MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-486

Epistemonikos: A Free, Relational, Collaborative, Multilingual Database of Health Evidence

Gabriel Rada[a, b], Daniel Pérez[b], Daniel Capurro[a, b]

[a] *Department of Internal Medicine and Evidence-Based Healthcare Program, Pontificia Universidad Católica de Chile, Santiago, Chile.*
[b] *Epistemonikos Foundation, Santiago, Chile*

Abstract

Epistemonikos (www.epistemonikos.org) is a free, multilingual database of the best available health evidence. This paper describes the design, development and implementation of the Epistemonikos project. Using several web technologies to store systematic reviews, their included articles, overviews of reviews and structured summaries, Epistemonikos is able to provide a simple and powerful search tool to access health evidence for sound decision making. Currently, Epistemonikos stores more than 115,000 unique documents and more than 100,000 relationships between documents. In addition, since its database is translated into 9 different languages, Epistemonikos ensures that non-English speaking decision-makers can access the best available evidence without language barriers.

Keywords:

Information systems, information storage and retrieval, systematic review.

Introduction

During the last 20 years, a wide consensus has been reached around the need for making health decisions informed by the best available research evidence, in order to assure quality and efficiency: maximizing the benefits while minimizing harms and costs.

For most health-related decisions, there is already a substantial body of research that may help inform the process. However, the total amount of health research has made it impossible for individuals to keep up with all that it is produced, even in relatively narrow health topics. For instance, PubMed, the largest database of health publications, has reached 22,000,000 records, and more than a million are being added each year [1]. Moreover, PubMed is only one of several hundred health databases [2]. It includes approximately 5,000 journals, when the total number of health journals is no less than 20,000 [3], and a significant proportion of information is not even published in journals (e.g. conference proceedings, thesis, institutional reports, web pages), or not published at all.

The principles of Evidence-Based Healthcare (EBHC) have allowed the establishment of 'hierarchies of evidence', where not all the information has the same importance. Within this hierarchy, systematic reviews play a key role. Systematic reviews are reviews conducted following a systematic approach to reduce error and bias, usually summarizing results in a meta-analysis [4]. They have the potential to provide an objective and comprehensive summary of what is already known for a specific health question, by identifying all relevant studies, selecting those that meet explicit criteria, appraising their quality, and synthesizing the results using a transparent process. Drawing on an existing systematic review constitutes a more efficient use of time for decision makers, enabling them to draw upon research literature without having to comb through it themselves. In addition, decision makers are less likely to be misled by results of a systematic review than by a single investigation and can be more confident about what can be expected. While syntheses are useful to inform decision-making, they are also useful to the research community to inform their own decisions regarding future research undertakings, and to research funding organizations to focus their investments.

Together with the benefits of using systematic reviews for health decision-making, three issues are known barriers for their full incorporation. First, as systematic reviews have been adopted as useful tools for health decision-making, their number has steadily increased, and now we are faced with many systematic reviews for the same question, not uncommonly reaching different conclusions. Second, systematic reviews – such as primary research studies – can be difficult to read for those not well-versed in the arcane language of academia, so there is considerable interest in structured summaries of evidence, research briefs and other user-friendly formats that allow users to understand and apply the results of research. Finally, although the academic community has agreed on the importance of making decisions using systematic reviews, the uptake of this concept is far from being complete for many health professionals and policy-makers.

The need to have easy access to high quality research—especially in the form of systematic reviews—, and to provide results in a readable format are the key ingredients that led us to develop Epistemonikos.

Background

Since the advent of the Internet, researchers and decision-makers around the world have been relying on search engines to retrieve relevant information. The main database containing biomedical literature is PubMed, a database maintained by the US National Library of Medicine. Other well-known databases include EMBASE, CINAHL, and LILACS, among many others [2]. This multiplicity highlights the first problem: a decision-maker needs to search in many databases to find all that is known in a specific subject. These databases also share one

common characteristic: they store and index raw data, which leads us to a second issue: they return results without giving the user a sense of quality or relevance of the results of a search.

Several approaches have been developed to address these limitations. Meta search engines have been developed to solve the issue of multiple sources of information. With a single user-generated query they search in multiple databases and return a single set of results. This is the case of SUMSearch, in which users can use a single search box to enter search terms [4]; however the results are not presented in a way that highlights their relevance or the quality of retrieved articles. In terms of relevance and quality of results, biomedical databases have adopted terminologies to standardize search terms as well as systems to focus searches on specific types of publications such as randomized controlled trials or systematic reviews. In the case of terminologies, PubMed uses a standard terminology to annotate individual database entries to facilitate searches using standardized terms that group biomedical concepts [5]. Finally, some databases specialize on specific types of publications or provide curated contents as methods to improve the relevance and quality of their results. For example, the Cochrane Collaboration maintains the Cochrane Database of Systematic Reviews, a highly curated database of systematic reviews produced by authors around the world according to the organization's strict guidelines, thus ensuring high quality [6].

