SBBL/CILEA a Metacrawler Service

P. Arvati(*), F. Clementi(**), G. De Vito(**), R. Ferrari(*), F. Ferrario(*), P. Mozzati Gemelli(**), V. Pistotti(**), E. Rodi(*)

(*) CILEA, Segrate
(**) SBBL

Abstract
SBBL/CILEA Metacrawler is an in-house developed service that proposes as integrator of database and distributed resources in the within of the Digital Libraries.
SBBL/CILEA Metacrawler service performs querying and using local or remote databases like if they were web services using their gateway interfaces; for example: Medline (NLM PubMed) as remote database, Cancerlit and Cinahl as information resources on CILEA servers.
SBBL/CILEA Metacrawler service performs suggesting to the final customer the greater number of link directed to the primary resources and it gives more relevance to CILEA digital library services: Elsevier catalog, Kluwer, ISI Current Contents.
It has been realized an original and personal interface application that allowed a professional retrieval on the available databases, the management of the queries done and a Linkout function to redirect toward the full-text resources.
SBBL/CILEA Metacrawler service is offering to Italian researchers and universities the possibility to use in a unified view different databases for searching and several mirrors for accessing to full-text resources.

Keywords: Beni culturali, Sanità, Medicina, Digital Library.

Foreword

The growing need for users to research and exchange information and the growing willingness of publishers to promote the distribution of their materials in electronic formats have created new requirements, which have led CILEA to formulate a single product both to provide electronic editorial services on the network and serve as an interface for scientific users.

CILEA & SBBL's motive for developing a new product for digital editing and research was to provide users with the opportunity to consult and interrogate different data banks contemporaneously, and interface with remote service providers.

Important incentives to devise the new applications that resulted in the creation of MetaCrawler, emerged from previous collaborations with the SBBL group, which has guided CILEA programmers in the creation of new professional and competitive solutions on the Digital Library market.

The primary aim was to provide personalized access to our services and all the Medline and PreMedline data banks made consultable through the PubMed service:

(http://www.ncbi.nlm.nih.gov)

placed at our disposal by the National Library of Medicine (http://www.nlm.nih.gov).
It was clear from the outset that there was a need to integrate PubMed with other services and data banks managed directly by CILEA. Subject to agreement, these are accessible from outside.
The project's objectives can be synthesized in the following bullet points:

- to offer added value to its users in comparison with existing services through providing a unique interface tailored to

the requirements of professional research;
- to offer managerial tools for research strategies and document recovery, alongside other personalized services;
- to provide users with various pathways through which they can access full text articles or obtain from the system information regarding the services at their disposal.

CILEA - http://www.cilea.it

CILEA (The Lombard Interuniversity Consortium for Automatic Data Processing) is a non-profit organisation that pools the resources of nine Lombard Universities:

1. *Università degli Studi di Bergamo*
2. *Università degli Studi di Brescia*
3. *Università Commerciale "L. Bocconi"*
4. *Università Cattolica del Sacro Cuore*
5. *Politecnico di Milano*
6. *Università degli Studi di Milano*
7. *Università degli Studi di Milano Bicocca*
8. *Università degli Studi di Pavia*
9. *Università degli Studi dell'Insubria*

Figura 1 - CILEA consorted Universities' sites

Despite not yet being members of the consortium, the IULM and LIUC universities take part in the initiatives of CILEA.
CILEA's activities are subject to the supervision and control of the Italian Ministero dell'Istruzione, dell'Università e della Ricerca (the ministry responsible for university education, research).
CILEA, which was established in 1974, provides Information Technology services on behalf of universities and related organisations, alongside other public organisations and enterprises. It also provides professional advice for both the planning and dissemination of advanced technologies in the following fields:

- High Performance Computing
- Networking Services
- Multimedia Data Bases
- European Research Projects
- Advanced Courses

In particular, CILEA is concerned with the Medicine and Digital Library sectors.

Medicine: with regard to the use of information technology (IT) in medicine, CILEA in cooperation with the Bioengineering Department of the Politecnico di Milano carries out activities in different fields, ranging from: sanitary information processing to workshop organisation and management, telemedicine and medical image processing.

Library Automation & Digital Library: another CILEA service relates to library automation, which is provided in cooperation with some of the most significant national and European projects. CILEA is a nodal point in the SBN (Servizio Bibliotecario Nazionale). Several libraries are directly connected to CILEA, which provides them with professional advice for the use of local library networks, as well as technical aid for planning them.
For some years CILEA has been running a new project called the CILEA Digital Library whose function is to make the world's most important scientific journals in electronic format available to the scientific community.

