MOPAC (Math OPAC) project : One query to all math library catalogues

Elizabeth Cherhal

▶ To cite this version:

MOPAC (Math OPAC) project : One query to all math library catalogues

Elizabeth Cherhal, July 2001

Abstract

This document presents the french mathematical libraries' "multibases" facility: its evolution over the years, the lessons learned from different projects who have endeavoured to do the same thing, and proposes an architecture for the future "MOPAC" service.

History

At the beginning of the 1990s, the RNBM (Réseau National des Bibliothèques de Mathématiques) decided to use the WAIS indexing software to make its catalogues available via the Internet. The data were processed as follows:

- Starting with the library management software (usually "texto" or 4D) an ascii file is exported, and when necessary copied to the machine hosting the WAIS server.
- This ascii file is reindexed by the WAIS software ("waisindex" command).
- The WAIS database is made accessible to the Internet by the WAIS server ("waisserver" commmand)
- The user can access all the WAIS databases via a special WAIS client installed on her machine (mac, pc or unix).
- The client and the server communicate via a special protocol (Z39.50 first version), use of this protocol enables simultaneous distributed queries (as does Z39.50 as we know it today).
- Heterogeneity of the indexed data makes a uniform presentation, and deduplication of results practically impossible. (If 5 libraries have the same document, we obtained 5 responses, all slightly different).

At this time, the web was still in its childhood.

In the mid nineties, the web is developing fast, with the CGI (common gateway interface) library, WAIS developers in Dortmund wrote the WAIS/web gateway SFgate. Thanks to SFgate, one could query multiple databases via a web navigator.

- SFgate is installed in different places: Grenoble (upon creation of MathDoc Cell in 1995), Jussieu, Orsay, Lille, etc…
- At the same time, the first web/database interfaces appear (the ZBW3 (ancestor of EDBMW3) software developed in Grenoble for Zentralblatt-MATH is an early example). Library software houses start developing their web interface (Texto/web Ever/web, etc…).
- As each web/database interface is a specific development, multibase distributed queries are not possible. The libraries who acquire (Lille, Strasbourg) or develop (Bordeaux, Nice) web interfaces keep the WAIS server running to remain compatible with the "multibase" SFgate-based system.
- New libraries set up the WAIS/SFgate system, and some old ones update their server, thus permitting field searches (which was impossible with the first versions of WAIS).

In the late nineties, the web explodes. Everyone has their server.

- The important libraries (central campus libraries, BNF, etc…) now have their web interface.
- The Z39.50 protocol, (much talked about, but less implemented) is implemented in certain campus (Valenciennes, Lyon 3) , town or county (Val d'Oise) libraries. The Z39.50 multibase distributed query facility starts being used.
- Mathematical libraries start to have their own web interface, and do not always see the point of a multibase distributed querying system.

The EULER project and its multibase distributed query system

Early 1998, MathDoc Cell is contacted for participation in a European project aiming to make available all sorts of heterogeneous documentary resources (library catalogues, but also online documents, such as preprints and articles from online journals) in mathematics. The EULER project partners decident to use the "Dublin Core" element set as a common denominator for bibliographical description, and the Z39.50 protocol for distributed queries. Software from indexdata (Denmark) was used for the Z39.50 server and client.

The technical working of this solution is very like the WAIS RNBM one, but with a few improvements:

- The libraries/information managers/project partners convert their records (either classical bibliographical records, or in the case of online documents, existing metadata) into the "EULER XML format", which is compliant with Dublin Core. The conversion programmes are generally quite simple and adaptable from one case to another.
- The "records" produced by conversion programs are then processed by a common postprocessor which uniformizes the accents and produces, with the values of certain fields, a "deduplication key".
- The result of this postprocessing is then indexed by the Z39.50 server ("zebraidx" command).
- The databases created by zebraidx are made accessible to the network by the Z39.50 server ("zebrasrv" command).
- The client (web/Z39.50 gateway) provides a uniform access to all the databases.
- As the data is pre-converted, the results can be written in a uniform manner.
- Thanks to the deduplication key, the result list is presented "deduplicated" to the user. (if 5 libraries have the same document, it will appear only once with a clickable list of holding databases).

Pros and Cons of WAIS/SFgate or "EULER" solutions

Let us first say that both solutions have a major drawback: the necessity of working with an exported copy of the database rather than accessing the database itself. Records will never be updated more than daily, and "real time" access to the database is impossible.

This drawback is in itself also a major advantage: the web access system is totally independent from the actual library management system. Practically nothing must be changed if the library acquires new software.

The following table summarises other pros and cons.

Problem	WAIS solution	EULER solution
Compiling and Installation	WAIS easy, SFgate quite difficult	Easy if one adopts the "out of the box" EULER configuration. Difficult if one wants to do something else
Configuration	Quite easy	Needs a data conversion programme (programmes converting popular "ajout piloté" or "tabular" formats have been written)
Result presentation	Heterogeneous Duplicates are inevitable	Homogeneous Deduplication works at 80%
Performances	Distributed queries are very slow Network protocol inefficient Indexing and searching not very efficient	Client/server system more efficient Indexing and searching very fast
Limits	A WAIS database is limited to about 25000 records	No limit
Compatibility	The protocol used by WAIS is not compatible with Z39.50. WAIS is only compatible with itself	Use of Z39.50 V3 facilitates use of the same client program to search EULER databases along with other ones from libraries.

Different other data models for "multibase" querying

"All Z39.50"

This solution, adopted by important libraries and combined catalogues, such as the "SU", and the "CCF", consists in having library management software frontended by a Z39.50 server, interacting directly with the database on the one hand, and the internet on the other (see the Val d'Oise website). As all the servers speak the same language (Z39.50), one can access simultaneously different databases, as one currently does with WAIS or EULER, but without needing to bother to export, convert and re-index data.

