

**Proposal for a Joint  
Data Resource Facility**

**Between**

**University of Guelph,  
University of Waterloo  
Wilfrid Laurier University**

**DRAFT**

**(for discussion purposes only)**

**Revised:**

**January 6th, 1998**

**Prepared by:**

**Representatives of Tri-University Data Groups**

## **CONTENTS**

<b>Introduction</b> .....	<b>-3-</b>
<b>Current Structure</b> .....	<b>-4-</b>
<b>Guelph</b> .....	<b>-4-</b>
<b>Waterloo</b> .....	<b>-4-</b>
<b>WLU</b> .....	<b>-5-</b>
<b>Proposed Structure</b> .....	<b>-5-</b>
<b>Guelph</b> .....	<b>-6-</b>
<b>Waterloo</b> .....	<b>-6-</b>
<b>WLU</b> .....	<b>-6-</b>
<b>Centrally Maintained</b> .....	<b>-6-</b>
<b>Summary:</b> .....	<b>-7-</b>
<b>Appendix A - Major Functions of a Data Library</b> .....	<b>-8-</b>
<b>Appendix B - Current Commitments by Institution</b> .....	<b>-9-</b>
<b>Guelph:</b> .....	<b>-9-</b>
<b>Waterloo</b> .....	<b>-10-</b>
<b>WLU</b> .....	<b>-11-</b>
<b>Appendix C: Joint Facility Tasks</b> .....	<b>-12-</b>
<b>Appendix D: Hardware Requirements and Estimated Costs</b> .....	<b>-15-</b>

## **Introduction**

This document summarizes a proposal for establishing a joint facility for the dissemination of electronic data for research and teaching between the University of Guelph, the University of Waterloo and Wilfrid Laurier University. Based on extensive discussions with interested parties at all three institutions it was agreed that the demand for the services associated with 'data resources' was high at all institutions and is expected to continue growing quickly<sup>1</sup>. . With the huge increase in 'electronic' information from such sources as the Data Liberation Initiative (DLI) and the Inter-University Consortium for Political and Social Research (ICPSR), institutions are finding it challenging to provide access.

Presentations and discussions have been made to the larger data community in Canada and there is strong interest in web retrieval models such as the model being developed at Guelph. We are beginning to see the emergence of other collaborative ventures among groups of Canadian Universities. A good example is the CANSIM data base at the University of Toronto. This has been extremely successful and students and researchers in Canada have benefitted greatly. The long-run objective is to have our service complement the other ventures in Canada, rather than duplicate what they are doing.

All three institutions face similar data needs. With the current fiscal situation a collaborative approach to providing access to this information is essential. The type of information involved and the introduction of new tools to deliver this information makes this service ideally suited for a collaborative venture. The WEB retrieval development, already well established, would constitute a good basis from which to build. A joint venture would also allow delivery of a far more comprehensive and useful service opening access to a larger community of researchers and students. It was noted that each institution had, or could develop, unique skill sets to contribute to such a project. It was also noted that the need for this service is immediate at all the institutions and as such we should proceed as quickly as possible.

---

<sup>1</sup> This is based on the experiences at the University of Guelph over the last year and to a lesser extent the experiences at Waterloo and Wilfrid Laurier. It is noted that the demands at each institution are different, but follows a trend at other Universities throughout Canada and the rest of the world.

## Current Structure

Each University is currently committing different amounts of resources to providing access to data. This is reflected in the level of service provided at each institution. Currently Guelph is committing the largest amount of resources and providing a more comprehensive service, relative to the other institutions<sup>2</sup>. The objective of this proposal is to initially bring the level of service at all three universities up to this level, without having to commit the same amount of resources at each institution. After this initial push, the objective will be to capture the efficiencies of this joint venture to increase the level of service beyond what is feasible at one university. Sharing hardware capacity and workload are the main areas where gains can be made.

The following sections briefly outline the current commitments at the three institutions. For more detail refer to Appendix B.

