Information & Services Integration

Filipe Manuel dos Santos Bento

Documentation Services
University of Aveiro
3810-193 Aveiro, Portugal
Tel: +351 234 370 346
E-mail: fil@ua.pt

Introduction

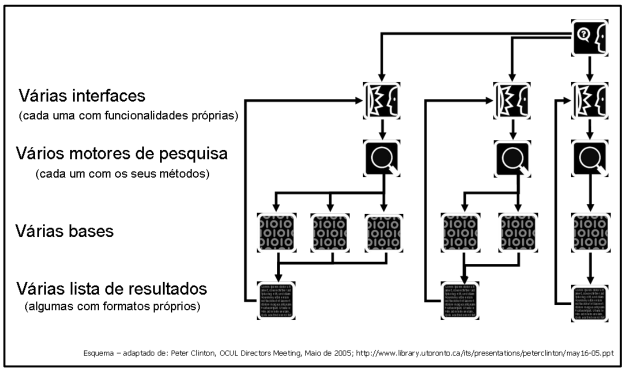
Web search engines have on their central database data harvested from the so called *visible web,* the set of static pages, for which their contents is always the same for a certain URL. A huge variety of contents that dynamic web pages can have, contents that change in response to a query or action submitted by the user, is not visible to the *web crawlers* that feed the search engines like Google or SAPO. These contents are said to belong to the *deep web* (also called the *invisible web*).

Federated / distributed search engines act as an *Information Integrator* by searching and retrieving the contents that lay down in the so called *deep web*, contents that are stored in some kind of warehouse (database) and that are only shown in response to query submitted by the end-user, displaying contents accordingly. These *federated search engines* do so without having previously harvested the records from the different sources, gathering data in real-time immediately after the user request.

To do so, these engines need to recur to standard protocols, such as z39.50 for the exchange of bibliographic records, and XML data exchange when consuming web services from the remote server. Exceptionally, when the remote server does not has a z39.50 server or any kind of web service to honour remote clients' requests, HTML scraping might be the only way to retrieve the desired data. The integrator analysed in the case study presented in session 1, ColCat, and for these last cases, it simulates a human interaction with the remote server, requiring such a training process in order to record the necessary steps so that the desired data is retrieved. Furthermore, additional training is required, so that it knows how to deal with successful searches but also with the ones that do not retrieve any record at all.

Ultimately, information integration via distributed search aims at a higher efficiency and quickness in the information search process, performing a cross search through several heterogeneous sources, without the need for special source related search skills: one interface, many sources, one result list.

The following schemes illustrate the search process flow for both methods: with and without federated / distributed search.


Várias interfaces
(cada uma com funcionalidades próprias)
Vários motores de pesquisa
(cada um com os seus métodos)
Várias bases
Várias lista de resultados
(algumas com formatos próprios)

Esquema – adaptado de: Peter Clinton, OCUL Directors Meeting, Maio de 2005; http://www.library.utoronto.ca/its/presentations/peterclinton/may16-05.ppt

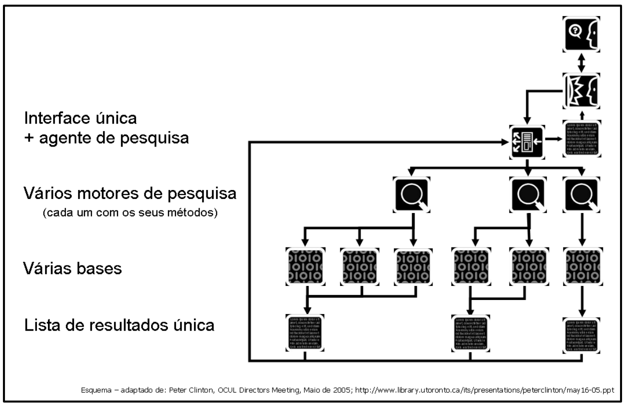
Figure 1: without federated search (adapted from Clinton, 2005)

No matter how efficient may be the ILMS (Integrated Library Management System) for the different OPACs, the huge diversity of options that a user may encounter when searching through several systems translates into a considerable set of difficulties that he/she as to overcome in order to retrieve the desired records.