CILEA'S RESOURCES IN ELECTRONIC EDITING

CILEA'S contribution to the realisation of a "digital library" for the national scientific community, has, for some time been, the CILEA Digital Library service.
The service allows for acquisition, memorization, conservation and facilitated access to digital documents (Electronic Editing), archives of preprints, technical reports and data banks.

By way of an overview we would like to take a brief look at the services CILEA can offer users interested in the Digital Library field so that they can better understand the background to the introduction of the CILEA METACRAWLER and the development opportunities it enjoys for the immediate future and further afield.

1. E-JOURNALS

One of the most significant services that CILEA has provided for several years relates to E-Journals; the service allows access to some of the most important journals in electronic format, published in Europe and throughout the world.

1.1 ELSEVIER

CILEA has entered into a contract with Elsevier to put on line its own server (mirror **SDOS ScienceDirect OnSite**) to access the contents of the entire Elsevier catalogue starting from 1995.
Over 1400 titles are available to CDL – Elsevier users. At the moment 32 organisations are members of the consortium.

1.2 KLUWER

CILEA has entered into a contract with Kluwer to access the publisher's entire catalogue (about 800 titles); 27 organisations are currently members. At present subscribers have online access to the Kluwer server:
http://www.kluweronline.com/.
We are currently loading Kluwer data onto the SDOS CILEA server, which means that Kluwer data will soon be accessible alongside Elsevier data.

1.3 ACADEMIC PRESS

Academic Press (Harcourt) was recently acquired by Elsevier, which means that over 170 of this publisher's titles will soon be part of the Elsevier package.
As of today there are 16 subscriber organisations that have temporary online access to the publisher, while they await CILEA to complete its loading onto the SDOS server.

1.4 BLACKWELL SCIENCE/PUBLISHERS

A contract was entered into with the publisher Blackwell Science to access the titles in their Science & Medicine Collection and/or their Social Science & Humanities Collections (500 titles in total). At the moment, there are 14 member organisations that can access via the remote server. The metadata will soon be loaded onto the SDOS server and CILEA will have the full text documents at their complete disposal.

1.5 WILEY

A contract is being drawn up to access the entire catalogue (It includes over 300 titles).
Access is available via a remote site until the metadata is loaded onto the SDOS server. 23 organisations participate.

1.6 OTHER E-JOURNALS

Service	Jou.	Org.
ACS - American Chemical Society	30	19
ACM - Association for Computing Machinery	--	14
IOP- Institute of Physics	--	9
JSTOR - Journal Storage Project	--	12

2. DATA BASE

In order to promote its Digital Library, CILEA's second point of emphasis has been the development of its databases in various scientific sectors. In the list we are about to present, databases from the field of medicine are highlighted.

2.1 CANCERLIT

CANCERLIT®, compiled by the United States National Cancer Institute from the cancer-related references in the National Library of Medicine's MEDLINE, is the world's foremost cancer literature database.
CANCERLIT® is a bibliographical database that contains more than 1.5 million references and abstracts from over 4,000 different sources, including biomedical journals, proceedings, books, reports, and doctoral theses. Areas covered include: diagnostic treatment and procedures; epidemiology; risk factors and prevention; molecular and cell biology; cancer virology and immunology; carcinogens and carcinagenesis; and anti-cancer drug development.

2.2 CINHAL

The CINAHL® database provides authoritative coverage of the professional literature in nursing and 17 allied health disciplines and also covers consumer health, biomedicine, alternative therapy, and health sciences librarianship.

The database has full texts from selected state nursing journals, standards of practice, practice acts, critical paths, research instruments, and government publications.

2.3 EMBASE

Still in the process of agreement. Installation on SDOS is anticipated in the near future.

2.4 EBSCO

Terms of collaboration agreement are yet to be defined.

2.5 ISI

- Web of Science (WoS)
- Journal Citation Report (JCR)
- Current Contents Connect (CCC)

CILEA has entered into a three-year contract with ISI for the data bases listed above.
WoS and CCC can be found on the CILEA server while the JCR data bank is on the ISI remote server.
Over 40 organisations have subscribed.

