The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors:
The QUARTERLY is published four times per year. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Word processing files may be sent via email to jstratford@ucdavis.edu. Manuscripts should be sent to Editor: Juri Stratford, Government Information and Maps Department, Shields Library, University of California, 100 North West Quad, Davis, California 95616-5292. Phone: (530) 752-1624.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press Manual of Style or Kate L. Turabian's Manual for Writers. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". Journal of the American Society for Information Science 30(2):77-82, March 1979. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

Editors
Karsten Boye Rasmussen,
Eckersbergvej 56,
5230 Odense M,
Denmark.
Phone: +45 6612 9811
Email:kbr@sam.sdu.dk

Juri Stratford
Government Information and Maps Department,
Shields Library,
University of California,
100 North West Quad,
Davis, California 95616-5292
Phone: (530) 752-1624,
Email:jstratford@ucdavis.edu

Production
William Block,
25 Blegen Hall,
269 19th Ave S,
University of Minnesota,
Minneapolis, MN 55455.
Phone: 612-624-7091
Email:block@socsci.umn.edu

Walter Piovesan
Research Data Library,
Simon Fraser University,
Burnaby, B.C
Canada V5A 1S6.
Phone: (604) 291-5869.
Email:walter@sfu.ca

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137 © 1999 by IASSIST. All rights reserved.

CONTENTS
Volume 23 Number 3 Fall 1999

FEATURES

4 Supermarket: Where Do Social Scientists Shop? Kirsti Nilsen
12 Metadata and Metainformation - Old Concepts and New Challenges Dusan Soltes
15 Sherlock: A Web Magnifying Glass for Microdata Files Gaëtan Drolet
19 Changing Boundaries: Gazetteers, Information Retrieval and Data Browsing Cressida Chappell
22 The Royal Statistical Society Working Group on Archiving Data by Hilary Beedham
Supermarket: Where Do Social Scientists Shop?

This paper presents some findings of research which examined the statistics and data sources used by Canadian social scientists, the formats in which they obtained the data, and their preferences with respect to data formats. Five disciplines were the focus of the research: economics, education, geography, political science, and sociology, based on a literature review which is summarized below. The research was part of a larger study which examined the effects of government information policy on Canadian social scientists. That research focused on policy-initiated price and format changes at Statistics Canada. (Nilsen, 1996, 1997, 1998). In order to monitor the effects of the policy it was necessary to determine which statistics and data sources were used and any changes in that use over a period before and after policy implementation. Using both bibliometric and survey methods to gather data, the study identified statistics sources used in published articles over the period 1982 to 1993, and supplemented those findings with a survey of authors in the Fall of 1995.

The terms “statistics” and “data” have unique definitions; however, for the purposes of this research, the terms tended to be used interchangeably as they are in everyday speech. In the survey, respondents were asked about their use of “statistical data (i.e. numeric information)

Literature Review

Research on social scientists’ use of and demand for materials has confirmed that social scientists do use statistics and raw data. Because governments collect, analyze and publish the largest amounts of data, social scientists will use government-produced statistics, along with other statistics sources. Obviously not all social science disciplines use published statistics and data sets to the same extent. In order to determine which disciplines should be the focus of this study, published research on social scientists’ use and demand for materials was examined. It provided the data needed to identify those social science disciplines which use published statistics. Where statistics were not specifically identified, use of government publications served as an indicator of use of statistics because, as Hernon had shown, social scientists use government publications to obtain statistics more than for any other purpose (Hernon, 1979, p. 10). No research was found which distinguished between use of statistical publications of governments versus those of other publishers.

Use of Statistics by Social Scientists

The first major study of users of social science materials was undertaken by the Investigation into Information Requirements of the Social Sciences (INFROSS, 1971) in the United Kingdom. With 1,089 social science researchers responding to the INFROSS survey, it has been described as the largest, most ambitious and influential study in the area (Slater 1989, pp. 1). No research on a comparable scale has been done in North America.

INFROSS provided extensive data on the use of a variety of types and physical forms of information, along with data on information demand, by discipline, and with comparisons among disciplines. It specifically addressed the question of the use and perceived importance of statistics by researchers in each of the disciplines covered (anthropology, economics, education, geography, political science, psychology, and sociology).

The INFROSS study found that statistical, methodological and conceptual information was used by almost everyone, while historical and descriptive information was least used (Line, 1971, p. 416). Statistical material was used by 91% of respondents and over half used it frequently in their research. When asked to rate the importance of types of materials to themselves, INFROSS found that 58% of respondents rated statistical material as very important, 20% rated it as moderately important, and 12% rated it as not very important (INFROSS, 1971, vol.1, pp. 48, 50, 52).

With respect to disciplinary differences in use of statistical materials, INFROSS found that economists were the heaviest users of statistical data, followed closely by geographers. When asked to rate the importance of statistics, economists and geographers were much more likely than any other researchers to rate statistical material as “very important” and historians and anthropologists less likely to do so (INFROSS, 1971, vol.1, pp. 43, 51).

In its analyses of statistics use, INFROSS did not discriminate between data which were self-collected and
data gathered and published by someone other than the researchers themselves. However, the type of raw data used (e.g., interviews, experiments) was correlated with discipline of respondents (INFROSS, vol. 2, table 20). The report noted that psychologists were more likely to use empirically derived data from experiments conducted by themselves than were other social scientists (INFROSS, 1971, vol.1, p. 57).

Use of Government Information by Social Scientists

In reviewing the literature on citation studies, Hernon and Shepherd determined that the percentage of citations to government publications ranged from 2% to 36% (Hernon & Shepherd 1983, p. 227). Weech found that in various citation studies a median of 17.5% of total references were to government publications (Weech, 1978, p. 179).

The largest citation study of social scientists’ use of materials was Design of Information Systems in the Social Sciences (DISISS, 1979), a follow-up study to INFROSS. DISISS collected data from 140 social science serials, published mostly in 1970, for an examination of social science literature via citation analysis. Out of 47,342 citations only 2.7% were to official (government) publications (DISISS, 1979, p. 75). The variability in the findings on use of government publications among social scientists can be accounted for by disciplinary differences in the choice of disciplines included in citation studies. Low percentages in general relate to the fact that statistical sources are often not cited in footnotes or reference lists (Hernon & Shepherd, 1983).

