Establishing a Data Resource Centre

Experiences at the University of Guelph

by Bo Wandschneider & Doug Horne *

Introduction:
The following paper outlines the process of establishing a Data Resource Centre (DRC)[1]. The paper documents the experiences at the University of Guelph, where such a service was established from scratch, and where gains have been made relatively quickly. Prior to the fall of 1996 Guelph was in a situation similar to many other research/teaching institutions. There were no formal procedures in place for acquiring, distributing and analyzing data in an electronic format. It was the responsibility of individual faculty, researchers and students to develop the necessary skills to make use of data, there was limited statistical support, and overlap existed in acquiring data resources. All of this resulted in duplication of effort with respect to the use of electronic information on campus.

It is hoped that an account of these experiences can be of use to others currently in the process of establishing a DRC, as well as those considering undertaking such an endeavour. To that end, this paper is written in an easy to follow manner, with limited technical details. Certain goals and objectives were set, and this paper looks at how these goals are being achieved and some of the obstacles encountered. Issues such as motivation, targeted audience, teaching needs, research needs, levels of service, staffing, hardware, software, security, and delivery tools will be discussed.

Establishing a data resource centre (DRC) can be a very complicated process. For the people involved in the front line delivery and use of electronic information, the needs and benefits have previously been laid out and shown to be substantial[2]. The challenge to the manager of a data centre is to express these benefits in such a way that the administrators who control funds see the need to commit scarce resources to this type of service.

Background:
The University of Guelph is a major research institution in Canada, with approximately $81 million dollars in research grants per year. The undergraduate enrollment is approximately 10,500 students with another 2,000+ graduate students. The University is broken up into six separate colleges including the College of Applied and Human Sciences, Ontario Agricultural College, Ontario Veterinary College, College of Biological Sciences, College of Physical and Engineering Sciences and the College of Arts. Recently several new remote campuses were added that deal specifically with agriculture. There are also significant ties with the Ontario Ministry of Agriculture and Rural Affairs. The OMAFRA head office was recently relocated to Guelph.

At the moment, the biggest users of DRC services seem to come from the first two Colleges, although clients are spread amongst all groups. The nature of DRC holdings and background of staff members has lead to heavier use from departments such as Economics, Geography, Sociology, Rural Planning and Development, Agricultural Economics and Consumer Studies. As the service grows and new contacts are made, and a new population of users is developed, it is expected that this will change. It has been found that one of the most important tasks, and a possible problem, is informing the user community about what is being done. Experience to this point has been that once contact is made with people, and there is an actual demonstration of the capabilities of the DRC, response is extremely positive.

The process of establishing a DRC at the University of Guelph actually began in the fall of 1993. The pilot project did not begin until December 1996. In order to get support for the project a detailed proposal was written, outlining all of the possible options for running a DRC. A great deal of this information was developed from a workshop offered during the summer program at ICPSR[3]. This was augmented with tours of the University of Toronto Data Library and CHASS facilities, University of Western Ontario's Social Science Data Centre, and input from the Canadian Association of Public Data Users (CAPDU). The proposal was very detailed, defining the users, the benefits, the possible levels of service, considerations for a suitable computing environment, the departments capable of managing the service, the costs, staffing, and what other institutions in Canada were doing. This document was used as background information to justify and explain options of how a DRC could be run.

Until this point in time, individual faculty members had been responsible for managing their own data needs. This usually entailed hiring a research assistant and spending time and resources getting these individuals 'tooled-up' to using whatever data they needed[4]. The net result was that there was frequent duplication of efforts, especially for major data sets, and there was a great deal of frustration for many researchers.

The ideas presented in the initial proposal were well received by various groups and individuals on campus. Generally the response was that such a service had been needed for a long time and would be of use and welcome on campus. The problem was to not only find the resources to get this project to run, but to find sufficient resources to make it function properly. After a long period of preparation and waiting for the right circumstances, it was clear to those involved in the project that initial impressions of the functioning service must be positive. If the service lacked support from the beginning, and did not manage to impress the various stakeholders, it would be very difficult to attract and maintain the client-base needed to develop a commitment in the long term. The pilot project, it was clear, was going to be an all or nothing situation.