A slow process, disorientation felt by the end-user (and to some extent, by the information professionals) should be on the top of the list of such difficulties, which is probably headed by such a simple thing as not have a general view, an updated reference of all entries where to search (URLs of the sources, for instances).

Performing a search using a distributed searc engine, by submiting the query in a single interface (and only once), the user can abstract himself from the diferent levels of complexity refered above and just focus on what to search for.

In this mediated process between the user and the different sources searched, a new role emerges: search agent. It is this agent that translates the user query, submitted in the user-friendly interface of the integrator, to the specific syntax that the remote sources need in order to retrieve the records that best suite the search submitted. Also this Agent, in the nest step in the process, deals with these records, some very different in structure and content, in order to present them in an integrated manner to the user.


Interface única
+ agente de pesquisa
Vários motores de pesquisa
(cada um com os seus métodos)
Várias bases
Lista de resultados única

Esquema – adaptado de: Peter Clinton, OCUL Directors Meeting, Maio de 2005; http://www.library.utoronto.ca/its/presentations/peterclinton/may16-05.ppt

Figure 2: with federated search (adapted from Clinton, 2005)

Having the records in a format that can be object of element extraction (title, author, year, keywords [subject], etc) enables the next level in integration: resource services' integration. This data can be used to feed another system, that shall perform some operation with it in a automatic manner (for instance, reference management software, ILL system for Libraries, full text / acquisition request, photocopy request, etc).

Bibliography quick review

Mark Hinnebusch and Charles Lowry (Hinnebusch ; Lowry, 1997) examine the Z39.50 protocol, as the standard for communications between library systems. They perform a comparison of traditional and current capabilities, with an interesting part regarding the effect of the insistence of librarians and library managers on use of this protocol on market development (software developers). Additionally a context background history is presented; features of the latest version, 3.0 (z39.50-1995), and of the various community profiles, enrol this article.

Clifford Lynch, a very renowned author in the maters of resource sharing and search integration, in his article "Building the infrastructure of resource sharing: Union", published in the specialized journal *Library Trends* (Lynch), focuses on two approaches to effective resource sharing, which permits users to locate materials of interest in both print and electronic format. The advantages of implementing Union catalogs and Z39.50-based distributed search systems are examined, especially the limitations to each approach. As a conclusion, he states that the two approaches should be considered complimentary rather than competitive.

In September 1998, Jim Rapoza reports for *PC Week* magazine (Rapoza, 1998) about one of the first commercial products ever to perform metasearches for websites: Infoseek's Express.

More recently, Jarogniew Rykowski, in the article "Management of Information Changes by the Use of Software Agents" published in *Cybernetics & Systems* journal (Rykowski, 2006) proposes a new approach to personalize mass-scale information systems. The approach is based on the technology of software agents, and the Agent Computing Environment (ACE). According to the author, these agents are programmed and managed by the end-users. Basic tasks of an agent are related with efficient, individualized access and monitoring of selected information sources. Once developed and moved to given location, an agent performs autonomously given tasks, pre-programmed according to its owner's needs. Each ACE agent may adjust its behavior to the environment where it is executed at a moment, as well as to some user-independent restrictions, e.g., communication costs, limitations of end-user hardware, etc. Several agents may be logically combined to create a complex agent able to fulfill more sophisticated tasks. Complex agents may be settled both statically, as a result of a user request, and dynamically, as a result of environmental changes. The approach may be used in such domains, as e-banking, sport and cultural news, health and telemedicine, shopping and e-commerce, logistics, e-learning, etc. Using ACE agents as external brokers to distributed systems makes it possible to personalize behavior of such closed and highly secured environments as bank databases, company's internal systems, etc.

Focusing on the user behaviours and needs and as they should be best served by Library professionals, Janet Balas (L. Balas, 2006) does an interesting analysis of the resemblance between the emergent one-stop, do-it-yourself shopping trend in the U.S. and patrons' information needs and how this trend affects libraries. According to the author, she noticed two contradictory trends in retail: self-service and individual customer service. For Library professionals the integrated library system appeared to be the ultimate in convenience for both the profession and their patrons. The appeal of one-stop searching is obvious, but it seems that it is not an easy goal to achieve. Se reckons that these professionals should continue discussing and searching for solutions and reference services to make sure their patrons always find what they are seeking. One interesting question she poses: "Can we ever build the perfect search engine?".