This solution, ideal in principle, encounters a few obstacles, the most important of which is economical. It has proved that developing a Z39.50 server adapted to such and such software (or more precisely the underlying database management system) is a tedious and not always profitable business, and software vendors either hesitate to develop it, or sell it for a good price. Clients tend to prefer the "straightforward" web interface, and all the nice little features that often go with it to the more complicated (normalised) Z39.50 one. Real examples for mathematical libraries are that CINCOM proposes no Texto/Z39.50 (or today no CINDOC/Z39.50) or that GB concept has been selling Alexandrie/web for several years, and has only just announced a Z39.50 server for 2001. In our opinion, it is unlikely that **all** french mathematical libraries will ever have the money to invest in a library management system supporting a Z39.50 server.

"All web" or "mixed"

Considering that each library likes to have its management software, and its web interface with its own particularities, why could we not, without Z39.50, develop a system only based on web software (httpd protocol, and cgi or java programs for instance) to access all databases with one query ? One of the best examples of this model today seems to be the KVK program[1] . However, it seems to us that development of such a system would be quite time consuming, and also, once developed, difficult to maintain. A site only has to change its web interface a little bit for the application to stop working without the developers realising it.
Another nice example of this model is the AskOnce[2] commercial software developed by xerox. However, apart from, or as well as, the initial cost, the same remarks apply as regards maintenance.

MOPAC , the catalogue search engine

Everyone is familiar with Internet search engines. They are based on a fairly simple idea: periodical visits to sites, harvesting and centralised indexing of the harvested files. Why not do something like this for mathematics library catalogues?
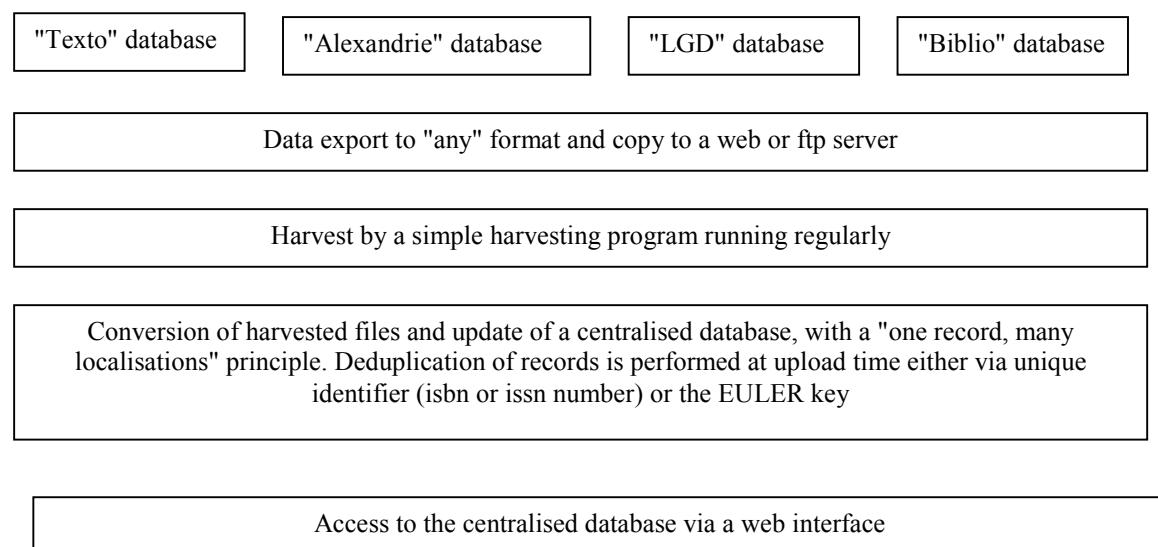On the one hand, our evaluation of the EULER project, and its pros:
- The "dublin core like" metadata model and its expression in XML
- Easiness of writing conversion programs for most structured data to this format
- The "deduplication key" creation algorithm

But also its cons:
- Heaviness of Z39.50 server and client software
- Difficulty installing servers everywhere

And on the other hand, the positive experience of the "CFPM" database, leads us to propose a schema like this:

"Texto" database	"Alexandrie" database	"LGD" database	"Biblio" database

Data export to "any" format and copy to a web or ftp server

Harvest by a simple harvesting program running regularly

Conversion of harvested files and update of a centralised database, with a "one record, many localisations" principle. Deduplication of records is performed at upload time either via unique identifier (isbn or issn number) or the EULER key

Access to the centralised database via a web interface

---

[1] http://www.ubka.uni-karlsruhe.de/hylib/en/kvk.html
[2] http://www.redoc-grenoble.org

We estimate the total number of "deduplicated" bibliographic records at about 150.000.
Currently, we are considering using MySql as the underlying database software for this application.

The advantages of such a system would be:
- Libraries have nothing to install. They only need to export their database every now and then to their web or ftp server.
- We are delivered from network problems. Harvesting is done in the background.
- Conversion programs are stored, and maintained at MathDoc Cell. This is easier to maintain than if we have multiple copies all round different sites.
- The software development and maintenance is easier than an "all web" solution.

The disadvantages would be:
- The data would not be totally up to date (but sufficiently we think)
- Libraries must trust a copy of their data to MathDoc Cell.
- Good cooperation is necessary when setting up the system (but we think this is easier than trying to get people to use "wais").

MathDoc Cell and RNBM (via UMS J Hadamard, Orsay) are planning to set up the "MOPAC" protoype late 2001.