## Guelph

Guelph established a pilot project in December of 1996. This is a joint venture between Computing and Communications Services and the Library. In April of 1997 this became a full service facility. There has been a large commitment of resources, including approximately 2.5 fte's (along with occasional work study and grant funded personnel). There has also been a commitment of approximately \$40,000 in local hardware and software for staff workstations and an NT server used to disseminate CD-Rom products. At present there is about a \$28,000 yearly commitment to the purchase of data distributed through the DRC. This is likely to grow as the service becomes more established and we centralize the acquisition and distribution of data. The web based retrieval system is housed on a shared central Unix system with approximately 12 GB of space currently dedicated to the DRC. There is a variety of software used on this system.

## Waterloo

Waterloo has established an *Electronic Data Service* (EDS). No staff have been formally committed, although 6 staff from different areas have been addressing some of the data needs, in addition to their full time responsibilities. There is an office on the 2nd floor of the library with a workstation for staff and users of the service. Waterloo has centrally committed approximately \$21,000 per year to the purchase of various data products. This data is distributed on a central

---

<sup>2</sup> The University of Guelph has established a comprehensive 'Data Resource Centre' over the last year. For a basic outline of the types of services that can be offered, refer to Appendix A. For a more detailed background on the costs and benefits of such a centre, along with information on the proposal from the pilot stage to the present, please consult the web site at : <http://drc.uoguelph.ca/whatis.html>. Many documents are available in 'pdf' format as listed in the left hand margin of the web page.

ftp server. Individuals, departments and faculties have also acquired data independently of each other. Currently these data are virtually inaccessible. There is also a collection of data purchased in the rest of the University community. Currently this is not being distributed centrally.

## **WLU**

WLU is meeting user demand on an 'as needed basis'. Limited support is provided by a librarian and a library computer support person in addition to their other responsibilities. There is also access to a central computer support person on an ad-hoc basis. There is currently no dedicated workstations or space allocated to this type of service. Centrally there is a \$3,000 yearly commitment to data and there is also an unknown collection of data being acquired outside of the library. Any centrally acquired data is stored on a central shared Unix system.

## **Proposed Structure**

The centralized portion of the service will be the development and maintenance of the web pages and the web retrieval system. The objective will be to have one web site available to users at all three universities linking to one copy of the data. Restrictions on access will be enforced based on the data each university subscribes to. The hope is that by working collaboratively we may be able to negotiate group purchases of some data.

In order to do this a centralized Unix system capable of handling all three universities will need to be purchased. Several options have been considered and the recommendation is to purchase an HP system to be housed and maintained at the University of Guelph. Staff from all the universities will contribute to the maintenance of the web site and the generation of data files.

It is recommended that each site maintain their own support/reference people to interact with their own user community. Each institution may develop areas of expertise and users would be encouraged to contact staff at one of the other institutions if the need arises. For example it is already apparent that Guelph may develop expertise in the areas of agricultural data, Waterloo in spatial data (GIS) and possibly WLU in the area of business/financial data. A management group would be established to coordinate the activities.

Based on the experience at the University of Guelph it is recommended that six groups be involved in this service. Computing Services (IST - Waterloo, CCS - Guelph, CCS - WLU) along with the Libraries at each institution. Commitments may vary among the institutions.

For a more detailed breakdown of tasks refer to appendix C. The following outlines the recommended commitments.

## **Guelph**

Continue as currently being done, aside from the movement to a central Unix system shared among the three Universities

## **Waterloo**

Waterloo would build on what currently exists. The biggest change would be a commitment of 2 fte's to perform tasks as outlined in Appendix C, such as, locating WWW resources, developing updating services such as the web retrieval system, preparing data sets, reference, maintaining local facility, and user education. Any local hardware/software needs are to be determined locally as under other TUG initiatives.

## **WLU**

Upgrades to staff commitments and hardware will be made but are yet to be determined at the local level.

## **Centrally Maintained**

The recommendation is that we purchase an HP D280/2 along with a separate disk array to increase performance of disk I/O. This system is designed to drive the web retrieval software.

In discussions it was emphasized that we should be prepared to meet increased demands. The above system has room to expand memory and hard disk capacity, but there is no room to expand processing power. Other configurations are available that allow for this expansion, but the value is not as good as this system. The best option may be to add another D280 if the need arises and split system load between the two machines. We also note that with the continual change in technology it may not be in our best interest to restrict ourselves to a certain system that can be expanded as demand grows. By this time the technology may be outdated.