An issue that has been insufficiently addressed in the literature is the provision of results in a readable and user-friendly fashion. Some databases have started to produce summaries tailored to health consumers but their coverage is still limited [6]. Moreover, one of the main contributors to limited readability is the lack of reliable translations, which hinders the dissemination of research findings to the non-English speaking world.

As much as these issues have been addressed individually in existing databases as we mention above, to our knowledge there is no one that addresses all of these issues, providing an integral solution for users.

To formally address the issues of a) multiple sources of information, b) information overload in which the decision-maker cannot easily assess quality and relevance of results, and c) readable results, we embarked in the creation of Epistemonikos, a relational, collaborative, multilingual database of health evidence. Epistemonikos should be able to process a query generated in multiple languages, retrieve high-quality results, and present them to decision makers ranked by relevance and linked to additional resources that might answer the same health question. Moreover, Epistemonikos should be able to present results in the language in which the query was initially formulated.

Methods

A medical doctor with Evidence-Based Healthcare expertise and five computer scientists comprised the Epistemonikos core development team. Through agile programming methodologies and weekly meetings, project priorities were defined and executed upon.

Design Principles

The first major decision was what to include in the database. To address the issue of *quality of included studies*, we decided that the fundamental unit of Epistemonikos would be systematic reviews. Systematic reviews are conducted following a well-defined methodology to ensure the inclusion of all clinical evidence that answers a particular clinical question [7]. As a consequence, they are currently considered the best available evidence [8]. Using systematic reviews as the fundamental piece of the database architecture, we also included articles linked to an included systematic review: overviews of reviews including evidence-based policy briefs and guidelines, primary studies included in systematic reviews, and structured summaries of that evidence. All primary studies included in systematic reviews that fulfill predefined inclusion criteria are entered into the database, independent of language or publication status. This structure—a systematic review and all its relations—defined the core structure of the data model, which will be discussed in the following sections.

The second requirement for Epistemonikos was that it should include *knowledge from multiple sources*. To achieve this, the database was populated and is currently maintained by systematically searching PubMed and 18 other sources for relevant systematic reviews and overviews of reviews. Examples of these 18 sources are the Cochrane Database of Systematic Reviews (CDSR), EMBASE, and the Database of Abstracts of Reviews of Effects (DARE). Each database is searched periodically and when new systematic reviews are identified, they are sent to previously selected domain experts so they can upload the information—including the systematic review's included articles—to Epistemonikos. We currently have over 250 collaborators who perform this task.

In third place, we needed to address the issue of *language*. Most scientific literature is published in English but most clinical decision makers speak other languages. In addition, if we wanted to include all available systematic reviews, we needed to include non-English sources of publications. As a consequence, we decided that Epistemonikos would support a broad variety of languages. This should enable users to search in multiple languages as well as to read results in multiple languages.

To obtain translations in multiple languages we used three different and parallel approaches. First, when available, we use official translations such as the ones provided by the Cochrane Database of Systematic Reviews in Spanish, French and Chinese. These translations are flagged in the database as *official translations*. Second, when domain experts upload new systematic reviews or other types of eligible articles, they have the chance to translate themselves the original titles and abstracts. Since certified translators do not perform them, and other collaborators can edit and correct them iteratively, these translations are flagged as *collaborative translations*. To translate articles in which there is no official or collaborative translation available, we resort to automated statistical machine translations using Google Translate. This provides a rough translation of the whole database. Since machine translations are not error-free, we highlight these translations as *automatic translations*. These can be subsequently edited, turning them into *collaborative*. Currently we have 99.6% of the total articles with automatic translation and 8000 articles with official translations in languages other than English, with Spanish being the most-represented language.

Finally we had to address the issue of *relevance* when presenting results to users. Currently, with the exception of Google Scholar, most databases present results with little consideration to relevance. For example, PubMed and EMBASE, the most widely used clinical databases, present results ordered by publication date, which has little to do with relevance. To

provide results ordered by relevance, we implemented a scoring algorithm based on the frequency and location where the searched terms appear:

- Exact phrase in the title (29 points)
- Exact phrase in the list of authors (21 points)
- Exact phrase in the abstract (15 points)
- Every term appears in the title (11 points)
- Every term in the list of authors (8 points)
- Every term in the abstract (6 points)
- Every term, anywhere—title, authors or abstract (4 points)
- At least two terms with a distance not larger than 10 words, anywhere—title, authors or abstract (1 point).

