The original publication is available http://www.springerlink.com/.
Stephan Kiefer, Jochen Rauch, Riccardo Albertoni, Marco Attene, Franca Giannini, Simone Marini, Luc Schneider, Carlos Mesquita, Xin Xing
An Ontology-Driven Search Module for Accessing Chronic Pathology Literature
OTM 2011: On the Move to Meaningful Internet Systems: OTM 2011 Workshops pp 382-391

Volume 7046 of the book series Lecture Notes in Computer Science (LNCS)

Meersman R., Dillon T., Herrero P. (eds)
2011,
ISBN: 978-3-642-25125-2
Doi: 10.1007/978-3-642-25126-9_50

An ontology-driven search module for accessing chronic pathology literature

Stephan Kiefer[1], Jochen Rauch[1], Riccardo Albertoni[2], Marco Attene[2], Franca Giannini[2], Simone Marini[2], Luc Schneider[3], Carlos Mesquita[4], Xin Xing[5]

[1]Home Care / Telemedicine- Fraunhofer Institute for Biomedical Engineering, St.Ingbert Germany
{stephan.kiefer,jochen.rauch}@ibmt.fraunhofer.de
[2]Institute for Applied Mathematics and Information Technology – C.N.R. Genova, Italy
{riccardo.albertoni,marco.attene,franca.giannini,simone.marini}@ge.imati.cnr.it
[3]Institute for Formal Ontology and Medical Information Science, Saarland University, P.O. Box151150, Saarbrücken, 66041, Germany
luc.schneider@ifomis.uni-saarland.de
[4]Link Consulting, Lisbon, 1000-138, Portugal,
carlos.mesquita@link.pt
[5]TZI - Center for Computing and Communication Technologies, Bremen, 28359, Germany
xing@tzi.de,

Abstract. This paper presents an advanced search module for bibliography retrieval developed within the CHRONIOUS European IP project. The developed search module is specifically targeted to clinicians and healthcare practitioners searching for documents related to Chronic Obstructive Pulmonary Disease (COPD) and Chronic Kidney Disease (CKD). To this aim, the presented tool exploits two pathology-specific ontologies that allow focused document indexing and retrieval. Besides the search module, an enrichment tool is provided to maintain and to keep up-to date such as ontologies. In addition link with the terms of the MeSH (Medical Subject Heading) thesaurus has been provided to guarantee the coverage with the general certified medical terms and multilingual capabilities.

Keywords: Ontology driven search, Health care literature, Chronic diseases.

1 Introduction

Chronic diseases are mostly characterized by complex causality, multiple risk factors, long latency periods, a prolonged course of illness and functional impairment or disability. Most chronic diseases do not resolve spontaneously, and are generally not cured completely. Chronic diseases may get worse, lead to death, be cured, remain dormant or require continual monitoring. It is then clear that reducing the severity of both the symptoms and the impact would mean significant benefit for both the individual and the society. This is possible in many conditions. Many physicians are very optimistic about the benefits that a proper disease management can bring to provide a

full and active life and recognize that the long-term health outlook for chronic disease has improved in the past decade, and most of them credit the improvement to better management and monitoring.

To support the management of persons at risk or with chronic health conditions, the project CHRONIOUS (EU Contract N. FP7-ICT-2007–1–216461) is developing an open framework for their remote monitoring and treatment. Chronic obstructive pulmonary disease (COPD) and Chronic kidney disease (CKD) are two case studies to set up a test bed in the project, but the CHRONIOUS architecture is suitable for any chronic disease.

Within this project an intelligent literature Search Module has been developed. The CHRONIOUS Search Module enables healthcare professionals to access well focused and up to date healthcare information in the form of scientific literature, guidelines, hospital-specific documentation and grey literature.

The CHRONIOUS search mechanism overcome Google Scholar, PubMed and GoPubMed by combining the MeSH potentiality with domain ontologies as topic-neutral representations of the domain of entities (objects, processes, qualities, dispositions, functions, etc.) related to the COPD and CKD on ontologies. Indeed, while MeSH encodes terminological knowledge, the ontologies encode expert domain knowledge. In this respect, MeSH and domain ontologies do not compete, but fulfill complementary tasks and the domain ontologies provide additional structure and depth as far as knowledge items are concerned; this additional representational power can be harnessed to improve annotation of and hence search over medical literature.