Details of the system and prices are available in appendix D. In summary:

Initial outlay in year 1: \$115,502

We also believe that increased growth in year 2 and 3 would suggest that we should budget for purchasing more disk capacity, CPU power and possibly memory. These figures are crude estimates and we do not want to forecast beyond 3 years. The exact timing of these additional expenditures will depend very much on the growth of the service.

Year 2:	\$57,500
Year 3:	57,500

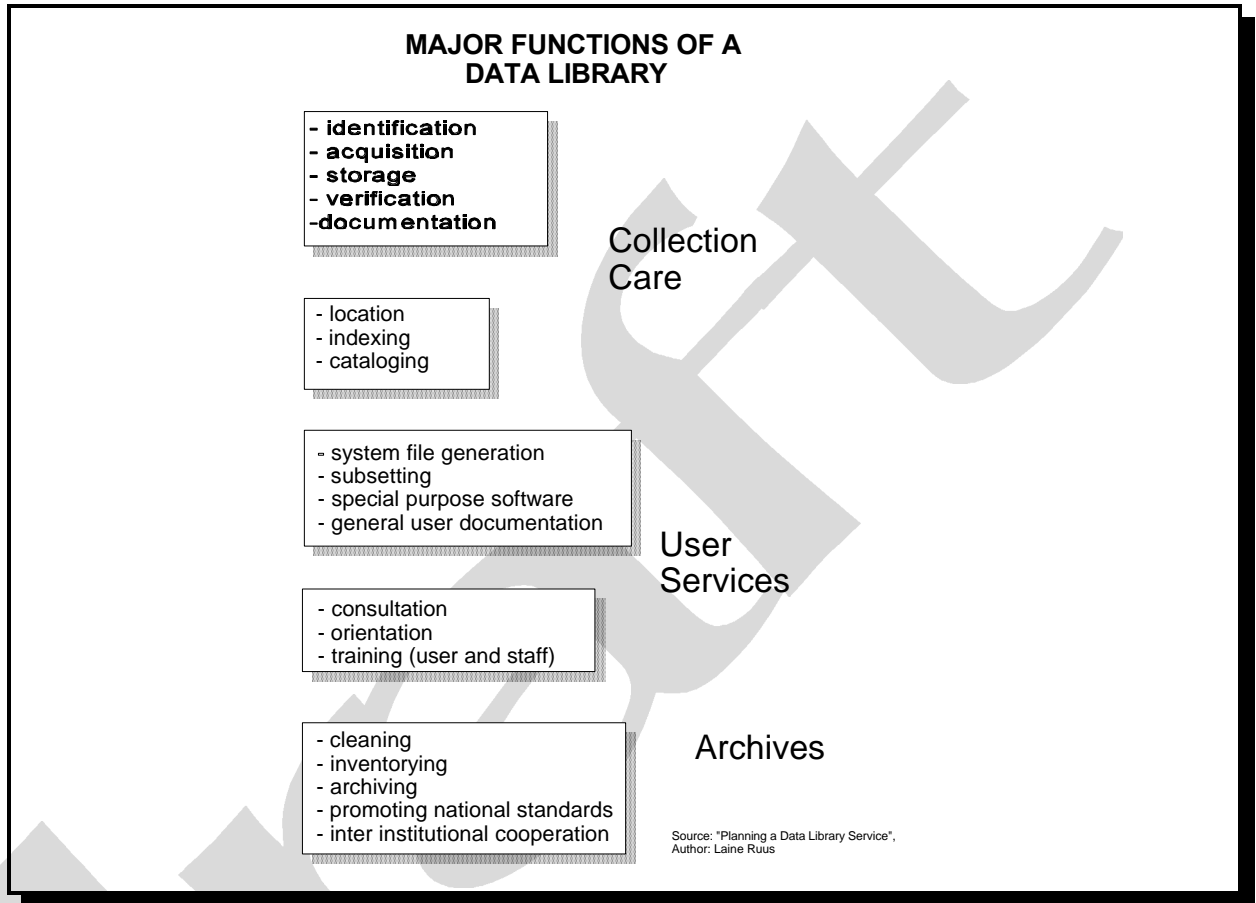
**Summary:**

The recommendation is that we provide a central service to store and give access to electronic information. This would be based around the web retrieval system already established. Individual sites would still maintain staff to provide reference and training and also to act as representatives to consortiums such as DLI and ICPSR. These staff would also help with web development and be capable of uploading and preparing data for the retrieval system.