KM World reviews on their article "Alfresco 2.0" (Alfresco 2.0, 2007) the content management solution Alfresco 2.0, an *Open Source Alternative for Enterprise Content Management* (ECM), providing Document Management, Collaboration, Records Management, Knowledge Management, Web Content Management and Imaging. Looking at it's features at this product web site, http://www.alfresco.com/, a relevant feature for the current study is found; it implements *OpenSearch*, that enables Alfresco to search across multiple Alfresco repositories and external wikis, blogs and news feeds.

Distributed search can be applied in many fields of science to bring the extra value that an integrated search can achieve. In the second issue of this year's edition of the *Journal of the Medical Library Association,* on their article "Evidence-based Medicine Search: a customizable federated search engine", Paul Bracke *et al.* (Bracke; Howse ; Keim, 2008) report on the development of a tool by the Arizona Health Sciences Library (AHSL) for searching clinical evidence that can be customized for different user groups. The AHSL provides services to the University of Arizona's (UA's) health sciences programs and to the University Medical Center. Librarians at AHSL collaborated with UA College of Medicine faculty to create an innovative search engine, Evidence-based Medicine (EBM) Search, which provides users with a simple search interface to EBM resources and presents results organized according to an evidence pyramid. One interesting outcome or a conclusion that may be withdrawn: informal and anecdotal feedback from physicians indicates that EBM Search is a useful tool with potential in teaching evidence-based decision making. Overall conclusion is that a tool such as EBM Search, which can be configured for specific user populations, may help lower barriers to information resources in an academic health sciences center.

Bringing back the focus to information search and resource discovery at the Libraries level, Steven Baule does the "technology Connection" on a concise but rather complete article about the state of the art, published in March 2007's edition of *Library Media Connection* (Baule, 2007). This article discusses the use by students of a variety of non-uniform search interfaces to find materials within a library collection. Baule affirms that Library Collections can compete with Google and other Web-based search engines, by implementing user interfaces which should be as simple and easy to use as Google. Furthermore, metasearching is identified as the solution to the problem of having too many databases with multiple search interfaces.

Relating the previous to the one before, Beth Evans' article "Library 2.0: The Consumer as Producer" for *Information Today*, this month edition (October 2008) (Evans, 2008), suggests to "let the public pave the way" as a way "how libraries can step into socializing".

On a more scientific approach of the "bits and bytes" that should be tweaked to retrieve pertinent records, starting with the right query construction and going all the way with ontologies, hierarchal classification, taxonomies, collection selection, search engines and data mining, John D. King *et al.* from the School of Software Engineering and Data Communications, Queensland University of Technology, Australia, in their article "Mining world knowledge for analysis of search engine content" published in *Web Intelligence & Agent Systems* (King [et al.], 2007) present an automatic learning method which trains an ontology with world knowledge of hundreds of different subjects in a three-level taxonomy covering the documents offered in their university library. In a next step they mine that ontology to find important classification rules, and then used these rules to perform an extensive analysis of the content of the largest general purpose internet search engines in use today. Instead of representing documents and collections as a set of terms, they represent them as a set of subjects, which they defend to be a highly efficient representation, leading to a more robust representation of information and a decrease of synonymy.

In fact, without these aids some authors argue that metasearch products or products that have the ability to search multiple resources simultaneously have severe limitations. Marshall Breeding relates that the recent debut of Google Scholar has convinced him that the architecture that underlies the traditional library approach toward search and retrieval cannot succeed as the sole system that librarians rely on to simultaneously search multiple electronic resources (Breeding, 2005). He exposes that current strategy of metasearch that depends on live connections casting queries to multiple remote information sources cannot stand up to search systems based on centralized indexes that were created in advance based on harvested content. He thinks of these competing approaches as distributed search and centralized search, respectively. In closing, the author does not discourage librarians from making good use of the metasearch products available as of 2005. While not perfect, according to him, they go a long way toward the goal of providing user-friendly ways to search the electronic resources provided by libraries. According to the author, it might be the right time by then to seriously reconsider how librarians approach the problem of creating a search environment for library-provided electronic resources.

