IHSN

INTERNATIONAL HOUSEHOLD SURVEY NETWORK

The Canadian Data Liberation Initiative

An Idea worth Considering?

Ernie Boyko and Wendy Watkins

IHSN Working Paper No 006
November 2011

www.ihsn.org

The Canadian Data Liberation Initiative
An Idea worth Considering?

Ernie Boyko and Wendy Watkins

November 2011

Abstract

The Data Liberation Initiative (DLI) is a Canadian program aimed at providing Canadian post secondary institutions affordable access to Statistics Canada data resources. It is a partnership between Statistics Canada and the academic sector. While it initially focused on the dissemination of public use microdata files it now encompasses all publically-available data. This paper describes the background of this project and some of the key success factors so that other agencies may be able to determine its applicability for their own situations. It was written in the hopes that other agencies may find the Data Liberation project as a useful model to consider. It was also written for a Canadian audience that is interested in the history of this project which has now been in operation for over 16 years.

About the Authors

Ernie Boyko is a former staff member of Statistics Canada where he held a number of Directorships, including Agriculture, Corporate Planning, Electronic Publishing, and Operations for the 1991 Census. It was during his time at Statistics Canada that Wendy Watkins and he co-founded the Canadian Data Liberation Initiative. He is an active member of the Canadian Association of Public Data Users and the International Association for Social Science Information Services and Technology. He is currently an Adjunct Data Librarian at the Carleton University Data Centre and occasionally works on projects with the International Household Survey Network. He can be reached at boykern@yahoo.com.

Wendy Watkins is the Manager of the Carleton University Library Data Centre. Her inspiration for Data Liberation came while working at Statistics Canada for a two year period. She is a founding member of the Canadian Association of Public Data Users and active in the International Association for Social Science Information Services and Technology. She can be reached at watkwen@yahoo.ca.

Acknowledgments

This document was developed for the International Household Survey Network (IHSN) with financial support from the World Bank Development Grant Facility, Grant no 401010-06, administered by the PARIS21 Secretariat at OECD.

It was prepared by Ernie Boyko and Wendy Watkins with contributions in the form of input or comments and suggestions from Rosemary Bender (Statistics Canada), Olivier Dupriez (World Bank), François Fonteneau (OECD-PARIS21) and Michel Sequin (Statistics Canada). The authors are indebted to the many pioneers who contributed to the establishment of this project over the years. Material was drawn from the work of the late Dr. Paul Bernard (Université de Montréal), Marcel Lauzière (Social Science Federation of Canada), Carol Martin (Social Science Federation of Canada), Dr. Raymond Currie (University of Manitoba), and Chuck Humphrey (University of Alberta). Discussions with many colleagues at Statistics Canada and the Canadian university community were invaluable.

The paper (or a revised copy) is available on the website of the International Household Survey Network at www.ihsn.org

Citation

Boyko, Ernie and Wendy Watkins. 2011. "The Canadian Data Liberation Initiative. An Idea Worth Considering?" International Household Survey Network, IHSN Working Paper No 006.

Table of Contents

Introduction

Statistics Canada has been producing public use microdata files (PUMFs) for many years. This was in response to the desires of researchers in the academic and government sectors whose needs could not be met with aggregate data and for whom access to the confidential microdata was not possible or feasible. Since 1971, PUMFs for the Census of Population and other surveys such as general social surveys and household income and expenditures surveys have been popular with a wide range of users. The availability of PUMFs has produced a wealth of analysis into Canadian society and has enabled students to experience research with Canadian data as opposed to American and other foreign data as had been the case in the past.

This paper describes the Data Liberation Initiative (DLI), the Canadian experience in ensuring that students and academics had full and unfettered access to public microdata files. The DLI story is being shared with other agencies in the hope that other NSOs can study the Canadian solution and perhaps implement similar methods.

Sharing this story comes at an opportune time. Many agencies are now embarking on "open data" initiatives. While implementing a new program is never easy, the technical impediments that existed 20 years ago have, in large part, been ameliorated.[1] Large organizations such as the World Bank support such programs, and standards, practices and tools have been developed and are widely accessible to agencies interested in liberating their microdata.

The Problem

The interest in PUMFs by the user community in Canada has grown steadily over the years and has been fuelled by three main factors:

- the availability of cheaper, more-powerful computers which could be used to create and analyze survey files;
- the development and spread of statistical packages such as SAS, SPSS and Stata, and
- the availability of disclosure-control procedures so that the confidentiality of the respondents could be protected.

Statistics Canada started producing PUMFs in the early 1970's. While PUMFs were always seen as useful and powerful outputs from surveys, they tended to be treated as 'special products' by Statistics Canada. This meant that some of the cost of producing them had to be recovered from users. Initial pricing for PUMFs was intended to cover the costs of duplicating and shipping a copy of the data file and accompanying documentation on round magnetic tapes (the standard technology of the time). Files were priced on an individual-user basis and, according to licensing rules, were not sharable among researchers. The prices charged were in the order of a few hundred dollars. While such pricing was regarded as minimal by Statistics Canada, most of these files were, in fact, too expensive for students. To make matters worse, Statistics Canada increased these prices exponentially (to fully recover the entire cost of producing the files) in response to severe budget pressures experienced during the 1980's. This had the effect of putting microdata out of financial reach to all but the few very well-funded researchers or those who had close ties with the data producers in the agency (and very often were able to receive free access to data files).

In addition to steep pricing for microdata, the appropriate data infrastructure in Canadian universities was limited as there were very few (9) data centers in Canadian universities. They were, for the most part, not well-funded by their universities and there was no national program to support such activities. Typically, they provided access to microdata files (only some of which were from Statistics Canada while others came from foreign sources) which had been purchased using ad hoc research grants. Indeed, universities in 1990 were entering their second decade of serious underfunding.

1 The context for this paper will be better understood by making reference to IHSN Working Paper 5(WP005) titled *Dissemination of Microdata Files: Principles, Procedures and Practices* and available at http://www.ihsn.org/home/index.php?q=focus/dissemination-microdata-files-principles-procedures-and-practices. It shows some of the choices that a NSO must make in preparing microdata files for dissemination. The principles in WP005 parallel those that are followed in Canada for the creation of PUMFs.

The CANSIM[2] University Base (Statistics Canada's only product offering designed for academia) was seen by Statistics Canada as a way to provide academics with access to data. The problem was that it contained only a limited number of time series and very little social data. More importantly, it was NOT microdata. It served some economists adequately, but left most quantitative social scientists longing for microdata files.[3]

This problem had two consequences. First, those researchers who wished to pursue quantitative work often turned to US data from the Interuniversity Consortium for Political and Social Research[4] (ICPSR) or to international data from one of the European data archives. In both cases they were addressing research questions from other countries while Canadian questions went unanswered. A second, and more alarming consequence was the large number of researchers who gave up using data altogether, causing a decade of lost capacity in quantitative expertise.

Paul Bernard[5] gives voice to this problem:

> *The genuine exercise of democracy increasingly requires that citizens get access to complex information and have the skills required to understand it ... Concerning such issues, the public must have appropriate knowledge and not only hypothetical access to the data. Paradoxically, indeed, contemporary societies offer a wealth of information, but workers and citizens can be totally mystified, surrounded as they are by data whose flow and codes they do not master. ... All institutions producing specialized information must involve themselves in the transformation of these data into knowledge actually usable in economic activity and democratic debates. (1991)*[6]

It was the influence of Bernard's writing that gave rise to the original proposal, *"Liberating the Data: a Proposal for a Proposal."*[7]

Initial Attempts at a Solution

An early attempt to remedy the situation was the formation of the Canadian Association of Public Data Users (CAPDU). CAPDU was an organization consisting mainly of members from university data centres and had a mandate to lobby federal parliamentarians for improved access to Statistics Canada's data. CAPDU's concerns were heard by the Canadian Association of Research Libraries (CARL) who acted as a consortium to purchase one complete set of the public data from the 1986 Census of Population. Thirty-six institutions took part in the purchase, paying just over $200,000 to Statistics Canada for the Census files. Later that year, the consortium made a second purchase of the five General Social Survey files. In this case, 28 institutions participated. The University of Toronto undertook to act as disseminator and made and distributed copies to each of the participating institutions.

While the consortium approach seemed at first glance to be an ideal solution, several factors limited its potential for expansion.

First, many of the institutions purchasing the data were unable to use them. This stemmed from a combination of the complexity and size of the files and a lack of infrastructure to provide professional and technical support necessary to deal with complex tape files in many of the institutions. While the universities had the data, in many cases the tapes sat unused on a shelf in the computing centre. Under those conditions, one would have to expect some hesitation on the part

2 The CANSIM University Base was a subset (25,000) of Statistics Canada publically available time-series database which contained approximately 2 million time-series.

