Sherlock: A Web Magnifying Glass for Microdata Files

by Gaëtan Drolet[*]

Context

In Canada, the Data Liberation Initiative (DLI) approved in 1996 by the Treasury Board of Canada has removed a significant obstacle to obtaining Canadian data in our universities.

With DLI, Canadian universities and Statistics Canada have solved the problem of obtaining Canadian data at an acceptable price. I remind you that in Canada, access to data is not free.

In the province of Quebec in particular, a number of universities obtained access to data but without really improving methods of consulting these data.

It is important that people realize that there are no full time data librarians in any of the Quebec universities. Most universities (except McGill, Montreal and Laval) are small in size, with equally small resources. We do not have a long-standing tradition of data use as in other Canadian universities. This is why, in order to render microdata files more usable, university libraries in Quebec have pooled their resources and expertise for the development of a common infrastructure to facilitate access and use of data.

What is SHERLOCK?

No, we aren't talking about the world-famous detective, Sherlock Holmes. According to the conference theme, SHERLOCK is a kind of regional bridge to data. Using Sherlock, the data user becomes a detective of sorts. SHERLOCK is a bilingual tool, designed by the numerical data file subgroup[1] of the CREPUQ (Conference of Rectors and Principals of Quebec Universities). At the Conference of Rectors and Principals of Quebec, we are a small but active group of four data librarians who are been working together since the beginning of the 90's. We organise data workshops for our colleagues. We share our experiences and expertise. All of us are here at IASSIST.

This paper is being read on behalf of the four of us. We are the designers and the managers of SHERLOCK in our different institutions. The CREPUQ provided the place where Quebec university libraries were able to initiate and discuss this co-operative project. We have a 30-year tradition of co-operation between libraries. SHERLOCK was developed mainly for members of the Quebec academic community to enable them to access and utilise the survey microdata of the DLI (Data liberation Initiative) and the ICPSR (University Consortium for Political and Social Research) data.

Project origin and description

The first document submitted by the subgroup on numerical data files was *Rapport de la consultation sur l'intérêt et la faisabilité d'une approche collective à la gestion des données numériques* (Report On Consultations Concerning the Value and Feasibility of a Collective Approach to the Management of Numerical Data,) CREPUQ, November 1996. Our colleague Chuck Humphrey of the University of Alberta acted as a consultant for this stage.

After approving this report, the heads of the Quebec university libraries asked the subgroup to conduct a preliminary analysis on a top-priority basis. The timing seemed to be right.

When the subgroup took stock of data extractors in operation at the time, the LANDRU system, developed at the University of Calgary, stood out as one of the best although it did not meet all the requirements of the system to be implemented in Quebec. We wanted a bilingual interface; a decentralized and distributed approach to encourage the sharing of expertise and responsibilities in many institutions; management of all survey files available in the Quebec university network; compliance with licences; etc.

Therefore the four data librarians, who are members of the CREPUQ subgroup, with the help of an analyst from the library of Laval University, conducted a preliminary analysis and designed a pilot project. In March 1997, the subgroup submitted its report, titled *Infrastructure collective pour la gestion des données numériques dans les bibliothèques universitaires québécoises* (A Common Infrastructure for the Management of Microdata Files in Quebec University Libraries) CREPUQ, March 1997. This report was subsequently accepted by library directors from eleven universities, and they asked to my library (Laval University) to undertake the task of implementing Phase 1 of the SHERLOCK project.

The pilot project
Phase 1 of the pilot project started in September 1997 and was completed in October 1998. The phase focused on developing all of the system's capabilities and setting up a first server centre.

Development team
The responsibility for implementing Phase I of the project was assigned to the library of Laval University, which established a development team made up of a project leader, the data librarian, a librarian and a computer analyst.

The team's mandate was to develop all the system's capabilities, with bilingual interfaces, set up an initial server for a limited number of surveys, and make corrections as needed during the trial period.