The INFROSS survey found that 34% of social science researchers used government publications “often”, while 23% never used this form of material (INFROSS, 1971, vol.1, p. 53; Line, 1971, p. 417). When use of government publications was examined by discipline, the investigation found that 53% of researchers in economics stated that they sometimes or often used them, followed by those in sociology (41%), education (29%), geography (22%), and political science (20%). Fewer than 10% of researchers in anthropology, history and psychology used government publications (INFROSS, 1971, vol.2, table 59).

Hernon (1979) investigated the use of government publications by faculty members from economics, history, political science, and sociology departments in American colleges and universities. He found a statistically significant difference among the four disciplines in frequency of document use, with economists and political scientists as the heaviest users of government publications, which was consistent with the INFROSS findings (Hernon, 1979, pp. 9,45).

Some research has shown which disciplines seek statistical information within government publications. Hernon found that the “top priority of economists and sociologists [in using government publications] is to gather census and normative data,” and that historians used government publications for historical data more, while political sciences use them equally for statistics and current events information (Hernon, 1979, p. 51). Other studies by Hernon and Shepherd (1983) and Hernon and Purcell (1982) corroborated Hernon’s earlier findings.

Determining the Disciplines for This Research

On the basis of the INFROSS and DISISS research, which has been substantiated by other research, a typology of use of statistics and government publications was developed, as shown in Table 1.

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Typology of Use of Statistics and Government Publications:</strong></td>
</tr>
<tr>
<td><strong>By Discipline</strong></td>
</tr>
<tr>
<td>(1)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(2)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(3)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(4)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(5)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>(6)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Based on this typology, and the research which supports it, five disciplines were identified which use primarily published statistics and sometimes or often use government publications. These five disciplines were economics, education, geography, political science, and sociology. Thus, these five disciplines defined the domain of this research.

Methodology

Two methods were used to gather data on the use of statistics sources. Bibliometric analysis provided objective evidence of use of statistics, while a survey supplemented the findings with more subjective data. A systematic,
stratified and proportionate sample of 360 articles was selected from a population of 5,414 articles in 21 Canadian social science journals in the five disciplines noted above. The source journals were published in English or French in Canada, covered primarily Canadian topics, focussed widely in the discipline, were peer-reviewed, and published over the entire period 1982 to 1993. All journals which met these criteria were included. Articles to be included in the population to be sampled were those listed in the tables of contents under “Articles” or “Research Notes” or similar headings. In the final sample, the disciplines were represented in proportion to the amount of publishing in the 21 journals: economics 26.9% (97 articles), education 22.5% (81 articles), geography 7.2% (26 articles), political science 18.1% (65 articles), and sociology 25.3% (91 articles).

The 360 articles in the sample were examined and data were collected from the text, tables, and citations. The bibliometric examination revealed the statistics sources used by the authors of the articles. All uses of statistics sources, whether documented or not, were recorded, whether governmental, nongovernmental, Canadian or foreign. More detailed information about the use of Statistics Canada was gathered for the policy effects aspect of the larger study. Data analysis dealt with the complete sample and, in more detail, with a subset of 207 articles which were identified as using published statistics and written with a Canadian focus or setting.

A survey questionnaire was sent in English or French to 163 authors (all who could be located) of these 207 articles. Ninety-seven responded (59.5%). The questionnaire asked for background information, extent of use of statistics, statistics sources used, formats used and preferred, means of obtaining data, and opinions regarding prices and formats of data.

Findings
The 360 articles sampled for the bibliometric component of the research were categorized as to discipline, type, geographical focus or setting (if any), and language. The categorization was by discipline of the journal, (which was not necessarily the discipline of the author or of the subject covered). Most articles (78.4%) could be categorized by type as either empirical (200, 55.6%) or descriptive (82, 22.8%), both of which were likely to use statistics. The remaining 78 articles (21.6%) were either historical, opinion, methodological, or theoretical, articles less likely to use statistics. The geographical focus or setting was Canadian in 269 of the articles (74.7%), and the focus was not Canadian in 34 articles (9.5%). The remaining 57 articles (15.8%) could not be categorized geographically, usually because of their methodological or theoretical focus. Two-thirds of the articles (239, 66.4%) were in English, 121 articles (33.5%) were in French.

As expected, not all of the articles used statistics. As Figure 1 illustrates, 70 articles (19.4%) made no use of any statistics, most of these were categorized as theoretical or methodological. Thirty-nine articles (10.8%) used only self-collected data derived by the author from experiments or other research methods. A few articles (13, 3.5%) used only unidentifiable published statistics which could not be categorized as to source. The remaining 238 articles (66.1%) used identifiable published statistics. More than 70% of the articles in each discipline (excepting education) used identifiable published statistics. Some of these also used self-collected data.

![Fig 1. Use of Statistics in Total Sample](image)

A subset of 207 articles was identified which had a Canadian focus or setting and used published statistics and this subset provided the data which follows.

Statistics Sources Used
Information was gathered on the use of the following broad categories of statistics sources: Statistics Canada, other Canadian federal and provincial/municipal governments, foreign governments, intergovernmental, nongovernmental. More detailed information was gathered on the use of Statistics Canada in terms of formats used. It was found that social scientists used a wide variety of statistics sources and many used multiple sources. Figure 2 illustrates the percentage of articles which used the various
sources.

As can be seen, Statistics Canada was used by 41.1% of the articles, and other Canadian federal sources were used by an almost equal number (40.6%). American sources (governmental and nongovernmental combined) were used almost as much as was Statistics Canada, which was somewhat surprising in these 207 articles with a Canadian focus or setting. Nongovernmental sources were used by the highest percentage of articles (71%). Nongovernmental sources include trade and scholarly books and journals, associations, universities, business, think tanks, polling organizations, etc.

The survey respondents indicated higher use of all statistics sources (except nongovernmental) than was found in the bibliometric research. This probably results from the fact that the bibliometric analysis was looking at one-time use in a single article while the survey questioned life-time use. For example, 86.5% indicated that they had used Statistics Canada at some time, but only 41.1% used Statistics Canada in the articles. However, only 41.5% of survey respondents indicated that they used Statistics Canada often or almost always (i.e. more than 50% of the time in the years between 1985 and 1995), which is more consistent with the bibliometric finding. They also indicated less use of nongovernmental sources than was found in the bibliometric analysis. This difference might result because respondents might have been thinking of major nongovernmental suppliers such as polling organizations, rather than their use of sources from which they might obtain single facts such as a book or journal article.