3 Even the Chief Statistician realized the limitations of aggregate data when he gave his 1991 annual address to the Statistics Canada senior employees and remarked that the analyses carried out by the Canadian Centre for Justice Statistics had been greatly enhanced as the Centre was now receiving microdata as opposed to aggregated files and was, therefore, finally able to focus on their own concerns rather than those of the data producers.

4 ICPSR is an international consortium of academic institutions and research organizations, which provides access to an archive of data. It also provides training in data access, curation, and methods of analysis for the social science research community. It is housed at the University of Michigan and has about 700 members worldwide. See http://www.icpsr.umich.edu/icpsrweb/ICPSR/

5 Paul Bernard was an internationally renowned Sociologist from the Université de Montréal. He was a member of the Social Conditions Advisory Committee at Statistics Canada and the National Statistics Council. He was passionate about the democratization of data access.

6 Bernard, Paul, "Discussion Paper on the Issue of Pricing Statistics Canada Products", February, 1991.

7 Watkins, Wendy. 1992

of these CARL members to participate in another consortium purchase of Statistics Canada data unless there was a remedy to the situation.

Second, the CARL members were reluctant (and in some cases, unable) to make ad hoc budgetary commitments on an ongoing basis. Under the previous arrangement, they were asked several times for funds which were outside their regular budgets. Given the large number of Statistics Canada's public microdata files, it must have seemed like a never-ending series of requests. The quote below, encapsulates the situation as it was at the time:

> *In summary, the situation as it currently stands is lamentable. Canadian academics are looking to the United States for data to use in their research and teaching activities. At the same time, Statistics Canada has a wealth of unmined data lying on the shelf for lack of an audience that can afford to purchase it. These data, collected at public expense, are largely under-analyzed. Thus, the public too is short-changed in this process. What is required is a solution that would see everyone win; Statistics Canada, the academic and research communities and the public at large.*[8]

An Idea is Born

The challenge was to devise a plan whereby Statistics Canada's data would be accessible to researchers in Canadian universities regardless of the availability of on-site support and without unduly burdening any one institution to act as a distributer. More fundamentally, it must ensure a continued flow of research files from Statistics Canada.

While there were a variety of funding and organizational models which could have been proposed, in order to be successful *any plan* must contain certain generic components:

- a broad definition of the data that are sought by the research community. In this case, the main focus was public use microdata files
- an inexpensive, reliable dissemination medium. Until this time, the main dissemination medium was computer tapes which were relatively expensive to produce and distribute.
- a delivery system that ensures access to the files regardless of the availability of on-site expertise and,
- a funding formula that was affordable to both universities and Statistics Canada

To accomplish this, Watkins proposed a four-phase pilot in 1993. Phase 1 was simply proof-of-concept that the internet and FTP (file transfer protocol) could be a cost-effective, reliable dissemination medium. Phases 2 through 4 proposed increasingly-sophisticated data access and manipulation technologies to enable those universities without the technical capability to offer the service. Phase 2 envisioned a data extraction system; Phase 3 sought to expand this into a menu-driven interface that could be used by non-specialists and Phase 4 proposed using a software such as NSDStat (the forerunner of NESSTAR) to introduce students to the power of microdata. While this grand plan did not materialize immediately, it is strikingly similar to the infrastructure of the Ontario, Canada's <odesi> project.[9]

From the outset of the idea, the members of IASSIST (the International Association of Social Science Information Service and Technology) provided a valuable sounding board. The proposal was presented at the annual conferences long before it came to fruition. IFDO (the International Federation of Data Organizations) was also very interested and asked Watkins to join the IFDO board even though Canada had no national data archive and thus, no national director.

Unfortunately, the proposal was, for Statistics Canada, a non-starter even though some members from the agency had assisted in the preparation of the proposal. Concerns about leakage and misuse of data made it unpalatable to senior management, even while the data producing divisions and two senior managers applauded the venture. The proposal also made its

8 Watkins, Wendy "Data Liberation Revisited: A Proposal for a Proposal", unpublished, March, 1993

9 Established in 2007, <odesi> (Ontario Data Documentation, Extraction Service and Infrastructure) is a digital repository for social science data, including polling data. It is a web-based data exploration, extraction and analysis tool that uses the Data Documentation Initiative (DDI) social science data standard. See http://search1.odesi.ca/

way to the Social Science Federation of Canada[10] (SSFC), where it was filed for future reference.

As time went on, further budget cuts mandated that certain efficiencies be found at Statistics Canada. The SSFC contacted the senior manager at the agency who was involved in assisting Watkins. He suggested a second look at the liberation paper. This resulted in a call for a meeting of interested parties—granting agencies, some government officials, professional association representatives and both of the paper's proponents. The idea was enthusiastically received and a task force, affectionately known as the Data Liberation Army, was established. The task force was facilitated by the Social Science Federation and had representation from the university research and library communities, Statistics Canada, the Treasury Board Secretariat and the Government's Library Depository Services Program.[11]

Building Support for the Proposed Initiative

The task force which developed the final Data Liberation proposal took into account the proposals made by Watkins and also broadened it to take into account the concerns of both Statistics Canada and the academic community. The broad principles of the proposal are outlined below:

- Statistics Canada and the academic community should work as a partnership to enhance the use of Statistics Canada's data, ensure that students have an opportunity to use the data for academic purposes and that researchers can enhance the knowledge of Canadians about Canada.
- All publically available data files (microdata, aggregate data and geography files) would be made available to all members of the partnership who could choose which files to access.
- The project would be supported financially by Statistics Canada and the universities. Large universities would be asked to pay $12000 and all others would be asked to pay $3000.
- Each institution would sign an agreement with Statistics Canada (see Appendix A) and would designate a contact point to provide support to their institutions and to liaise with Statistics Canada. See http://www.statcan.gc.ca/dli-ild/contact-eng.htm
- The data were to be used for academic purposes (teaching, research and academic planning) and were not to be redistributed or used for commercial purposes.

While the members of the task force were in unanimous agreement with the proposal, it could not be implemented without support from both government and academia.

Building Support within Statistics Canada

The most difficult challenge for the initiative was to convince Statistics Canada of the merit of the proposal. The agency had two major counter-arguments: the first was loss of revenue and the second, and related, was 'leakage' of data outside the academic sector to commercial users who would normally pay for their files.

These were addressed in a number of ways. First, the successful use of CA*Net (the Canadian Internet backbone at the time) to distribute Statistics Canada's 1989 Survey of Literacy Skills Used in Daily Activities (LSUDA) via FTP gave proof of concept that dissemination could be done for little marginal cost. To ensure a proper test, this file was made available to anyone in the world even though the implementation of Data Liberation was limited to Canada only.

Changing the culture from cost-recovery to cost-avoidance was a major hurdle. Statistics Canada was accustomed to gaining revenue through the sale of individual items. The more units they distributed, the more revenue they expected to receive. Applying the same model to the distribution of survey files to a large number of universities resulted in the expectation of potential revenue amounting to millions of dollars.

Since this method of determining the monetary implications of the program could not be justified to the Treasury Board, Statistics Canada was asked to estimate the 'cost' (as opposed to estimates of revenue) of providing a data service to universities. This included

10 The SSFC represented over 25,000 academics in the social sciences, across Canada. They lobbied government on behalf of the interests of the community.

11 This latter group was involved because the task force wanted to establish the principle of ensuring that publically funded data were available in depository libraries (often referred to as the DSP program) in the same way that government publications were. This idea was not accepted but the DSP manager agreed to assist the group with their proposal. http://publications.gc.ca/site/eng/programs/dsp.html).

personnel and infrastructure costs, but did not cover the costs incurred by the producing divisions to provide support.

The leakage issue was dealt with by developing a license agreement that would be signed by each University Librarian and the designated DLI contact. This license approach was in keeping with the types of licenses that were being signed by other PUMF users and other data users. It was argued that if this type of license were sufficient to support Statistics Canada's sales activities, then it would be adequate for universities as libraries take the responsibility of managing licenses very seriously. This license has been remarkably effective. The penalty for a breach of the license would be the loss of data access to the offending institution. As a result DLI contacts are extremely diligent in applying the conditions and have a number of tools at their disposal to help in this regard.

The final push for accepting Data Liberation as a project came from the political side. Various groups and individuals had been lobbying the government of the day to encourage Canada's central government agencies and Statistics Canada to relax their cost-recovery policies. The specific Data Liberation proposal was seen by one of the ministers of the cabinet as a good way of improving data access and strengthening teaching and research. His support was signalled to Statistics Canada and the agency jumped on board at that time. This launched serious discussions about funding and implementation. The decision was made easier by the fact that a government-wide task force had noted that many of the policy departments, who used data to support evidence-based decision-making, were short of skilled staff to support quantitative analysis. The obvious answer was to encourage universities and colleges to make use of Canada's rich data resources for teaching and research and to build a community of highly qualified quantitative analysts. Not co-incidentally, it also meant that Statistics Canada would have a cadre of trained recruits who were familiar with the agency's data.

