://nlm.nih.gov/LOINC_to_UMLS_Enrichment>
from <http://nlm.nih.gov/ICD10_to_UMLS_Enrichment>
from <http://nlm.nih.gov/MeSH_Enrichment>
WHERE {
?icpc2_id rdfs:label ?label .
?icpc2_id UMLS_MT:hasCUI ?cui .
?code UMLS_MT:hasCUI ?cui .
filter(?icpc2_id = ICPC2E:N01)};

RESULTS:
ICPC2 Concept | label | UMLS CUI | Concepts in other sources
ICPC2E:N01 | Headache | UMLS_CUI:C0018681 | SNOMED CT:25064002
ICPC2E:N01 | Headache | UMLS_CUI:C0018681 | LOINC:LP74908-2
ICPC2E:N01 | Headache | UMLS_CUI:C0018681 | ICD10:R51
ICPC2E:N01 | Headache | UMLS_CUI:C0018681 | MeSH:D006261
```

5. Results

A total of 66,769,781 unique triples were created, and the corresponding graphs were loaded into Virtuoso. More specifically, 97,457 triples derive from ICD10; 18,650 from ICPC2E; 2M from LOINC; 1.8M from SNOMED CT; 16.6M from MeSH; and some 50M from UMLS.

Overall mapping. Results show that 77% of the ICPC2 concepts (587 of 760 concepts) are integrated in the UMLS. Of these 587 ICPC2 concepts integrated in UMLS, 251 are specific to ICPC2 and 336 are common to other terminologies. The 587 ICPC2 concepts mapped to 1773 UMLS CUIs and 1420 UMLS LUIs.

Mapping to other terminologies. We found a total of 1189 mappings, distributed as follows: 257 mappings to ICD10, 663 mappings to SNOMEDCT, 201 mappings to MeSH, and 68 mappings to LOINC. Detailed results are shown in Table 2.

Mapping to several terminologies. Among the 336 ICPC2 concepts mapped to the other resources, 90 map to two terminologies (e.g. *A71 - Measles*), 130 map to 3 terminologies (e.g. *A73 – Malaria*), 33 map to all four terminologies (e.g. *A03 - Fever*) and 83 map to one terminology (e.g. *A18 - Concern about Appearance* maps only to SNOMED CT).

Multiple mappings. We observed a large number of multiple mappings between ICPC2 and other terminologies, i.e., when a concept from ICPC2 maps to several concepts in another terminology. In particular, 40% of the ICPC2 concepts map at least to three SNOMED CT concepts and two ICD10 concepts. The greatest number of mappings for an ICPC2 concept is for “B72”, Hodgkin's disease/lymphoma (12 SNOMED CT concepts, 2 ICD10 concepts, 1 MeSH descriptor and 1 LOINC concept). Multiple mappings come in part from the fact that multiple concepts from a given source (SNOMED CT, MeSH or ICD10) are collapsed in the same UMLS concept, despite the fact that they are not considered synonyms in the original source.

Additional synonyms. The list of synonyms for ICPC2 concepts can be enriched through the mappings. We found 906 synonyms for the 587 ICPC2 concepts mapped to the other terminologies. The ICPC2 concept with the greatest number of synonyms is *Incontinence Urine*, code “U04” (14 synonyms). The number of additional synonyms is compounded by multiple mappings. For example, through the UMLS, the

Table 2. Mapping of 587 ICPC2 concepts to other resources

Mapped Entities	SNOMED CT	MESH	ICD10	LOINC	UMLS
Unique Resource IDs	663	201	257	68	(1773)
UMLS CUIs	321	201	189	54	1773
UMLS LUIs	1197	363	156	62	1420
Preferred Terms	703	149	257	68	--
Synonyms	824	0	0	0	--
ICPC2 concepts	321	201	189	54	587

ICPC2 concept “Malaria” maps to 4 distinct SNOMED CT concepts: “Malaria”, “Disease due to Plasmodiidae”, “Malaria, unspecified”, and “Malaria fever”. Finally, 739 additional lexical variants were found. Most of these lexical variants were contributed by SNOMEDCT.

Term usage. Of the 587 ICPC2 terms mapped to the other resources, 311 are also used as preferred term in the other terminologies. Most of the time these are concepts related to the “social problems”, “skin”, “musculoskeletal”, “female genital”, “general” symptoms and diseases.