Disciplinary differences
There were statistically significant disciplinary differences in use of most statistics sources in the articles as is shown in Table 2. Note that variation by discipline was statistically significant for all sources but provincial government and foreign government sources. Table 2 shows the importance of other Canadian federal government sources in articles from economics and political science journals. Those writing in education and political science journals were more likely to use other Canadian federal government sources more than they used Statistics Canada. Geography and sociology journal articles used Statistics Canada more often, and economics used the agency's
statistics at approximately the same rate as they use the statistics of other federal sources. Nongovernmental sources were important for all disciplines, particularly political science, which made heaviest use of both Canadian and American nongovernmental sources.

Use of Computer Readable Products
In the survey (conducted in the Fall of 1995), most respondents (81%) indicated that they had used computer readable formats at some time. There was statistically significant variation by discipline in these responses, with 100% of those who had published in economics and geography journals indicating prior use, while 77.8% of those in sociology, 73.7% of those in political science, and 56.3% of those in education indicating such use. However, when asked how they normally obtained statistics most still used paper (print) formats more than computer readable files. Of the 97 respondents, 74 (76.3%) indicated that they obtained data in paper format, while 59 (60.8%) used computer readable formats, or both formats, as seen in Table 3.

**TABLE 3**
Formats in Which Survey Respondents Normally Obtained Statistics (N=97)

<table>
<thead>
<tr>
<th>Formats</th>
<th>No</th>
<th>Percent*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Print (i.e., paper)</td>
<td>74</td>
<td>76.3</td>
</tr>
<tr>
<td>Microform (microfiche, microfilm)</td>
<td>13</td>
<td>13.4</td>
</tr>
<tr>
<td>Computer readable files**</td>
<td>59</td>
<td>60.8</td>
</tr>
<tr>
<td>Special tabulations***</td>
<td>36</td>
<td>37.1</td>
</tr>
<tr>
<td>Collect my own data</td>
<td>60</td>
<td>61.9</td>
</tr>
</tbody>
</table>

* Numbers exceed 100% because respondents could use more than one format
** Computer readable files were defined as "any 'off-the-shelf' computer files created for public dissemination or for limited dissemination within business or commerce, etc.
*** Special tabulations are created in response to specific request for a specific purpose, whether in print or computer readable form

Questioned as to how they normally acquired the data they used, responses are shown in Table 4:

Respondents were then asked to rank their first preference and their first three preferences of the various means of acquiring data, as shown in Table 5.

Note that where 47% in Table 4 used a library to acquire paper copies, for only 20% was that a top three preference. A larger percentage ranked purchasing computer readable files as a top three preference than had indicated normally acquiring data in this way. Also, more preferred to use a data library. Fewer preferred to collect their own data than actually did so.

The bibliometric analysis provided objective data on the actual use of paper and computer readable formats. The determination of use of products by format focussed on the 85 articles with a Canadian focus or setting which used Statistics Canada as a statistics source. Using various Statistics Canada catalogues and other sources where necessary, the researcher determined the formats of the Statistics Canada issues used in the articles, if the author had not provided this information. Seventy-one (84%) of these 85 articles used paper issues, while 29 (34%) used computer readable "issues", with some articles using both formats. The format of some issues could not be determined in 11 articles. The ratio of number of articles using paper issues to the number using computer readable issues was 2.5:1. Table 6 illustrates variation in the number of articles using issues by format.

---

**TABLE 4**
How Survey Respondents Normally Acquired Statistics (N=97)

<table>
<thead>
<tr>
<th>Means of Acquiring Statistics</th>
<th>No</th>
<th>Percent*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect My Own Data</td>
<td>56</td>
<td>57.7</td>
</tr>
<tr>
<td>Purchase Computer Readable Files</td>
<td>50</td>
<td>51.5</td>
</tr>
<tr>
<td>Use a Library for Paper or Microform</td>
<td>46</td>
<td>47.4</td>
</tr>
<tr>
<td>Use a Data Library</td>
<td>37</td>
<td>38.1</td>
</tr>
<tr>
<td>Purchase Paper Copies</td>
<td>35</td>
<td>36.1</td>
</tr>
<tr>
<td>Purchase Special Tabulations</td>
<td>30</td>
<td>30.9</td>
</tr>
<tr>
<td>Use the Internet</td>
<td>19</td>
<td>19.6</td>
</tr>
<tr>
<td>Use a Departmental Collection for Computer Readable Data</td>
<td>15</td>
<td>15.5</td>
</tr>
<tr>
<td>Use a Departmental Collection for Paper Copies</td>
<td>10</td>
<td>10.3</td>
</tr>
<tr>
<td>Purchase Microform Copies</td>
<td>9</td>
<td>9.3</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>2.1</td>
</tr>
</tbody>
</table>

* Totals exceed 100% because respondents could indicate all means of data acquisition which they used
TABLE 5
Ranked Preferences for Acquiring Statistics (N=79)

<table>
<thead>
<tr>
<th>Means of Acquiring Statistics</th>
<th>Percent Indicating First Choice</th>
<th>Percent Indicating One of Top Three Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purchase Computer Readable Files</td>
<td>22.8</td>
<td>68.4</td>
</tr>
<tr>
<td>Collect Own Data</td>
<td>22.8</td>
<td>68.4</td>
</tr>
<tr>
<td>Use a Data Library</td>
<td>20.3</td>
<td>51.9</td>
</tr>
<tr>
<td>Purchase Paper Copies</td>
<td>10.1</td>
<td>26.6</td>
</tr>
<tr>
<td>Use a Library for Paper Copies</td>
<td>10.1</td>
<td>20.3</td>
</tr>
<tr>
<td>Purchase Special Tabulations</td>
<td>5.1</td>
<td>19.0</td>
</tr>
<tr>
<td>Use the Internet</td>
<td>2.5</td>
<td>22.8</td>
</tr>
<tr>
<td>Use a Departmental Collection for Computer Readable Files</td>
<td>2.5</td>
<td>16.5</td>
</tr>
<tr>
<td>Use a Departmental Collection for Paper Copies</td>
<td>1.3</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Variation over the two time periods 1982-1987 and 1988-1993 in the number of articles using these formats was statistically significant for paper issues.

It should be noted that while the year-to-year variations are not statistically significant, the percentage of articles in the sample, which used computer readable formats as a statistics (or raw data) source, increased in the last two years studied. These formats were used in 20% of the articles in 1992 and 36% of the articles in 1993. Figure 3 illustrates year to year variation.

This might be an indication of a trend which might have been evident in a larger sample and which could be examined in further research.