To ensure the long-term viability of the project, a submission was made to the Treasury Board (TBS) for formal approval. While the submission only sought start-up funds, the step gave it visibility and status and ensured a documented decision at 'the centre'. Treasury Board approval also formalized the agreements by the seven organizations to support it for up to five years. [12]

Once the project finances were dealt with, the challenge fell to Statistics Canada and the universities to develop their plans.

Building Support within the Canadian Academic Community

As might be imagined, gaining support within the academic community was not as difficult. The research community was onside and excited at the prospect of being finally able to conduct their research using Canadian data. That said, there were still some hurdles.

University libraries were not as enthusiastic. Expertise to support data services within libraries was in very short supply. Only a few of those running data services were librarians; many were social scientists located in research centres in various university departments. Academic librarians had little, if any, experience with quantitative analyses and less in dealing with data. Yet, if the project were to be successful, it had to have a permanent home in a service organization and the university libraries were the logical choice. After an intensive lobbying campaign by both the 'army' team and groups such as the Association of Deans of Graduate Studies, the university librarians accepted the challenge.

Preparations at Statistics Canada

Once the formal decision by the Treasury Board approved the DLI as a five year pilot project starting in January 1996, a number of steps had to be taken very quickly. Because the project required the rapid establishment of the technical infrastructure, it was housed in the Dissemination Division which had responsibility for online data dissemination and the Internet servers. A senior manager undertook to establish the necessary technical environment and the key staff. One of the

12 The organizations and the amounts pledged were:
- Statistics Canada - $100,000 per year (this has continued and was later increased to $175,000 per year)
- Human Resources and Skills Development Canada - $25,000 per year for 5 years
- Health Canada - $25,000 per year for 5 years
- Social Science and Humanities Research Council - $25,000 per year for 5 years
- Industry Canada - $200,000 one time contribution
- Treasury Board of Canada - $100,000 one time contribution
- Medical Research Council - $15,000 one time contribution

challenges was to make the investments as quickly as possible as the 'one time contributions' were due to expire after a few months.[13] The majority of the soft funds were spent on technical infrastructure (internet and network servers, workstations and software - SPSS, SAS, and internet tools such as FTP and Listserv[14]). A brief profile of the infrastructure used at Statistics Canada and the universities is shown in Appendix C.

An External Advisory Committee (EAC) was established and held its first meeting in February 1996. It was comprised of data specialists from large and small institutions, a University Librarian, a researcher and various staff from some of Statistics Canada's data producing divisions. The chair of the committee was an external researcher. Since that time, the EAC has been meeting twice a year.

The project staff included a project manager, a liaison person with training in library science, an information technology specialist and a support person. The operations were handled inside the Dissemination Division and the project guidance was provided by the EAC and the Director of the Library and Information Centre (LIC). A Board of Management (BoM) which involved representatives from each of the major stakeholder groups was established as requested through the approval decision and was tasked with overseeing a formal program evaluation before the end of year five.

Preparations at Universities and Colleges

While selling Data Liberation to the academic community was far easier than it was with Statistics Canada, the same cannot be said about the implementation of the project. Statistics Canada had a fully-formed, technically-savvy Dissemination Division who were able to get off the ground running in a matter of weeks.

The university libraries had, by and large, no such infrastructure for receiving the data, i.e. for being data repositories. Until the introduction of Data Liberation, university researchers in institutions without data libraries (there were only 9 data libraries at this time) dealt directly with data producers such as Statistics Canada. Data Liberation represented a new model where the libraries would act as repositories and play an intermediary role. Consequently, the libraries that were in the process of establishing data libraries, did not have the equipment to handle large files requiring, what was for these institutions, massive amounts of RAM, CPU and storage space. To add to the problem, they had no pot of money to draw on and, unlike Statistics Canada, were spread over more than 50 institutions.

It was evident that something needed to be done, and quickly to ensure that the universities had the capacity to provide data services. The first year was largely spent getting the physical infrastructure in place. This was a major hurdle, and resulted in the better-equipped universities sharing their infrastructure with their poorer colleagues.

Lack of trained personnel was also evident. To address this problem, the External Advisory Committee undertook the preparation of a comprehensive 'how to' manual and the development of a 3-day 'data boot camp' to give new data library recruits, who were being asked to be the DLI contacts, the basics of organizing and providing a data service. This became known as the 1997 bible and is still a part of the DLI reference collection. A train-the-trainers session was held in Ottawa where the manual was expanded and revised. Training teams ran four regional sessions in the West, Ontario, Quebec and Atlantic Canada. Although this didn't mean everyone was suddenly a data expert, it did build a bond and sense of community as contacts found that many of their colleagues at other institutions were facing the same problems. Additionally, the trainers were able to form a cadre of knowledgeable people who could provide support.

Launching Data Liberation

One could say that the January 1996 launch of the project was the 'soft launch' as the official launch did not take place until October of that year. By this time it was evident that the project was successful and it was a time to celebrate. A launch ceremony was held at Ottawa's Carleton University which was celebrating the 30th anniversary of its data centre. The Minister of State for Science and Technology (who had sponsored the Treasury Board submission) spoke as did the Chief Statistician and the President of the Humanities and Social Science Federation of Canada.

13 Because the fiscal year ended on March 31, 1996, $315,000 had to be spent in a period of just over 60 days.

14 Listserv was an early example of electronic mail list software to support communication among a group of people.

A major problem in the beginning was the difference in cultures. Statistics Canada was a production shop. The job was done when the file was out the door. If there were problems with the documentation, they were not dealt with at source. Consistency and completeness checks were left up to the recipients, most of whom had no idea how to remedy the situations. The university libraries, on the other hand, were into collection development and service. Ways had to be found to bring these two disparate views into line so both partners would benefit from the project.

Also, because there were over twice the number of participants than had been expected, the project risked drowning in its success. Statistics Canada had never operated anything remotely like DLI and the universities were equally virginal in the process. What was needed was a planning and management strategy that would ensure the success of the project but this did not seem to emerging from Statistics Canada.

After one of the early meetings, the university members of the EAC met for refreshments and mapped out a number of subprojects and activities as a starting point for discussions on establishing production-level procedures. This document became known as the "Bistro Manifesto".[15]

Bit by bit and piece by piece some of the EAC's plan for implementing the project was put into place. It was not, however, until after the 5-year evaluation that a transition document was written and the project relocated into the library at Statistics Canada, that the ideals expressed in the Bistro Manifesto were finally realized and the project moved into a service and collection development model.

Transition from Pilot Project to an Ongoing Program at Statistics Canada

Evaluation

The DLI project was initially approved as a pilot project for the period 1996-2001 which was to be evaluated in order to determine its future. Accordingly, part of the agreement with the TBS when the project was approved was that it would undergo a formal program evaluation after five years of operations. The task of overseeing the program evaluation fell to the Board of Management and was operationalized by the EAC co-chairs and the Director of the Library and Information Centre (LIC). The evaluation was framed by four key dimensions of the DLI, including:

- The financial structure and affordability of the DLI;
- Research and teaching using data obtained through the DLI;
- The program management and operations of the DLI and DLI team; and,
- Statistics Canada's role and participation in the DLI.

Appendix D summarizes the evaluation themes and the issues that were identified by each of the stakeholder groups.

The main findings of the evaluation were positive and paved the way for DLI to become part of Statistics Canada's ongoing program. The conclusions of the study highlighted three main findings:[16]

- The project had much greater support and was more successful than had been anticipated.
- The DLI approach to distributing data made access to Statistics Canada data for teaching and research more equitable across universities in terms of price.
- There was a wide gap in terms of the available expertise at some university data centers to provide a sufficient level of service.

In short, the evaluators felt that: "The long term viability of the project, in terms of service, training and infrastructure, depends on documenting a strategic plan for the future."[17]

This report was positively received and resulted in steps being taken to fill the gaps identified by the evaluators. The main steps to strengthen the project focused on governance and management and training.

Governance and Management

Obtaining approval, budgets and organizing the DLI were major tasks for the project. Its continued smooth operation depends on the governance and operational management. The project continues to be guided by the

15 See Appendix D for the text of the Bistro Manifesto.

16 Goss Gilroy Inc., Management Consultants, Evaluation of the Data Liberation Initiative: Final Report Ottawa, 1999.

17 Goss Gilroy Inc., ibid

EAC which, along with Statistics Canada was charged with ensuring that the project was on a steady course. The Internal Advisory Committee was disbanded after the completion of the project evaluation at the end of year five and key members were invited to join the EAC. A strategic planning session was held with senior management, staff and the EAC to set priorities for the continued development of the initiative.