Project co-ordination
To ensure that the project went smoothly, the CREPUQ data librarians subgroup on data files was assigned the role of advisory committee.

Funding
The funding for the pilot project was provided through contributions from Quebec's university libraries. Twelve institutions participated in the funding of Phase 1 out of a total number of 14. According to a complex formula, small universities invested less money than big institutions.

Institutions as clients
All users of Quebec universities, called client institutions, have access to SHERLOCK, but the use of the actual survey data requires that the institution's library be a member of the DLI or the Inter-University Consortium for Political and Social Research (ICPSR). In addition to being user institutions, a few libraries will become server institutions.

Institutions as servers
The management of the surveys and their files is a responsibility shared by different server centres. Each institution (server centre) that has taken on responsibility for managing surveys in SHERLOCK has designated a local manager who is responsible for the management and follow-up of these surveys in SHERLOCK. These managers will be the only persons authorised to complete, to modify or delete a survey. A survey management module has been developed to facilitate these operations. For the implementation of Phase 1 of the pilot project, only the Laval University library acted as a server centre.

Surveys included
For Phase 1, fourteen surveys from Statistics Canada and one from ICPSR were installed on the system's first server centre. Five of these support all the system's capabilities (including extraction by variable and statistical analysis), while ten others support the basic level of use (retrieval, consultation of documentation and block files transfers, with no extraction). Under access licences, owing to the number, diversity and breadth of the surveys, some files can only be downloaded as a block (ftp), with no data extraction, while others have limited access, specifically to member institutions of the ICPSR. Some local surveys (e.g., a survey of Quebec public service retirees) could be loaded into SHERLOCK and be accessible only to certain universities. It is the case for a survey on political attitudes done by a graduate students' class in my institution last semester.

The system
Access to SHERLOCK is based on a bilingual Web interface (French and English) offering a single and universal gateway to all the surveys. SHERLOCK is accessible in Quebec university libraries at the following URL address [http://sherlock.crepuq.qc.ca].

The general purpose of the system is to provide for the management and optimum use of all microdata files available in the Quebec university network. Sherlock is not a teaching tool with a set of exercises, but it is easy to use by professors in undergraduate classes.

Capabilities
The main capabilities of the public module are:

- to provide access to the inventory and description of surveys by means of a retrieval module;

- to provide the user with documentation (survey metadata) on data files (guides, user manuals or codebooks, SAS or SPSS statements, record layouts and description of variables) when available;

- to enable users to extract subsets of data files in different formats for later processing at a local workstation. Intermediate and advanced users who can handle large sets of variables can download the complete dataset;

- to enable users to obtain simple statistical results such as a frequency distribution, cross-tabulation, mean, median or regression analysis on a variable in using the module of analysis.

More specifically, SHERLOCK can be used

- to ensure the compatibility of and access to information systems in twelve member institutions;

- to make the greatest number of surveys available;

- to promote the sharing of resources for data preparation, storage and use;

- to promote the development and sharing of expertise in the use of data among both the clientele and the reference staff of our libraries.

Computer infrastructure
SHERLOCK is a decentralized system made up of two modules: a public module and a management module.

The public module is used to access Web pages, conduct searches and access the forms used for retrieval and analysis. Searches are conducted on a UNIX main server (sherlock.crepuq.qc.ca) located at the Laval University library.

The programs needed by the user are Netscape, an e-mail software, WINZIP, Acrobat Reader, Excel/SAS or SPSS.