2.6 OTHER Data Bases

Data Base	Host	Org.
WSS Worldwide Standards Service Plus	OnSite	7
CAS-Chemical Abstracts Service - SciFinder Scholar	OnLine	19
EI – COMPENDEX: Engineering Information Compendex	OnSite (soon)	6
LEXIS-NEXIS	OnLine	23
CIC - Consorzio Italiano Crossfire	OnSite	35
CSA - Cambridge Scientific Abstracts	OnLine	6
Gale- LRC - Literature Resource Center	OnLine	1

METACRAWLER: ARCHITECTURE

The CILEA MetaCrawler is among the services offered by CILEA for the Digital Library. Its object is to combine all services into a single information access.
The project was made possible due to developments in the editorial sector and the professional contribution of SBBL.

The basic concept that the CILEA Metacrawler team wanted to realize was to enable users to perform enquiries and use local or remote databases as if they were web services accessed via gateway interfaces.
In addition to the NLM-PubMed Medline service, which is a Metacrawler component considered to be one of the best sources of information within the medical sector, CILEA has also acquired the Cancerlit and Cinah data bases to add to the other Metacrawler information services on the CILEA server.
Before describing the service in more detail and listing all its uses and functions, we shall emphasise the 3 most important elements that characterise the structure of the CILEA Meta-Crawler:

- an interface for searches and interrogations modelled on the technical and functional characteristics of the information system of the service;
- handling and customisation of surfing with regard to search results and the saving of the enquiry strategies executed;
- connection to multiple on line services that can provide access to full-texts.

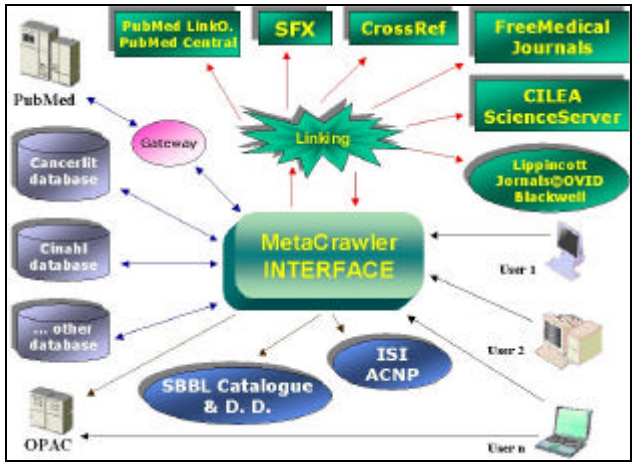

Figura 2 - CILEA MetaCrawler Schema

1. THE INTERFACE

The heart of the service is the interface, which includes a series of functions, applicable modules and customisations. It allows the user to formulate searches within variously structured services provided by the electronic editing.
Consequently, there is no apparent distinction between the services provided by the consortium and those provided from outside.
Research into the interface was carried out on two fronts: on the one hand by learning from

the widespread experience already acquired on the network. On the other, through direct feedback from our users, that has allowed us to tailor the service to user's requirements and expectancies.
Another fundamental characteristic of the interface structure is its ability to send a particular search to different services and databases.
The user is provided with the necessary tools to carry out an interrogation using ordinary syntax, which is translated, in sequence, into the specific languages of the various services.
Interrogations using Metacrawler must accord with certain syntactical rules: the field of research must be defined and rules must be applied to the use of logical connectors and brackets in the establishment of computational precedence. Certain criteria and values are used for a more accurate search (Limits) and combining individual searches (Combine), as well as other options. Every search is recorded in a "history" table and the system transforms the syntax corresponding to the user into a more generalised and abstract meta-syntax.
At this point the system utilizes the appropriate parsing function, which is compatible with any service or database that a search is directed towards, translating the "abstract" syntax into the specific languages of the search engines (for example: PubMed syntax or syntax to interrogate databases using SQL language).

2. HISTORY

One of the most important characteristics that the research structure is founded upon is the memorisation and reuse of interrogations built up in an incremental manner by users.
The logic of an incremental build-up of searches revolves around the use of a history that is memorized for each user by the server.
In order to manage their own histories users have various functions at their disposal:

1) *Save Session*: allows one to save the current history under an assigned name.
2) *Load Session*: allows one to load a previously saved history while working on a current session.
3) *Delete old Sessions*: allows one to eliminate the history saved by the user.
4) *New Session*: allows one to cancel a history or set it to zero.
5) *Logoff*: allows one to exit the service, eliminate the current client server session or return the present history to zero.