It took approximately four years to move from the initial ideas to the start of the actual pilot, during which there was a great deal of lobbying done by the interested parties. In the mid 1990's the University was developing a detailed strategic plan and the need to address electronic information was included as a small paragraph in the plan. This was an important step, as it became clear that the idea of a Data Centre was becoming central enough to the University's plans that it was being discussed in various high-level committees. Partly in response to this, a beta version of the web retrieval system was developed over a few days[5]. This was extremely useful for presentations, and faculty were able to clearly see the potential and the possible applications of this system. With a working prototype in place, it was much easier to generate enthusiasm about what was being proposed, and a number of presentations to potential users where very successful during this period. This was also about the time that the Data Liberation Initiative was being established by Statistics Canada, with the basic idea of making the data more available and universally accessible. It was very clear that the University needed something like a DRC to take advantage of the opportunities that were being presented. At the same time there was a movement at other institutions in Canada to establish data centres. All of these factors helped to make the Data Centre seem like a feasible, and particularly timely, project.

The initial proposal was a for a collaborative effort between the Library, Computing and Communications Services and the College of Social Science. The basic resources being committed included the secondment of staff, infrastructure money (which was very limited), and physical space to house the centre. Central computing facilities such as a UNIX system and software, already in place, were also used. The collaboration between the Library and Computing Services (The College of Social Sciences dropped out early on) was very important in that it brought together a diversity of skills that is still reflected in our current staff. Our planning had always taken into account that useful skills could be drawn from the computing and library fields, and that input from the user community is vital. The latter includes the group that uses the information, be it researchers or teachers. As pointed out by Kroeker (1997) you need to know your patrons and what their needs are.

Figure 1 gives an example of the environment that existed prior to the establishment of the DRC. The University can be simplified by dividing everyone into 4 groups. Assume students gain access through any of the four groups. These 4 groups communicate in an ad-hoc fashion, as depicted by the dashed lines. Note that there is no direct communication between type A and type B researchers-teachers. The distinction between the two groups is that type B have direct access to incoming data. This may be due to certain skills they posses, or their access to resources needed to acquire this data. The problem lies in the fact that data flows into either the library, or type B researchers/ teachers. It is not clear where data will finally reside and the information flow related to data is poor. This is especially true between group A and B.