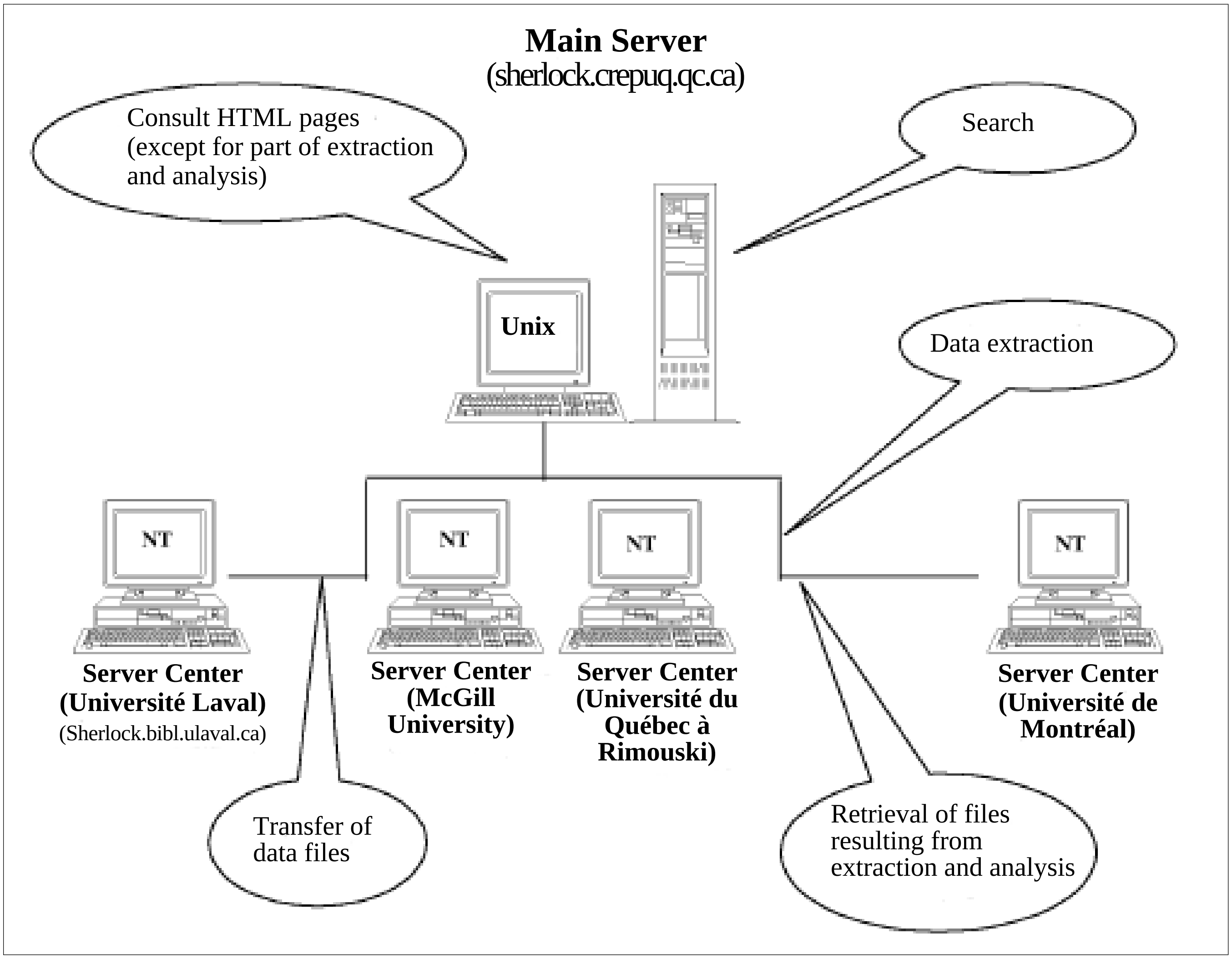
Whereas the documentation is accessible to the general public, access to data (transfer of complete file, extraction, analysis) is controlled by IP numbers, ensuring observance of licences governing use. Access to metadata is public but access to file transfers, extraction and analysis is controlled.

The html pages, for searching the description, the list of surveys resides on the main server (UNIX).

The survey metadata (codebook, record layouts, SAS and SPSS files, etc.) and data files reside in the different server centres on NT servers.

The data extraction and analysis is also done on the different NT servers. Extraction and analysis operations use Perl procedures.