One of the first changes was to address the organization and management of the project. It was moved from the Dissemination Division to the Library and Information Centre (LIC) at Statistics Canada[18]. There were a number of reasons for this move. The operations of the DLI had evolved from being a dissemination activity (where products are passed through to users) to one which involved building and managing a collection[19]. In other words, rather than simply sending files out, as is, on request, the DLI unit received the files and made certain they were useable[20] before sending them on. Once the file had been handled by the DLI unit, it remained in the collection ready for use by other members of the community.

While the move to the library also gave the LIC staff an opportunity to learn about data operations, the main impetus for relocation was that the director of the LIC (Boyko) was the co-founder of the initiative and this gave him an opportunity to build a long-term plan for the project.

Another change had to do with the appointment of a project manager who had experience with survey operations and the dissemination of PUMFs. By this time, there were over 65 universities and colleges in the project, a DLI staff of six and a steady flow of enquiries on the listserv. The web site was being developed as was a program of quality control for assuring the quality and the integrity of the surveys. There had been no systematic checks on the quality of PUMFs at the corporate level prior to DLI as the importance of standardized metadata had not yet been realized at Statistics Canada. While DLI disseminated PUMFs to the academic community, the files continued to be prepared by the individual data producing divisions and disseminated to non-DLI users.

What the Data Liberation Unit Does

The work of the DLI unit expanded over time due to the increasing number of universities that were part of the project, the number of files to be loaded and supported, the number of enquiries to be responded to and the nature of special projects such as documenting the data files. There are currently 92 different survey titles in the DLI collection for a total of more than 1300 survey files. There are over 50,000 files in the collection when one includes census and other tables and databases.

Messages on the Listserv currently average over 1700 per year. In addition, an equal or greater number of messages are received 'off list' (directly to the DLI unit) by members who choose not to make their enquiries public. The disadvantage of 'off list' messages is that they can only be answered by the DLI unit whereas public enquiries can be answered by anyone who is on the message list.

While DLI had been in a steady state for some time now, some factors were in flux. The External Advisory Committee had, in large part, been there since the beginning—early 1996. This had been one committee where members felt the need to 'fire' themselves in order to pass the torch to another generation. No one wished to quit, but the time had come for a transition. This will be complete in 2012.

In addition to the EAC renewal, the project had recently been moved from the Library to the Microdata Access Division (MAD). DLI is now co-located with the Research Data Centre project, a program that developed as a direct consequence of DLI. Once researchers were

18 However, as will be seen below, the evolution of Statistics Canada's microdata access program recently led to a further change for the project when it was moved from the LIC to Microdata Access Division.

19 Collection development ensures a focus on content curation. A description of collection development can be found at http://en.wikipedia.org/wiki/Library_collection_development

20 "Useable" in this context means consistency checks (e.g., to ensure that things such as 'pregnant males' etc. are not present) and ensuring understandable variable and value labels as well as the declaration of missing values where appropriate. This process ensures that the files can be used with SPSS as this was the most common software used by the universities.

aware of the breadth of the DLI microdata holdings, some required access to confidential data in order to pursue their research questions. Thus, a secure network of Statistics Canada sites (data enclaves) has been instituted at a number of universities across the country.[21] Finally, funding has stabilized with about a 70% university and 30% Statistics Canada commitment. In fact, the subscription price has not increased since the start of the project in 1996.

Training and Development

From its initial 'boot camp' in 1997, DLI has developed a strong training program for data librarians and other contact persons.[22] In fact, the training program has become a cornerstone of the Data Liberation Initiative. The findings of the evaluation *("There was a wide gap in terms of the available expertise at some university data centres to provide a sufficient level of service")* led to an even greater emphasis on training. Following is an outline of the DLI training program.

Each region holds a 2-3 day training session annually. Every four years when the International Association of Social Science Information Service and Technology (IASSIST) meets in Canada, DLI holds a national training session to encourage contacts to experience the cutting-edge work of the international data community.

To encourage participation, the project funds travel for one member from each institution to attend the regional and national meetings. Boot camps are held on an irregular basis and new contacts are encouraged to attend at least one. Travel for one boot camp for each contact is also covered by the program.

The training program is planned by the Education Committee which includes a Chair (an EAC member) and two Training Coordinators from each of the four regions. This committee has just completed the process of renewal and it is expected that this will only strengthen the function. Training retreats are held approximately every five years to expand the cadre of available instructors from within the community and plan the training program for the next five years. A good overview of the program may be found in the article *Introducing data into Canadian academic libraries: The straw that didn't break the camel's back*.[23]

In addition, a Training Repository has been set up and currently holds more than 350 presentations and exercises.[24] The original training manual has been converted to a Survival Guide which is updated on a regular basis.

User Support: Usability and the Evolution of Metadata

Having an active user support program has been an essential element of the project right from the outset. In addition to the training program outlined above, the DLI members need to be able to ask questions on a regular basis. A major lifeline in this regard is the DLI listserv which also serves as a means of providing appropriate services at all institutions. Contacts are encouraged to post whatever questions arise and the community and/or the DLI team provides answers. These range the gamut from fairly simple 'how do I?' to very technical, methodological questions.

The complexity and quality of the questions posted on the list have evolved from the early days of the project when DLI contacts had difficulty in finding the right surveys to respond to the needs of students and researchers. The reason for this was quite simple: The data were not supported by adequate metadata. The DLI experience in Canada underlines the crucial role metadata play in data dissemination.

As the quality of metadata improved, researchers became more self-sufficient and mediated access to find appropriate data became less of an issue. Nevertheless, as the number of files on the website increased, questions continued to be posted on the list.

Initially the only metadata distributed with the data consisted of codebooks, often on paper. This quickly led to the realization that further work by the DLI unit was necessary to properly document the files. The first stage involved ensuring that the data files could be used with SPSS and SAS.

During the second stage, the paper documents were scanned and placed on the web site so that they could be searched and downloaded. Prior to the establishment

21 For a full description of the RDC program, please see: http://www.statcan.gc.ca/rdc-cdr/index-eng.htm

22 Training of students and researchers is left to the universities to look after. In many cases, the DLI data librarians are involved in supporting this process through an outreach program and by offering statistical and other support services.

23 Humphrey, C and W. Watkins, Paper presented at ICOTS7, Salvador, Bahia, Brazil, 2006.

24 https://ospace.scholarsportal.info/handle/1873/69

of the DDI[25] as a documentation standard, some of Statistics Canada's data producers chose to use a tool called DDMS (Data Documentation Management System) a tool developed by Health Canada who used it to document their survey files and shared it with a number of other organizations. It was compatible with DDI version 1.

As the DDI standard evolved and the NESSTAR[26] Data Publisher became available, DLI started to document its files using this standard and tool. Following this standard had the advantage of allowing other groups (such as Ontario's ODESI project) to exchange metadata with Statistics Canada and with each other to support local initiatives and distribution. It was further refined by the development of a "Best Practices" document that was co-authored and shared between Statistics Canada and the universities. This document ensured that everyone marking up the files was adhering to a common interpretation of the tags.

The work in documenting files was also supported by the creation of MARC records (Machine-Readable Cataloguing - a data format and set of related standards used by libraries). This enabled the universities to include records for data files in their online catalogues to aid discovery.

Benefits of Data Liberation

Data Liberation has proved beneficial to not only its partners but to the Canadian population as well. Some of these benefits were recognized at the beginning and led to the development of the proposal. Others have been unintended consequences. All have been positive. Any worries about the loss of confidentiality have been found to be misplaced.

For Statistics Canada

Having an academic partnership is important for Statistics Canada's credibility. This is a large and important sector and when it is not well-served it can be quite vocal. It was embarrassing for Canada to have Canadian researchers using US data, especially when Statistics Canada had such a vast collection of available public use microdata. Establishing Data Liberation greatly reduced the criticism of Statistics Canada that publicly funded data were priced out of range for all but businesses and the well-heeled and that social data were underexploited.

Market research has shown that students continue to be large users of Statistics Canada's data and its website. Having knowledgeable people at 75 Canadian universities and colleges who can answer questions about Statistics Canada is an asset to the agency's dissemination program. The DLI partners essentially function like a network of mini satellite offices.

Data Liberation has added value and quality control to Statistics Canada's extensive collection of PUMFs. Statistics Canada continues to serve its commercial and government clients through the sale of PUMFs. As Data Liberation started to undertake data documentation activities it introduced the DDI standard to the work of Statistics Canada. By using standardized metadata based on DDI, the Data Liberation team has developed a collection of PUMFs that are now being marketed to the world. As well, as the standard evolves, it may enable Statistics Canada to pursue its desire to become a metadata-driven organization.

