MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

A Multilingual Medical Thesaurus Browser for Patients and Medical Content Managers

Georg Göbel, Stefan Andreatta, Joachim Masser, Karl Peter Pfeiffer

Institute of Biostatistics and Documentation
University of Innsbruck, Austria

Abstract

This paper introduces a user-friendly browser interface which integrates multilingual search and browsing functionalities within medical thesauri via the internet. The tool is being developed as part of the GIN Austria Patient Information System and is based on an adapted datamodel of the MeSH thesaurus. A prototype offers the possibility to build up queries and export lists of MeSH main headings collected during browsing the relevant MeSH trees. The thesaurus browser can be used both by patients and citizens to build queries based on a controlled vocabulary to match them with existing documents within GIN and by medical information managers to find out appropriate keywords for interactive tagging or indexing of medical contents. A key component of this tool is the flexible choice of different languages of the MeSH datasource as well as of the user interface. Both can be changed independently at any point during a session. Another central aspect is the use of the UMLS Metathesaurus in combination with localized Thesaurus versions due to existing international character set problems.

Keywords

Internet, Patient Information System, Subject Headings, Unified Medical Language System, Information Services

Introduction

Although internet health information systems for consumers and patients have become more and more popular, internet information retrieval (IIR) is "still a very hard task, even for IR experts" [1,2]. The difficulties range from the "lost in the hyperspace syndrom" to problems with medical sites without any systematic quality assurance (QA) programs [6,7,14]. The QA problem is addressed by some initiatives that are working successfully on QA guidelines (e.g. HON Initiative [4]) and they are attracting increasing worldwide attention from medical site managers and providers [23]. On the other hand, conventional internet retrieval environments (search-engines) are not very helpful to users who are not used to specify Boolean queries containing appropriate keywords based on a controlled medical vocabulary. These consumers can be supported by active use of domain knowledge during query specification and the user will not be limited to his (known) set of key terms. Sometimes this type of query is called conceptual query [22]. Next generation web search engines will make heavy use of semistructured web data to improve crawling and indexing performance [5,16]. Therefore, tools for indexing or tagging sites during (not after) a creation or maintaining process will be necessary [8]. Flexible usability of different languages may thereby be a major task.

Objectives

The scope of the entire GIN Austria project (Gesundheitsinformationsnetz Austria) is to define a methodology for the management of patient oriented medical contents / websites [9,10]. Within this project the Medical Thesaurus Browser has three core objectives:

- To provide medical site content managers with interactive access to multilingual medical thesauri via internet.
- To assist patients and citizens with building queries based on controlled vocabularies for the successful retrieval of information from the internet.
- To provide a program interface for international indexing tools and search engines.

Methods and Materials

The interactive thesaurus browser is based on technologies standardized by the World Wide Web Consortium (W3C) [24] and international medical datasources to ensure platform interoperability. It is currently a system of dynamically generated webpages, which allow the interactive use of the thesauri.

Sources of information

The Medical Subject Headings (MeSH) are published by the National Library of Medicine (NLM). Several institutions in other countries are translating the English sources and release localized versions [11,18]. The Unified Medical Language System (UMLS) which is also edited by the NLM is a Metathesaurus comprising many different controlled vocabularies within the medical field. Among these sources are the English MeSH Thesaurus and seven

translations (Table 1). Therefore this Metathesaurus was chosen as a consistent datasource the for the reference implementation of the multilingual thesaurus browser.

Table 1 – Numbers of preferred terms and synonyms in the English MeSH vocabulary and seven translations

language	preferred terms	synonyms
English	19768	20905
Finnish	19285	0
French	19768	9007
German	19647	25993
Italian	10711	411
Portuguese	19768	18774
Russian	19626	20160
Spanish	19768	18629

Data storage

The data for the Browser are stored in a relational database management system in four parts:

1. The MeSH source data contain the preferred terms and synonyms of the MeSH in eight languages. Furthermore it stores the language independent logical structure of the vocabulary, i.e. the relations between preferred terms and their multihierarchic organization [15].
2. User preferences, which are kept beyond individual sessions, such as the default thesaurus language.
3. All language specific information for the browser interface such as text or button captions are also stored in a database. They are referenced by a unique identifier within the webpages which are replaced by the actual language elements prior to delivery to the user. Thus, multiple language versions of the browser interface can easily be kept consistent.
4. A log database records user accesses and delivers information on the use of the system. Entries are labelled by random session identifiers. No personal information on the user is gathered or stored.