News from the actual worldwide panorama, Josh Hadro reports in a one column "Academics Add Federated Search" article for the specialized periodical publication *Library Journal*, June 1, 2008's edition (Hadro, 2008), that the University of Oxford, Stanford University and the University of Cambridge have announced that they will adopt federated search products in the hopes of improving their academic communities' access to electronic research materials. Oxford and Cambridge will implement Metalib from Ex-Libris and WebFeat and Stanford has selected the Explorit Research Accelerator from Deep Web Technologies.

At a national level, the author submitted an article to the IX National Congress of Librarians, Archivists and Documentalists entitled "ColCat: integrate to ease" (Bento, 2007), a follow-up of the poster presented at the 2004's VIII National Congress of Librarians, Archivists and Documentalists, "ColCat: Metabibliographic Distributed Search" (Bento, 2004). In this article, motivation, context, development, dissemination, features, demo, future and curiosities, are the main topics, advantages and disadvantages of each method are presented in comparative analysis "Integrated Search versus Integration of Records". A third method, between the two previous, integration of recordings via Metadata Harvesting, is shown.

Finally, an important endnote for the an entire edition of *Internet Reference Services Quarterly* that was dedicated to Federated Search putting the focus of the first issue on user experience, perceptions, designing for users and overall usability (Federated Search: Solution or Setback for Online Library Services 2007). This issue was complemented with the follow-up edition (Federated Search: Solution or Setback for Online Library Services, part II, 2007) focusing on Learning and Teaching aids provided by federated search and all the sub-systems that can be built around it.

References

Alfresco 2.0. - KM World. 10998284. Vol. 16, n.º 4 (2007). p. 1/5p.

Baule, Steven - Data, Data Everywhere, and How Do You Sort Through It? Library Media Connection. 15424715. Vol. 25, n.º 6 (2007). p. 3p.

Bento, Filipe M S - ColCat: integrar para facilitar. IX Congresso Nacional BAD – Bibliotecários, Arquivistas e Documentalistas. (2007).

Bento, Filipe M S - ColCat: Metabibliographic Distributed Search". VIII National Congress of Librarians, Archivists and Documentalists. (2004).

Bracke, Paul J.; Howse, David K.; Keim, Samuel M. - Evidence-based Medicine Search: a customizable federated search engine. Journal of the Medical Library Association. 15365050. Vol. 96, n.º 2 (2008). p. 6p.

Breeding, Marshall - Plotting a New Course for Metasearch. Computers in Libraries. 10417915. Vol. 25, n.º 2 (2005). p. 3p.

Clinton, Peter - Federated Searching: Extending our Reach. (2005).

Evans, Beth - Library 2.0: The Consumer as Producer. Information Today. 87556286. Vol. 25, n.º 9 (2008). p. 3p.

Federated Search: Solution or Setback for Online Library Services - Internet Reference Services Quarterly. Vol. 12, n.º 1/2 (2007). p. 1-236.

Federated Search: Solution or Setback for Online Library Services, part II. - Internet Reference Services Quarterly. Vol. 12, n.º 3/4 (2007). p. 237-430.

Hadro, Josh - Academics Add Federated Search. Library Journal. 03630277. Vol. 133, n.º 10 (2008). p. 1/5p.

Hinnebusch, Mark; Lowry, Charles B. - Z39.50 at ten years: How stands the standard? Journal of Academic Librarianship. 00991333. Vol. 23, n.º 3 (1997). p. 5p.

King, John D. [et al.] - Mining world knowledge for analysis of search engine content. Web Intelligence & Agent Systems. 15701263. Vol. 5, n.º 3 (2007). p. 21p.

L. Balas, Janet - Does One-Stop Searching Really Serve All? Computers in Libraries. 10417915. Vol. 26, n.º 9 (2006). p. 3p.

Lynch, Clifford A. - Building the infrastructure of resource sharing: Union. Library Trends. 00242594. Vol. 45, n.º 3 p. 14p.

Rapoza, Jim - Express: Power searching. PC Week. 07401604. Vol. 15, n.º 38 (1998). p. 3/4p.

Rykowski, Jarogniew - Management of Information Changes by the Use of Software Agents. Cybernetics & Systems. 01969722. Vol. 37, n.º 2/3 (2006). p. 31p.