Implementation

The first prototype was developed between August 2009 and February 2010, which allowed adding documents using the RIS format, automatically translating documents into 9 languages and a simple search interface. Since the fundamental component of the database was systematic reviews and their relations, we stored data using the Resource Description Framework (RDF) schema, a semantic web representation standard, which allows storing resources as a combination of nodes and relations. The combination of nodes and relations led to the conformation of a RDF graph.

Interestingly, when the database is populated, some clusters of systematic reviews and their linked articles start to appear. A cluster occurs when multiple systematic reviews answer the same—or a closely related—clinical question. Those clusters have a deep clinical significance: since they are built around a common clinical question, they can be later visualized to give the user a sense of the amount of published research around a single clinical question. This visualization is currently under development.

Initially, we used RDF and a standard relational database to store and query the document database. However, to improve query times we finally switched to MongoDB, a non-relational database specially designed to store documents [9] and the Apache Solr search engine [10]. This combination greatly improved response times.

User Interface

Epistemonikos has two kinds of users: *collaborators*, who load and translate documents into the database, and *end users*, who use the main search page to find health evidence.

Collaborators interact with the system through three interfaces:

Document upload panel: this panel allows the direct upload of a new document in RIS format or entering an identifier to the original database where the document is stored.

Metadata and relations panel: once the document is uploaded, this panel allows entering document metadata such as the document's category (primary study, systematic review, overview), and the document's relations.

Translation panel: once the document is uploaded, it is automatically translated into nine different languages. Collaborators can then generate community translations by correcting these automated translations.

End users—people that visit Epistemonikos to search for health evidence—have a simple user interface, available in multiple languages, in which they can enter search terms and see the results ranked by relevance. When the results are presented, users can also see a panel with a summary of all the types of documents available—overviews, systematic reviews, structured summaries and primary studies (Figure 1). When the user selects a document to explore in detail, the interface presents the abstract and a figure—the dolmen—linking to all the additional articles linked to the currently selected one (Figure 2). Using the dolmen, users can explore the literature relevant to that publication and, as a consequence, relevant to the clinical question.

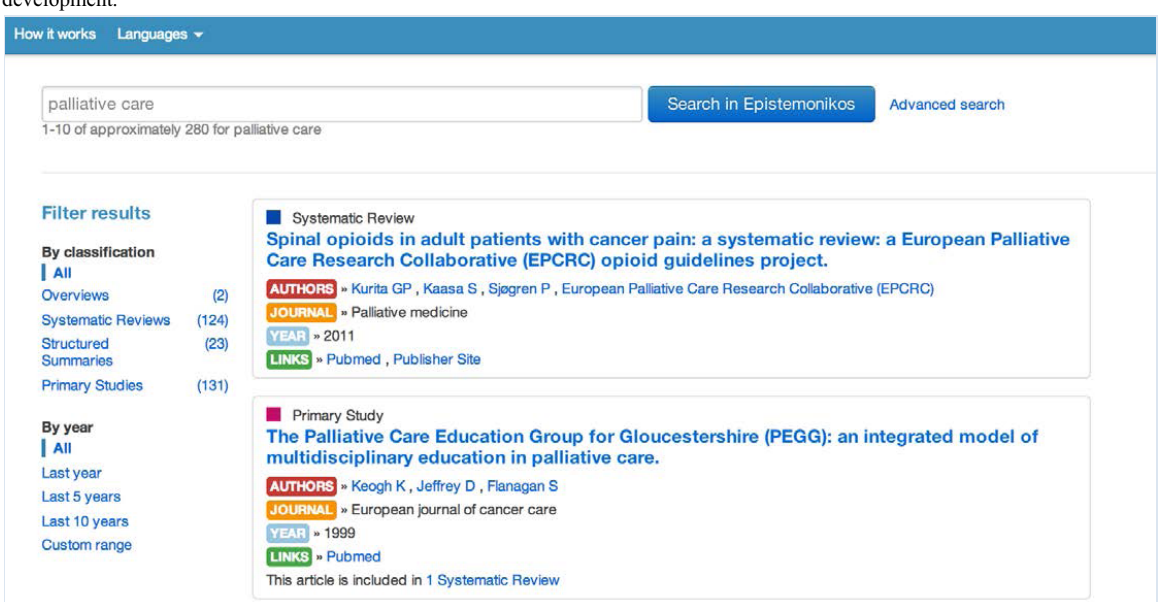

Figure 1- Results displayed after a search for 'palliative care' using Epistemonikos. Left pane shows result categories.