3. LINK TO FULLTEXT (LINKOUT)

The LinkOut service was one of the most highly developed sections when the CILEA Metacrawler was assembled.
A truly self-sufficient service that offers on one page hypertext links (titles, authors, journals, etc.) as well as the largest number of WEB search paths through which one can trace a full text article. It is a content delivery service offered by CILEA that can be configured in accordance with the user's requirements.
In order to search successfully using this function, one only needs to provide the minimum information on a document (reference): ISSN of journal, volume, issue and first page of the article.
Here is a brief list of the services currently connected to the LinkOut function:

1) Distributor of full-text listed by Pubmed (http://www.ncbi.nlm.nih.gov)
2) Link to PubmedCentral, a new full text distribution service set up by the National Library, which is constantly being expanded (http://www.pubmedcentral.nih.gov)
3) FreeMedicalJournals accessable (http://www.freemedicaljournals.com). At the moment there are 130 publications.
4) Direct connection to the full-text article on ScienceDirect, which is obtainable through CILEA or the Elsevier site (http://scienceserver.cilea.it).
5) Direct connection to the full-text articles of Lippincott W&W journals from "Journals@OVID".
6) Direct connection to the full-text articles of Blackwell journals through Sinergy and Swetsnet.
7) CROSSREF (http://www.crossref.org) offers the possibility of being redirected in a single operation to the main distributor of the article via DOI (http://www.doi.org/) when article references are provided. CILEA presently has a Consortium Library type contract with

CrossRef that provides access to retrieval services.

8) Connection to user's particular resources. For example: connection to the SBBL collective catalogue of periodicals and Document Delivery functions or connection to the ACNP service (Italian Periodicals Catalogue: http://acnp.cib.unibo.it) of the CIB group (Centro InterBibliotecario Università di Bologna).
9) Connection to other available resources through CILEA: an example is the connection to the ISI Web of Science database.

Other market products and services are under consideration. Among the most likely candidates for inclusion is the SFX product (http://www.sfxit.com) provided by Ex Libris.
The LinkOut function will be expanded to interface with dozens of full text distribution services.

TECHNICAL SOLUTIONS

The technological solutions selected for the realisation of the SBBL/CILEA MetaCrawler came from SERVLET, which provided the server technology and JAVA that supplied the language. They proved most appropriate in addressing the needs of analysts and programmers.
The JAVA programming language was chosen primarily because it could satisfy certain specific needs of the network protocol as well as guarantee an exchange of information between WEB services.
A secondary consideration was that the software incorporated desirable features such as reusability, maintenance and portability.
In the case of this project there were specific objectives, namely: the facility to reuse parts of the written code for other projects deploying similar network technologies (reusability) and access to data banks; the possibility to alter an already written code rapidly and cheaply (maintenance) in order to correct eventual testing-errors and to move directly from one release code to implement the next without having to rewrite the modules.
Not least, care was taken over the installation of the service onto machines equipped with different hardware and software (Portability) so that they could provide:

- the option of using the product for a short period at low cost and later moving over to more advanced systems using different technologies.
- the option of simultaneously running installations, which are slightly different from each other (perhaps for different users) on different machines.

From a general IT perspective, choices were made during the evolution of the project that were unrelated to the underpinning hardware systems.
At the moment, Orion Application is being used as a server by SERVLET (http://www.orionserver.com/).
It is well regarded in its sector and was chosen by Oracle as the platform for the JAVA/SERVLET/JSP.
The possibility of making the service available on other platforms conforming to the same standard and, in particular, the IBM WebSphere, is being considered.
SBBL/CILEA MetaCrawler uses DBMS MySQL 3.23 for its own control of users and researches, installed in SUN Solaris 2.7.
The Cancerlit and Cinahl databases however were derived from the DB2 Universal Database (DB2 UDB) 7.2 version.
The DBMS is equipped with some of the most important Extenders. Of particular note are the DB2 Net Search Extender (full-text search module for bibliographical databases) and the DB2 XML Extender (module for information processing in XML-format). JAVA is also used for these databases (servlet, jsp.).

