

Searching the CRISses

*Wolfgang Sander-Beuermann, Michael Nebel, Wolfgang Adamczak**

Leibniz University of Hannover, *University of Kassel

Summary

The CRIS-systems of the World Wide Web are of extraordinary high value of content because of their “handpicked” data. The disadvantage of CRISses however is the sheer number of such systems, requiring the user who wants to know about research projects of dedicated themes to question all those systems. This situation is well known from other areas of data mining and demands for a more intelligent solution. The solution should sustain the benefits of current systems, which is the “handpicking” of data by local human editors and database managers, and should on the other hand open these treasures of data via a single point of easy access. It will be discussed, what different kind of searching are used what a role the concept of meta-searching can play in these efforts.

Introduction

The development of comprehensive CRIS-systems, like required for ERA, often neglects the searching functionality. Last time it was topic of the 2002 CRIS conference at Kassel, Germany (Hennig and Sander-Beuermann).

EuroCRIS promotes CERIF as a unique standard for all CRISses as the solution of the searching problem (Jeffery 2007). But CERIF is far away from being a standard for all systems (Adamczak and Jacobs 2006). Furthermore the software for a unique CRIS solution is just in a state of a vision of web services applications. Additionally such solution would require the user being familiar with CERIF, which certainly holds just for a small community.

What does a user wish to find at his/her local CRIS and also, other CRISses (Jeffery 2007):

- persons (experts for cooperation or refereeing)
- teams (orgunits) in their context
- projects
- publications, products and patents
- facilities and equipment
- events
- funding sources
- ideas for innovation to wealth-creation

CRIS Systems

The CRIS systems are of extraordinary high value of content because of their handpicked data the user wants to see. But the efforts of building a DRIS-system (Directory of CRIS) show, that real world CRIS systems actually have a very large bandwidth of any kind of solution sometimes very far away from CERIF-standards.

The disadvantage of CRISses however is the sheer number of such systems, each one containing a relatively small amount of data. The provisional list of Marika Meltsas, responsible for DRIS in euroCRIS board contains the URLs of 44 systems in a first step. So the user who wants to know about research projects of dedicated themes has to question all those CRISses. This situation is well known from other areas of data mining and demands for a more intelligent solution. The solution should sustain the benefits of current systems, which is the handpicking of data by local human editors and database managers, and should on the other hand open these treasures of data via a single point of easy access.

This is in contrast to current automatic crawler driven search engines, which often are overloaded with data of questionable substance.

Considering these requirements it seems not suitable to harvest all these data into one (new) search engine or grabbing the data into one (new) catalogue. The local know-how, expertise and responsibility must stay where it is. Furthermore the content of the hidden web, which is a considerable part of CRISses, can by principle never be explored by harvesting. Therefore the only solution which fulfils the requirements for a widely accepted single point of access is the metasearch.

The different kinds of searching

In order to clearly define about which kind of searching we will talk in the following, we distinguish three kinds of search instruments.

The directory

The oldest instrument of organizing information to enable its retrieval is the directory: documents are categorised and stored under the headline of that category. This is the way the Internet Search of Yahoo started many years ago. Web documents had been categorised by humans and were stored in appropriate folders or directories. To retrieve these documents one has to follow the structure of such categories until the level of document listing.

Yahoo has left this approach long ago, because of the high costs of manually categorizing documents and their unmanageable number. Nowadays only the „OpenDirectoryProject/ODP“ (<http://dmoz.org>) still follows this method, based on thousands of voluntary editors who are doing the categorising job. Nevertheless ODP is very far away from achieving a relevant portion of the billions of web documents.

The search engine

Therefore the next instrument which has been developed to organize the information of the Internet was the use of well known database technology, combined with a new device called „crawler“ (or „spider“, „robot“, „harvester“, „web wanderer“, which are all synonyms). The crawler software wanders over the webpages, like a human would do by mouse-click-

ing, but very much faster. During this wandering the crawler transfers each webpage to a central server, which runs the database. When a webpage approaches, the database software does a full text indexing of the text of that webpage, so gradually building a very huge database of webpages. In this way a substantial part of the whole Internet can be stored. To make this database available to the users of the Internet an additional web-interface is connected which transfers data to the end-user by means of the http-protocol.