"The Data Liberation Initiative (DLI) is a highly valued partnership between Statistics Canada and postsecondary institutions all across Canada. Statistics Canada's contribution in this partnership continues to generate major returns in the form of better understanding and use of relevant and quality information in training and academic research for the benefit of all Canadians. The DLI has been a driving force in Statistics Canada's development of a comprehensive repository of all our microdata files. Through the strength of the DLI-Statistics Canada partnership, we have improved our products, upgraded documentation and continue to make progress towards new standards in access and dissemination. My appreciation goes out to all those who have helped form and strengthen this partnership, a foundation that should serve us well for many years to come."

Anil Arora, Assistant Chief Statistician, Social, Health and Labour Statistics Field (DLI Annual Report, March 2009.)

25 DDI stands for Data Documentation Initiative and is used as a model for documenting files such as those from household surveys and censuses. See http://www.ddialliance.org/

26 NESSTAR stands for Networked Social Science Tools and Resources. It was developed and is distributed the Norwegian Social Science Data Services. It comprises a suite of tools for data publishing and online analysis. One of these tools is called the Data Publisher which can be used to markup data according to the DDI 2 standard. See http://www.nesstar.com/index.html

Like many Canadian institutions, Statistics Canada is busy recruiting new staff to replace those who are retiring. Today's recruits have all had an opportunity to use Statistics Canada data while in university. This makes it much easier to find recruits who understand the work of the agency and are inclined to choose it as a career path.

For the Academic Community

Prior to Data Liberation, Statistics Canada's extensive data resources were available for university research on a very limited basis and were not routinely used for teaching. This has now changed completely. Data use is now spreading to Canadian colleges where they have not been a traditional resource. Data services are now available in the universities in every region of Canada.

In addition, the number of trained analysts graduating from Canadian post-secondary institutions has increased dramatically. As a result, students with strong quantitative skills are finding jobs in government and business, even in a tight employment market.

> "DLI was the start of a major national movement to free data and promote its use within the academic community in Canada. The collaboration and advocacy that DLI demonstrated and developed in our community led to other important data projects like the network of Research Data Centres. So thank you to all of you in DLI; you have done an amazing service for researchers, students, faculty and librarians across Canada."
>
> Margaret Haines, University Librarian, Carleton University
>
> (DLI Annual Report, March 2009.)

Data Liberation also led to the creation of data enclaves (Research Data Centres) which provide a rich research environment. The availability of detailed research files has led to the possibility of a student/researcher having training in sophisticated analysis through the RDC summer schools. Researchers are now able to pursue research questions using Canadian data. Graduate students no longer have to scrounge for grant money to purchase data to do their thesis research. University administrators now have up-to-date data to use for planning and marketing purposes.

For Canada

Canadian researchers are now addressing Canadian problems with Canadian data. They no longer face the situation where they use US data adjusted for differences in population, or simply find another research question that does not require the resource.

> "The DLI has been at the heart of profound transformations on campuses across Canada since the mid-1990s. In fact, the availability of Canadian data is helping us move beyond the false dichotomy of teaching and research by offering both undergraduate and graduate students the chance to describe, analyse and interpret key features of Canadian society. "
>
> Chad Gaffield, President, Social Sciences and Humanities Research Council of Canada
>
> (DLI Annual Report, March 2009.)

A study conducted by Hamilton and Humphrey[27] has documented the large flow of research output from the exploitation of a population-based health survey. Canadians are better informed about social and economic issues and changes in society. The analysis carried out by Canadian researchers adds to the analysis provided by Statistics Canada. Published outputs from statistical offices are typically descriptive in nature. Academic researchers can add insights by portraying causes and implications from the situations described by the statisticians.

> "….. DLI has significantly enhanced the quality of the student experience while increasing knowledge and understanding of Canada as a complex and diverse country. We are finally beginning 'to know ourselves'!"
>
> Chad Gaffield, President, Social Sciences and Humanities Research Council of Canada
>
> (DLI Annual Report, March 2009)

27 Hamilton, E and Humphrey, C *Data and the Life Cycle of a Survey: National Population Health Survey Outcomes, 1994-2002*, DLI Research Paper Series, La série de documents de recherche de l'IDD, Data Liberation, Statistics Canada, see http://dspace.hil.unb.ca:8080/bitstream/handle/1882/231/Research%20paper%201_life%20cycleR5.pdf?sequence=3

One of the aims of Data Liberation was to enable researchers to provide Canadians with evidence to guide policymakers and for students to acquire valuable skills which are marketable. This has increased the numeracy skills of new entrants to the workforce and has resulted in a better informed citizenry.

Unanticipated Consequences of Data Liberation

The Data Liberation project has now settled into a predictable and steady state but it is instructive to think back about the things that transpired that were not anticipated. Some of these have been previously mentioned, their importance bears a brief repetition.

Number of Participating Institutions

DLI has grown from the original 50 university members in 1997 to 75 universities and colleges in 2011. While it was not surprising that the university community would be interested in joining, it was not obvious that Canadian colleges would share that interest. Colleges and small universities now make up the bulk of the membership. As time goes on, more and more colleges are expected to join DLI. This is occurring for a number of reasons:

In addition to their teaching duties, college professors are now expected and encouraged to conduct research. Thus they need data.

The status of many colleges changed from 'college' to 'university-college' or 'university'. The DLI collection was required to support their new status and offer their professors and students the resources required to maintain their university status.

More specialized courses (e.g., GIS, Justice, Social Work, and Business Marketing) require data as part of the program. The tools to access and use the data are now readily available and this equips the graduates to be able to perform more effectively in the community they will be working with.

Some college professors would like to make the jump from college to university. In order to do that, they need to publish in peer-reviewed journals and often require data for their research.

The participation of colleges has provided a challenge for Data Liberation. Many of the programs being offered by colleges require that students become involved with the community in which they will be working and this may require the sharing of some data. This often requires a very careful application of the data license agreement to ensure that the data use conditions are not violated. This is can be done by ensuring that it is the research results that are shared rather than the data themselves.

Expansion of the Collection

While the original DLI proposal focused on microdata, the needs of smaller universities and colleges were quite different. Practical programs in GIS, social work, criminology, justice and health often have data requirements which require aggregate data or for which microdata are not available. These were welcome additions to the DLI collection that enriched the work of all its members.

The Emergence of Research Data Centres

The use of PUMF files in DLI has led to the establishment of an access method for more detailed confidential data files in the Research Data Centre network. See http://www.statcan.gc.ca/rdc-cdr/index-eng.htm. The RDCs have been particularly valuable as a means to access surveys for which an anonymized file cannot be produced (e.g., longitudinal surveys) and for conducting analysis which cannot be supported without access to the more detailed files.

The Central Role of Training and the Evolution of the Data Services Staff

The training component of the project has been one of its most valued and enduring services to support the members of the network. This increases the strength of the network and bodes well for the future of data services in Canada and the guidance of the project in the future. Library schools are now beginning to see "Data" as a professional stream.

The Wide-Spread Acceptance and Use of the Term 'Data Liberation'

Fifteen years ago, 'data liberation' was the name attached to a data project that was proposed to the Canadian Government by the academic community. After its Canadian success and numerous presentations by the co-founders and other Canadians at international

conferences, the project is well known and the term 'data liberation' is used in many instances where the data user community is attempting to make priced data more available. This is especially true when participants in the 'Open-Data' community discuss their desire to have more access to local government data.

Interest in Canadian Data

There have been a limited number of researchers from other countries using Canadian data for a number of years, but having the DLI web site open to the world has brought added interest to some of Canada's more innovative surveys. While non-members are able to browse the metadata they are not able to access the DLI data files. Instead, they have the option of approaching the data producers to buy the files or to negotiate some form of access. Numerous requests from US based universities led Statistics Canada to tailor a service to suit these requests. See http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=11-625-X&lang=eng

Lessons Learned

The Canadian experience with Data Liberation has demonstrated the strength of partnerships between statistical agencies and academia. Fears about the unauthorized redistribution of data should not pose a problem if there is a clear license. It was found that academics will police themselves, especially if they run the risk of losing access to the data as a result of license violations. Questions regarding authorized use are reviewed by a subcommittee of the EAC and their decisions have been respected by the players involved. A common resolution of a request for use of data that could not be sanctioned by the review group was for the user to approach the author division at Statistics Canada to obtain permission or to purchase the file for their own use.

Current dissemination models used by statistical offices may need to be reassessed and adjusted to accommodate initiatives such as DLI and Open Data. For example, the DLI represented a shift from a revenue-generation to a cost-avoidance model within Statistics Canada. At the same time, Statistics Canada was able to continue to use its traditional model for other sectors such as government and business.