Figure 2- Details displayed after selecting one document. Right pane shows the dolmen with the articles directly linked to the currently selected one. In this example, the user entered the record corresponding to a primary study on multidisciplinary education in palliative care. On the right side the dolmen shows that this article is included in one systematic review and zero overviews. Moreover, the systematic review that included this primary study also includes 15 additional primary studies, which, presumably, answer the same or a very similar clinical question. In this way the user can easily navigate through the studies addressing a single clinical question.

Results

Epistemonikos was officially launched in Spanish on April 20th, 2012 and internationally on August 14th. Epistemonikos has over 250 active collaborators that continuously upload and translate documents. As a result, the database now contains over 115,000 documents, of which 22,649 are systematic reviews, 76,834 are primary studies and 17,144 are structured summaries of evidence. More importantly, the database currently stores 105,707 relations between documents, providing an extensive description of the available literature.

The database is available in nine different languages: Arabic, German, English, Spanish, Italian, Dutch, Portuguese, Chinese, and French. It contains 7,266 official translations, 642 collaborative translations and 899,022 automated translations.

Since its launch, visitors to the site (www.epistemonikos.org) have increased consistently. During November 2012 the site received 43,169 unique visitors and processed 69,437 clinical queries.

Discussion

Accessing the best available research evidence is increasingly complex for people making health decisions or trying to summarize existing knowledge. Failing to do so has a detrimental effect on quality and efficiency.

Epistemonikos tackles many barriers in achieving this objective, through a combination of approaches, both conceptual and technical.

First of all, it provides a one-stop shop for the most relevant information for health decision-making. Multiple databases are searched and screened by a network of collaborators. Keeping track of the several million records that have been and will be assessed requires purposely-developed software. The Epistemonikos project addresses this issue with specialized software to conduct searches, retrieve information, allow collaborative screening and classification of information, and to perform the tasks needed to connect information. Although this software solves several of these tasks, intensive use of human resources is still needed. A collaborative model of multiple volunteers performing simple tasks has resolved many steps that cannot yet be replaced by computers.

Additionally, Epistemonikos takes advantage of a key concept of Evidence-Based Healthcare decision-making—that some evidence is more reliable than others—by putting strong emphasis on systematic reviews and structured summaries using a relational database architected to reflect this fundamental issue. Linking research using systematic reviews as the pivotal element constitutes a proxy for the *health question* the user is trying to answer. This approach relieves pressure on users, who must develop complex search strategies to obtain what they are looking for when using other databases. With our approach, starting from any single record answering the question of interest, users can browse through the relations, or follow the suggestions proposed by the site, to reach the best availa-

ble evidence, even using very simple queries as the starting point.

Finally, Epistemonikos' stepped approach to translations offers several advantages. On one hand, users can search in different languages, since at least an automatic translation is stored in the database. On the other hand, users can read titles and abstracts in their own language. As more official or collaborative translations are collected or generated, this aspect will become more prominent.

To summarize, the Epistemonikos project is addressing several barriers in knowledge translation of health research information, using a combination of novel and established approaches.

References

[1] PubMed [Internet] cited December 10, 2012. Available from: http://www.ncbi.nlm.nih.gov/pubmed

[2] NLM: Electronic Databases & Directories [Internet] cited December 10, 2012. Available from: http://www.nlm.nih.gov/psd/ref/databases.html

[3] Crom V. SUMSearch. J Med Libr Assoc2007;95(4):471-3.

[4] Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc 2000;88(3):265-66.

[5] The Cochrane Library [Internet] cited March 20, 2013. Available from: http://www.thecochranelibrary.com.

[6] Report of the Conference to Promote International Cooperation among Medical Journal Editors, March 13-17, 1995, The Rockefeller Foundation Bellagio Study and Conference Center, Bellagio, Italy. [Internet] cited December 10, 2012. Available from: http://www.wame.org/the-bellagio-report/report-of-the-bellagio-conference

[7] Egger M et al. Systematic reviews in health care. London: BMJ Books, 2001.

[8] Moher D, Tetzlaff J, Tricco AC, Sampson M, and Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007 Mar 27;4(3):e78

[9] Mongo DB [Internet] cited December 10, 2012. Available from: http://www.mongodb.org

[10] Apache Solr [Internet] cited December 10, 2012. Available from: http://lucene.apache.org/solr/

Address for correspondence

Dr. Gabriel Rada
Lira 63
Santiago, Región Metropolitana
Chile. 8330074
radagabriel@epistemonikos.org