The Search Module combines two domain ontologies (one for each chronic disease treated) built on top of the Middle Layer Ontology for Clinical Care (MLOCC) [1] with the linguistic capabilities provided by the Medical Subject Heading thesaurus (MeSH) [2].

The use of the ontology for Clinical Care (MLOCC) guarantees the exploitation of well-funded and formalized medical concepts and the integration with other medical domain ontologies (e.g., ACGT Master Ontology [3], Open Biological and Biomedical Ontologies (OBO) foundry [4, 5]). Based on medical guidelines, the two domain-specific ontologies have been defined and properly validated by clinical experts coordinated by an international medical board. Within CHRONIOUS, ontologies and MeSH provide complementary benefits: for each specific disease, indeed, an ontology provides a fine-grained knowledge that MeSH is not supposed to reach; on the other hand, MeSH provides well-established generic medical terminology, multilingual representations, synonyms, narrower/broader/related concepts which can be exploited by the clinicians for the composition of expressive queries in a simple and effective way. The combination of the domain specific ontologies and the MeSH thesaurus allows CHRONIOUS to provide more focused results when compared to traditional search engines for the medical literature. Besides the two domain-specific ontologies, the literature Search Module consists of components for processing and indexing the scientific literature, as well as the components that allow the user to formulate appropriate queries. Moreover, to deal with the problem of medical knowledge evolution, the developed system provides tools for supporting the experts in upgrading the pro-

vided ontologies to new emerging concepts in the concerned documentation. This is even more important when considering remote patient monitoring, with wearable sensors, as in CHRONIOUS, since the availability of new signals and information is expected to introduce new practices and new insight for patient care which turns in more evolving concepts.

2 Architecture

The architecture of the CHRONIOUS Search Module, depicted in Fig. 1, is based on different sub-modules including:

Upload Tools to load documents related to CKD and COPD in the CHONIOUS System for their indexing. It provides two different capabilities: one supporting the clinicians to upload their own documents, the second offering capabilities to automatically treat resources directly coming from publishers upon specific agreements;

Transformation Module converts the imported documents, which might be in different formats, to normalized plain text format;

Natural Language Processing (NLP) Tool is based on the GATE Framework [6], allows to specify processing pipelines (e.g. consisting of Sentence Splitter, Tokeniser, Part-of-speech Tagger and Morphological Analyser) as well as filter parameters (e.g. for word kinds and categories). The NLP module is also exploited by the Ontology Enrichment tool for the extraction of candidate concepts from the documents;

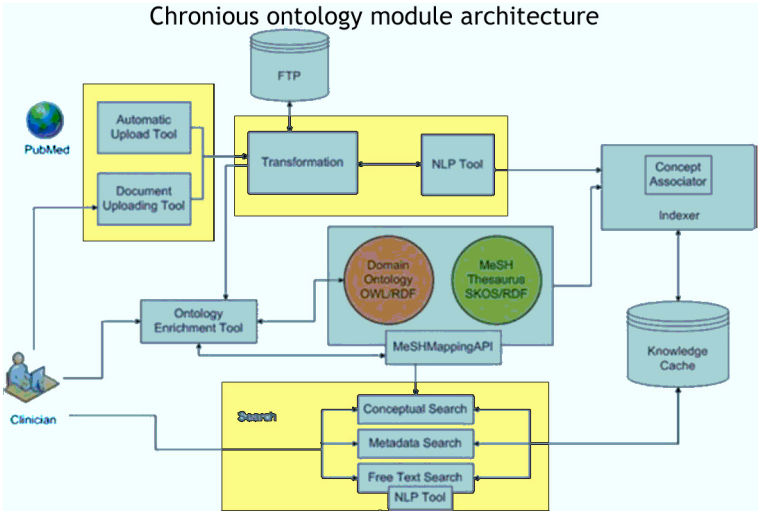


Fig. 1. The CHRONIOUS Search Module components.

Knowledge Cache maintains a graph database to store the generated indices and connections;

Search Module offers search methods as well as several tools that facilitate the user's interaction with the system;

Ontology/Thesaurus module incorporates the domain knowledge. It includes the CHRONIOUS Ontologies, MeSH and its translations in different languages, as well as the mapping between MeSH and Ontologies;

Ontology Enrichment module is responsible for the maintenance of the ontology in a semi-automatic way. It uses NLP tool to pre-process the imported documents and to find word groups and linguistic patterns that are related to the domain.

In the following paragraphs, the most important modules are detailed.