SBBL: THE LOMBARD BIOMEDICAL LIBRARY SYSTEM

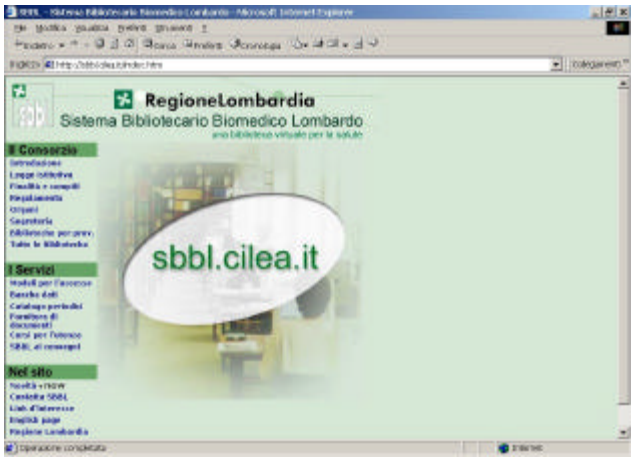

Figura 3 - SBBL http://sbbl.cilea.it

1. PRESENTATION

SBBL (The Sistema Bibliotecario Biomedico Lombardo) is an advanced system for the sharing and distribution of information as well as the formation and production of new knowledge for the biomedical and hospital sectors.

It was set up under regional law in 1994 (L. R. n 41-12.12.94) as the Virtual Library for Health and comprises:

- 16 libraries (Polo SBBL) that deliver the service and have the task of keeping abreast of the continuous developments in the health sector and making the information available (Fig. 4).
- 121 organisations throughout Lombardy already avail themselves of the services (Hospital Departments, Universities and Research Institutes, Zoological Institutes for preventative medicine, The Italian National Health Service, Scientific Foundations, Medical Councils and Nursing Schools) (Fig. 4).

2. DISTRIBUTION AND STRUCTURE OF SBBL

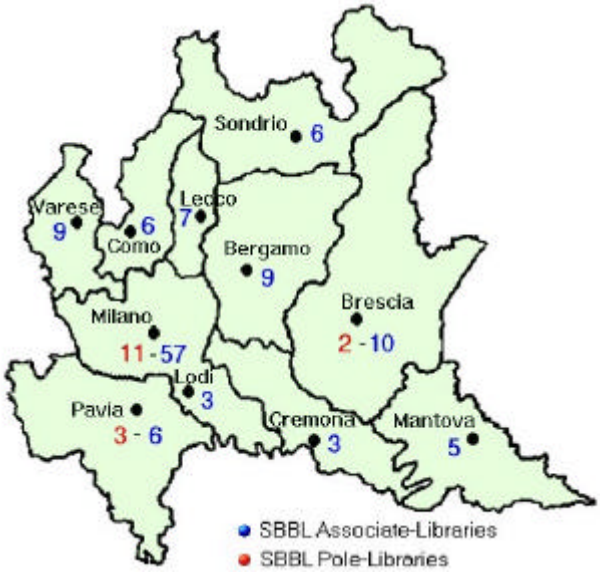

Figura 4 - The SBBL distribution

3. OBJECTIVES

The first objective reached by SBBL was the creation of a **Collective Catalogue of Biomedical Periodicals** (on paper), including about 6000 journals owned by the 16 Lombard libraries with the richest bibliographical resources:

- in 1996 these were on paper;
- in 1997 on CD-ROM;
- and by 1999 on-line.

The principal aim was to create a **free exchange of published articles** among the participating Lombard libraries. Subsequently:

- the most important international data banks were acquired (Medline, Embase, Cinhal, etc.) (1998)
- a network was set up so that they could be consulted by accredited Lombard organisations - Hospitals, Research Institutes and the Italian National Health Service (1998)
- contracts were entered into with about 2000 electronic journals with access to complete articles (1999)
- the Metacrawler Cilea/SBBL portal was set up to connect the collective catalogue to the data banks and electronic articles so that the resources could be shared comprehensively (2001).

Via the portal and the Internet all SBBL users such as doctors, researchers, health workers (and hopefully also patients in the near future), are able to pursue their specific interests from their own desks. They can:

- utilize interactive research on updated information,
- request and
- acquire texts.

