Combining different standards and different approaches for health information retrieval in a quality-controlled gateway

Lina F. Soualmia, Stéfan J. Darmon

1. Introduction

The amount of health information available on the Internet is considerable. Information retrieval remains problematic: users are now experiencing huge difficulties in finding precisely what they are looking for, among the tons of documents avail-
able online. Generic search engines (for example Google\(^1\)) or generic catalogues (for example Yahoo\(^2\)) cannot solve this problem efficiently because they usually offer a selection of documents that turns out to be either too large or ill-suited to the query. Free text word-based (or phrase-based) search engines typically return innumerable completely irrelevant hits requiring much manual weeding by the user, while missing important information resources. Free text search is not always efficient and effective: the sought page might be using a different term (synonym) that points to the same concept; spelling mistakes and variants are considered as different terms; search engines cannot process HTML intelligently. We propose in this paper to combine two knowledge-based methods (natural language processing and knowledge discovery in databases) in the KnowQuE (Knowledge-based Query Expansion)\(^3\) prototype meant to focus and expand a user query and cope with the problems of free-text based search into the catalogue CISMeF\(^4\) (acronym of Catalogue and Index of French-speaking Medical Sites). CISMeF has been developed since 1995 to help health professionals, as well as students and the general public, with their search for electronic health information. All the resources (documents and Web sites) indexed in the CISMeF catalogue are described by the librarians using the vocabulary of a structured terminology that may be assimilated to the concept of ontology in the medical domain, and a set of metadata based on the Dublin Core\(^5\). The first KnowQuE module is composed by a morphological knowledge base that has been built according to the terminology. Recent works\(^4\)–\(^6\) present the contribution of morphological processing to information retrieval in French. For example for a query on ‘asthmatic child’ the module should return documents on ‘children with asthma’. Lexical resources (in French) are needed for the medical vocabulary. The second module is founded on association rules\(^7\) between terms, extracted from the indexed resources by a data mining technique. These association rules are used in the information retrieval process. For example the association rule ‘prevention of breast cancer → mammography’ is extracted because ‘prevention of breast cancer’ and ‘mammography’ are frequently used conjointly to index the resources. Applying the association rule, a query on ‘mammography’ should return documents on ‘prevention of breast cancer’. The resources stored in the CISMeF catalogue are indexed with a set of metadata and according to a terminology, founded on the MeSH thesaurus, which has the same structure and use as an ontology of the medical domain. The structure of the catalogue allows us to place the project at an overlap between the present Web, which is informal, and the forthcoming Semantic Web\(^8\). The paper is organized as follows: The section “Towards a medical Semantic Web” describes the modeling of the CISMeF metadata and terminology and “Ontology exploitation” the exploitation of these two standards in the Web site of the catalogue, mainly information retrieval and navigation. “Enhancing information retrieval” will detail two approaches to enrich the terminology and hence to improve information retrieval. “Conclusion and future work” includes the conclusions and outlines some future directions for work.

2. Towards a medical Semantic Web

The Semantic Web\(^8\) is an infrastructure that has to be built. It aims at creating a Web where information semantics are represented in a form that can be understood by humans as well as machines, better enabling computers and people to work in co-operation. One of its advantages is to bring sufficient information on the resources, by adding annotations in the form of metadata and to describe formally and significantly their content according to an ontology. This infrastructure must be formalized. The current Web is informal: It is mainly composed of HTML pages, hand-written or generated automatically, for human treatment only. Ontologies and metadata are two major components for the construction of the Semantic Web. Ontologies are powerful tools that may remove ambiguity: they provide a controlled vocabulary of terms and some specification of their meaning and are very useful for interoperability, browsing and searching. Metadata describe Web information resources enhancing information retrieval and enabling accurate matches to be made while being totally transparent to the user. Many projects and tools using ontologies have been developed for information retrieval but also for classifying and indexing.