2.1 Natural Language Processing tool

Within the CHRONIOUS Search Module, the NLP tool is applied to the text documents provided by the transformation component to automatically extract structured information. The specific annotations provided by the NLP tool are used by the Document Indexer and Concept Associator as well as by the Ontology Enrichment Tool for proposing new concepts. The NLP Tool runs on a server and can be accessed through a web service. Transformation tool provides methods for receiving documents in different formats and responds with processed data. These operations also allow defining filter parameter (for annotation types and features) in order to focus on the actual required data and to reduce the number of delivered annotations.

For implementing the NLP algorithms the GATE framework [6] has been used that allows developing and deploying software components that process human language. The NLP Tool extracts the headwords of a text using different modules: the *Tokenizer* splits the sentences of the text obtained by the *Sentence Splitter* into very simple tokens such as numbers, punctuation and words of different types; the Part-of-speech (POS) Tagger produces part-of-speech tags (e.g. noun, proper noun, preposition) as an annotation on each word or symbol. Finally the *Morphological analyser* identifies the lemma (headword) and affix (e.g. “runs”, “ran” and “running” has all “run” as lemma) by considering one token and its part-of-speech tag, one at a time.

In addition to the above standard processing sequence, in order to satisfy the requirements of the Ontology Enrichment Tool and the Indexer, in the NLP Tool some additional resources of the GATE framework have been integrated:

- OntoRoot Gazetteer [7]: A GATE plugin that produces ontology-aware annotations for extracted terms (ontology matching using names and labels of the ontology concepts).
- Shallow Parser: analyzes the sentences to identify word groups by linguistic patterns (e.g. “chronic disease”, “lung function”). For the prototype implementation, which is focused on the English language, we extracted and utilized a JAPE transducer, executes a linguistic pre-processing in terms of ontology learning, from the Text2Onto project [8].
- RegEx-Pattern Matcher: matches the lemma of a token with word patterns defined as regular expressions. It has been implemented as JAPE resource, which expects an input list with entries coded as regular expressions.

- Dictionary Matcher: matches the lemma of a token to a (common) dictionary. In our implementation it uses the WordNet dictionary [9] and Java API for searching.
- Thesaurus Matcher: matches the lemma of a token to a (domain) thesaurus. For the target medical domain we have implemented a JAPE resource, which uses the MeSH-Mapping API to access MeSH terms and the mapped ontology concepts.

2.2 Search

To better support users various search methods have been provided together with tools that facilitate the interaction with the system and with functionalities for the ontology/thesaurus browsing and the query building. Search is then possible as:

- Metadata Search, which is a keyword-based search looking into the metadata information associated with a document and indexed in the Knowledge Cache.
- Conceptual Search, which allows the user to submit queries at a higher level of abstraction as compared to keyword based search by retrieving content that does not contain keywords from the query. The conceptual search consist of three components the user can exploit to build and refine their queries:
  - An Ontology Browser panel displaying a clinical view, which is a simplified view of the ontology structure enabling the user to exploit relations with connected concepts.
  - A Concept Finder tool that provides on-type suggestions of concepts included in the Ontology and/or MeSH and that exploits the multilingual capabilities of MeSH.
  - A Query Building tool where queries are build combining Ontology and MeSH concepts through boolean expressions.
- Free-text search: this is also a conceptual-based search allowing the user to submit queries in natural language by exploiting the NLP capabilities. The NLP Tool processes the sentence, keeps the valuable information and the search investigates the associations stored in the Knowledge Cache. This search exploits the multilingual capability of MeSH allowing the user to use concept terms in their languages (currently in Italian, Spanish and Portuguese).

To avoid copyright infringement, while guaranteeing a sufficient coverage of the search capabilities, full papers are used only for indexing purpose. The search is performed on the metadata and indexed concepts, and displays the results' metadata which also include for published document the related DOI. Therefore access to the document is performed by dereferencing the HTTP URI of the paper's DOI respecting the paper access constraints and users' subscriptions. The solution has been verified to be acceptable for getting access to the bibliography provided by the main publishers.

2.3 Ontology/thesaurus module

This module provides the access to all the domain and linguistic resources used in the annotation and retrieval of medical reference documents enhancing the search functionalities therewith.

Through an ad hoc developed API, the Ontology and thesaurus module provides access to: COPD [13], CKD [14] and MLOCC (middle-layer) [15] ontologies; MeSH (English) [2] thesaurus version 2010; MeSH multilingual lexicographic representations (currently the Italian, the Spanish and the Portuguese versions have been provided); mapping between MeSH concepts and correspondent Ontologies classes.