4. THE GROWTH OF SBBL FROM 1996 TO 29TH JULY 2002

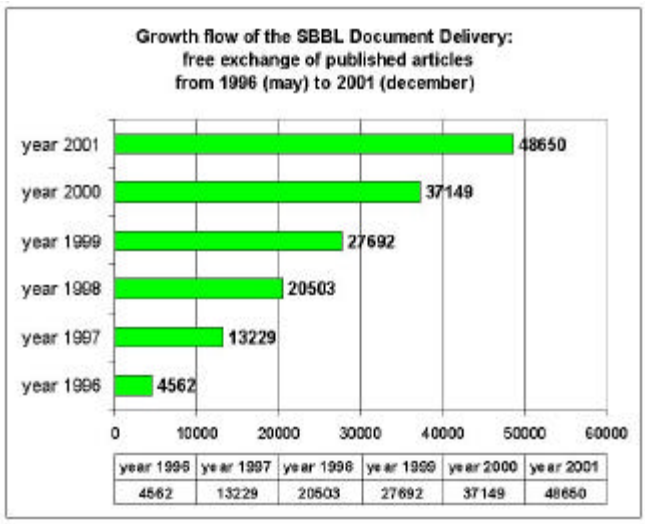

Figura 5 - Flow of Document Delivery requests

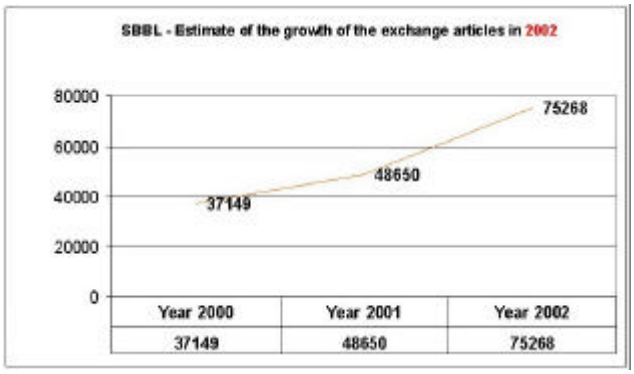

Fig 6 - Document Delivery in 2002: an estimate

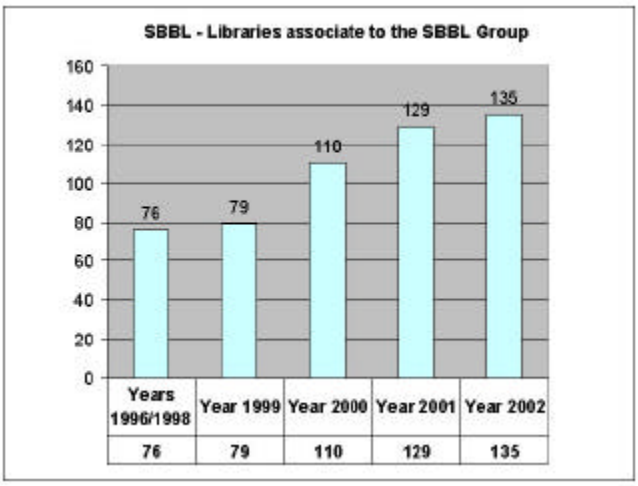

Figura 7 - The growth of the SBBL group

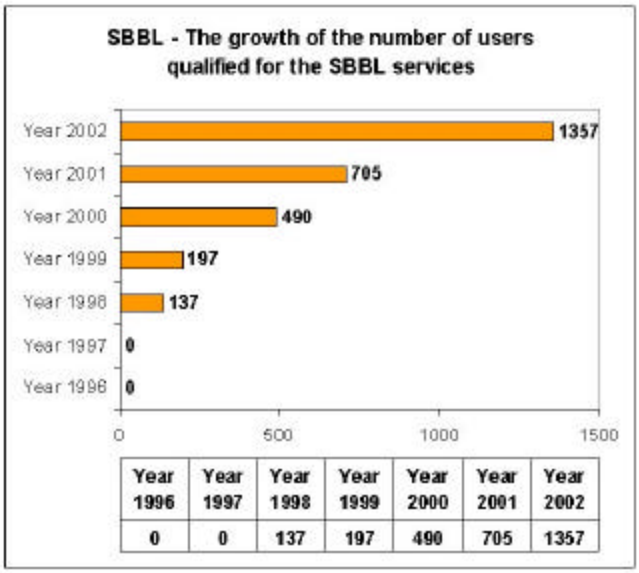

Figura 8 - The growth of the SBBL users

SBBL/CILEA METACRAWLER

The SBBL/CILEA MetaCrawler project came to fruition within the Library Automation services of SBBL.

The project's fundamental components are:

- a "research engine" which can access any type of accredited on-line information source, provided that a linkage can be established comparatively straightforwardly
- an interface that can be customised in accordance with the graphic and functional requirements of the individual SBBL user
- results administration
- LinkOut.