The CISMeF catalogue describes and indexes a large number of health information resources\(^6\) (n = 13,452). CISMeF references high quality information resources. A resource can be a Web site, Web pages, documents, reports and teaching material: any support that may contain health information. CISMeF and DocCISMeF, its associated search tool, take into account the diversity of the end-users and allow them to find good quality resources. These resources are selected according to strict criteria by the team of librarians and are indexed according

---

Combining different standards and different approaches for health information

The user type is also taken into account. CISMeF, oriented ontology [12] used in the catalogue. A set of metadata elements and the terminology to CISMeF: extend the DC standard. Eight elements are specific to CISMeF: institution, city, province, country, target public, access type, sponsorships, and cost. The user type is also taken into account. CISMeF defined two additional fields for the resources intended for the health professionals: indication of the evidence-based medicine and the method used to determine it. In the teaching resources 11 elements of the IEEE 1484 LOM (Learning Object Metadata) "Educational" category are added.

In 1995 the metadata format was HTML. In 2000, in order to allow interoperability with other platforms the XML language became the metadata format. Since December 2002 RDF, a basic Semantic Web language, has been used within the EU-project MedICIRCLE framework [11] in which CISMeF is a partner. This project was initiated to qualify the quality of health information and to guide consumers to trustworthy health information. The vocabulary of the HIDDEL metadata is included in an ontology (represented in RDF Schema) and the resources are described in RDF according the concepts of the HIDDEL ontology.

2.1. The CISMeF metadata

The notion of metadata appeared before Internet but its interest has grown with the number of electronic publications and digital libraries. "The Semantic Web dream is of a Web where resources are machine understandable and where both automated agents and humans can exchange and process information". The solution proposed by the W3C is to use metadata to describe the data contained on the Web and to add semantic markup to Web resources, thus describing their content and functionalities, from the vocabulary defined in ontologies. Metadata are data about data or in the Web context, data describing Web resources. When properly implemented, metadata shall unambiguously describe resources, so enhance information retrieval.

In CISMeF we use several sets of metadata. Among them there is the Dublin Core (DC) metadata set, which is a 15-element set, intended to aid discovery of electronic resources. The resources indexed in CISMeF are described by 11 of the elements of Dublin Core: author, date, description, format, identifier, language, editor, type of resource, rights, subject and title. DC is not a complete solution; it cannot be used to describe the quality or location of a resource. To fill these gaps, CISMeF uses its own elements to extend the DC standard. Eight elements are specific to CISMeF: institution, city, province, country, target public, access type, sponsorships, and cost. The user type is also taken into account. CISMeF defined two additional fields for the resources intended for the health professionals: indication of the evidence-based medicine and the method used to determine it. In the teaching resources 11 elements of the IEEE 1484 LOM (Learning Object Metadata) "Educational" category are added.

In 1995 the metadata format was HTML. In 2000, in order to allow interoperability with other platforms the XML language became the metadata format. Since December 2002 RDF, a basic Semantic Web language, has been used within the EU-project MedICIRCLE framework [11] in which CISMeF is a partner. This project was initiated to qualify the quality of health information and to guide consumers to trustworthy health information. The vocabulary of the HIDDEL metadata is included in an ontology (represented in RDF Schema) and the resources are described in RDF according the concepts of the HIDDEL ontology.

2.2. The CISMeF terminology

The catalogue resources are indexed according to the CISMeF terminology, which is based on the MeSH (Medical Subject Headings) [13] thesaurus of MEDLINE and its French translation. The MeSH was selected because it fulfills the aims of medical librarians and it is well known by the health professionals. Approximately 22,000 descriptors (e.g.: abdomen, hepatitis) and 84 qualifiers (e.g.: diagnosis, complications) compose the MeSH thesaurus in its 2003 version. These concepts are organized into hierarchies going from the most general on the top of the hierarchy to the most specific in the bottom of the hierarchy. For example, the descriptor hepatitis is more general than the descriptor hepatitis viral A. The qualifiers, also organized into hierarchies, allow to specify which particular aspect of a descriptor is addressed. For example the association of the descriptor hepatitis with the qualifier diagnosis (noted hepatitis/diagnosis) restrict the hepatitis to its diagnosis aspect. The specializations relations between concepts are extracted from the MeSH text files to define the subsumption relationships in the CISMeF descriptors hierarchy (n = 9765).