TABLE 6
Number of Articles Using Statistics Canada Products: By Format
In Articles Which Used Statistics Canada (N=85)

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Paper (p =.045)</th>
<th>Computer Readable (p =.290)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982-1987</td>
<td>36 (92%)</td>
<td>11 (28%)</td>
</tr>
<tr>
<td>(n=39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988-1993</td>
<td>35 (76%)</td>
<td>11 (28%)</td>
</tr>
<tr>
<td>(n=46)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There were statistically significant disciplinary differences in use of Statistics Canada’s paper ((r = .009) and computer readable formats (r = .014) among the 207 articles written with a Canadian focus or setting. These are illustrated in Figure 4.

These data apply only to Statistics Canada formats, and as the discussion below indicates, other sources of computer readable information were used by these disciplines as well. Political science for example, showed little use of Statistics Canada overall, but was a heavy user of nongovernmental materials, and made some use of computer readable sources, such as the National Election Studies.

The decline in the number of articles which used paper formats might be attributed to the declining publication of
paper formats at Statistics Canada, rather than any absolute preference. However, the survey responses suggest that computer readable formats are preferred. It should also be noted that of the surveyed respondents who began to do research after 1980 (younger researchers?) 85% indicated that they had use Statistics Canada computer readable files at some time, while of those who began to do research before 1970, only 46.5% had used them. This suggests that in the future data users will rely on the computer readable files to an even greater extent.

Machine Readable Data Files Used
When computer readable files were used as major sources of data in articles, the titles of the MRDFs were recorded. Because authors tended to cite these materials incompletely, if at all, the following discussion should be interpreted cautiously.

Statistics Canada computer readable files were used by articles in education, economics, geography and sociology, with articles in economics journals using the greatest variety of files. Special tabulations were used by economics and geography authors for census data, and by economics authors for family expenditures, manpower, manufacturing, and agriculture data. An education article used the Labour Market Activity file; a Justice database was used by one sociology article. Public Use Sample Tapes were used by one article from economics and two from sociology. CANSIM was mentioned by only one author.

Other Canadian federal MRDFs were used in economics and political science articles. Three economics articles used Labour Canada files, and the International Trade Data Bank was used by a political science article. Quebec provincial health databases were used by two sociology articles. One US government database was used by an economics article (Dept. of Agriculture CRIS), and two sociology articles used US government data obtained from ICPSR. Canadian universities were an important source for data for sociology articles, and to a lesser extent for political science and economics articles. Here, the National Election Studies were used by an article in economics and one in political science. Both York University’s Quality of Life Survey and the University of Western Ontario’s Canadian Fertility Study were each cited by one economics article and one sociology article. Two francophone sociology journals cited SOREP data on Quebec population, while a third cited a database created at the École des hauts études commerciales. Other databases used include one use of the FAO trade tape, one use of the data from the Correlates of War Project (US university), and proprietary databases were cited by one economics article.

The above information suggests a rather limited use of MRDFs by Canadian social scientists. However, as noted above, authors do not cite these sources with any consistency. Additionally unless an item could be clearly identified as an electronic file, it was assumed to be a paper product if such a product was available in print. Thus, it is possible that some items that were recorded as paper products were in fact electronic files. Bibliometric analysis of the use of electronic files suffers from inconsistencies in citation practices. This was noted as early as 1982 (White), but the situation has not improved.

Discussion
The findings of this research are consistent with the findings of earlier studies cited in the literature review. Social scientists do indeed use a wide variety of sources to obtain statistics and raw data. There are statistically significant variations in the sources used among disciplines. If any agency such as Statistics Canada wishes to expand its market, analyses by discipline can assist in identifying target consumers, or areas where its products are not meeting the needs of researchers.

At the time period covered by the bibliometric research (1982-1993) and the survey (1995), social scientists still used paper products more than computer readable products to obtain statistics and data, but there was a statistically significant decline in the use of paper products.
Additionally, respondents to the survey were enthusiastic about computer readable formats. These findings suggested that computer readable formats would be used more heavily in the future. Indeed, in 1999, we see much more availability of information on computers. It is highly likely that future research will show a much stronger shift to electronic formats for data access. The findings of this research can provide baseline data for future comparisons.

Bibliography


Footnotes:

1 Based on INFROSS results and including only those in which 60% or more of respondents indicated that they use statistics.

2 Economists, social geographers, sociologists and some political scientists also use experimentally derived data, but to a lesser extent.

3 Based on INFROSS results and including those in which 20% or more of researchers claim they use government publications.

* Paper presented at the IASSIST Conference, May 19, 1999, Ryerson Polytechnic University, Toronto, Ontario. Kirsti Nilsen, Faculty of information and Media Studies, University of Western Ontario, London, Ontario N6A 5B7, Canada
Metadata and Metainformation - Old Concepts and New Challenges

Introduction
Since the very beginning of computerized data processing there has been a tendency for ever growing amounts of data to be processed and stored by computers. Probably not by accident, the modern computerized data processing was also referred to as mass-data processing. Especially in the environment of the so-called large-scale information systems, as e.g. statistical ones, there was an ever growing necessity to find the ways and means how to handle these rapidly expanding amounts of statistical data. The technological advancement and users needs finally led not only to introduction of very-large data bases and their distribution to the data base networks but also to the necessity to invent and introduce the particular tools for handling especially their content, i.e. data and information in the form of data and information on another - source and/or object data and information, which started to be referred to as metadata and metainformation.

Brief History of Such “Old” Concepts as Metadata, Metainformation and Metainformation System
Since their introduction in the 1970s, metadata, metainformation and metainformation systems have been the object of systematic research and development at international as well as national levels: first as part of a cooperative network program of European statistical offices and later on in 1981-84 as an object of the joint work of an inter-country joint group of experts of national statistical offices under the Statistical Computing Project of the United Nations Economic Commission for Europe. The main results of this international joint work, in the form of a pilot Users Guide to Metainformation Systems in Statistical Offices (1) and Selected Chapters for Designing METIS in Statistical Offices, defined the following basic concepts: metadata as a physical representation of metainformation, metainformation as a semantical contents of metadata while metadata is a description of (statistical) data and metainformation informs about (statistical) information. The data and information as objects of “description” by metadata and metainformation are, for terminological clarity, also referred in this paper as object data. Metadata and metainformation at the same time represent a content of metainformation system (METIS) in the form of its metadata base. Basic relations between these fundamental concepts of metainformation were defined for the first time by B. Sundgren of the Central Statistical Office of Sweden. His work became a basic theoretical framework of the particular international joint work.