This arrangement, consisting of the well known database technology, the crawler and the web-interface has been called „search engine“ in the early 1990's.

The metasearch engine

After several of the above described search engines had been developed and were operated on the Internet the problem aroused that the user often had to query several search engines to find what he was looking for. No single search engine is able to store the whole Internet in its database (although nowadays Google tries to come close to this). Because of the necessity to query many single search engines the next step in the development of searching the Internet was the metasearch: instead of manually querying several engines one after the other, this cumbersome task was overtaken by a program, called the metasearch software. This program replaces the manual job of querying and composing the different results from lots of search engines into one interface.

The first metasearch service on the Internet was the metacrawler, developed at the Computer Science Department of the University of Washington. Shortly after that and independently developed the first metasearch engine in Europe went into operation at Hannover University, Germany. It is the German metasearch named MetaGer (<http://metager.de>), which is still one of the top 10 mostly used search engines in Germany.

This technique makes use of already available data sources on the web by transferring the user queries to these data sources, collecting their results, merging duplicate hits, and presenting the combination of all the results in HTML via a familiar browser window. In the case we consider here, the data sources on the web will be the already available CRIS-systems. In this way any relevant system can be included, even if it otherwise is not available to search engine crawlers and harvesting (part of the hidden web). To exclude any misunderstandings we explicitly state, that metasearching has nothing to do with the technique, which in this community is commonly called „harvesting“:

Metasearching means that we do not harvest/collect/gather the original data of web documents, but we collect the results of already available web search engines (here: CRISses). Hierarchically seen, this is one layer above the original data. Metasearching can be transferred and adopted to the CRISses, which we consider here. The technique is based purely on the http-protocol, HTML and (eventually) XML only. The user and the CRIS system maintainers can stay in their familiar or special environment and act in known ways with knowledge management.

Co-operations

The workshop will take place in co-operation with Elly Dijk (Royal Netherlands Academy of Arts and Sciences) and Adrian Price (The Faculty of Life Sciences, University of Copenhagen, Denmark) to show different solutions of searching.

Structure of the workshop

- Short introduction by Wolfgang Adamczak
- Elly Dijk will present searching concept in NARCIS (NARCIS is a search service which gives free access to academic research output and research information in the Netherlands – www.narcis.info). In NARCIS the national CRIS (the Dutch research Database) is being harvested by using the OAI-PMH protocol.
- Adrian Price will talk about data, reporting and searching regarding to PURE (PURE is a Repository and CRIS platform, which is used by research and higher education institutions in Denmark)
- Wolfgang Sander-Beuermann and Michael Nebel will introduce the concept of meta-searching
- Discussion is opened

What should be discussed in the workshop:

- Does searching can help to establish a one-stop-shop?
- Disadvantages of searching - could we overcome them?
- How to build up catalogues
- How to get statistics by this way
- No access to hidden web?! (who will hide something or who does not know how to organize access to search-engines?!)
- How to compare data from different CRISses (e.g. research documentation and funding documentation)
- How topical searched data can be
- Does the CERIF-standard can help exploring CRISses?
- What are the advantages and disadvantages of meta-searching, harvesting by using the OAI-PMH protocol and other kind of searching?

Literature

Adamczak, W., Jacobs, N. (2006): *Results of the membership survey*, euroCRIS Members Meeting, Copenhagen.

Hennig, D., Sander-Beuermann, W. (2002): Data Collectors meet Data Suppliers on the Internet, in Wolfgang Adamczak and Annemarie Nase (Eds.), *Gaining insight from research information, 6th International Conference on Current Research Information Systems (CRIS 2002)*, Kassel.

Jeffery, K. G. (2007): *CRIS Architectures for Interoperation*, workshop at euroCRIS Members Meeting, Vienna.

Contact information

Dr. Wolfgang Adamczak

Gottschalkstr. 22, 34109 Kassel, Germany

e-mail: adamczak@uni-kassel.de