The Development of a Canadian Union List of Machine Readable Data Files (CULDAT)

This article is an abridged version of the final Report, "Pilot Project for the Development of a Canadian Union List of Machine Readable Data Files (CULDAT)," prepared by Edward H. Hanis, Social Science Computing Laboratory, University of Western Ontario for the Machine Readable Archives, Public Archives of Canada.

A survey of Canadian social scientists, undertaken in 1982, indicated that a need existed for an inventory or union list of data files available for secondary analysis. In the mid-seventies, the Data Clearing House for the Social Sciences (DCHSS) had spent considerable time and effort in the development of an automated inventory. The loss of DCHSS, due to lack of funding, unfortunately also involved the physical loss of the magnetic tape which held the descriptions of these files. The results of the 1982 survey indicated strongly that the research community continued to feel that a union list of data files would be a valuable resource. In response to this need, the Machine Readable Archives Division (MRA) [of Public Archives Canada] established a contract with the Social Science Computing Laboratory of the University of Western Ontario to develop an online inventory describing computer files held by Canadian data archives and libraries. The overall purpose was to develop organizational, technical and informational foundations for maintaining and disseminating a computerized inventory. Specific objectives involved: the establishment of a standard for describing MRDF for entry into the data base; the design and implementation of the pilot data base containing a partial inventory; and the definition of the organizational roles and mechanism to effect routine and cost-effective flow of descriptive information from data archives and other organizations to the union list beyond the conclusion of the pilot.

The pilot project was carried out over a fourteen-month period. In January of 1985, a committee of data archivists and data librarians established a list of elements which were to be used to describe the holdings of the institutions. These elements were taken from those defined in the MARC format for data files. A data dictionary was developed to aid participants in the entry of descriptive information. The Social Science Computing Laboratory was involved in six major activities: the creation of the pilot data base; the set-up of online access with Basis on the lab's VAX11/785; the set-up of DATAPAC and standard dial-up communications; conducting an evaluation of the online system; a survey of potential contributors; and production of a hard copy reference document.

Contributors to the data base were from the university-based archives and libraries and included: Data Library, University of British Columbia; the Institute for Social Research, York University; Data Resources Library, University of Western Ontario; Institute for Social and Economic Research, University of Manitoba. The MRA also contributed descriptive entries. In all, 753 records were entered into CULDAT. Evaluation of the data base was extended to more participants than those listed above and included both frequent users of online systems as well as infrequent users. Although a number of suggestions have

been made as tohow to improve the online inventory, the general consensus was that the data base was very useful and should be continued.

It is not surprising that the most crucial component of the data base was the description of the data file. A number of difficulties were experienced with the lack of consistent terminology used and the detail of the description itself. The problems encountered are summarized in the following paragraphs. The resolution of these difficulties have formed the basis of the CULDAT work plan for 1986–87.

The choice of data elements to be included in CULDAT was based on the fields in the MARC format for data files. A limited number of elements was chosen as it was felt by the committee that the intention of the data base was to include only sufficient information to identify a unique data file, to aid researchers in selecting files of interest, and to locate archived copies of the file. The resulting CULDAT Data Element Dictionary contained the field names and a brief description. During the pilot project, it was noted that in some cases the data dictionary did not provide sufficient guidance to the archivist or librarian to allow him to adequately describe data files, and presumed a knowledge of the MARC format and Anglo-American Cataloguing Rules II. This created some difficulty in mapping out the information received for input into CULDAT. The consequences of a weak data element dictionary are inconsistent presentation of the information which can make the descriptions difficult for the end user to interpret. Weak data descriptions yield inefficient indexes, which, in turn, require that the user anticipate all possible variations of a term in order to find all relevant records in the data base. Specific problems were found in the following data elements.

I. Investigators: The differentiation between principal investigator and other investigators caused some difficulties for both cataloguers and the users. The determination of principal investigator for a data file is difficult, if not impossible, at times. The separation of these fields requires searching two fields rather than one for the user wishing to browse the index. The distinction between investigator (personal) and investigator (corporate) was considered essential. The lack of authority control in the corporate investigator field was a problem which could be overcome through the use of Canadiana to control the use and spelling of names.

2. Producer; Generator, Distributor: A tendency to repeat the same data in these fields was found. This may have been due to the inadequacy of the data dictionary. Abbreviations and acronyms were used. The adoption of an authority file for corporate names would apply to these fields as well.

3. File Size; Number of Cases: Some difficulty was experienced in the data provided in this field. Again, this was due to lack of guidance in the data dictionary.

4. Access Restrictions: As all institutions have their own access regulations, it was felt that this field should only be completed when the distributing organization has contributed the record.

5. Abstracts: Information contained in this field was found at times to repeat information found in other fields. The vocabulary used varied widely which made control of the field extremely difficult. The types of variables used in a data file is vital information for the prospective user. In order to provide improved access to this field, it would be

preferable to separate the abstract from the variable list. Variables could then be left unindexed. Such a change would significantly reduce the indexing overhead and improve the quality of the printed keyword index by using variable names instead of individual words. The online system could continue to index variables as individual words as well as expressions.

6. Geographic Coverage: The pattern adopted by the pilot was as follows: site, city, region, territory, province, state, country (qualifier) continent. The pattern worked well in most cases and ensured that the user interested in data about a particular province could retrieve information on a file which covered only a city in that province. The only records which do not conform to this pattern are physical data where orbital coordinates are submitted.

7. Chronological Coverage: The format of the dates recorded in this field was inconsistent, rendering the retrieval of data ineffective. The data dictionary should prescribe one acceptable format to which all dates would be converted. A standard format will provide the possibility of performing systematic retrieval on time periods by scanning the text, even though every unit of time within a range is not actually recorded in the field.

The difficulties which have been encountered will provide valuable information to allow us to improve the quality and guidance required for the data dictionary. The second version should improve the consistency of the descriptive entries. The contributions made by the data archives and libraries were extremely useful in building the pilot data base and allowing us to identify specific areas for improvement in the data dictionary.

User Evaluation and Potential Contributors

The original project design called for online testing and evaluation of the pilot CULDAT data base by project participants and constructing a list and contacting potential contributing organizations in order to learn about their holdings and interest in submitting entries into CULDAT in the future. Three important additions were made to enhance the project. The establishment of a DATAPAC Service reduced usage costs and significantly improved convenience to remote users. In addition, the survey of contributors was expanded to include questions on evaluation as they were potential users as well. The third activity was to include three local University of Western Ontario groups (students in the School of Library and Information Science, the University's reference librarians, and social science researchers who use the Lab's support services). These additions increased the use of CULDAT during the pilot phase.

The evaluation of the data base was very favourable and many respondents expected to benefit from the availability of CULDAT in the future. Considerable information from prospective contributors and users was acquired. This information and experience provide a sound foundation for the design and planning of the next stages in the development of CULDAT. The major activities planned for 1986/87 will include: 1) the revision and expansion of the CULDAT Data Dictionary in order to provide more guidance on the description of holdings for entry into CULDAT; 2) continued support to university based data archives and libraries to ensure their holdings are included in the inventory; and 3) the redesign of the formatted hardcopy version to make it available as a reference document at less cost.□