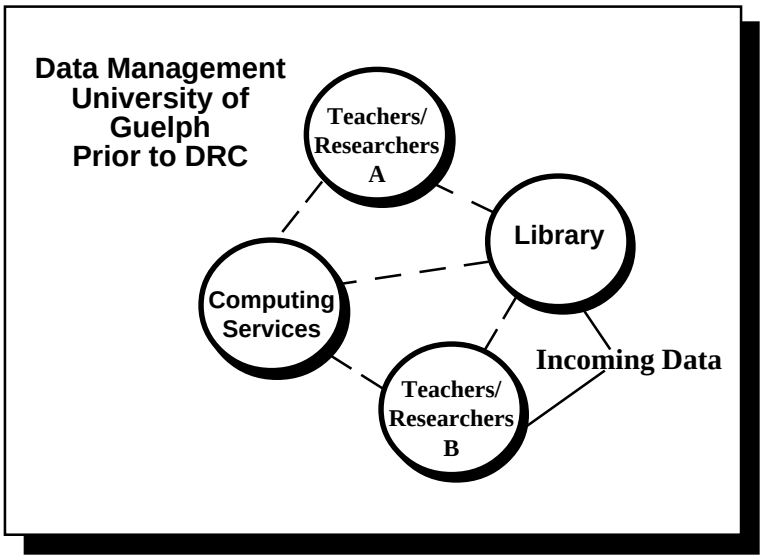

Figure 1

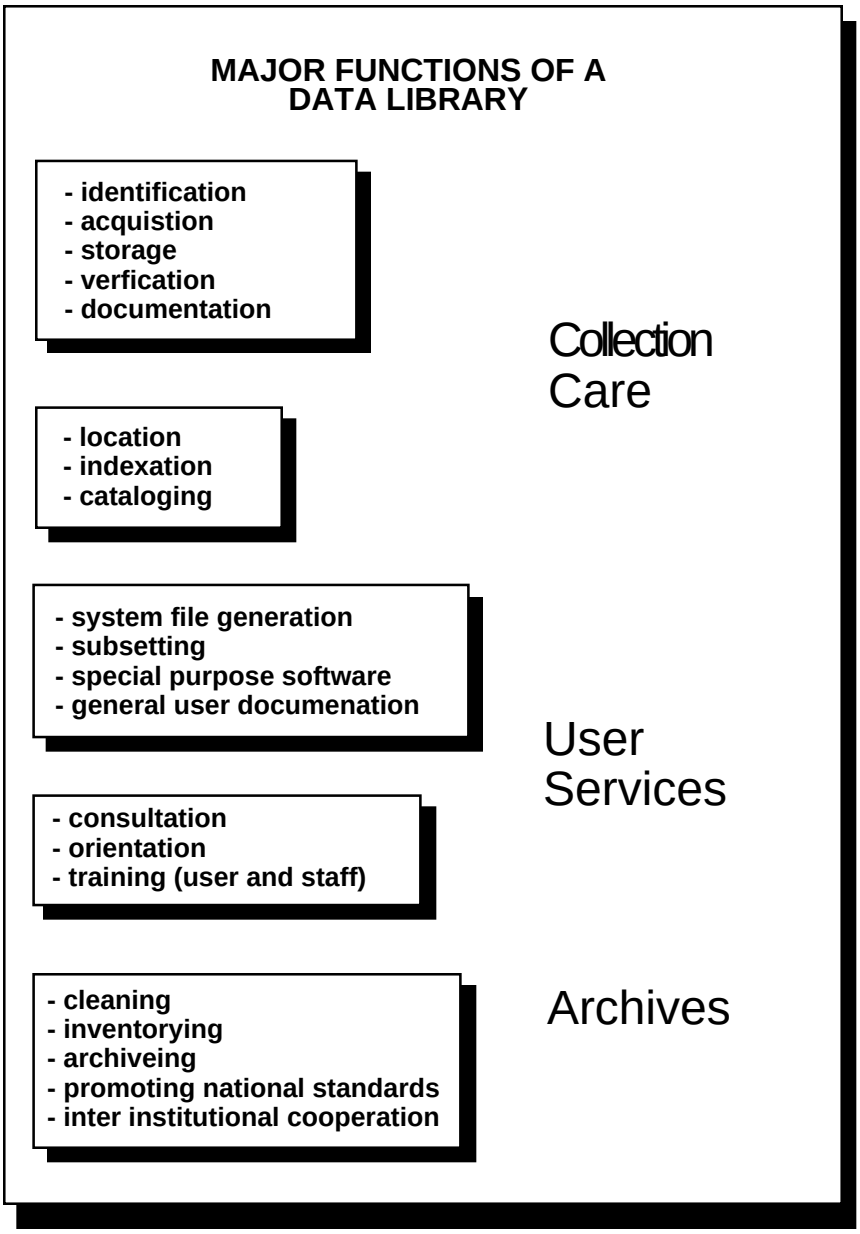

figure 2

Services

A paper by Jacobs (1991) gives a very good outline of the needs and ways to deliver data. Jacobs lists users expectations and the services associated with these expectations, breaking it down into general library services, references services and computing services

Early on in the proposal stage there was a need to clearly define the services that were going to be offered at Guelph. Figure 2 outlines different levels of service associated with a data library (see Ruus (1990)). Basically, the DRC was prepared to undertake most aspects associated with collection care and user services. However, there was no commitment to archiving services. Over the first year this decision was reconsidered. It has been found that there is a demand for these services and in the summer of 1997 there was a grant to hire a graduate student to begin archiving historical census records from 19th century Canada. It is believed that the DRC will continue to evolve in this direction.

Another interesting development that occurred was related to the user community. Initially it was believed that the heaviest users would be faculty, researchers and graduate students. Applications related to teaching, particularly in applied undergraduate courses such as statistics and upper year research courses, have made use of the DRC. This tied in nicely with the services being provided in the Government Documents section of the library, where the traditional paper-based statistical sources are housed, and where many of the supporting documents that DRC users would require could be found. It was suspected that there would still be a large number of 'one-off' type questions that would be more efficiently answered using the traditional hard-copy sources. For example, questions like the population for a given CD, CSD, or EA might best be addressed using traditional sources. In these cases the user was often looking for one number, or even just a few numbers. However, the ease and flexibility of the web retrieval system has opened up the DRC for these types of questions. One of the challenges has been deciding when users should refer to the DRC in order to get the quickest and most efficient response and when they should refer to traditional sources.