MeSH descriptors and qualifiers are organized into hierarchies that do not allow a complete view concerning a specialty. The descriptors and qualifiers in CISMeF are brought together according to metaterms (e.g.: Cardiology). Metaterms (n = 87) concern medical specialties and it is possible to know the sets of the MeSH descriptors and qualifiers that are dispersed in several trees and are semantically related to the same specialty. In addition to

---


---
the set of metaterms, the CISMeF team has modeled a hierarchy of resource types \((n = 127)\). They describe the nature of the resource (e.g.: teaching material, clinical guidelines) and are a generalization of the MEDLINE publication types. In fact metaterms have been created to optimize information retrieval in CISMeF and to overcome the relative restrictive nature of MeSH descriptors. For example a search on "guidelines in cardiology" or "databases in virology", where cardiology and virology are descriptors and guidelines and databases are resource types, will yield few or no answers. Introducing cardiology and virology as metaterms is an efficient strategy to obtain more results because instead of exploding one single MeSH tree (information retrieval and navigation), the use of metaterms will result in an automatic expansion of the queries by exploding other related MeSH or CISMeF trees besides the current tree.

Fig. 1 The CISMeF terminology structure.

3. Ontology exploitation

3.1. Resource indexing and classification

The CISMeF ontology is exploited for several tasks: resource indexing (manually and automatically), visualization and navigation through the concept hierarchies and information retrieval using the Doc'CISMeF search engine. Each catalogue resource is indexed according to the vocabulary of the ontology (terms being descriptors, qualifiers, and resource types). Using heuristics and a rule-based classification algorithm [15], the related specialties to a resource are deduced due to the existing semantic links between (metaterm, descriptor), (metaterm, qualifier) and (metaterm, resource type) and are ranked according to their level of importance.

3.2. Information retrieval and navigation

The navigation through the ontology, based on alphabetical and thematic indexes, allows the user to know the terms that represent the concepts used in the domain and also their positions in the different hierarchies. Each term has its own Web page and a set of links, which represent preformatted queries, enabling the user to retrieve all the resources that are related to this term. S/he can also restrict the search according to his category: resources intended for health professionals, students, patients and the general public. The other and main utility of the ontology is its exploitation by the search engine. Different modes are possible. "Simple search" is done via an interface in which the user can tape in queries in natural language (French or English, with or without accents, capital letter or not). "Advanced search" is a more precise search: it uses frames and drop-down lists and different attributes (keywords, titles, year) can be combined with Boolean operators (And, Or, NOT). "Logical search" done with Boolean operators and a specific query language with particular characters. Currently, the simple search is based on subsumption relationships. If the query (a word or an
Combining different standards and different approaches for health information

expression) can be matched with a concept of the ontology, then the result of the query is the union of the resources that are instances of the concept, and the resources that are instances of the concept it subsumes, directly or indirectly. In all the hierarchies it belongs to (explosion). For example a query on "hepatitis" will return as answer all the resources related to hepatitis but also those related to hepatitis A, hepatitis B, etc. If the query cannot be matched to a concept of the ontology, the search is done over the other fields of the metadata elements set. In the worst case, a full-text search is carried out. Doc'CISMeF is interoperable with PubMed. A query is transformed automatically into its correspondent in English and by clicking a generated link a new query into the PubMed syntax is performed, only if the query corresponds to a MeSH descriptor or a qualifier.

3.3. CISMeF-patients: a sub-catalogue dedicated to the general public

Due to the important quantity of information for the patients and the general public, written by health professionals, medical institutions and associations of patients, the sub-catalogue CISMeF-patients [16] was designed in 1997. CISMeF-patients and CISMeF share the same terminology. CISMeF-patients is a specific view that corresponds to the metaterm Patient in the CISMeF catalogue. Popularized synonyms were associated to the terms at each level of the model. The synonyms are terms used in the current language (e.g. 'mania' is a popular synonym of 'bipolar disorder') and were determined thanks to a collaboration with the patients’ associations. The navigation into CISMeF-patients can be done through an index of medical specialties (n = 34). A general index of all terms used in CISMeF-patients is also available (n = 343). This avoids the layer-sen to type its queries on the search interface. Preformatted queries are applied on Doc'CISMeF and are generated automatically when a user clicks on a keyword link. Therefore, the user is neither obliged to know the MeSH descriptor, nor the query language used in the Doc'CISMeF search tool. The corresponding query generated is: patient [metaterm] And [descriptor]. For example, in the case of clicking on 'mania' the generated query is: patient [metaterm] And bipolar disorder [descriptor]. The patient resource types are the following ones: hot lines, associations of patients, news group and discussion list for patients, and patient and health consumer information.