It was also shown that well-documented data files lead to more use and require less support than poorly-documented data files. Proper data documentation standards and well-supported metadata are an important part of this process.

Having a single contact at each university and a liaison officer at Statistics Canada is important in reducing support costs. Carrying out user support through open communication systems such as listservs, keeps everyone informed and reduces costs even further.

Data and data services are complicated. The importance of a well-designed training program cannot be over-emphasized as it is one of the keys to the Canadian success. These programs should be designed and delivered in partnership between the data users and the data producers.

Statistical organizations should not be afraid of differences between the official numbers that they publish and those found in the public research domain. This is bound to happen but it must be remembered that only the statistical organizations have the complete data files. Data files released to academia have had data reduction techniques applied to them to ensure confidentiality. Thus the versions held by the statistical organization will always be authoritative. Differences in results that may arise due to errors in the dataset should cause the dataset to be corrected. Since most of the material published by researchers is peer reviewed, there is ample opportunity to detect errors. Above all, it must be stressed that data files must be well-documented to ensure that users know the proper meaning and limitations of the data.

In addition, the more that the data are used the more errors will be found in the documentation and even in the data (coding). This should be viewed as a positive as it leads to better data quality and analysis in the long run.

Looking Back and Ahead

If DLI were to be implemented in today's environment instead of 15 years ago, it would be much easier. A few reasons are noted below.

- The idea of making more data available more easily or even freely available is being fuelled by initiatives such Open Data and changing policies in data producing agencies.

- Using the internet for transferring files is now taken for granted and there are a number of tools now available to support this work.
- The broad acceptance of the DDI standard provides a robust way to document data files. Using this standard also facilitates the use of search engines to support searching at the variable level. Any NSO that is using DDI-compliant data documentation and cataloguing tools has much of the necessary infrastructure to support an initiative like Data Liberation.

In looking ahead, it is important to maintain the momentum of the project. Many of the players who have guided the work of Data Liberation have been involved in the project since its inception 15 years ago. Groups like the EAC have remained relatively intact during this time but are now starting to undergo a 'changing of the guard'. New players will undoubtedly bring different perspectives to this work. That said, it is important that the academic component of the EAC continues to have strong analytic capability as well as strong library skills to ensure a high level of professional service. Without quantitative research experience, it will be difficult to develop both effective collections policies and training programs.

The DLI financial model has been stable since the beginning of the project. The fees paid by universities have not changed during this time. However certain costs (for example staff costs) have been increasing. It is unclear whether the project can continue to generate enough efficiency to avoid fee increases. If the fees must be increased, how can this be done in an equitable fashion?

A question that arises from time to time has to do with the possibility that the government of Canada may change its policy regarding charges for data. If this were to happen, what would be the impact on DLI members? The current fees pay for the service and not for the data. And yet, if the DLI collection were to be opened up to the whole world, there would undoubtedly be increased support costs. The question remains as to whether or not Statistics Canada would be able to obtain funding to offset the increases in demand and maintain current service levels.

One of the most important assets of a statistical agency is its reputation for producing quality information and the trust of the population that it will maintain the privacy of respondent data. Microdata files are valuable resources that agencies should learn how to produce, utilize and share. These files have proven to be excellent tools for expanding the knowledge that the country has about itself and for training future researchers and analysts.

Public use microdata files can be created from confidential master files by following well-documented procedures for the anonymization of data. Proper documentation and quality control for these files is also essential for a successful dissemination activity and can be shown to reduce support costs significantly.

The academic sector should be viewed as an important ally to be served by the national statistical organization either through a partnership arrangement or some other means. Affordability of data is a key ingredient to successful partnership with this sector as they are generally poorly funded and have little in the way of discretionary budget. Viewing academia as a partner rather than a 'customer' can result in a win-win situation. Greater use of data files increases the return on the investment made by the funders, the quality of the files is improved, more knowledge is created and numeracy and statistical literacy skills are improved. Numerate students are more likely to find employment and the talent pool for statistical agency recruitment is increased. In an ideal world one would hope that greater access to data should lead to more informed policy-making. This would indeed be a win-win-win for government, academia and the country.

Appendix A: Data Liberation License [28]

Data Liberation Initiative

DATA ACQUISITION AND USE AGREEMENT

between

Statistics Canada

and

Educational Institution

(Name) Address

as represented by:

(Name)

(Title)

I, the undersigned, do certify that my educational institution is an accredited Canadian post-secondary educational institution, and has committed to be a financially contributing member of the Data Liberation Initiative (DLI).

I understand that:

1) via the Data Liberation Initiative (DLI), Statistics Canada will offer my educational institution, timely access, on a subscription basis, to standard Statistics Canada data products, such as public use microdata files (non-identifiable datasets containing characteristics pertaining to surveyed units), standard files and databases (containing aggregate data as defined and determined by Statistics Canada) and geography files, in available electronic formats.

2) the Government of Canada is the owner, or the licensee, of the intellectual property rights (including copyright) in these data products, and this Agreement is only a license to acquire and use these data products. No title or other rights are conveyed by this Agreement.

3) the data products are provided "as is", and the owner makes no representations or warranties, either expressed or implied, as to the appropriateness and fitness for a particular purpose.

4) the data products are provided to my educational institution for the exclusive purposes of teaching, academic research and publishing, and/or planning of educational services within my educational institution, and may not be used for any other purposes without the explicit prior written approval of Statistics Canada.

5) the data products are to be made available only to educators, students and other staff members of my educational institution (referred to herein as "authorized users") and only while they have such status with my educational institution.

6) authorized users associated with my educational institution are prohibited from using the data products in the pursuit of any contractual or income-generating venture either privately, with other government departments, or under the auspices of my educational institution, without the express written permission of Statistics Canada

28 Retrieved from http://www.statcan.gc.ca/dli-ild/caselaw-jurisprudence/license-licence-eng.htm on June 17, 2010.

7 authorized users shall not attempt to re-identify the records on the microdata files so as to relate the particulars to any individual person, business or organization.

8) the distribution of any data obtained under this agreement outside my educational institution through sale, donation, transfer or exchange of any portion of the data in any way is strictly prohibited, with the exception of distribution to bona fide participants in the Data Liberation Initiative. The DLI Contact of my institution must agree to such distribution.

9) the publishing of analysis and results from research using any of the data products is permitted in research communications such as scholarly papers, journals and the like. The authors of these communications are required to cite Statistics Canada as the source of the data, and to indicate that the results or views expressed are those of the author/authorized user and are not those of Statistics Canada. Permission to include extracts of these data in textbooks must be obtained from the Licencing Section of Statistics Canada's Client Services Division.

10) security measures must be implemented to prevent unauthorized access to the data products.

11) my educational institution will be invoiced by Statistics Canada once a year, during the second quarter (April-June) for access to DLI products and files, and payment shall be remitted within ninety days of receipt of invoice. Exceptions to this procedure must be approved by the Director of Statistics Canada's Communications and Library Services Division.

12) this agreement shall remain in effect until it is terminated by either party after giving one year's notice.

13) should my educational institution choose to withdraw from this agreement or not respect its financial commitment to DLI as per clause 11, this agreement will be terminated.

14) unless my educational institution has negotiated with Statistics Canada conditions under which it could continue to provide access to data products that it obtained under this Agreement, it shall:

    a. stop providing access to the data products and inform users accordingly;

    b. immediately take measures to terminate all uses of data obtained through DLI by all authorized users within the institution and in any partner institutions;

    c. destroy all copies of the data and data products;

    d. provide written certification to Statistics Canada that destruction of data and termination of access and notification have occurred.

15) should my educational institution choose to withdraw from the DLI and decide to rejoin within a period of five years, I understand that it will be subject to a penalty equivalent to the annual fees that my institution would have paid during the time it had withdrawn from the DLI.

16) any repeated or unremedied violations of this Agreement may result in the termination of this Agreement by either party upon written notice to the other party.

17) authorized users shall be made aware of the conditions of use of the data by being provided with a copy either of the data use license which constitutes Appendix 1 to this document, or of a document or notice containing similar information.

I acknowledge that I have read and understand the terms and conditions under which the data products are supplied. I agree to abide by these conditions and to take all reasonable measures required to enforce and administer them within my educational institution.

Representative of the Educational Institution

Name Signature

Witness - DLI Contact Name

Signature Date

Representative of Statistics Canada

Name Signature

Date

Data Liberation Initiative

MODEL DATA LICENSE

The Government of Canada is the owner, or the licensee, of the intellectual property rights (including copyright) in the data products offered under the Data Liberation Initiative, and this license is only a license to use these data products. No title or other rights are conveyed by this license.

