Data liberation, bridges to cross

by Richard Boily *

Abstract
In Canada, the use of statistical data (micro data files and major databases) for teaching and research is an important phenomena that does not seem to be losing strength in the near future. This situation is a major consequence of the Data Liberation Initiative (DLI), established in 1996 as a partnership among Statistics Canada, other federal departments and Canada's academic community. The idea of providing affordable access to Canadian information results from a co-operative effort among the Humanities and Social Science Federation of Canada (HSSFC), the Canadian Association of Research Libraries (CARL), the Canadian Association of Public Data Users (CAPDU) and the Canadian Association of Small University Libraries (CASUL). Less than two years after its inception more than 50 universities have joined the consortium, which is a clear indication of a true willingness to make data more available.

This illustrates the fact that the high cost of buying data was an obstacle to its availability, especially in small universities where the absence of a minimum number of students results in a higher cost/benefit ratio related to data acquisition. However, there are still many obstacles to free numerical data use. If some Canadian universities have a long history in data services (Carleton University's Data Centre celebrated its 30th anniversary in 1996), such a tradition does not exist everywhere, especially in small universities.

To maximise use of data files, increased education at the reference staff level and at the consumer level, including professors, must occur. Data usage requires a good knowledge of data extraction and associated analytical instruments. How can these tools be made accessible to customers who are not able to manipulate data files, but who have a definite need for the information? How can we satisfy different needs for different types of users? How can data be included in the academic curriculum? How can data librarians play their educational role and how can this role be balanced with professorial responsibilities? Fortunately, interesting answers are unfolding.

Introduction
Numerical data collected from various surveys conducted throughout the nation by organisations such as Statistics Canada constitute an information source that is both important and extremely powerful in understanding social phenomena. Access to these data is necessary for teaching and academic research. In fact, this access is essential to intellectual freedom and democracy.

In Canada, the conditions of accessibility to numerical data have undergone major changes within the past two past years, due to the implementation of the Data Liberation Initiative (DLI) by Statistics Canada.

The Data Liberation Initiative is a management framework that modifies the conditions of data access. These changes coincide, and certainly not by accident, with the arrival of new technologies: the development of both powerful personal computers and their increased data storage capacity and of user-friendly software (Excel and SPSS), in addition to the advent of the Internet. Such events bear directly on the theme of this conference, notably Global Access and Local Support. With new parameters defined by DLI, Canadian data are potentially more accessible than ever to Canadians.

Even if DLI is successful, the fact remains that widespread numerical data use in Canadian universities is uncommon and many obstacles exist that would allow the situation to change. The objective of this presentation is to examine problems of accessibility to numerical data within the context of DLI.

This presentation is composed of three parts. First, a brief history of the origins of DLI and its role will be given. This will also entail a review of the objectives. Then, the problems of developing numerical data use within the context of DLI will be addressed. Finally, it will be shown that there are elements of DLI that represent an opportunity to improve the democratisation (accessibility) of data in Canada.

1.DLI

Origins

Traditionally, Statistics Canada publishes the statistical information that it has collected in the form of aggregated data tables. These documents are largely distributed to libraries via the government publication deposit program. However, numerical data files that have been excluded from the deposit program and until recently, were available only at a very high price. Such a situation has been strongly discredited by the research community, notably by Professor Paul Bernard, professor of sociology at the University of Montreal and a member of the National Statistics Council. In 1991, Professor Bernard asserted that, "the genuine exercise of democracy increasingly requires that citizens get access to complex information and have the skills required to understand it".

In 1993, following the opinions voiced by Professor Bernard and others, several individuals representing the Social Sciences and Humanities Research Council of Canada, the Association of Universities and Colleges of Canada, the Canadian Association of Research Libraries and the Canadian Association of Public Data Users united under the auspices of the Social Science Federation of Canada. Their specific objective was to develop a strategy to render Canadian survey data more accessible to the research and teaching community. The work of this group led to a proposal that rapidly passed through the various levels of the federal government and thus, was accepted by Statistics Canada. The DLI received official recognition from the Treasury Board of Canada in February 1996. It was subsequently included as part of the Canadian government's Science and Technology Strategy in March of the same year.