If we also try to define the METIS itself then we come to conclusion that it is again a specific information system which as an object of its representation has an another i.e. object information system, e.g. in this particular case a statistical information system. Then on a more detailed level as the objects of METIS we can define various other individual components of the particular information system, not only its data and information but also:

- statistical surveys
- statistical forms and/or questionnaires
- statistical populations and/or files of statistical units
- statistical indicators
- data files/data bases
- publication tables and publications themselves
- programs, methods and procedures

All these objects are needed one way or another for any proper handling of the content of the particular object information system i.e. its data and information. If any user of these data or information wants to properly use, analyse, interpret them he/she always needs to know not only the quantitative values of the particular data or information but needs also to know many other accompanying information on this data or information i.e. metainformation. Without knowing which statistical survey it has produced these data and when and how, and how the objects of a survey i.e. statistical units and their populations have been defined, etc, it is almost impossible to utilize them properly or even to utilize them at all.

It is quite evident that not only for the convenience of users, but for any systematic handling of these specific accompanying data i.e. metadata they have to be organized as any other data into some organizational units - records, files, databases, etc. In general all these forms of metadata organization and storage are identified as a metadata base. The metadata base itself is organized as a system of mutually related individual metadata files or holdings.
which can have different forms. The basic forms of metadata files and/or holdings in a metadata base are e.g. as follows:

- catalogues
- dictionnaries
- directories
- registers
- glossaries
- thesauri.

Some other authors include into the content of a metadata base also some other rather specific „metadata” and their files and/or holdings as e.g:

- classifications
- nomenclatures
- code-lists.

The inclusion of these specific data - also sometimes referred to as service or auxiliary data - into the metadata base is based on the fact that their primary function is not to „describe” objects of the real world as (statistical) data do but to assist in more precise specifications of object data/information themselves. The content of the metadata base itself is created by formalized descriptions of particular objects of formalized descriptions such as e.g. (statistical) data i.e. indicators, but also surveys, units, files/populations etc. The most important and at the same time also the most voluminous part of any metadata base is its part regarding the core part of the data component of the particular object information system i.e. its operational data which in the case of statistical information systems are represented by (statistical) indicators. In this case the formalized descriptions contain descriptions of such attributes which help to properly interpret quantitative values of statistical information as e.g.:

- indicator name
- type of indicator
- code of time characteristics
- periodicity of collection
- measurement unit
- origin
- semantics and/or definition
- cross-sectional classification characteristics
- acronym and/or code (identifier) of indicator
- etc. depending upon the specifics of the particular object information system.

On the basis of its metadata base and its contents of formalized descriptions, METIS is then able to fulfill and/or at least assist or support such various functions regarding the object data and information as e.g.:

- interpretation
- navigation
- localization
- retrieval
- etc. regarding possibly also some other functions depending upon the objectives of the particular information system and/or its metainformation system.

On the basis of that METIS then can serve in several possible operational and functional modes e.g. as a SILS i.e. a simple information/interpretation and location system which assists users in proper interpretation and localization of object data/information without any further functions towards accessing them directly. This function is approximately on the level of catalogue systems in libraries which inform users about the books and their basic characteristics and location but without possibility of any direct retrieval. The higher function of METIS is its function DIRS - detailed information and retrieval system which in combination with the particular database and retrieval system enables also a direct access to the object data.

Challenges of Contemporary Information Highways vis-a-vis Metadata, Metainformation and Metis

If we compare the above concepts, elements and functions of metadata, metainformation, metadata base, METIS with the challenges of the contemporary world wide web and in general with information sources on contemporary „information highways” we may see their almost absolute inevitability, relevance and direct utilization especially in a case when users worldwide have access to practically unlimited sources of various data. Under such conditions it is sometimes almost impossible to secure any kind of proper information, identification, interpretation, comparability, consistancy, etc. between these data coming from very different methodical environments if there is not at the same time available also some kind of accompanying metadata and/or metainformation. If we take as an example statistical data on education and only on the level of basic and/or elementary schools we have almost unlimited possibilities for various interpretations of this relatively easy and commonly very well known concept in case that the data are coming from different data/surveys sources. They are as follows:

- The first problem is with the proper interpretation what it is an elementary and/or basic school or education. In different parts of the world it varies from 5 years e.g. in many developing countries up to 8, 9, or even 10 years of by the law obligatory „basic” education.

- The second problem is with the structure of this kind of education. In some countries it is a continuous education through the above 5 to 10 years. But in some
other cases it has been divided into two subsequent levels as e.g. 1 through 4 and then 5 through 8 or 9, 10, etc. But in some other cases it is organized in several parallel options as e.g. 1 through 4 and then 5 through 9 but there is also an alternative that enables for the best pupils after completing 1 through 4 grades to continue in an uninterrupted education from 5 through 12 grades what is already automatically combined basic education with the 8 year high school education completed by leaving examinations. But there still exist also a possibility to complete first 8 or 9 year basic or elementary school and then to proceed to a four year high school.

- The third problem is with the interpretation of the age. In some countries the basic education starts at the age of 5 but in many other countries it is at 6 and there are also countries where this age has been defined as 7 years.

Even this simple example demonstrates that not having the proper metadata to identify and interpret the data or information on elementary education could provide us with results covering a variety of populations including: 7 through 12, 5 through 17 or even 7 through 19 years of age. These differences or variations in expression of data are too great for some analyses and international comparisons. Therefore the existence of accompanying metadata and/or metainformation in the conditions of a direct access to these data by any user in the world in the conditions of an information highway becomes an absolutely inevitable and objective necessity.

Unfortunately, this is not always the case at present. The data and/or information are available, but their interpretation in many cases is left up to the users. Hence, analogically as in the case of large statistical information systems in the past also in the current www, it is possible to expect that one of their further main development trends will be towards some kind of normalization, standardization, unification and finally towards legally required necessity for providing some accompanying metadata and metainformation. Sooner or later we can expect that in addition to the existing information highway there will be necessity to have either its parallel metainformation (sub)highway and/or what seems to be more practical and less technically demanding, that all data and information will have to have a some kind of accompanying metadata sector which will contain basic (meta-)information describing all particular data/information. As we have already demonstrated also in this paper, the basic methodology for such accompanying metadata and metainformation has already been at disposal for a while, the new and still newer metadata handling systems are becoming ever more common. What is „just” missing for the time being and what is at the same time also the main challenge in this field for the future it is to find the ways and means how to standardize all this metadata, metainformation, METIS concepts and mainly tools into a world wide accepted standards. If it is too much and/or too demanding we will have to see.