Management module
The management module is used for the capture of data (description of surveys, metadata, and data files) from surveys that can be retrieved using the public module. The management module can be used only by the institutions who are server centres. The management


Main Server
(sherlock.crepuq.qc.ca)
Consult HTML pages (except for part of extraction and analysis)
Search
Unix
Data extraction
NT
NT
NT
NT
Server Center (Université Laval)
(Sherlock.bibl.ulaval.ca)
Server Center (McGill University)
Server Center (Université du Québec à Rimouski)
Server Center (Université de Montréal)
Transfer of data files
Retrieval of files resulting from extraction and analysis

module has different functions. Using html forms, it is possible to work on the surveys, the files (metadata, data sets) and the variables. Only the French version of this module is available at this time.

At the survey level, the data librarian can add a survey, modify it or delete it. The data librarian also decides the treatment level (E/T), server address where the files will be loaded, which universities will have access to the survey. The data librarian enters the description and the abstract in both languages.

Once Inside a survey at the files level, you enter the files (metadata and data), giving a title to each file.

Inside a survey at the variables level, you can also add, modify or delete variables.

The module includes technical notes which are like an online manual. They are guidelines and procedures to facilitate the entry of metadata information.

SHERLOCK also collects statistics on usage (monthly/ annual) by surveys, and by universities. With these statistics we can determine whether the users consult only the description, whether they transfer the complete dataset or whether they perform an extraction or an analysis.

Promotion

Now that the development of SHERLOCK is complete, institutions participating in the project are responsible for promoting this collective tool among data users in their respective universities.

To facilitate the marketing of SHERLOCK, the CREPUQ data subgroup organized two SHERLOCK information and familiarisation workshops. The first one took place at McGill University on October 15, 1998 and the second one, at Laval University (Québec City) in December 1998. These activities drew more than 50 participants (data librarians and staff serving the public). The introduction of SHERLOCK was supported by a press release and a presentation to the heads of university libraries.

In the Quebec universities network, library heads voted unanimously to continue the SHERLOCK project.

Accordingly, Phase II was developed from November 1998 to May 1999. This phase had a two-fold objective: to install SHERLOCK in three server centres (Université du Québec à Rimouski, Université de Montréal and McGill University) and to increase the number of surveys in the SHERLOCK collection, because we have gathered around 40 surveys in our collective tool. In addition to maintaining the system, the development team of the Laval University library has assisted the institutions with installation procedures.

For the year 1 starting next month, a Board of management has been created. This group will establish an annual program and will report to library directors. The users will be represented on the group.

More recently, the SHERLOCK project won a second prize among fifty projects presented at the CAUBO (Canadian Association of University Business Officers) as an academic initiative and development increasing productivity and effectiveness in higher education. The development team is very pleased with this recognition.

Conclusion

Among Quebec university libraries' the collective approach to the management of microdata files is two-fold : first, to "liberate" access to data, and secondly, to liberate their use.

SHERLOCK is also an active participant in the Data Liberation Initiative in Canada, which concerns the development of a data culture in our universities.

In jointly supporting the development of this research infrastructure, Quebec university Libraries are **1)** encouraging the analysis of the statistical information available in the Quebec university network, **2)** promoting student learning, **3)** supporting the work of professors and researchers, and **4)** participating in the demystification of data among library staff.

I would especially like to thank my three data friends (les trois amis des données). These data friends are not the same as the "Los tres data amigos", well known at the ICPSR Summer Institute. I invite you to meet SHERLOCK in person at the poster session.

1 Consisting of Richard Boily (Université du Québec à Rimouski), Jerry Bull (Université de Montréal), Gaëtan Drolet (Université Laval) and Anastassia Khouri (McGill University).

*Paper presented at the IASSIST Conference, May 19, 1999, Ryerson Polytechnic University, Toronto, Ontario. . Gaëtan Drolet Université Laval.