Additionally, session data which is used during one individual user session is serialized and stored in a file system rather than a database.

Browser technology

Oracle 8i (Oracle, Redwood Shores CA) and MS Access 2000 (Microsoft, Redmond WA) are used as relational database management systems during development and operation of the browser. For maximum flexibility all data is accessed using standard Structured Query Language (SQL) instructions [2].

Web pages are generated dynamically using the PHP Hypertext Preprocessor version 4 [20] as a server side scripting language. PHP code can be used together with the Hypertext Markup Language (HTML) in the same files, allowing for an integration of functionality and design of dynamic web pages. PHP 4 runs as a module under the Internet Information Server (IIS) in a MS Windows NT environment. In its recent version, PHP offers advanced functionality for automated session management. Variables and objects can be serialized and reused across different requests during within user sessions. Sessions are recognized by unique identifiers stored as cookies on the user side. Although PHP is not a genuine object oriented language, it offers basic support for classes. Therefore, the largest part of the MeSH browser could be realized using object oriented concepts.

The output of the thesaurus browser delivered to the user are standard webpages using HTML 4.0, Cascading Style Sheets (CSS 1.0) and basic Javascript routines. All output is validated to meet W3C standards. Furthermore the system is being tested with the most common WWW-browsers on different operating systems (MS Windows: Internet Explorer 5, Netscape Communicator 4.7, Opera 4.x; Linux: Netscape Communicator 4.7, Opera 4.x).

Results

Reference implementation

The interactive thesaurus browser enables users to search preferred terms and synonyms in several language versions of the MeSH vocabulary and receive further information on the terms found. Furthermore, it assists in the formulation of a query based on preferred terms [5,12], which may then be submitted to other independent sources of information on the internet. The database is also prepared for the integration of other datasources such as the International Classification of Diseases (ICD 10).

Functionalities

The function of the thesaurus browser is outlined in Figure 2. Numbers in brackets refer to the items in this diagram.

Initially, the browser asks the user to enter a query term (1). A search for this string is performed within the main headings and synonyms of the MeSH database. As a result, a list of preferred terms is returned, which either themselves or in a synonym contain the requested string (2). Each entry within this list links to a page with detailed information (3). There, synonyms, related terms and the hierarchical context of the particular preferred term are shown. Again, search results and hierarchy entries, which are preferred terms themselves, link to detailed information pages.

Search results and detailed information pages are shown alternatively in a tabbed view within the browser window. Independent accounts are kept for both types of information. Thus, the user can move forward and back

within the history of results of a session (4). New search strings can be entered at any time. The language of the MeSH database to search can be changed at the beginning or during a session. Upon a change of database language existent results will be translated, so that gathered information is not lost. The language of the browser's interface is completely independent of the searched database and can also be changed during a session.

Preferred MeSH terms of interest can be selected from search results or detailed information pages into a separate list (5). These terms can be then be submitted as a query to common internet search engines (e.g. AltaVista, Google), or to more specialized sources of information such as PubMed or local medical information networks (e.g. GIN).

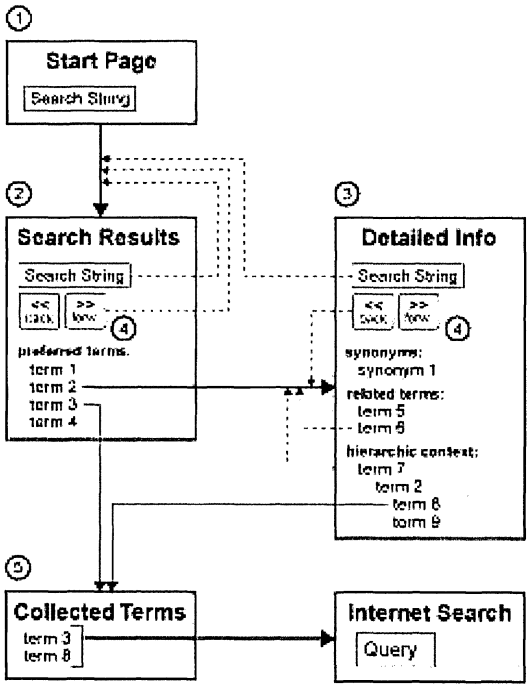

Figure 1- Diagram of the Thesaurus Browsers Functions. Numbers in circles refer to descriptions in the text.