Reference:

* Paper presented at: International Association for Social Science Information Service & Technology, Building Bridges, Breaking Barriers: the future of data in the global network, Toronto, May, 1999. Dusan Soltes, Faculty of Management, Comenius University, Bratislava, Slovakia

Dusan Soltes is a Senior Lecturer for MIS at the Faculty of Management of the Comenius University of Bratislava. Before joining the faculty he was working as a researcher in the area of computerized statistical information systems not only in his home country but also under the cooperative and development programmes of the United Nations – Economic Commission for Europe and the Conference of European Statisticians. Among others, in 1981-84 he has led a particular international joint group of experts for METIS under the ECE/UNDP Statistical Computing Project which has produced two basic publications viz. Users Guide to Metainformation Systems in Statistical Offices and Selected Chapters for Designing METIS in Statistical Offices.
Sherlock: A Web Magnifying Glass for Microdata Files

Context
In Canada, the Data Liberation Initiative (DLI) approved in 1996 by the Treasury Board of Canada has removed a significant obstacle to obtaining Canadian data in our universities.

With DLI, Canadian universities and Statistics Canada have solved the problem of obtaining Canadian data at an acceptable price. I remind you that in Canada, access to data is not free.

In the province of Quebec in particular, a number of universities obtained access to data but without really improving methods of consulting these data.

It is important that people realize that there are no full time data librarians in any of the Quebec universities. Most universities (except McGill, Montreal and Laval) are small in size, with equally small resources. We do not have a long-standing tradition of data use as in other Canadian universities. This is why, in order to render microdata files more usable, university libraries in Quebec have pooled their resources and expertise for the development of a common infrastructure to facilitate access and use of data.

What is SHERLOCK?
No, we aren’t talking about the world-famous detective, Sherlock Holmes. According to the conference theme, SHERLOCK is a kind of regional bridge to data. Using Sherlock, the data user becomes a detective of sorts. SHERLOCK is a bilingual tool, designed by the numerical data file subgroup1 of the CREPUQ (Conference of Rectors and Principals of Quebec Universities). At the Conference of Rectors and Principals of Quebec, we are a small but active group of four data librarians who have been working together since the beginning of the 90’s. We organise data workshops for our colleagues. We share our experiences and expertise. All of us are here at IASSIST.

This paper is being read on behalf of the four of us. We are the designers and the managers of SHERLOCK in our different institutions. The CREPUQ provided the place where Quebec university libraries were able to initiate and discuss this co-operative project. We have a 30-year tradition of co-operation between libraries. SHERLOCK was developed mainly for members of the Quebec academic community to enable them to access and utilise the survey microdata of the DLI (Data liberation Initiative) and the ICPSR (University Consortium for Political and Social Research) data.

Project origin and description
The first document submitted by the subgroup on numerical data files was Rapport de la consultation sur l’intérêt et la faisabilité d’une approche collective à la gestion des données numériques (Report On Consultations Concerning the Value and Feasibility of a Collective Approach to the Management of Numerical Data,) CREPUQ, November 1996. Our colleague Chuck Humphrey of the University of Alberta acted as a consultant for this stage.

After approving this report, the heads of the Quebec university libraries asked the subgroup to conduct a preliminary analysis on a top-priority basis. The timing seemed to be right.

When the subgroup took stock of data extractors in operation at the time, the LANDRU system, developed at the University of Calgary, stood out as one of the best although it did not meet all the requirements of the system to be implemented in Quebec. We wanted a bilingual interface; a decentralized and distributed approach to encourage the sharing of expertise and responsibilities in many institutions; management of all survey files available in the Quebec university network; compliance with licences; etc.

Therefore the four data librarians, who are members of the CREPUQ subgroup, with the help of an analyst from the library of Laval University, conducted a preliminary analysis and designed a pilot project. In March 1997, the subgroup submitted its report, titled Infrastructure collective pour la gestion des données numériques dans les bibliothèques universitaires québécoises (A Common Infrastructure for the Management of Microdata Files in Quebec University Libraries) CREPUQ, March 1997. This report was subsequently accepted by library directors from eleven universities, and they asked to my library (Laval University) to undertake the task of implementing Phase 1 of the SHERLOCK project.
The pilot project
Phase 1 of the pilot project started in September 1997 and was completed in October 1998. The phase focused on developing all of the system’s capabilities and setting up a first server centre.

Development team
The responsibility for implementing Phase I of the project was assigned to the library of Laval University, which established a development team made up of a project leader, the data librarian, a librarian and a computer analyst.

The team’s mandate was to develop all the system’s capabilities, with bilingual interfaces, set up an initial server for a limited number of surveys, and make corrections as needed during the trial period.

Project co-ordination
To ensure that the project went smoothly, the CREPUQ data librarians subgroup on data files was assigned the role of advisory committee.

Funding
The funding for the pilot project was provided through contributions from Quebec’s university libraries. Twelve institutions participated in the funding of Phase 1 out of a total number of 14. According to a complex formula, small universities invested less money than big institutions.

Institutions as clients
All users of Quebec universities, called client institutions, have access to SHERLOCK, but the use of the actual survey data requires that the institution’s library be a member of the DLI or the Inter-University Consortium for Political and Social Research (ICPSR). In addition to being user institutions, a few libraries will become server institutions.

Institutions as servers
The management of the surveys and their files is a responsibility shared by different server centres. Each institution (server centre) that has taken on responsibility for managing surveys in SHERLOCK has designated a local manager who is responsible for the management and follow-up of these surveys in SHERLOCK. These managers will be the only persons authorised to complete, to modify or delete a survey. A survey management module has been developed to facilitate these operations. For the implementation of Phase 1 of the pilot project, only the Laval University library acted as a server centre.

Surveys included
For Phase 1, fourteen surveys from Statistics Canada and one from ICPSR were installed on the system’s first server centre. Five of these support all the system’s capabilities (including extraction by variable and statistical analysis), while ten others support the basic level of use (retrieval, consultation of documentation and block files transfers, with no extraction). Under access licences, owing to the number, diversity and breadth of the surveys, some files can only be downloaded as a block (ftp), with no data extraction, while others have limited access, specifically to member institutions of the ICPSR. Some local surveys (e.g., a survey of Quebec public service retirees) could be loaded into SHERLOCK and be accessible only to certain universities. It is the case for a survey on political attitudes done by a graduate students’ class in my institution last semester.