The DRC is centered around the WWW[6]. Expectations are that users will become self-sufficient in finding and extracting information. This is similar to the objectives of other Data Centres (see Kroeker (1997)). The feeling was that if we established a DRC, demand was so high that we could easily spend all of our human resources answering requests without expanding the information available. Initially the DRC did not have any public hours for walk-in consultation, and once it did, these hours were limited. This allowed staff to get a jump on providing self-help information for users, get comfortable with the DRC's services themselves, and have some time for some early fine-tuning of the service before declaring the service fully

functional.

The DRC web site has gone through several iterations since it appeared in the first week of operations having, at that point, been created with minimal content behind it. It has developed "on-the-fly" in response to demand, feedback, and experiences in a live web environment, with the guiding principle being that web sites should be dynamic, and regularly edited to fit the changing demands or the latest ideas of users or staff. Recently there was a major overhaul to simplify the site. Most of the changes were a direct result of user input, and studying other sites on the net. The main page links the user to 5 major areas:

The first area takes users to the on-line data holdings, which includes access to the web retrieval system (discussed later), CD-Rom products available over the net, access to GIS data from the Census and any on-line services that are subscribed to, such as CANSIM.

The second area takes the user to information and links to all the CD-Rom holdings. If they have not been made available over the net then there are instructions on how to access the information.

The third area links the user to information on data from consortia agreements which include ICPSR and DLI.

There are direct links to sites where the user can search holdings. Only a small fraction of this data is stored locally and is usually obtained on a request basis.

The fourth area deals with external data sources. As time permits staff gather links to sites that provide free access to data (some commercial sites are also linked) and logically order these links to help users find what they need. This is also a valuable resource for reference staff within the DRC. More often than not this list is expanded as sites are discovered on routine reference questions. The page is divided into categories such as Economics, Agriculture, Science and others. The final area links the user to other data centres and data related sites that may be useful.

The office of the DRC is located adjacent to the Government Documents section of the Library.

Staff in the Centre have workstations and desk space within this office, aside from the Librarian who has his own office in this area. There is also one additional workstation that users have access to, if needed. The idea is that this is a point of reference within the library where users can contact staff in person or by phone. There is also a work area to hold meetings and discuss data related problems. There are also computer pools within the library to which users can be directed to work on their problems, and recently three more workstations were added just outside the DRC that can be used by clients (they double as library catalogue machines).