What is the DLI?

The DLI is a five-year project among universities, Statistics Canada and several federal departments. Under the agreement, participating universities pay a known and affordable yearly fee ($12,000 for CARL members or $3,000 for CASUL members), that gives them access to all standard data products provided by Statistics Canada. FTP on the Internet is the primary method of accessing these files. If files exist only on CD-ROM each participant is entitled to a copy. However, if they exist in both forms, users may choose one or both types of files. Participating libraries must make acquired data available to their users while, at the same time, insuring that they do not use it for commercial purposes.

Objectives of the DLI.

As stated on the DLI Web site, itself, "timely access to data is essential if researchers are to focus on Canadian problems and students are to learn to analyse Canadian information. Without affordable data for research and training, Canada risks producing innumerate graduates and basing its policy decisions on incomplete information. Independent analyses enhance public debate and policy making on questions relevant to all Canadians. The federal government invests large amounts of public money in data collection. The DLI can ensure a valuable return on this investment by distributing data to the university community, which will encourage analysis and put more information in the public domain".

Before the inception of DLI, we experienced the embarrassing situation where Canadian researchers, who needed to develop methodological expertise or study a particular social phenomena, had to work with American data because Statistics Canada data were either too expensive or not available at geographically specific levels. Unfortunately, this situation still exists.

2.Numerical data use in the context of DLI, or, how do things happen now?

Although DLI has been in existence barely two years, it is showing positive results that are measurable and it seems to be fulfilling expectations.

Table 1. University participation to DLI

Universities offering data services before DLI	Participating universities in the Data Liberation Initiative (DLI)		
	1996	1997	1998
Between 15 and 20	50	59	61

There is an excellent participation among Canadian universities in DLI that exceeds even the most optimistic expectations. Before DLI's inception, barely 15 to 20 universities offered any form of numerical data service. This obviously does not take into account individual professors and researchers who ordered files from Statistics Canada. It is even probable that the acquisition of these data has been made through the library. It does not take account either of the various numerical data files, notably the 1986 and 1991 Canadian census data distributed on CD-ROM, that were made available by several libraries well before the inception of DLI. For the past several years, we have offered training sessions on CD-ROM census data search at our library. However, the fact remains that these transactions were not integrated within a real data service.

Today, more than 60 institutions participate in the DLI consortium, which is nearly 80% (61/78) of Canadian

universities. Fifty of those affiliated have been with DLI since its first year. However, it does not follow that all 61 participating institutions offer a numerical data service. This number simply indicates that there is, at minimum, a DLI representative in each of these universities and that this person is involved, to one degree or another, in data-related activities that may lead to the implementation of a data service.

Table 2. Numerical data use in the context of DLI.

CD-ROMS delivered to paricipating univesities to DLI	Files downloaded from the FTP site of Statistics Canada		
1035	1995	1996	1997
	887	10173	24384

We have discussed institutional participation, but what happens to the data use level?

In March 1998, more than 1,000 CD-ROMs had been delivered to participating universities. The number and growth of FTP file transfers rose from 10,000 in 1996 to 25,000 in 1997. Obviously, these transfers were not carried out exclusively on data files, but also on command files and text files (e.g., code books, readme files, etc.). Also, transfers do not focus only on micro data files. Aggregated census data accounts for a high proportion of transactions.

Looking at the previous data about DLI, one can advance some observations:

1. Perhaps some data would not have been ordered due to cost, even by libraries that already have data use experience;

2. One can suppose that the global cost would have been far more important if all these data had been acquired individually;

3. It is probable that several files have been downloaded or that CD-ROM products have been ordered by libraries that, until now, had little used numerical data.