The data retrieval system would continue to be developed at the University of Guelph, taking input from all members of the group. A new hardware system would be purchased capable of handling demand from all three institutions, with the option to expand as demand increases. There is also the option of providing access beyond the three institutions if there is demand. The new system would be housed at the University of Guelph, to best take advantage of the existing software and hardware support/licensing.

Ongoing costs would include human resources dedicated by each institution as well as software licensing and upgrades. These resources will be determined on a local level by each institution. The bulk of the centralized staff to maintain and develop the web retrieval system would be centered at the University of Guelph. This would also apply to the majority of the software requirements that can fit under current site license agreements at the University of Guelph.

Appendix A - Major Functions of a Data Library



---

## Appendix B - Current Commitments by Institution

### Guelph:

#### *Human Resources*

1.0 fte from library  
1.5 fte from CCS

#### *Location*

1 - office for 4 workstations, meeting space and shelving for codebooks (library)

*Workstations* ( approx. \$40,000 has been committed to hardware and software)

3 - workstations for staff (windows 95)  
1 - workstation for part-time/contract workers (windows 95)  
3 - workstations for users (on library floor - shared resource - DRC priority)  
1 - NT server with CD-ROM tower and HD capacity to disseminate CD-ROM products  
1 - HP printer for staff  
1 - access to distributed printing (cost recovery)

*Data (approx. yearly \$ costs for purchases through library)*

DLI - 12,000  
CANSIM - 2,000  
ICPSR - 7,000  
TSE/ Western - 4,500 (8,500 initial)  
OECD - 2,600

- there may be more commitments made over the next while as we gauge demand and there is also an unknown amount of data being acquired outside of the DRC.

#### *Central hardware/software*

HP 735 - Unix system  
- 12 GB disk space  
SAS, SPSS, DBMS, ORACLE (all site lic. used in other places)

---

**Waterloo**

*Human Resources*

0 -fte - No staff have been formally committed, although 6 staff, members of the UW-EDS Group, from different areas have been addressing some of the data needs, in addition to their full time responsibilities.

*Location*

1 - office on 2nd floor of the Dana Porter Library

*Workstations* ( approx. \$3,000 has been invested in hardware and software)

Dana Porter Library - EDS office

1 - workstation for staff/users (windows 3.1)

1 - HP colour printer for staff/users

University Map and Design Library

1 - workstation for accesing geophysical data (windows 3.11)

1- HP colour printer

*Data (approx. yearly \$ costs for purchases through library)*

DLI - 12,000

CANSIM - 2,000

ICPSR - 7,000

- there is an unknown amount of data being acquired outside of the EDS

*Central hardware/software*

Access to shared central computer and disk space to distribute data through ftp.

**WLU**

*Human Resources*

- part of a librarian and a support persons time, plus ad hoc support from CCS.

*Location*

no dedicated space

*Workstations*

no dedicated workstations

*Data (approx. yearly \$ costs for purchases through library)*

DLI - 3,000

- there is an unknown amount of data being acquired outside of the library

*Central hardware/software*

- shared space on a central unix system

## Appendix C: Joint Facility Tasks

Each institution will still maintain 'DRC' staff, in the sense that there would still be people who provide front-end consulting to the user community. Whether these staff are dedicated or not would be up to each institution, although there are clear advantages to having dedicated staff. It is foreseeable that reference staff in the library could be trained in the use of a DRC system and provide support in this fashion. As well, each institution would have one or more staff members who are responsible for handling inquiries and orders for data from consortiums such as ICPSR and DLI. Ideally these staff would be capable of loading data onto the retrieval system. The skills required are minimal and experience suggests that extensive training is not necessary. Although not difficult, some staff will be better suited for this task. There should also be staff who can promote the resources and provide instruction to users through regular seminars or classes.