Status quo

The specifications of the complete system were developed. A basic monolingual MeSH browser has been completed and tested (http://biostatistik.uibk.ac.at/gin/mesh). The multilingual MeSH browser incorporating all described features is in an advanced state of development and scheduled for release in March 2001.

Discussion

Technology

The tools chosen for the MeSH browser are clearly adequate for the task. Furthermore, the core technologies on the server side PHP and SQL are widely platform independent. On the output layer HTML 4.0 and CSS 1.0, which were recommended by the W3C, have become widely accepted standards for some time now [24]. Still, compatibility with different WWW-browsers had to be thoroughly checked and was a constant challenge during development.

Database

Controlled medical vocabularies are readily available from a range of sources [21]. However, several severe problems were encountered during the development of this browser.

The International Organization for Standardization (ISO) released guidelines for monolingual thesauri in 1986 [13]. However, the MeSH vocabulary cannot be mapped directly to this standard. The hierarchical position of each preferred term in the MeSH is determined by one or more tree numbers which uniquely identify each superordinate term up to the top of the hierarchical tree [15]. On the other hand the ISO standard identifies hierarchical relationships by specifying one level of superordinate (broader) and subordinate (narrower) terms. This may recursively lead to more hierarchic trees for one term than are explicitly specified in the MeSH thesaurus. This lack of an authoritative and generally applicable standard particularly impedes the development of tools working on more than one source of controlled vocabulary.

The UMLS is a very effective datasource as it provides data from different sources in a consistent format [18]. However, the MeSH data is not completely equal for different languages (see Table 1). While, except for the Italian version, nearly all preferred terms are translated, the number of synonyms available for each language varies between zero (Finnish) and 25 993 (German). These numbers comprise translations of the English MeSH synonyms and additional language specific synonyms. Generally, synonyms within MeSH do not have unique identifiers but are referenced by the corresponding preferred term. Furthermore several categories such as descriptions or alternative word forms are only available in the English MeSH thesaurus. Consequently, we had to limit the functionality of our multilingual MeSH browser to those categories which are generally available.

A special problem with multilingual MeSH data in the UMLS is that all strings are 7-bit ASCII encoded. Special characters such as German umlauts or Cyrillic characters are represented by their ASCII transcriptions, even if they are available in localized versions from the translating institutions. This severely impedes the usability of a multilingual thesaurus. To overcome these problems, we started by using only the UMLS data including all logical information. Localized language data, i.e. preferred terms

and synonyms, could then replace the language terms from the UMLS by matching their unique MeSH identifiers. Nevertheless, in the long term, it would be useful to have a Unicode encoded version of the UMLS.

Future prospects

As mentioned above, the next step will be the complete implementation and an international field test. For this task different license agreements must be fixed.

Other multilingual datasources (e.g. ICD) will be integrated. This concerns part 1 of the database, because parts 2,3,4 and the file system are already designed for other thesauri.

Further we are working on a specification of a SOAP interface (Simple Object Acces Protokoll [24]) for automatic tagging of XML documents with medical contents.

In connection with other standards (e.g. XML structured documents, adapted Dublin Core Index Scheme [17]) this tool could also be used in an integration process of electronic patient records and patient oriented information systems [7].

Acknowledgments

This work was supported by grant 741/1 of the "Jubiläumsfond der Österreichischen Nationalbank"

References

[1] Adelhard K, Obst O. *Evaluation of medical internet sites*. Methods Inf. Med. 38, 75-79. 1999.

[2] American National Standard Institute. *ANSI Homepage.* http://www.ansi.org

[3] Baud RH, Lovis C, Rassinoux AM, Scherrer JR. *Alternative ways for knowledge collection, indexing and robust language retrieval.* Methods Inf Med. 1998 Nov;37(4-5):315-26.

[4] Baujard V, Boyer C, Griesser JR, Scherrer JR. HONselect: A multilingual and intelligent search tool integrating heterogenous Web ressources. Proc. Medical Informatics Europe '2000, IOS Press, 273-78.