We will present concisely the most interesting functional characteristics of the SBBL/CILEA MetaCrawler

1. VERIFICATION OF USER

Access to the research engine requires verification through the provision of username and password (Fig. 9).

Password accreditation enables access to MetaCrawler not only from workplaces within affiliated organisations, but also from the user's office or home.

An appropriate database gathers and controls the service users, and activates a series of customised facilities:

- it recognises the individual user's research methodologies (history);
- it also recognises users' rights to download full text documents based on the particulars of their organisation's subscriptions;
- it allows to choose the database.

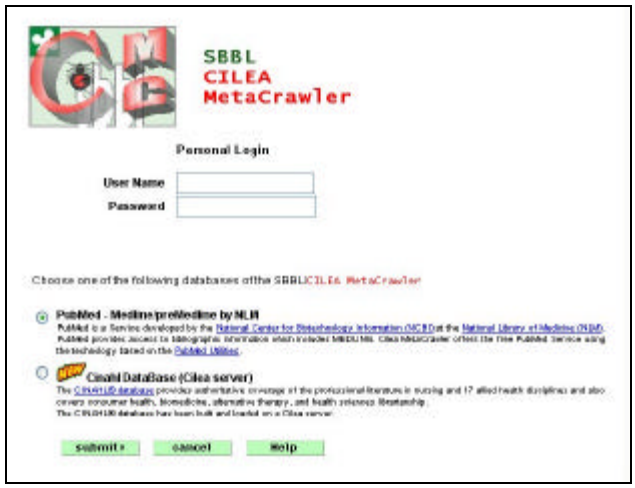

Figura 9 - MetaCrawler login page

2. RESEARCH

Searches can be carried out in different ways; mainly by: author, title of article, title of journal, other looser terms or subject headings (Fig. 10).

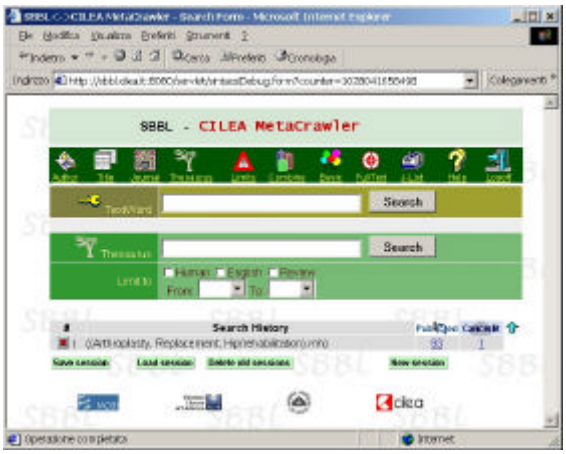

Figura 10 - MetaCrawler working page

Subject Headings

MeSH Terms (Medical Subject Headings): 19.000 biomedical terms constitute the actual vocabulary of the Metacrawler, which is used by PubMed and Cancerlit (Fig. 11).

The selection of one or more Mesh terms during a search is provided by a search engine supplied by Pubmed and controlled by our application.

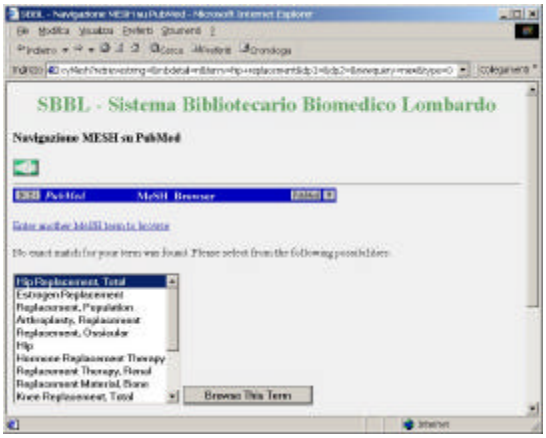

Figura 11 - MetaCrawler MeSH tree

Cinahl is preparing a customised thesaurus for the database that will preserve many MeSH characteristics yet will provide a more specific dictionary.

3. MEMORIZING SEARCHES

Every search generates a number of results-documents found which are saved in a table that is unique to every user: a research table or history.

If we call **search** every single request for information sent to the data banks and **inquiry** the sum of the individual searches for documentation regarding a given topic, we could say that an inquiry evolves and is refined by a succession of searches, which are memorized in the history.