Another type of access is available via the 'life periods’. It is similar to the 'life events' of HealthInsite. These periods are: birth, child, adolescence, adult and aged. By clicking on the links, it is possible to obtain automatically the related information resources of these periods. For example, the query generated automatically on Doc'CISMeF for Birth is the following one: new born [descriptor] And patient [resource type]. The major difference between MedlinePlus and CISMeF-patients is the structure of the terminology. CISMeF-patients and CISMeF share the same terminology whereas MedlinePlus and MEDLINE do not: another terminology has been built for MedlinePlus. CISMeF-patients and CISMeF also share the same search tool Doc'CISMeF. The benefit of sharing the same terminology and search tool is the possibility to extend the patient query to another one using another resource type than patient. For example ‘clinical guidelines' for evidence based medicine resources, or 'education' for teaching resources. A patient searching information about 'leukemia' will have access to 'patient' resources. There are two links to other preformatted queries to find other resource types: leukemia [descriptor] And clinical guidelines [resource type] and leukemia [descriptor] And educational material [resource type].

3.4. Other features: major topics and search strategies

Any search could be limited to Major Topics of each level of the CISMeF terminology. Major Topics exist in the MEDLINE database for descriptors and qualifiers. In CISMeF Major Topics are extended to resource types and metaterms. This task is manually performed by CISMeF medical librarians for resource types, and automatically performed for metaterms: a metaterm is Major for a CISMeF resource if and only if at least one descriptor, qualifier or resource type, which is semantically linked to this metaterm is Major for the same CISMeF [resource type] (minor if not).

Another kind of preformatted queries has been modeled: the search strategies. A search strategy is a medical concept defined by a Boolean expression composed by several other concepts of the ontology. For example: dental surgery = (dentistry, operative [descriptor] Or tooth extraction [descriptor]) Or (tooth replantation [descriptor] Or (identification [descriptor] Or tooth diseases [descriptor]) And surgery [qualifier]) urological surgery = urology [metaterm] And (surgical procedures, operative [descriptor] Or surgery [qualifier]).
4. Enhancing information retrieval

The submitted queries over the search engine are seldom matched to the vocabulary. We have extracted and analyzed the kind of queries of the http server and their associated number of answers between the 15th August 2002 and 6th February 2003. 1,552,776 queries were extracted. Among them 892,591 (58.62%) were submitted via the simple search interface and 365,688 (40.97% of the simple queries) had no answer. To enhance this kind of information retrieval, which is largely used over the catalogue, we have developed the KnowQuE [1] prototype founded on morphological processing and association rules mining. The details of the results obtained are presented in [17,18].

4.1. Morphological processing

A major task in information retrieval is to match a query with a document. It may be achieved by query normalization (lemmatization by reducing a word to its lexeme, stemming by reducing a word to its root form) or by query enrichment (by adding inflexions or derivations). In the Web, the user queries are frequently composed by few words. Many works have addressed "terminological variation", which can be processed at different levels: characters [19] (spelling and accenting mistakes, case variants), words and their morphological variants [20–22], syntax [22], or concepts with general-language derivations. Phonemic matching is yet another method to match the exact expression of each word of the query with the terminology. For example 'accident' will be matched with the term 'circulation accident' (but not with 'accidents' and 'chutage accidentelle'). Morphological knowledge: we replace each word by its root in the morphological family. The root is generally the simple form of the term. A morphological family of a term is composed by its inflexions (for example {accident, accidents}) and derivations (for example {probability, probabilistic}). If the user query is "Children with asthma" it will be replaced by the Boolean query "child [descriptor] AND asthma [descriptor]". The problem is that this kind of knowledge base does not exist yet for the French medical language [24].