The system
Access to SHERLOCK is based on a bilingual Web interface (French and English) offering a single and universal gateway to all the surveys. SHERLOCK is accessible in Quebec university libraries at the following URL address [http://sherlock.crepuq.qc.ca].

The general purpose of the system is to provide for the management and optimum use of all microdata files available in the Quebec university network. SHERLOK is not a teaching tool with a set of exercises, but it is easy to use by professors in undergraduate classes.

Capabilities
The main capabilities of the public module are:

- to provide access to the inventory and description of surveys by means of a retrieval module;

- to provide the user with documentation (survey metadata) on data files (guides, user manuals or codebooks, SAS or SPSS statements, record layouts and description of variables) when available;

- to enable users to extract subsets of data files in different formats for later processing at a local workstation. Intermediate and advanced users who can handle large sets of variables can download the complete dataset;

- to enable users to obtain simple statistical results such as a frequency distribution, cross-tabulation, mean, median or regression analysis on a variable in using the module of analysis.

More specifically, SHERLOCK can be used

- to ensure the compatibility of and access to information systems in twelve member institutions;

- to make the greatest number of surveys available;

- to promote the sharing of resources for data preparation, storage and use;
- to promote the development and sharing of expertise in the use of data among both the clientele and the reference staff of our libraries.

**Computer infrastructure**

SHERLOCK is a decentralized system made up of two modules: a public module and a management module.

The public module is used to access Web pages, conduct searches and access the forms used for retrieval and analysis. Searches are conducted on a UNIX main server (sherlock.crepuq.qc.ca) located at the Laval University library.

The programs needed by the user are Netscape, an e-mail software, WINZIP, Acrobat Reader, Excel/SAS or SPSS.

Whereas the documentation is accessible to the general public, access to data (transfer of complete file, extraction, analysis) is controlled by IP numbers, ensuring observance of licences governing use. Access to metadata is public but access to file transfers, extraction and analysis is controlled.

The html pages, for searching the description, the list of surveys resides on the main server (UNIX).

The survey metadata (codebook, record layouts, SAS and SPSS files, etc.) and data files reside in the different server centres on NT servers.

The data extraction and analysis is also done on the different NT servers. Extraction and analysis operations use Perl procedures.

**Management module**

The management module is used for the capture of data (description of surveys, metadata, and data files) from surveys that can be retrieved using the public module. The management module can be used only by the institutions who are server centres. The management
module has different functions. Using html forms, it is possible to work on the surveys, the files (metadata, data sets) and the variables. Only the French version of this module is available at this time.

At the survey level, the data librarian can add a survey, modify it or delete it. The data librarian also decides the treatment level (E/T), server address where the files will be loaded, which universities will have access to the survey. The data librarian enters the description and the abstract in both languages.

Once Inside a survey at the files level, you enter the files (metadata and data), giving a title to each file.

Inside a survey at the variables level, you can also add, modify or delete variables.

The module includes technical notes which are like an online manual. They are guidelines and procedures to facilitate the entry of metadata information.

SHERLOCK also collects statistics on usage (monthly/annual) by surveys, and by universities. With these statistics we can determine whether the users consult only the description, whether they transfer the complete dataset or whether they perform an extraction or an analysis.

Promotion

Now that the development of SHERLOCK is complete, institutions participating in the project are responsible for promoting this collective tool among data users in their respective universities.

To facilitate the marketing of SHERLOCK, the CREPUQ data subgroup organized two SHERLOCK information and familiarisation workshops. The first one took place at McGill University on October 15, 1998 and the second one, at Laval University (Québec City) in December 1998. These activities drew more than 50 participants (data librarians and staff serving the public). The introduction of SHERLOCK was supported by a press release and a presentation to the heads of university libraries.

In the Quebec universities network, library heads voted unanimously to continue the SHERLOCK project.

Accordingly, Phase II was developed from November 1998 to May 1999. This phase had a two-fold objective: to install SHERLOCK in three server centres (Université du Québec à Rimouski, Université de Montréal and McGill University) and to increase the number of surveys in the SHERLOCK collection, because we have gathered around 40 surveys in our collective tool. In addition to maintaining the system, the development team of the Laval University library has assisted the institutions with installation procedures.

For the year I starting next month, a Board of management has been created. This group will establish an annual program and will report to library directors. The users will be represented on the group.

More recently, the SHERLOCK project won a second prize among fifty projects presented at the CAUBO (Canadian Association of University Business Officers) as an academic initiative and development increasing productivity and effectiveness in higher education. The development team is very pleased with this recognition.

Conclusion

Among Quebec university libraries' the collective approach to the management of microdata files is two-fold: first, to “liberate” access to data, and secondly, to liberate their use.

SHERLOCK is also an active participant in the Data Liberation Initiative in Canada, which concerns the development of a data culture in our universities.

In jointly supporting the development of this research infrastructure, Quebec university Libraries are 1) encouraging the analysis of the statistical information available in the Quebec university network, 2) promoting student learning, 3) supporting the work of professors and researchers, and 4) participating in the demystification of data among library staff.

I would especially like to thank my three data friends (les trois amis des données). These data friends are not the same as the “Los tres data amigos”, well known at the ICPSR Summer Institute. I invite you to meet SHERLOCK in person at the poster session.

I Consisting of Richard Boily (Université du Québec à Rimouski), Jerry Bull (Université de Montréal), Gaëtan Drolet (Université Laval) and Anastassia Khouri (McGill University).

Changing Boundaries: Gazetteers, Information Retrieval and Data Browsing

This paper examines the role which historical gazetteers can play in web-based catalogues and data delivery systems. A gazetteer is a list of geographic names, which includes locational and other descriptive information. In this paper, the term ‘historical gazetteers’ is used specifically to describe gazetteers that incorporate both historical and modern geographical perspectives. In order to handle changed and changing geographical boundaries these gazetteers need to hold a wide range of information about geographic names, units, and hierarchies. This paper explains why gazetteers of this type are crucial for effective information retrieval and data browsing. In particular, it uses the History Data Service as a case study to describe how gazetteers of this type can be used to improve access to data via web-based catalogues and data delivery systems. This paper does not aim to describe the actual process of constructing and populating gazetteers (see Harper 1997, Hill et al. 1999, Moss et al. 1998).