As a consequence, the history is a function of fundamental importance that the user can control through a series of pre-programmed functions, such as:

- **Save Session:** allows one to save the current history under an assigned name.
- **Load Session**: allows one to load a previously saved history while working on a current session.
- **New Session:** allows one to cancel a history or set it to zero.
- **Purge**: allows one to selectively clean out unimportant or poorly formulated searches.
- **Logoff:** allows one to exit the service, eliminate the current client/server session and clean the present history.
- **Display**: allows one to reuse a search, simply by clicking on the history table;
- **Limits**: allows one to refine a search by using a selective filter to pinpoint interesting documents by imposing the following limitations: Publication types, Journal subsets, Languages, Age groups, Human/Animal, Gender, Publication date (Fig. 12).

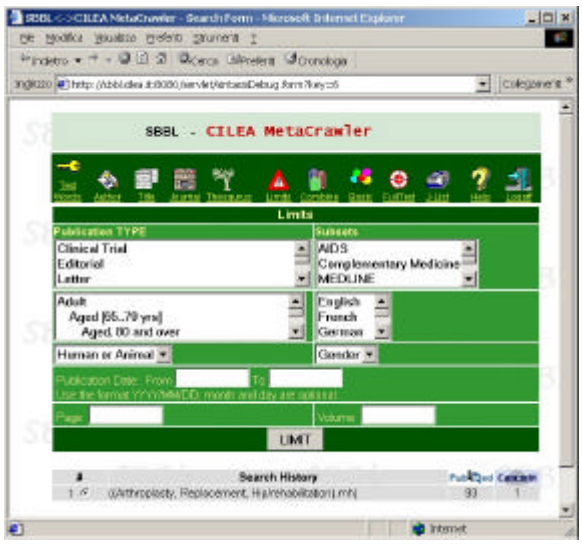

Figura 12 - MetaCrawler: "Limits" page

- **Combine**: allows one to combine one or more searches located in the history by using common logical connectors (AND, OR, NOT). It is a complex function, developed for Metacrawler, which is memorized in "Search History" alongside all other searches.

4. RESULTS

Visualisation of results is obtained by clicking on the result numbers of the search listed in the "History" table (Fig. 13). The system displays the references of the articles found, 20 at a time.

Figura 13 - MetaCrawler: Citation Manager

It is possible:

- to save one's own research in various formats (Medline, Abstract, Brief, XML and others) alongside the history of the research methodology;
- to import the saved work as a text file (save), or as a HTML page (display);
- or send the document via e-mail.

5. DETAILS

Once the user has identified an article of interest he can click on the title and proceed to a more detailed session phase where a more complete description of the article is displayed, up to abstract level (when available) The following functions can be activated from this page (Fig. 14):

- link to SBBL Collective Catalogue to request an article through the ad hoc procedure programmed for the automated compilation of request forms (SBBL Document Delivery procedure);
- link to the LinkOut function with possible connections to the "Full Text" of the article itself;
- link to the Related Articles page offered by PubMed or other databases (if available);
- link to the reference of the article on the PubMed service of the National Library.

Figura 14 - Document Detail

6. LINKOUT

As of the beginning of the detail session, this function provides a list of links that connect the user with one or more sites and the full text article (Fig. 15).

It is important to note that the procedure for the creation of links, as well as the access typology, is consistent with the specifications of the user's profile.

The rules for personalisation are based on the following factors:

- subscriptions to electronic journals and distribution services from providers with whom **SBBL** has contracts and agreements: Elsevier, Blackwell, Lippincott W&W; in addition to eventual free services on the web/network: Free-Medical Journals, PubMed Central.
- Subscriptions to electronic journals and written agreements with distribution services, which the **individual** organisations affiliated to SBBL (hospital libraries, documentation centres, etc) have entered into.
- Electronic inventory (existing and missing journals) if, obviously, such information is provided by the distributor-publisher.

Figura 15 - LinkOut Page

Moreover, the LinkOut function is connected to the CrossRef service (when searching the full-text of an article) and to the ISI service for the Current Contents of an article.

IN CONCLUSION

The combination of resources illustrated (even if in a rapid and concise form) bear witness to some of the fundamental choices made by SBBL:

- the sharing of resources both in acquisition and distribution,
- the procurement of new technologies,
- training for operators and users.

Metacrawler combines all the above, allowing the system to be expanded so as to include every accredited source of biomedical information. This, in turn, will enable the system to produce ever more significant results contributing to effective health management in Lombardy.
Thus demonstrating that the libraries will continue to contribute ever more in determining the advanced status of our social fabric.