4.1.2. Extracting derivations and expanding queries

To build a morphological knowledge-based according to the CISMeF terminology, we have used a French general-domain lexical resource [25]. It is not specific to the medical domain but it allowed us to obtain 2732 morphological families of descriptors (total of 9401; 3388 are composed by one word), 55 families of qualifiers (total 84; 55 of one word) and 28 families of resource types (total 127; 28 of one word). In this first step we have only considered the terms that are composed by one word. The root of a morphological family is the matched descriptor with the terminology (qualifier and resource type respectively) even if it is not in its simple form. By analyzing the other terms composed by two or more words, we have found that 1935 terms (1899 descriptors; 8 qualifiers; 22 resource types) are semi-matched. We consider that a term is semi-matched when at least one of the words that compose it is matched. For example the descriptor "accidents" has as family: {accident, accidents, accidente, accidentees, accidentel, accidental, accidentelle, accidentelles, accidentelelement, accidenter}. Therefore, the descriptor "accident circulation" is semi-matched because accident is matched and not circulation. The semi-matching is useful for the exact expression step (Table 1).

We have implemented the algorithm in Java with an ODBC connection to the CISMeF Oracle 8.1 database. The different functions of the algorithm (Segmentation, Normalization, Stop Words,
Combining different standards and different approaches for health information

Table 1

<table>
<thead>
<tr>
<th>Coverage of the vocabulary</th>
<th>Keywords</th>
<th>Qualifiers</th>
<th>Resource types</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of terms matched</td>
<td>2732</td>
<td>55</td>
<td>28</td>
<td>2815</td>
</tr>
<tr>
<td>One word matching (%)</td>
<td>80.64</td>
<td>100</td>
<td>100</td>
<td>81.33</td>
</tr>
<tr>
<td>Semi-matching</td>
<td>4631</td>
<td>63</td>
<td>50</td>
<td>4750</td>
</tr>
<tr>
<td>Total (%)</td>
<td>48.80</td>
<td>75</td>
<td>39.37</td>
<td>49.11</td>
</tr>
</tbody>
</table>

Exact Expression and Morphological Knowledge were expressed using SQL queries and regular expressions.

For preliminary results, we have tested the algorithm on a set of 77,382 queries with empty answers, which correspond to 48,255 distinct queries. The size of the queries is small. The 12,974 queries (26.89%) are composed by one word; 16,347 (33.88%) by two words; 10,972 (22.74%) by three words; 4360 (9.03%) by four words; 3602 are composed by more than four words (7.46%).

The 48,255 null queries were segmented into 121,958 words. By applying the different steps of our algorithm, a total of 92,887 terms (76.16%) were matched with the terminology and the morphological base, but 29,071 terms remained unknown (23.84%) (Table 2). Many of the unknown words were spelling errors but, in addition to morphological knowledge, semantic knowledge is necessary, for example, heart and cardiac are semantically related and a syntactic analysis is not adapted.

We have performed a quantitative analysis to match the null queries of the users with the terminology enriched by a morphological knowledge base. As in [4] the ongoing qualitative evaluation is done by the medical librarian.

4.2. Data mining

These association rules are used in the information retrieval process. For example, the association rule ‘prevention of breast cancer → mammography’ is extracted because ‘prevention of breast cancer’ and ‘mammography’ are frequently used conjointly to index the resources. Applying the association rule, a query on ‘mammography’ should return documents on ‘prevention of breast cancer’.

The second KnowQuE module is founded on Association Rules mining. We apply this data mining technique on the CISMeF database. Our first goal is to extract knowledge in the form of new and interesting association rules between couples of (descriptor/qualifier) from the indexed resources. The second goal is to exploit these association rules in the query expansion process. Association rules were initially used in data analysis and in data extraction from large relational databases [7]. We are interested in the discovery of Boolean association rules.