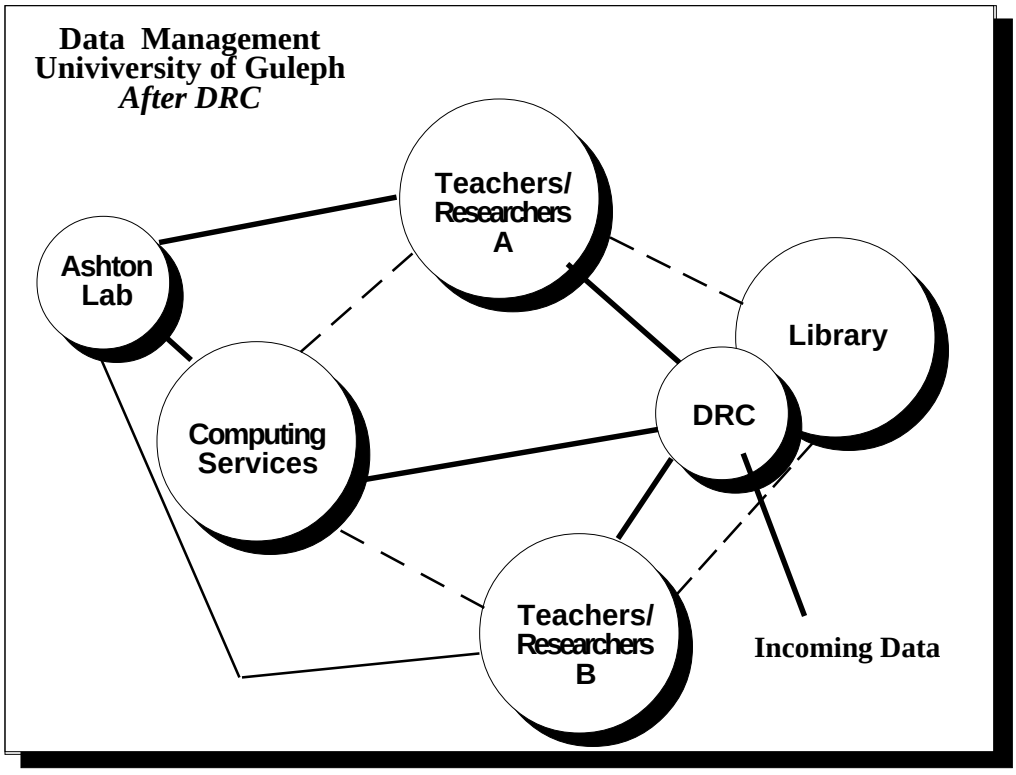

figure 3

The web server for the DRC is split into two components. The main portion runs on an NT server located in the DRC office. This server stores data that come with a pre-written PC/windows interface[7]. This server also stores CD-Rom's that can't be loaded onto the WWW. There is a CD-Rom tower attached to the server.

The bulk of the data holdings, in terms of both size and number of data sets resides on a central Unix server that is used for running statistical applications. This server was already in existence,

and the DRC web retrieval system runs along with other services. This makes it very convenient for more experienced users to by-pass the web retrieval system and work directly on the data with centrally maintained software. This will be discussed more later.

Pilot project

Under the pilot project, a format similar to the one in figure 3 was undertaken. Essentially, the DRC would be a service offered through the library, and as such, was physically located in the main library. There was direct communication with all researchers, CCS and the library. All data would be channeled through the DRC, so that everyone was aware of what was available and where to find it.. The Ashton Lab already existed, providing advanced consultation with researchers dealing with data collected in the field. Clients also have access to SAS/ SPPS help through Central Computing Services. The DRC does not provide these services.

Staff for the DRC were assigned from both the library and CCS. For a more detailed break-down of tasks, refer to Appendix A. Currently, the DRC has a full-time Systems Analyst assigned by CCS as Project Leader. This person is essentially responsible for coordinating the project and participates in all aspects of the DRC, including interaction with researchers using larger data sets available through the DRC. CCS has also supplied a 0.5 fte Systems Analyst, whose main task is managing and writing the web retrieval system. The library has assigned a 0.5 fte Librarian who has experience in Government Documents and working with CD-ROMs. This person coordinates the DRC within the library and handles all CD-ROM issues. The library has also assigned a 0.5 fte library associate with experience in the Government Documents section. This person functions as a reference person for clients coming into the DRC, and helps to bridge the gap between the traditional collection and the DRC collection. These assignments are best case situations. There is rarely 2.5 fte's available in any given week. On occasion funds are secured for students to work on specific applications such as 1991 Census GIS files, Historical Census, and HIFE files .

Web Retrieval System

The System

A large portion of the efforts in the DRC are centered around the development of a web retrieval system . A perl script has been developed to provide a web-based interface with SAS. This allows an enormous variety of data to be easily mounted, distributed and analyzed on-line. In the 14 months since the first iteration of the script, over 200 surveys have been mounted and made available.

There are many objectives in running a WWW interface to the data. The interface allows simple point and click access to a variety of data sets. Users are able to select a subset of variables, draw a sample based on conditions of certain pre-defined variables[8], output to approximately 30 different formats[9] and perform simple statistics on their subsets. At this point this includes frequencies, crosstabs, means and simple regressions. However, most importantly the interface is consistent across all the data sets. Several data suppliers are developing very good interfaces to their own data. One of the problems with this arrangement is that there is always a cost, no matter how intuitive, to learning a new interface. Users seem to appreciate the consistency and the ability to easily move from one data set to another that is provided with our single web interface.

In addition to the above features the retrieval system also gives users access to electronic codebooks, record layouts, users guides, SAS contents files[10], and sample SAS and SPSS programs. These programs can be transferred to the user's own PC or central UNIX account. These programs can be used as a template on the user's own UNIX account to run and read the data as defined by them. The system also points the user to the raw data files, the SAS datasets, the associated format and index files.

All of the data is stored in compressed SAS datasets that are fully readable by anyone with a central UNIX account. Variables in these data sets normally have labels and many have associated value statements. The degree of completeness depends on what is available from the supplier of the data. There is a large variety in DLI and ICPSR data. Staff are currently in the process of indexing the larger data files to significantly improve retrieval times. As with the subset variables it is not efficient to index by every possible term , so an attempt is made to try and choose the 2 to 6 variables that satisfy 90% of the requests. Currently the raw data file is also stored with the SAS data set for users who prefer to write their own programs. Disk space is relatively cheap, but there may be constraints in the future.