The web site would be common among all the institutions. Maintenance and design could be performed jointly by staff at each of the institutions, or resources could be centrally dedicated to this function. The web retrieval system would be maintained centrally on one system, dedicated to this function. This would technically be the easiest option, allow for seamless extraction of different data sets and open up the possibility of expanding the service beyond the three 'core' Universities<sup>3</sup>.

We could divide the tasks into two different segments, technical and user support functions. The technical support functions are well suited to be performed centrally and the services delivered would be identical among the institutions. The user support functions remain the responsibility of each institution, although training and goals could be set centrally to develop consistency and assure that each institution's needs are met.

### Projects

In the next section we list some of the tasks that need to be done and how they may be implemented.

#### Technical Projects

1. Develop web retrieval system.

---

<sup>3</sup> It is foreseeable that other 'regional' data services would evolve. We may find that they specialize in data that we have not yet implemented in our model. In these situations we may want to 'trade' access to our data for access to their data. Nationally this would be an excellent model.

- continue to refine the system currently in place
- modify for use by different institutions
- expand statistical functionality
- add more data sets
- this is to be performed centrally on a system dedicated to the DRC.

2. Develop an inventory of in-house electronic data resources.

- develop an electronic list of resources.
- collect hard copy information on data resources and make as much as possible available electronically.
- collect all the data and store centrally, giving basic access to the user community.
- this portion of the project could be shared by staff at all institutions. There will clearly be overlap for the majority of our holdings but there will be items unique to each institution. Researchers and students should benefit from the combining of resources.
- one person per institution is assigned to collect this information and disseminate to a designated individual. The bulk of responsibilities of maintaining and updating this collection will move to this individual after the initial period when we are bringing together all the 'old' information.

3. Collect an inventory of data resources available over the Internet and make available on a master web site.

- this portion could also be maintained jointly and would benefit from being compiled by different individuals at different institutions. One searchable 'page' should be maintained that can be administered by an individual from each institution.

4. Develop an inventory of usage.

- This is needed to help identify areas of demand as well as help identify the types of data resources required.

5. Publicize services

- continue with the newsletter to let people know what we are doing and what other people are doing with the service. This is particularly important during the startup. The idea of a central newsletter is appealing as we can use it to promote research and teaching across all the institutions. Currently we include a section on what individual researchers are doing with the DRC to help expand joint projects and inform the community of possible options.

6. Approach other institutions, to develop long range agreements to share resources.

### **Locally Maintained**

1. Develop the CD-ROM databases

- Allow local access to CD\_ROM products like ESTAT, SABAL..... This may be possible to manage centrally.

### **User Support (local)**

1. Provide front-line consulting for the user community.

- helping users identify sources of information
- assisting in the use of the tools we develop
- provide information on how to get statistical support.
- see original DRC proposal for more detailed information

**Appendix D: Hardware Requirements and Estimated Costs**

The prices are in Canadian dollars adjusting for a 25% academic discount and approximately estimates for taxes.

HP D280/2

- 2 - 180 MHz CPU (PA-8000)
  - 384 MB ECC memory
  - 2 x 4 GB FWD SCSI-2
  - 100base-T LAN
  - 700/96 console
  - 12x CD-ROM SCSI-2
  - 8-9x faster than current machine
- \$ 60880

Disk System:

HP Model 20 High Availability Disk Array

- SP620 Controller
  - 16 MB write Cache
  - battery back-up
  - 5 x 8.8 GB disk drives (@2550 ea.)
  - expandable and reconfigurable disks
- \$ 54,622

Software Costs

- standard work station software
- SAS lic. - already covered by UofG
- DBMS Copy lic. - possibly covered by UofG
- perl 5.0 - N/C
- possibly ORACLE (database and web server) - already covered by UofG
- Apache web server - N/C
- Network connection - already covered by UofG
- backup strategy - already covered by UofG

**TOTAL COST**

**\$ 115,502**