4.2.1. Definitions

A Boolean association rule (AR) is expressed as:

\[ AR : i_1 \land i_2 \land \cdots \land i_j \Rightarrow i_{j+1} \land \cdots \land i_n \]  

This formula states that if an object has the items \{\(i_1, \ldots, i_j\}\) it also tends to have the items \{\(i_{j+1}, \ldots, i_n\)\}. To evaluate an association rule, objective measures (based on statistics) and subjective measures (based on human expertise) exist.

The \(AR\) support (2) represents its utility. This measure corresponds to the proportion of objects that contains at the same time the rule antecedent and consequent

\[ \text{support}(AR) = \left| \{i_1, i_2, \ldots, i_n\} \right| \]  

The \(AR\) confidence (3) represents its precision. This measure corresponds to the proportion of objects that contain both antecedent and consequent among those containing the antecedent

\[ \text{confidence}(AR) = \frac{\left| \{i_1, i_2, \ldots, i_n\} \right|}{\left| \{i_1, i_2, \ldots, i_j\} \right|} \]  

4.2.2. Knowledge extraction process

The knowledge extraction process is achieved in several steps: the data and context preparation (objects and items selection), the extraction of the frequent itemsets (compared with a minimum support threshold), the generation of the most informative rules using a Data Mining algorithm (compared with a minimum confidence threshold), and finally the interpretation of the results.
An extraction context is a triplet $C = (O, I, R)$, where $O$ is the set of objects, $I$ the set of all the items and $R$ a binary relation between $O$ and $I$. An itemset is frequent in its context $C$ if its support is higher than the minimal threshold initially fixed (by the user). The extraction problem of frequent itemsets has an exponential complexity in size of $n$, the number of the potential frequent itemsets is $2^n$.

The itemsets form a lattice [26] which construction is time and space consuming. The most known algorithm used to extract frequent itemsets is Apriori [7]. In our case we use the A-Close algorithm, which calculates the closed frequent itemsets [27] using the semantic based on the closure of the Galois connection [28], reducing by that itemsets space size studied. The algorithm calculates the generators of the frequent closed itemsets. The generators of a closed itemset $I_{close}$ are the itemsets of maximal size which closure is equal to $I_{close}$. New bases for association rules are deduced from the closed frequent itemsets and their generators. These bases consist of minimal non-redundant association rules [27].

In our case, the extraction context is the following: the objects are the annotations used to describe the indexed resources $O = \{\text{annotations}\}$. The relation $R$ represents the indexing relation between an object and an item with $I = \{(\text{Keyword}/\text{Qualifier})\}$, the couples of (descriptor/qualifier). We have implemented the A-Close algorithm in Java and tested it on several sets of resources by fixing the support to 10 documents and the confidence to 100% (exact rules) to have an objective measure (Table 3).

### 4.2.3. Evaluation

All the extracted rules (260) were evaluated by an expert (medical librarian) (Table 4). An interesting association rule is one that confirms a hypothesis or states a new hypothesis. In our case, there may be several cases of interesting association rules in function of the existing relationships between the terms of the terminology.

It could associate:
- a (in)direct son and its father in the hierarchy (FS),
- two sibling terms that belong to the same hierarchy (same (in)direct father) (B),
- a See Also relationship that exists in the terminology (SA),
- a new relationship considered interesting (NW).

Among the 260 rules, 142 (54.61%) were considered interesting by our expert:

- breast cancer/diagnostic $\Rightarrow$ mammography (support = 25 documents, confidence = 1),
- aids/prevention and control $\Rightarrow$ condom (support = 10 documents; confidence = 1).

Among the interesting rules we have obtained:
- 68.31% new rules (NW),
- 14.49% See Also relationships (SA),
- 10.56% Brother relationships (B),
- 05.63% Father–Son relationships (FS).