The Users

Essentially the script serves two types of clients. The first are those individuals who simply want a table or even a single number and are not willing to wait or perform a very complex procedure to get output. In many instances it may be faster to look up the table or number in a hard copy source. It must be kept in mind, however, that the web-retrieval system has a seemingly infinite number of custom tables available, whereas in hard copy the number, and nature of tables made available is decided upon by the publisher.

The second type of user is the researcher, who is looking for data on which to perform some analysis[11]. Experiences at Guelph suggest that many researchers have a problem when dealing with empirical work. There is an initial cost to getting a feel for the data, figuring out the record lay-out, writing a program to read in relevant data, and then performing some simple summary statistics on the data.

Once this is done, and the data is sufficiently massaged into a format they can use, the more detailed analysis begins. The DRC concentrates on helping with the first part of this process. The web interface makes it extremely easy for anyone to go in and 'play around' with the data before they decide what they want. Essentially there are now decreased costs, which allows more time for the more complicated analysis, as well as expanding what the user may attempt. Once the detailed analysis begins, staff forward problems to the Data Analysis Support group or the Ashton Statistical Laboratory.

The Administrator

From an administrative point of view the system is extremely simple and flexible. One script handles all the different formats of data. When a data set is added to the system a series of form files are created containing information on available variables, labels, variables to subset by, and their possible values. These files can be easily generated from the SAS program and contents file using a variety of simple editor commands. Once created, information such as directory location, data name, weights, whether to use scroll, input, or pull downs, and how to place boxes on the screen is quickly added. There is no modification of the perl script necessary to deal with data sets. Data such as the 1992 FAMEX survey have been added in as little as 1.5 hours. This includes downloading the data, creating the SAS dataset, and mounting it on the web.

In terms of security, access to the system is protected by IP address at the directory level. The use of Apache software to drive the web server allows us to place .htaccess files in any directory to limit or give access to users. The perl script is also capable of checking IP addresses based on the data being downloaded. These options allow us to open and close access to different data sets as the need arises. This will be particularly useful when we move to a shared system among Universities with differing licencing arrangements.

An area that is still being developed is the search capabilities. This was intentionally avoided during the first year of development as the priority was to give access to as much data as possible. Recently, the ability to do searches by keyword and strings was enabled on the contents files. The results of the search are presented in a tabular format where the user is linked to the contents files and the associated 'readme' file for that data set. Shortly there will be a link directly to the web retrieval forms for this data set. The files that are searched will also be expanded.

Integration with Library

One of the major objectives of the DRC was to integrate data identification and retrieval services with other services already in the library. The web-based nature of this service means that every workstation in the library (and on campus) is a potential contact point with the DRC, and library staff will be faced with data that is integrated with the other, traditional, library services. The fact that many reference staff in the library have had little or no experience with data means that training will be a time-consuming process, and developing a reasonable level of comfort with this type of information will have to occur gradually. This is an entirely new resource for the reference staff member to consider, and education will involve not only instruction in statistics, but more basically the understanding of the appropriate uses of data. The obvious place to start training was with staff in the Government Documents section, where staff had experience with the paper-based version of the data. There have been a few general sessions for library staff on what happens in the DRC and what data is available, with more detailed sessions planned for the future. A great deal of emphasis has been placed on the notion that we are trying to make users self-sufficient, and to minimize the need for consultation for relatively straightforward queries. Training sessions tend to consist of giving general outlines, asking staff to try to use the DRC web pages, and then to come back to us with questions. With lots of hand-holding this seems to be working, although, as is often the case, the technology itself rather than the nature of the data causes problems.

The feed-back from these staff members is extremely useful in helping set a direction. In other words, we get the users to tell us what works and what doesn't work. Training will slowly move into more detailed and specific sessions as services are better defined. Up until this point the DRC has been evolving rapidly and changes are frequent. The general view is that there will always be a need for specialized assistance that probably will not reside in the general reference staff of a library.

The First Year - Success and Areas for Improvement

Overall the DRC has been extremely successful and in most areas we have progressed well beyond expectations. There was a lot of uncertainty with respect to how we were going to do things, but everything seemed to fall into place very well. This can be attributed to a few things. The first was the level of technology in terms of software. Perl, SAS, DBMS Copy and Apache delivered what was needed.