The CHRONIOUS ontologies encode expert knowledge relevant to the envisaged domain (COPD and CKD) such as to provide a rich science-driven repository of concepts for annotating and searching medical literature in the scope of the project, as well as a wealth of relations to refine domain-specific literature search. The COPD and CKD ontologies are built on top of a Middle Layer Ontology for Clinical Care (MLOCC) whose purpose is to provide a link between the leaves of Basic Formal Ontology [12], a foundational ontology used in the Open Biological and Biomedical Ontologies (OBO) Foundry [5], and the top-nodes of the domain ontologies. MLOCC represents, as it were, the common core of biomedical and clinical knowledge shared by the COPD and CKD Ontologies. Furthermore, we have expanded the ontology providing reference for selected terms from MeSH. MeSH is a certified and very rich thesaurus, that covers a large part of the concepts pertaining to the medical domains. It is already used to index medical articles, but

- it does not deepen specifically the two diseases considered in the project, so it lacks of the required specificity to support in complex search pertaining the considered diseases;
- it is not compliant with the newest web standard suggested by the W3C, so it cannot be easily combined with the OWL [11] domain ontologies.

In the project, the former limitations are overcome by ontologies about COPD and CKD. However, searching for scientific papers may require to move from concepts strictly related to the aforementioned diseases to concepts less specific and vice versa, thus some connections between the two diseases-specific ontologies and the MeSH concepts are required. For this reason, a MeSH-Ontologies mapping linking the concepts provided by MeSH to the ontologies classes has been provided.

To ensure that the thesauri and COPD/CKD ontologies can be fruitfully connected, we have SKOSified MeSH and their Multilingual versions: we have mapped\encoded the MeSH content into Simple Knowledge Organization System (SKOS) model following the rules discussed in [16] and serialized the SKOS in RDF. SKOSyfied MeSH thesaurus, its translations and its mapping to concept described in the Diseases Ontologies are accessible by API developed in CHONIOUS which relies on the JENA framework [10]. Ontologies are accessible through standard OWL 2 API [17].

2.4 Ontology enrichment tool

As mentioned, since in the project we aim at supporting care givers in remote monitoring and managing chronic patients, for which knowledge and literature are still at their infancy, it is crucial to have tools for helping the upgrade of the ontology when new concepts arise. Therefore, we decided to exploit the possibility of deriving new concepts from pertinent literature in a semi-automatic way. Thus, the so-called ontology enrichment tool has been developed allowing the identification of gaps in the ontologies with respect to newly inserted documents to support the ontology curator. In particular, the tool combines ontology learning techniques and submission functionalities. Methods for ontology learning are based on linguistic, statistical and machine learning approaches. Thus the Ontology Enrichment Tool uses the NLP Tool described in section 2.1 for a linguistic processing of the transformed document data and annotate the text with several features. Based on these annotations candidate concepts are extracted in a second step. After the pre-linguistic processing the extracted candidate concepts are rated concerning their relevance by points based on several criteria:

- Corpus relevance: the relevance of each candidate concept is determined by computing its average Term Frequency Inverted Document Frequency (TF.IDF) value with respect to the whole document corpus.
- Domain relevance: matchings with a common dictionary, a domain thesaurus and with regular expression patterns is used to determinate the domain relevance of a candidate concept. The assignment of synonyms results in a higher relevance. This means that the candidate concept has been identified as a synonym of an existing concept (by dictionary synonyms or ontology mapped MeSH thesaurus terms).
- Subclass-of relations: the relevance of the candidate concept is raised, if it is possible to assign a subclass-of relation to it. Such an assignment can be determined by the extraction of vertical relations, linguistic patterns or dictionary hypernyms.
- Concept co-occurrences: the co-occurrences of the candidate concept with concepts in sentences can be considered as an indicator of a possible relation extraction. The average distance in the text between the candidate concept and a concept within the corpus is calculated as a benchmark.