Therefore, it seems that DLI contributes largely to data distribution and that the program replies to a real need. The objective to make numerical data accessible at a reasonable price is, therefore, partially attained. On the other hand, it is important to remember that DLI was not conceived merely to reduce the cost to those already using data. The real objective is to increase data use by the whole community, to see a real expertise developed in data use, and to enable a greater number of studies that focus on Canadian society to be conducted. This is the real meaning of accessibility and the democratisation of data.

Context of numerical data use

Even if Statistics Canada survey data are potentially available (data can be downloaded at any given time by whoever needs it) and the price no longer constitutes an obstacle to use, both the intrinsic complexity of the data and the means of exploiting it remain major constraints for its use. The majority of users are, in fact, incapable of manipulating raw data. Their need for a data service before the arrival of massive data sets occurs is more important than ever. However, the costs associated with setting-up and maintaining such a service are considerable and largely exceed the cost of the data. These costs constitute a major obstacle to the democratisation of data.

Variety of resource persons

The participation of new institutions in DLI and, the consequent arrival of new representatives are promising events for the development of new data services. On the other hand, all those who have accepted to be responsible for data files do not have, necessarily, the same level of competence. Not all have the same interest or desire to develop the expertise required by this new function.

The DLI is a young program that has experienced accelerated development (50 members the first year), probably due to a copy-cat effect. As we have previously stated, it is necessary to remember that statistical information distributed by Statistics Canada is traditionally published first as working documents and then, these documents are often integrated into governmental publications. For librarians who specialise in governmental publication reference, the responsibility of numerical data management constitutes a sizeable challenge. In many cases, and I have to recognise that it was true with me, new DLI representatives came to the job previously unaware of what was entailed with numerical data files.

Numerical data exploitation and use, especially micro data, supposes a knowledge of computers and statistics that is often deficient in both the users (students at all levels and good number of professors) and the library personnel. In terms of helping the clientele, it seems apparent that consultation services must be collaborative efforts involving both the computer service personnel and the professors and researchers. Among them are specialists in computer science and statistics. But, if one believes that users in search of numerical information are better served in a library (and do not forget that these are DLI participating libraries), it appears problematic to me to

think that data information specialists must always cater to other professionals. Collaboration is essential, but total dependence should be limited. Training needs are an important, yet considerable cost of rendering data more accessible.

Training

To satisfy their training needs, DLI representatives can count on a continuous training program put in place by the DLI External Advisory Committee. In addition to this national program, local organisations (such as the Council of Prairie and Pacific University Libraries' (COPUL) Consortium of Library Electronic Data Services (ACCOLEDS), or the Working Group on Data of the Conference of Rectors and Principals of Quebec Universities) also organise training activities.

Since the inception of DLI, several training activities have already taken place. At the initiative of the advisory board, 4 DLI initiation workshops (in fact, the same workshop repeated 4 times) have taken place in four Canadian cities during 1997. These workshops brought together 120 individuals who until then knew almost nothing about numerical data. In evaluating the workshop, participants stated their preference for the organisation of additional workshops on more specific themes.

As a result, new workshops will take place this spring on the use of SPSS for data processing and numerical analysis. In fact, one of these workshops was held two weeks ago in Montreal.

Conducting training sessions raises both challenges and opportunities for librarians interested in promoting numerical data use. It is a challenge because it is necessary to develop a certain level of competence before being able to teach. But challenge aside, the possibility of teaching users represents a privileged opportunity to assume our role as information specialists. Instead of waiting to be asked for information, we can create a demand for information. How can users ask to use numerical data if they do not know it exists or if they can not use it?

Far from me to suggest that librarians replace other professionals or professors. However, I believe that a solid knowledge of data and of the tools of exploitation - from which arises a need for training - is necessary. First, it allows us to exercise the educational aspect of our job and, second, it enables us to become knowledgeable spokespersons alongside professors and other professionals. It is only with a solid knowledge of data that we can become counsellors for users, orienting them towards the best data sources or advising them on the best manner of exploitation. The competence of data librarians is as necessary to establishing bonds with professors and researchers, as is identifying the main research areas for which data are required.