The second related to the centralized computing environment that existed at Guelph. Although there were some rough spots associated with system security and access to configuration information by DRC staff, the resources were again well suited for what was needed. Delivering the data from a centralized UNIX system, accessible by everyone on campus, is extremely efficient.

The final point was related to the staff involved. As mentioned earlier, it is important to have individuals from the computing fields, the library, and the user community, working together. The DRC was lucky to get a group of

people who not only technically complimented each other, but also worked very well together.

One of the biggest surprises has been how easy it is for administrators to add data to the retrieval system. The way the perl script has been written allows for data in a wide variety of formats to be easily mounted for retrieval. At best, expectations were that a few dozen data sets could be mounted during the first year and it would be difficult to determine which ones. Currently there are well over 200 data sets on the system and the effort necessary to mount these diminishes as staff comfort levels increase . It takes between 1 hour and 2 days (large multiple-file data sets) to prepare data. As we begin moving back in time and mounting 'old' data, that lacks SAS/SPSS code and doesn't have electronic codebooks, the process gets more difficult. It is so easy, we are starting to train library staff with limited SAS and UNIX skills to mount data.

Another surprise has been the speed of extraction. The current system runs on a fairly slow UNIX server. However, by indexing the data the response time can be improved significantly. The functionality of the retrieval system, in terms of output formats, and summary statistics, is also beyond expectations.

Some of the areas we still need to improve are related to publicity. It is still difficult to get people to understand what is being done in the DRC. As soon as we get 1 on 1 contact, or demonstrate the system to a class, it becomes easy, and the users become largely self-sufficient. A goal is to get out to classes more often, and to make the DRC a standard tool for the completion of assignments. We are also continuing to publish a newsletter each semester outlining developments and what people are doing.

The service is also not without it's detractors. Some researchers who are experienced with working on large data sets, and using SAS, SPSS do not always see the benefits. It is very difficult getting these users to understand that the structure of the whole system can help even those who want to do their own extractions[12] . In some cases we spend more time with these experienced users than with 'new' users who readily accept the system.

What is Next?

As mentioned above, we are progressing much faster than we initially expected, but there are several things that still need to be worked on. Some of these were mentioned in section 7, related to publicity and staff training. Other areas include increasing the functionality of the web retrieval system, expanding on-line statistical options, particularly related to graphing, and possibly interfacing better with GIS systems. Work also needs to be done with on-line keyword searching for information. The current system works well if you know what data you are interested in. One area in which much consultation is still needed, however, is in the identification of data to suit a query. It is hoped that an efficient search capability for the system may also make the selection of data possible for the inexperienced user. We have started in this direction and hope to make progress over the summer.

Possibly the biggest project to date will be the sharing of this resource with other institutions. The web-based system is ideal for taking advantage of the potential efficiencies of a joint service. Discussions are well under way to develop this service into a seamless, shared resource between the University of Guelph, University of Waterloo and Wilfrid Laurier University. The end result will be a better service for all parties involved. It his hoped that eventually other such centres will appear in Canada and the workload and overhead can be shared between even more institutions.

A**PPENDIX A - brief job descriptions**

Project Leader - 1 FTE - Systems Analyst, CCS

Tasks:

Coordination of project; liaison between CCS, Library and user community; participate in management group; report to SAC; periodically report to College IT Committees; liaison with outside groups such as CAPDU, DLI, ICPSR, Statistics Canada; coordinate joint ventures with WLU and Waterloo; control inflow of data from outside sources - ie download from DLI and ICPSR; participate in data purchase agreements and consortiums (work with other data centers on national issues); provide user support and consulting for staff, researchers and students; participate in development of front end applications; participate in production of newsletter and annual report; assist other DRC staff as needed.

WWW resource person - .5 FTE - Systems Analyst, CCS

Tasks:

Incorporate, develop and maintain web retrieval interfaces for electronic data resources; implement a process for controlling access to data; implement a process for measuring usage; limited user support and consulting services for end users; manage NT server and various data products

Librarian - .5 FTE -Library

Tasks:

Overall coordination of "library specific" side of project; user support and consultation; addition of data to web site, coordinate CD-ROM products and acquisition; establish and develop communication between DRC and the rest of the library; participate in planning of layout and design of service point; develop and participate in training of library staff (classes and production materials); work with library staff on publicity and information; work with Acquisitions staff to bring data acquisition in line with acquisition of other library materials; work with Cataloguing staff to develop workable method of cataloguing electronic data sets; keep up to date with development of data resources available on the Internet.