6. Conclusions

In this work we presented a method for creating an integration framework for consumer-oriented and professional medical terminologies, leveraging existing mapping from the UMLS and taking advantage of Semantic Web technologies, namely RDF. Using an RDF triple store for integration purposes, we were able to perform this study through simple SPARQL queries rather than *ad hoc* programming, reaching our goal of mapping ICPC2 to other medical terminologies. OWL could have been used as well, but we would not have been able to take full advantage of its expressiveness, due to the underspecification of most source vocabularies.

Such an integration framework could enable consumer health information systems to link medical information from different sources, such as Electronic Health Records (EHRs); bibliographic databases; decision support systems; and other applications. This would help consumers and laypersons in different scenarios: 1) searching healthcare information (e.g., it would facilitate automated mapping of consumer-entered queries to technical terms – searching a bibliographic database indexed with MeSH would produce better search results if the query term used is mapped to MeSH); 2) translating and interpreting clinical notes or test results, which frequently contain jargon (e.g., mappings between a medical vocabulary, such as LOINC or SNOMED CT used in EHRs to a consumer-oriented vocabulary could be useful in providing consumer-understandable names to help patients interpret these documents); 3) describing their clinical history and their complaints (e.g., in Online medical consultations, patients entering consumer expressions, such as "sudden hair loss", could receive appropriate help from health professionals after translation of their query into the corresponding technical concept).

A future perspective for this work is to analyze the hierarchical aspect of this integration framework, considering the hierarchical relations from the UMLS, in order to

find inconsistencies in terms of classification of symptoms and diseases among the terminologies; and to integrate the mappings found with the ICMV.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and by the TreC Project, eHealth unit of the Fondazione Bruno Kessler (FBK), Italy.

References

1. Zeng Q., Tse, T.: Exploring and Developing Consumer Health Vocabularies. J. of Am. Med. Inf. Assoc., 13, 24--29 (2006)
2. Cardillo, E., Serafini, L., Tamilin, A.: A Hybrid Methodology for Consumer-oriented Healthcare Knowledge Acquisition. In Knowledge Representation for Health-care: Data, Processes and Guidelines, LNCS(LNAI), vol. 5943, pp. 38--49. Springer, Heidelberg (2010)
3. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies, Chapman & Hall/CRC (2009)
4. Soergel, D., Tse, T., Slaughter, L.: Helping Healthcare Consumers Understand: An ``Interpretative Layer'' for Finding and Making Sense of Medical Information. In: the International Medical Informatics Association's Conference,IMIA2004, pp. 931--935 (2004)
5. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., et al.: Advancing translational research with the Semantic Web. BMC Bioinformatics; 8 Suppl 3:S2 (2007)
6. Cheung, K.H., Kashyap, V., Luciano, J.S., Chen, H., Wang, Y., Stephens, S. J., Biomed Inform. Oct; 41(5):683-6 (2008)
7. Bodenreider, O.: Comparing SNOMED CT and the NCI Thesaurus through Semantic Web Technologies. In the 3rd International Conference on Knowledge Representation in Medicine (KR-MED2008) R. Cornet, K.A. Spackman (Eds) (2008)
8. Bodenreider, O.: Issues in Mapping LOINC Laboratory Tests to SNOMED CT. In AMIA Annual Symposium, AMIA2008, pp. 51--55 (2008)
9. Okkes, I.M., Jamoulle, M., Lamberts, H., Bentzen, N.: ICPC-2-E: the electronic version of ICPC-2. Differences from the printed version and the consequences. Family Practice, 17, 101--107 (2000)
10. Cardillo, E., Eccher, C., Tamilin, A., Serafini, L.: Logical Analysis of Mappings between Medical Classification Systems. In: the 13th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, AIMSA2008, pp. 311--321. Springer Berlin (2008)
11. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, vol. 32 (2004)
12. McDonald, C. J., Huff, S.M., Suico, J. G., Hill, G.J., Leavelle, D., Aller, R., Forrey, A., Mercer, K., De Moor, G., Hook, J., , W., Case, J., Maloney, P.: LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. In Clinical Chemistry, 49: 624--633 (2003)