There certainly is not unanimity among colleagues on the way in which these new responsibilities will be handled nor, consequently, on the usefulness of advanced training. Some will never be able (nor want) to develop statistical expertise. The problem is complex and there is certainly no correct reply, but it will have to be considered and a great deal of progress will have to occur.

No matter what others decide to do, some libraries, such as the Carleton University Data Centre, have already specialised in numerical data and offer complete data services that include computer and statistical assistance. Such levels of service are not widely found. Even large universities do not always offer complete data services. All the new libraries that now play a role in data specialisation have not and will not develop such an expertise. But is there a middle road and where is it situated? If data accessibility is essential to democracy, the inequality of services offered constitutes an important obstacle to exercising our rights.

Equality of access

Within the DLI framework, the choice has been made to insure numerical data development in libraries. Considering information needs in general, researchers in small or regional universities are no longer penalised with regard to information accessibility. With the development of new technologies, such as the Internet and the emergence of periodicals and other electronic publications, the availability of large bibliographical data bases on either CD-ROM or the Internet (UNCOVER) and finally, with the development of increasingly specialised inter-library loan services (Ariel), the disparities have lessened between the large and small universities, between urban universities and those in the regions. This is true even if the level of document availability remains variable at the local level.

The situation with regard to numerical data availability is entirely different. It is obvious that students at universities that offer well structured data services are in a favoured position over their colleagues who do not have such access. The other libraries simply do not offer the students the same level of support.

In fact, one can assume that regional disparities were even greater before the inception of DLI and that the program will attenuate these differences. In this regard, the question of training for librarians and their perceived role in offering these new services is an important consideration.

Regional disparities

Another aspect related to the disparity of services offered has to do with the regionalisation of the data. Regional universities are often located in areas that are characterised by a low population density. Research work related to regional problems require data over geographical areas that are not comparable to data that defines large metropolitan

regions. This is the problem experienced at the University of Quebec at Rimouski, which has Masters and Ph.D. programs in regional development. The users from these groups need not just numerical data, but numerical data at a specific geographical level. Unfortunately, the data available to these researchers are often over too wide a geographical level. Even if more specific data exists, they are not available for their use. This is the problem, for example, with the large Survey of Consumer Finances and with the National Population Health Survey.

If the objective of DLI is to increase data access in all universities no matter where they are located, the question of geographical data specificity is important, even though we recognise that solutions will not come directly from DLI. Indeed, the mandate of DLI is to give access to standard data products provided by Statistics Canada. The notion of standard products is also linked to the question of data confidentiality. As a citizen, one can only rejoice in observing that Statistics Canada respects norms of strictest confidentiality and that these principles should never be challenged. From the academic viewpoint, the problem is not less important. On the other hand, it is probable that the successes of DLI will increase demand for this level of data. The question of confidentiality is also important from the standpoint of the installation of data services. It is not sufficient to simply initiate users to the data and to analytical instruments, such as SAS or SPSS, if one cannot also provide data that satisfies their research needs. There is a risk of losing hard-gained credibility if users do not have access to data that they know exist, especially after they invest considerable energy in learning complex instruments.

Conclusion

Previous statistics have shown that DLI has had real success; a success that exceeds the hopes of many. But beyond all the figures relating to file transactions and considering the current context of data use described here, one of major successes of DLI has been to enlarge data access to a greater number of individuals interested in using numerical data in Canadian libraries. These people share considerable amounts of information (via the two list servers that help in administration of the program). Several of them have had the opportunity to meet during workshops. The community is, therefore, in the process of widening and there exists a tangible willingness among colleagues who are more experienced to share their expertise.

Finally, I would like to mention that universities in Quebec are full participants in this process. All participate in DLI. Members of the sub-workgroup on numerical data files are presently working on the development of an identification and data extraction system that will facilitate and stimulate data use.

*Paper presented at the IASSIST Conference, May 21, 1998, at Yale University, New Haven, Connecticut.

Richard Boily, Université du Québec à Rimouski, and member of the Conference of Rectors and Principals of Quebec Universities (CREPUQ) Working Group on Data