Library Associate - .5 FTE - Carol Perry - Library

Tasks:

Link between 'hard-copy' reference and DRC collection; user support and consultation; adding data to web site; publicity (newsletter); backup strategies; web design and graphics; WWW searching and inventory (collect resources, data sites, and useful sources of related information); participate in training of library staff; maintain hard copy collection of codebooks; keep statistics of patron traffic and data use.

References

Horne, D., and McCaskell, P., and Wandschneider, B. (1996), University of Guelph Data Library/Centre - Proposal, mimeograph, University of Guelph, September 1996. (http:drc.uoguelph.ca/

Jacobs, J. (1991), "Providing Data Services for Machine-Readible Information in an Academic Library: Some Levels of Service, Public-Access Computer Systems Review, vol. 1, issue 2, 144-160.

Kroeker, B. (1997), Data Services Assists Teaching and Research: Delivery of Data Services via the World Wide *W*eb, IASSIST Quarterly, vol. 21, number 1, Spring 1997.

Lubanski, A. (1996), " Social Science Data Services During the Last Five Years of the Millenium", IASSIST Quarterly, vol. 20, number 4, November 1996.

Ruus, L.,G.,M. (1990), "Planning a Data Service Facilty", mimeograph,University of Toronto.

[1] In this paper we define a DRC as a point of service where users are able to get access to, and assistance with the use of, electronic information such as the census, general social survey, survey of consumer finances and so on. We do not include resources such as electronic journals and books. Experience suggests that users are frequently confused about this distinction.

[2] It is felt that the benefits are well defined and understood, and as such will not be discussed in detail. The emphasis will be on the process rather than the justification. However, it is essential to have a thorough understanding of them and for more detail see Horne et al. (1996). Lubanski (1996) gives some justification for increased funding, highlighting the increases in empirical research.

[3] The workshop was put on by Diane Geraci (Suny Binghampton), Chuck Humphrey (University of Alberta) and Jim Jacobs (UC San Diego).

[4] During a presentation of the web retrieval system at another University there was a very positive response in this direction. The faculty member was elated because they could now sit down at their workstation and retrieve subsets of the data in a matter of minutes. Up to this point they would spend valuable resources on graduate students and many times they ended up with nothing at the end of the project.

[5] The basic ideas for this were obtained from sample perl scripts and SAS programs written at Kansas. They were available over the WWW. This was a very crude implementation of a interface between SAS and the WWW.

[6] See http://drc.uoguelph.ca

[7] Many of Statistic Canada's data products now come in a format that uses Ivision's Beyond 20/20 browser. We find this a very useful interface for many queries, but stress it is only a compliment to our web retrieval system. It is not an efficient format to deal with may research type questions that are posed in the academic environment.

[8] Staff decided early in the process to keep the retrieval forms as simple as possible. As such, each data set has a small pre-determined set of variables to subset by. In the case of time series, this would include months and years, whereas in the case of most cross-sections something like age, region, education and gender are chosen. The objective is to try and satisfy the most possible requests with the smallest, most commonly used subset variables. If a user determines they would like to subset by some other variable, such as income, this can easily be added. The way the script is written it takes about 2 minutes to add another subset variable. The necessary 'form' file is created from the 'values' section of the SAS program.

[9] Examples include: SAS, SPSS, SPPS portable, Gauss,

STATA, Lotus, Excel, Quattro, Dbase, ASCII and many others.

[10] These files are used to easily access information on variable names, labels and formats, as well as sample size.

[11] This researcher could be very experienced and not need much assistance or they could be a student in an applied course who has never worked with this type of data before. In the later case the web retrieval system allows the user to easily (and painlessly) obtain the data and move on to more important tasks, central to the course.

[12] For example; codebooks, user guides, record layouts, SAS and SPSS code ready to read the data and even the data already in SAS data sets with formats and labels are all available through the web site.

[*] Paper presented at IASSIST 1998 Yale University May 1998. Bo Wandschneider Systems Analyst Data Resource Centre University of Guelph, Doug Horne Librarian Data Resource Centre University of Guelph.