[5] Bodner RC. *Knowledge-based approaches to query expansion in information retrieval.* Lecture Notes in Computer Science 1081, 146-158. 1996.

[6] Eysenbach G, Diepgen T. *Towards quality management of medical information on the internet.* BMJ Vol 317, p1496. Nov. 28, 1998.

[7] Eysenbach G. *Consumer health informatics.*BMJ. 2000 Jun 24;320(7251):1713-6.

[8] Florescu D, Levy A, Mendelzon A. *Database Techniques for the World -Wide Web: A Survey.* .ACM SIGMOD 1998 Rec. 27:3, 59-74.

[9] Göbel G, Pfeiffer KP. *GIN AUSTRIA Assuring quality and relevance on Internet-Health-Informations for patients.* Proc. Med.Inf.Europe'99, IOS Press, 562-567 1999.

[10]Göbel G, Masser J, Pfeiffer KP. *A MESH based intelligent search intermediary for Consumer Health Information Systems.* Proc. Medical Informatics Europe '2000, IOS Press, 673-77.

[11]Hersh WR, Price S, Donohoe L. *Assessing thesaurus-based query expansion using the UMLS metathesaurus.* Proc AMIA Symp. 2000;(20 Suppl):344-8.

[12]Humphrey S. *Indexing biomedical documents: from thesaural to knowledge-based retrieval systems.* Artif.Intell.Med. 4(5), 343-371. 1992.

[13]International Organisation for Standardization (ISO). *Guidelines for the establishment and development of monolingual thesauri.* ISO 2788-1986(E).

[14]Lawrence S./Giles C.L. *Searching the WWW.* Science 1998 April 3; 280 (5360):98.

[15]Lowe,H.J. & Barnett,G.O. *Understanding and using medical subject headings (MeSH) Vocabulary to perform liturature searches.* J.Am.Med.Inform.Assoc. 271 (14), 1103-1108. 1994.

[16]Lowe HJ, Lomax EC, Polonkey SE. *The World Wide Web: a review of an emerging internet-based technology for the distribution of biomedical information.* J. Am. Med. Inform. Assoc. 3, 1-14. 1996.

[17]Malet,G., Munoz,F., Appleyard,R. & Hersh,W. *A model for enhancing Internet medical document retrieval with "medical core metadata.".* J. Am. Med. Inform. Assoc. 6, 163-172 1999.

[18]National Library of Medicine. *UMLS Knowledge Sources: Methathesaurus - Semantic Network - Specialist Lexikon.* NLM 11th Ed., 01/2000.

[19]Petkoff B. *Wissensmanagement.* Addison-Wesley Publishing. Bonn 1998.

[20]PHP Hypertext Preprocessor. *PHP Homepage.* http://php.net.

[21]Ruch P, Wagner J, Bouillon P, Baud RH, Rassinoux AM, Scherrer JR. *MEDTAG: tag-like semantics for medical document indexing.* Proc AMIA Symp 1999;:137-41.

[22]Wiesman F., Hasman A. , van den Herik H.J. *Information retrieval: An overview of system characteristics.* Int. J. Med. Inf. 47, 5-26. 1997.

[23]Winker MA, Flanagin A, Chi-Lum B, White J, Andrews K, Kennett RL, DeAngelis CD, Musacchio RA. *Guidelines for medical and health information sites on the internet: principles governing AMA web sites.* JAMA 2000 Mar 22-29;283(12):1600-6

[24]World Wide Web Consortium. *W3C Homepage.*

[25]http://www.w3c.org.

Addresses for correspondence

Georg Goebel , Univ.-Ass. Mag.rer.nat.
Institut of Biostatistics and Documentation
University of Innsbruck
Schoepfstrasse 41/1
A- 6020 Innsbruck
Austria
Email: georg.goebel@uibk.ac.at
URL: http://biostatistik.uibk.ac.at/curigoebel_en.htm

Coauthors

Stefan Andreatta, Mag. rer. nat
Joachim Masser
Karl Peter Pfeiffer, Univ.-Prof. Dr.
Institut of Biostatistics and Documentation
University of Innsbruck
Schoepfstrasse 41/1
A- 6020 Innsbruck
Austria