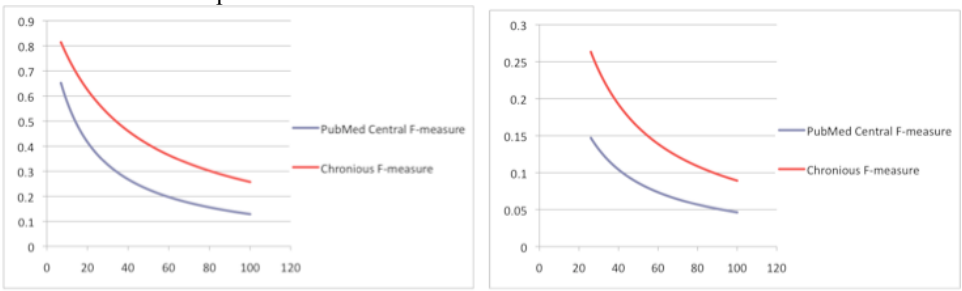
Candidate extensions need to be validated by clinicians which are supported by the submission facility of the Ontology Enrichment Tool, i.e. a GUI that allows the user review the extracted data. Furthermore the user has the possibility to administrate the candidate concepts in terms of a workflow state (“new”, “to validate”, “postponed”, “accepted” or “rejected”). In this way the submission function of the Enrichment Tool provides suggestions for new concepts (possibly with subclass-of relation assignment) or new labels of existing concepts, but modifications of the ontology must still be done by the ontology expert with an external tool (e.g. Protégé).

3 Evaluation

Building domain-specific ontology requires large involvement of experts and close cooperation with technical partners. Both the CHRONIOUS ontologies developed by computer engineers and the mapping relations between MeSH concepts and ontology classes have been evaluated by medical experts. Questionnaires checking the agreement among medical experts on the correctness of the developed ontologies and of the MeSH mapping relations have been created by test engineers and filled in by COPD and CKD clinicians. The evaluation result shows that most of the ontology classes have been defined correctly, a few ontology classes needed to be modified (eight of 964 COPD classes and nine of 972 CKD classes), and together 150 new classes needed to be added (80 new classes for COPD ontology and 70 for CKD ontology). That is less than 7.8% of the total number of defined ontology classes. Furthermore, 42 of 120 mapping relations between MeSH concepts and COPD ontology classes, as well as 43 of 138 mapping relations between MeSH concepts and CKD ontology classes have been modified. The CHRONIOUS search mechanism overcome Google Scholar, PubMed and GoPubMed by combining the MeSH potentiality with the two specific domain ontologies. Indeed Google Scholar supports the search of documents by providing a simple search by keyword mechanism. Such a mechanism is based on a pure syntactical search of the documents, without taking into account any semantic interpretation of the proposed query. As a consequence, two queries having different syntax but equivalent semantics yield different search results. PubMed improves Google Scholar by exploiting the synonymous provided by MeSH, where the thesaurus is used to produce a semantic interpretation of the query keyword mechanism. The GoPubMed search engine, improves PubMed by considering the concepts of Gene ontology (GO) during the query refinement step. In this case, the clinicians are facilitated during the query composition by exploiting both the terms of MeSH and the terms carried by the Gene ontology. Neither PubMed nor GoPubMed are supported by the specific knowledge related to the specific domain of the COPD and CKD ontology that provide additional structure and depth as far as knowledge items are concerned; this additional representational power is exploited to improve annotation of and hence search over medical literature.

In order to complete the previous qualitative comparison of the CHRONIOUS Conceptual Search, a quantitative evaluation of the performance of the system, i.e., the exactness and completeness of its search result, has been evaluated. Exactness (or *Precision*) measures the number of correctly found documents – documents which are relevant to the search query – as a percentage of the number of documents found, whereas completeness (or *Recall*) measures the number of correctly found documents as a percentage of the total number of existing relevant documents. Usually, Precision and Recall scores are not discussed in isolation. Instead, both are combined into a single measure – the F-measure [18], which is the weighted harmonic mean value of Precision and Recall. To quantitatively evaluate the search performance of the CHRONIOUS Conceptual Search, seven commonly used COPD terms and five CKD terms have been defined by medical experts as search queries. The F-measure of search results of each search query with the CHRONIOUS Conceptual Search has

been calculated and compared with one of the most popular text-based search engines – PubMed Central. To have meaningful comparison of the two systems we executed the test with same document database. Thus, documents meeting with certain limitations (e.g., published in a specific year, contain specific terms in their text body, open access, etc.) have been downloaded from PubMed Central and uploaded into the CHRONIOUS document repository. The evaluation result shows that the overall search performance of CHRONIOUS Conceptual Search with defined search queries is in most cases better than PubMed Central. Fig 2 shows the F-measure comparison results with two search terms: "Inhaler Device" and "PostBronchodilator Spirometry".[1] Furthermore, the overall end user satisfaction of the CHRONIOUS Search Module has been investigated with the medical experts by a questionnaire. The evaluation has shown that 90% of them generally accept the Conceptual Search development and 60% of them gave positive evaluation ("Somewhat satisfied") to the Free-text Search option.