The Father–Son relationships are already used in the information retrieval process. The other types of interesting association rules could be used to expand the users’ queries. A set of elements (descriptors and couples (descriptor/qualifier)) deduced from the association rules are proposed (Fig. 2) to the user who may expand the query by choosing and adding some of them. For example the query ‘mammography’ may be replaced by ‘(mammography [descriptor] OR breast cancer/prevention and control [descriptor/qualifier]). The CISMeF team has found these results to be interesting and has included the validated association rules in the CISMeF database in the form of couples (antecedent, consequent).
5. Conclusion and future work

We have discussed in this paper some of the problems of information retrieval on the Web. We have presented particular aspects of the CISMeF project, which has been developed to assist health professionals, medical students, patients and the general public in their search for health information on the Web. All the available tools of the catalogue have been developed to be useful for all users. We have also proposed to use two methods to enhance information retrieval. The natural language processing is used to build morphological knowledge base and data mining enables association rules discovery between concepts. Query expansion is now possible in CISMeF thanks to morphological processing and association rules. The third method we are experiencing is terminological reasoning. It is founded on a formal terminological knowledge base built by automatic translation of the terminology into the OWL DL-based language [29] and the resources of the catalogue into instances (or individuals) of the OWL concepts and roles (or classes and relations). As in [30–32], reasoning mechanisms are involved here to verify the knowledge-based consistency but in our case it is mainly to answer queries. By this formalization of the terminology [33] and thus the MeSH, such a formal ontology issued from the MeSH is promising and may be exploited in many applications, based on the MeSH thesaurus, mainly bibliographic databases such as MEDLINE and health gateways, also contributing to the Semantic Web.

References

S.J. Darmoni, B. Thirion, S. Platel, M. Douy`ere, P.
L.F. Soualmia, S.J. Darmoni, Combining knowledge-based
methods to refine and expand queries in medicine, in: Pro-
ceedings of the Sixth International Conference on Flex-
ible Query Answering Systems, FQAS’ 2004, LNAI3055,
Springer-Verlag, pp. 243–255.
19. C. Loisits, R. Baud, Fast exact string pattern-matching algo-
rithms adapted to the characteristics of the medical lan-
20. A.T. McCoy, S. Srinivasan, A.C. Browne, Lexical methods
for managing variation in biomedical terminologies, in: Pro-
ceedings of the 18th Annual Symposium on Computer Ap-
piled to Medical Care, 1994, pp. 235–239.
21. C. Loisits, P.A. Michel, R. Baud, J.R. Scherrer, Word segmen-
tation processing: a way to exponentially extend medical
dictionaries, in: Greimes, Peterson, Protti (Eds.), Proceed-
ings of the Eighth World Congress on Medical Informatics,
1995, pp. 28–32.
22. C. Jacquemin, E. Tzokakis, NLP for term variant ex-
traction: a synergy of morphology, Lexicon, Syntax, in:
Szczylikowski (Ed.), Natural Language Processing and Infor-
23. T. Hamon, A. Hazarene, C. Gros, A step towards the de-
tection of semantic variants of terms in technical docu-
498–504.
24. P. Zweigenbaum, R. Baud, A. Burgan, et al., Towards a uni-
25. R. New, C. Pallier, L. Ferrand, R. Matos, Une Base de
Données Lexicales du Français Contemporain sur Inter-
26. B.A. Davey, H.A. Priestley, Introduction to Lattices and Or-
27. H. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Efficient mining
of association rules using closed itemsets lattices, Inform.
28. B. Gaunt, R. Wiie, Formal Concept Analysis: Mathematical
29. J. Harrocks, P.F. Patel-Schneider, F. van Harmelen, From
SHQ and RDF to OWL: the making of a Web ontol-
30. S. Schulz, U. Hahn, Medical knowledge re-engineering —
converting major portions of the UMLS into a terminol-
ogical knowledge base, Int. J. Med. Inform. 64 (2–3) (2001)
207–221.
31. R. Cornet, A. Abu-Hanna, Usability of expressive description
logics — a case study in UMLS, In: Proceedings of the AMIA
32. V. Kastha, A. Borgia, Representing the UMLS Semantic
Network using OWL, In: Proceedings of the Second Interna-
33. L.F. Soualmia, C. Gobbe, S.J. Darmoni, Representing the
MeSH in OWL: towards a semi-automatic migration, In:
Proceedings of the First International Workshop on For-
mal Biomedical Knowledge Representation IR-Med, IR-MED