These data products are provided "as is", and the owner makes no representations or warranties, either expressed or implied, as to the appropriateness and fitness for a particular purpose.

The data products are to be used only by educators, students and other staff members of this educational institution (referred to herein as "authorized users") and only while they have such status with this educational institution.

These data products are provided for the exclusive purposes of teaching, academic research and publishing, and/or planning of educational services within this educational institution, and may not be used for any other purposes without the explicit prior written approval of Statistics Canada.

Authorized users are prohibited from using these data products in the pursuit of any contractual or income-generating venture either privately, or under the auspices of the educational institution.

Authorized users shall not attempt to re-identify the records on the microdata files so as to relate the particulars to any individual person, business or organization.

Copies of the data products can be retained by authorized users for the period necessary to conduct the research or teaching, including their use as evidence of research methodologies and results, and subsequent research and analysis.

The distribution of any data obtained under this agreement outside this educational institution through sale, donation, transfer or exchange of any portion of these data in any way is strictly prohibited, with the exception of distribution to bona fide participants in the Data Liberation Initiative. The DLI Contact of my institution must agree to such distribution.

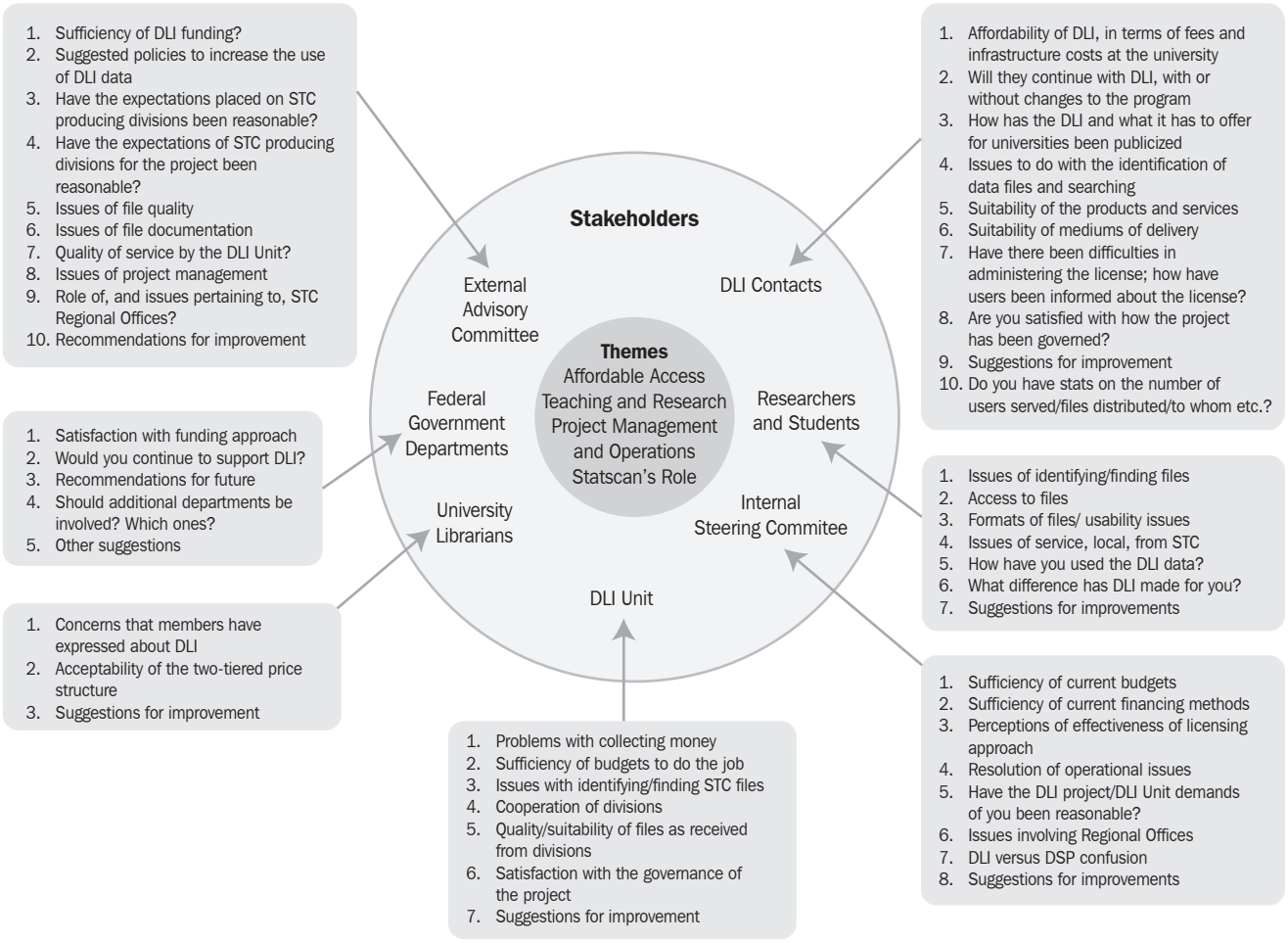
The publishing of analysis and results from research using any of these data products is permitted in research communications such as scholarly papers, journals and the like. The authors of these communications are required to cite Statistics Canada as the source of the data, and to indicate that the results or views expressed are those of the author/authorized user. Permission to use extracts of these data in textbooks must be obtained from the Licencing Section of Statistics Canada's Client Services Division.

Further information on the conditions of use can be obtained from <<DLI CONTACT>>

Appendix B: Evaluation Themes, Stakeholders and Issues

The following exhibit[29] presents the priority evaluation issues that were identified by the client-consultant project team. Each set of priority issues was tailored to align with the types of interaction each stakeholder group had with the DLI.

Exhibit A-1 - Primary DLI Evaluation Themes, Stakeholders and Issues



29 Goss Gilroy Inc., Management Consultants, Evaluation of the Data Liberation Initiative: Final Report Ottawa, 1999.

Appendix C: An Overview of the Data Liberation Technology

Over the past 15 years, the technology that supports Data Liberation has evolved both at Statistics Canada and among the universities but the basics have remained the same. Virtually everything is done using the internet, thus appropriate access is essential. As computers have become cheaper, and faster with increased CPU and memory, upgraded machines have been added. As the size of the collection has grown, more storage has been added. However, the biggest changes have occurred in the metadata content of the files and the manner in which the universities have banded together to serve each other to reduce duplication of effort.

Statistics Canada:

Networks: Statistics Canada has two computer networks only one of which can communicate with the outside world. Thus, the DLI project has two environments to support and requires three servers: one on the internal network where the files are prepared, a second one on the external network where access to the files is provided to the universities and a third one that provides FTP access to the data files.

Fibre Optic Backbone: The DLI has used various different networks to serve its users. For a time period, it used the fibre optic network supported by Canada's Advanced Research and Innovation Network (CANARIE) as it was faster than the one used by Statistics Canada. After Statistics Canada upgraded its internet access, DLI is now using the agency's network.

Software: The main software required was SPSS, office work station suites, FTP and utilities for compressing files (e.g. Zip). Listserv was chosen as the electronic mail software and is still being used. A DDI mark-up tool from NESSTAR is also being used now. Stata and SAS are now also in demand.

Content: The DLI content has evolved more than the rest of the technology has. Surveys are being added on a continuous basis and now include more than 1,300 public-use microdata files from 92 surveys, including surveys such as the Canadian Community Health Survey, the General Social Survey, the Labour Force Survey and the Census of Population. It also includes databases such as the Social Policy Simulation Database and Model and the database of Inter-Corporate Ownership.

The main tool for finding and accessing public use microdata files has been a database which provides access to survey details via an alphabetic list of titles which can be browsed or searched using a locally developed search engine. See http://www.statcan.gc.ca/cgi-bin/spider/dli.cgi

The content that the user finds after having selected one of the surveys varies but has continued to be enriched over time. For example, selecting the National Longitudinal Survey of Children and Youth - Cycle 1 (see http://www.statcan.gc.ca/dli-ild/data-donnees/ftp/nlscy-elnej/nlcsy-elnej-cycle1-eng.htm) brings the use to a page that contains access to the questionnaires, codebooks, user guides and the data. The data files are locked and available only to DLI members but the rest of the documentation is open to all users.[30] Initially, the SPSS and SAS files were very basic but over time, and with the help of the DLI team and the DLI members, the SPSS files (especially the variable labels) have been enriched.

While the Statistics Canada User Guides and Codebooks have always been essential tools for users, there was no way in which search this collection at the variable level. This was not possible until the last few years when the DLI project decided to start marking up its files by following the DDI standard and serving them to users using NESSTAR technology.[31] The metadata and other survey information are open to the public (see http://www62.statcan.ca/webview/) but once again, the data are not. Until this service became available, users had to browse codebooks in order to find variables of interest. (In some cases users would download the documentation and search it using their own tools in their own environments.) Since these data files are DDI compliant, the DLI could choose to migrate these data and metadata to other software environments (that are also DDI compliant).