(a) (b)

Fig. 2. F-measure comparison between CHRONIOUS Conceptual Search and PubMed Central with search query (a) "Inhaler Device" and (b) "PostBronchodilator Spirometry". The X-axis presents the estimated total number of existing relevant documents in the repository for the corresponding search query (increased incrementally).

4 Conclusions

In this paper, we presented a literature search module developed within the FP7 IP Project CHRONIOUS to complement out of shell search modules such as Google Scholar, PubMed and GoPubmed by improving the search capabilities within the COPD and CKD pathologies, and by providing the possibility to index and retrieve hospital-specific internal documentation. The adopted ontology-thesaurus approach aims at overcoming limitations of pure syntactical document search, in which two queries having different syntax but equivalent semantics yield different search results. The search mechanism combines MeSH terminological power with formal knowledge

[1] The search results of COPD terms have been retrieved from 930 documents that were published from January 1st to December 31st, 2008 and contained the term "COPD" in their text (incl. title, abstract, text body, figure/table caption, etc.).

of the domains specified by the ontologies. MeSH and domain ontologies do not compete, but fulfill complementary tasks. MeSH provides well-established generic medical terminology, multilingual lexical representations and synonyms while domain ontologies provide additional representational power can be harnessed to improve medical literature annotation and retrieval.

Acknowledgements. This research has been supported by the IP project CHRONIOUS (EU Contract N. FP7-ICT-2007–1–216461). The authors want to thanks all the partners for their support. This paper was written as a joint contribution; authors are listed in alphabetical order grouped by their affiliations, besides they have equally contributed to this paper.

References

1. Schneider, L., Brochhausen, M.,: The CHRONIOUS Ontology Suite: Methodology and Design Principles. Proceedings of the International Conference on Biomedical Ontology, University at Buffalo, NY July 26-30, 2011
2. Medical Subject Headings (MeSH), http://www.nlm.nih.gov/mesh
3. Brochhausen, M., Spear, A.D., Cocos, C., Weiler, G., Martin,V., Anguita, A., Stenzhorn, H., Daskalaki, E., Schera, F., Schwarz, U.,. Sfakianakis S., Kiefer, S., Doerr, M., Graf, N., Tsiknakis M.,: The ACGT Master Ontology and Its Applications - Towards an Ontology-Driven Cancer Research and Management System. Journal of Biomedical Informatics. E-published ahead of print. (2010), DOI 10.1016/j.jbi.2010.04.008
4. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C.J., Neuhaus, F., Rector, A., Rosse C. : Relations in Biomedical Ontologies, Genome Biology, 2005, 6:R46
5. The Open Biological and Biomedical Ontologies Foundry, http://www.obofoundry.org/
6. The GATE framework, http://gate.ac.uk/
7. OntoRoot Gazetteer, http://gate.ac.uk/sale/tao/splitch13.html#sec:gazetteers:ontoRootGaz
8. Text2Onto project, http://code.google.com/p/text2onto/
9. WordNet, http://wordnet.princeton.edu/
10. JENA – A Semantic Web Framework for Java http://jena.sourceforge.net/
11. OWL 2 Web Ontology Language Primer W3C Recommendation 27 October 2009 http://www.w3.org/TR/owl2-primer/
12. Spear, A.: Ontology for the Twenty First Century: An Introduction with Recommendations, BFO manual, 2006
13. Chronic obstructive pulmonary disease (COPD) ontology, http://www.ifomis.org/chronious/copd
14. Chronic kidney disease (CKD) ontology, http://www.ifomis.org/chronious/ckd
15. Middle Layer Ontology for Clinical Care (MLOCC), http://www.ifomis.org/chronious/mlocc
16. Van Assem, M., Malaisé, V., Miles, A., Schreiber, G.: A method to convert thesauri to SKOS, The Semantic Web: Research and Applications, ESWC 2006 LNCS, Vol. 4011, Springer Berlin/Heidelberg, (2006)
17. The OWL API, http://owlapi.sourceforge.net/
18. van Rijsbergen, C.J.: Information Retrieval (2nd ed.) Edn. Butterworth, 1979