30 It should be noted that changes in the Government of Canada's web access policies for the public at large, may cause DLI to move from open-access to its metadata to moderated access such as IP recognition.

31 Mention that the Nesstar Publisher is a free specialized DDI 2.n editor. Provide the link to the Publisher.

The Universities:

The situation at universities was quite varied. Those with existing data services had well-established computing facilities and continued to use them. This often comprised UNIX computers with lots of storage.

For those that were just starting up, they had to make choices as to what level of service to provide. As a minimum they were able to provide a 'pass through' service where they downloaded files from the DLI site using FTP and passed them to researchers and students on an 'as is' basis.

As time went on, decisions were made to provide a higher level of service but rather than having each university develop its own tools, they purchased or developed systems and chose to work together.
Some examples are as follows:

- The University of Toronto's Computing in the Humanities and Social Sciences (CHASS) developed a process for downloading Statistics Canada's CANSIM time series and making them available to other for a small fee. http://dc.chass.utoronto.ca/
- The Ontario University Libraries developed the Ontario Data Documentation, Extraction Service and Infrastructure (ODESI) which uses the DDI standard and includes large numbers of files in addition to those from Statistics Canada. It uses NESSTAR technology to provide data services. See http://search2.odesi.ca/
- The University of Western Ontario developed their Equinox system which is available on a subscription basis to others. As it has a bilingual (English French) interface, it is widely used by Quebec universities, most of who operate in French. See http://equinox.uwo.ca/EN/ProjectOverview.asp
- A number of universities have acquired NESSTAR technology (see http://www.nesstar.com/) while others use Survey Documentation and Analysis (SDA) from the University of Berkeley. See http://sda.berkeley.edu/

In short, the universities have cooperated in developing systems that suits their diversity and that of their clients.

Other options available to statistical agencies

Without a doubt, the internet is an essential component of the service, but even more important is the need to follow a data documentation standard such as DDI and have the tools to search and download the data. The work of the IHSN is exemplary in this regard. The Microdata Management Documentation Toolkit (see http://www.ihsn.org/toolkit) is an excellent way to document data files and the National Data Archive (NADA) open source software application developed by the IHSN solves the problem of searching for and downloading data. See http://www.ihsn.org/nada.

Appendix D: Bistro Manifesto

Reproduced with permission from an email sent by Chuck Humphrey after a three day meeting of the External Advisory Committee)

On Monday, 24 March 1997, Chuck Humphrey wrote:

Following Friday's EAC session, a few of us went for refreshments and discussed the range of topics that had been covered at Thursday and Friday's meeting. The consensus was that some detailed project planning is critical at this stage of the DLI pilot.

The first year of DLI required fast-tracking a number of things, such as getting the ftp site up and operating, creating a quick email order list, enlisting author divisions within Statistics Canada, getting recent titles onto the ftp server, settling the DLI license, and preparing to train university contacts in the use of DLI. For the most part, we have been quite successful in launching DLI using this fast-track approach.

Reflecting on the first year, the overwhelming response by universities to DLI far exceeded everyone's projection of involvement. Managing DLI would be quite different if there were only around 15 to 20 institutions involved, which was an early estimate of the number of subscribing members. DLI now has close to three times this number of participants; and consequently, there are far greater demands and expectations of this pilot. How do we deal with these growth pains of success?

Some of us think project planning and management are very important at this stage. Let me begin by outlining seven subprojects with some of the activities needing plans and coordination among these subprojects. I'm not suggesting that these seven projects are exhaustive of the planning that is needed, but offer them as a beginning point for further discussion. Also, not all of these subprojects need to begin from scratch. Most of them have some level of development. The challenge now seems to be one of charting the transformation of some of these subprojects into production level procedures.

Seven DLI Subprojects:

The Order Process

Activities:

a. Processes for Demand Orders (DLI Contact requests)

   i. define procedures for demand orders

b. Electronic order form (Web form)

   i. migrate from email ordering to e-form ordering

c. Order Database (records of all orders)

   i. create an order database with linkages to e-form

d. Processes for Standing Orders (CD ROM titles)

   i. identify titles on standing order

   ii. determine procedures for filling standing orders

Data Accession Procedures

Activities:

a. Linkage to the Order Procedures

   i. define linkage between order dbase and initiating an accession

b. Collection procedures to gather files within STC

   i. define procedures for acquiring files

c. Verification and Completeness Procedures

i. initially, develop standards for file products received, including naming conventions

ii. convert current collection to DLI standards

iii. procedures for receiving STC files and generating DLI standard products

d. Procedures for passing files to FTP server

e. Procedures for passing along cataloguing information

f. Procedures for updating the information on the Web site

Data Management Procedures

Activities:

a. Linkage to Accession Procedures

i. Directory structure and naming procedures

b. Security and authentication procedures

i. Access to the FTP server

c. Backup procedures

d. Mirror site procedures

Institutional Liaison

Activities:

a. Governance and oversight plan

b. Communications plan

i. Email lists and their purpose

ii. DLI Update -- communications with DLI Contacts

iii. Communications with STC author divisions

c. Problem-solving procedures

d. Licensing and Billing procedures

Training

Activities:

a. DLI staff training plans

b. DLI contact training plans

c. Developing a Data Culture

Historical Files Project

Activities:

a. Identification of files

b. Retrieval of files and linkage to accession procedures

Evaluation Project

Activities:

a. Setting the evaluation process and criteria

b. Setting measurement landmarks and/or benchmarks

Within each of these subprojects, the activities need to be clearly stated and then staff, timelines and resources need to be identified. Here is an example of one possible plan for electronic form ordering.

Time in Days

Design Order Database Elements |-----|

staff: 2

Design Order Form |----|

staff: 1

Test Order Form |-----|

staff: 6

Migrate Order Process from Email |----|

staff: 1

Putting together plans for subprojects may require bringing someone onto staff who can dedicate her or himself to this enterprise since it will require full-time attention for a fixed period of time. Once plans have been developed, implementation could be passed to current management for enactment and supervision.

Appendix E: Strategic Plan

The main elements of the strategic plan were:

- The development of the collection and continued improvement to its accessibility
    - Expand DLI data collection to include data from other data producers. PUMFs had already been augmented with all publicly-available databases and geo-spatial files. There are more than 50,000 of the latter in the current collection.
    - Improve quality and standardization of metadata. Use of DDI was becoming more widespread and the decision was taken to use this standard to mark up the microdata files in the collection in for the same reasons as those undertaken by the IHSN.
    - Improve the search and download capabilities.
    - Work with the group that was developing the Integrated Metadata Base (IMDB) to ensure that DLI metadata goals were in harmony with Statistics Canada's. Since the IMDB is a metadata registry based on concepts, it was felt that there had to be a clear cross-walk between it and the elements in the DDI.
- Strengthening relationships with partners
    - Communications and promotion of DLI.
    - Inform all major stakeholders about the project and its directions.
    - Attempt to provide data providers of some measure of use of their products.
    - Continue the training program as it had emerged as one of the key activities for the project.
    - Enhancing the DLI section and service to members.
    - Expand DLI membership to include colleges.
- Enhancing the technological infrastructure supporting DLI.
- Continue with the DLI governance process involving the EAC and preparing for membership renewal.
- Conduct research and development activities to support access and use of DLI data.

This outline became a standard item on each fall EAC agenda so that progress could be reviewed and new action items established. It became part of the work plan of the DLI section.

About the IHSN

In February 2004, representatives from developing countries and development agencies participated in the Second Roundtable on Development Results held in Marrakech, Morocco. They reflected on how donors can better coordinate support to strengthen the statistical systems and monitoring and evaluation capacity that countries need to manage their development process. One of the outcomes of the Roundtable was the adoption of a global plan for statistics, the Marrakech Action Plan for Statistics (MAPS).

Among the MAPS key recommendations was the creation of an International Household Survey Network. In doing so, the international community acknowledged the critical role played by sample surveys in supporting the planning, implementation and monitoring of development policies and programs. Furthermore, it provided national and international agencies with a platform to better coordinate and manage socioeconomic data collection and analysis, and to mobilize support for more efficient and effective approaches to conducting surveys in developing countries.

The IHSN Working Paper series is intended to encourage the exchange of ideas and discussion on topics related to the design and implementation of household surveys, and to the analysis, dissemination and use of survey data. People who wish to submit material for publication in the IHSN Working Paper series are encouraged to contact the IHSN secretariat via info@ihsn.org.

www.ihsn.org
E-mail: info@ihsn.org