The History Data Service (http://hds.essex.ac.uk) is funded by the Joint Information Systems Committee (http://www.jisc.ac.uk) of the UK Higher Education Funding Councils to collect, preserve, and encourage the re-use of digital resources which result from or support historical research and teaching. The History Data Service is part of the UK Data Archive and is the Arts and Humanities Data Service (http://ahds.ac.uk/) service provider for the historical disciplines.

The History Data Service collection covers a wide range of historical topics, and brings together over 450 separate data collections transcribed or compiled from original sources. The data collections cover a time period from the late tenth century to the mid-twentieth century, and the vast majority of data collections are either explicitly or implicitly geographically referenced. It is for this reason that the History Data Service is interested in developing and using gazetteers.

Explicitly and implicitly geographically referenced data correspond to a maze of complex geographies, which include administrative, electoral, census, and ecclesiastical geographies. These geographies are composed of a multiplicity of geographical unit types, which include amongst many others counties, wards, registration districts, and parishes. Because of this complexity, gazetteers are crucial for effective information retrieval and data browsing. This holds true both in the context of an historical service provider like the History Data Service, and in the context of the wider social sciences and humanities community.

Gazetteers are needed to make sense of this maze of complex geographies for three main reasons. Firstly many geographic names have a number of variant forms; secondly there are many incompatibilities between different geographies which means that boundaries do not align; and thirdly geographic names, units and hierarchies have changed in the past, and will continue to change. These problems are greatest with historical data, which are often associated with geographic names that have changed, or with geographical units that no longer exist, or with geographical units whose boundaries have changed significantly. It hardly needs saying that the disparity between modern and historical geographies increases with time.

Gazetteers improve information retrieval and data browsing by standardising geographic names and providing a controlled vocabulary of current and historical names within a system of preferred and non-preferred names. By linking disparate and changing geographies, gazetteers can help to integrate geographically referenced data collections, and deal with some of the incompatibilities when boundaries do not align. For example, gazetteers can make it easier to construct time series and other comparative data series by helping to identify those geographical units which, to a greater or lesser extent, correspond in different geographies.

If gazetteers are to be used to improve information retrieval and data browsing, it is essential that we understand the needs and requirements of users. The History Data Service has an active and ongoing policy of consulting with actual and potential users, and we have established that many users from the historical community require web-based catalogues and data delivery systems which will allow them to perform sophisticated geographical searches in an fairly automated manner. Users would like to be able to search
for data that cover a given place at a sufficient level of detail. For example, a user searching for the county of Essex would like to recover not only data that are indexed by the geographic name Essex, but also data collections that contain Essex county-level data but which are indexed by a higher level geographical unit such as England. They might also wish to extend the search to include data that are indexed by geographical units within Essex. Users would also like to be able to search for any data that can be analysed at the level of a specified geographical unit. It is self-evident that a reasonably complex gazetteer, which holds information about geographical units and hierarchies, would be required if these types of geographical searches were to be supported.

The History Data Service is working to improve and enhance access to its collection and a comprehensive UK historical gazetteer will be central to this work. Historical gazetteers are attracting an increasing amount of interest from data providers, research projects, and traditional archives. In consequence, the History Data Service would like to develop a comprehensive UK historical gazetteer in collaboration with other services and projects.

The History Data Service would use a comprehensive UK historical gazetteer both in web-based catalogues and data delivery services. It would use the gazetteer in web-based catalogues to support the types of geographical searches that users would like to be able to perform. Information about the History Data Service collection is made available through three different catalogues, the UK Data Archive’s information retrieval system BIRON, the CESSDA Integrated Data Catalogue and the Arts and Humanities Data Service Gateway; however, of these only BIRON even adequately supports geographical searches.

In BIRON geographical searching is facilitated by the geographical hierarchies in the Humanities and Social Science Electronic Thesaurus, HASSET (Data Archive, 1998). The geographical hierarchies in HASSET have been built up over time by the UK Data Archive and the History Data Service, but they are not by any means comprehensive; the historical hierarchies in particular have been developed only as when they have been needed. The geographical hierarchies in HASSET handle changing geographical boundaries by including geographical units in multiple hierarchies where necessary. The UK Data Archive and the History Data Service have increasingly recognised that the geographical hierarchies in HASSET cannot fully support the types of geographical searches that users would like to be able to perform, and that in consequence a more complex and comprehensive UK historical gazetteer is needed.

The History Data Service would also use a comprehensive UK historical gazetteer to help users to browse a web-based tree-structure, which will provide users with an alternative means of accessing information data. This will allow users to adopt a drill-down approach to locating data in addition to the more sophisticated geographical searching offered by web-based catalogues.

In web-based data delivery services the History Data Service would use a comprehensive UK historical gazetteer to support geographical data subsetting. A geographical subsetting service has been developed for a large collection of nineteenth and twentieth century statistics assembled by Humphrey Southall as part of the Great Britain Historical GIS Programme (Southall and Gregory, 1998). The Great Britain Historical Database Online (History Data Service, 1998) allows users to search across 30 tables simultaneously to retrieve a geographical subset. Users can select which variables are included in the subset, and they can also access online documentation. Because the data collection included all the necessary gazetteers it was easier to develop a geographical subsetting service as part of the Great Britain Historical Database Online. However, a comprehensive UK historical gazetteer is essential if the History Data Service is to extend this type of service to a wide range of other geographically referenced data.

The History Data Service would also like to use a comprehensive UK historical gazetteer in web-based data delivery services to provide integrated access to historical data and appropriate digitised boundary data, which users could then utilise in a GIS. The History Data Service and the UKBorders service, located at the Edinburgh Data Library, have been discussing the possibility of developing a joint interface which would provide integrated access to digitised boundary data held by UKBorders and attribute data held by the History Data Service (such as the Great Britain Historical Database Online). It hardly needs saying that it would not be possible to develop this type of service without a fairly comprehensive UK historical gazetteer.

The History Data Service is confident that a comprehensive UK historical gazetteer can be developed in collaboration with other services and projects. We believe that it will enable us to respond to user needs and develop web-based catalogues and data delivery services which allow users to perform sophisticated geographical searches, and we believe that its use will result in improved information retrieval and data browsing.

References


* Paper presented at the IASSIST Conference, May 19, 1999, Ryerson Polytechnic University, Toronto, Ontario. Cressida Chappell, History Data Service, UK Data Archive, University of Essex
The Royal Statistical Society Working Group on Archiving Data

Abstract
The Royal Statistical Society has recently established a working group to create standards for the collection and preparation of data in readiness for preservation. The working group consists of members of key organisations that are involved in both the collection and preservation of statistical material. The working group includes representatives from the private sector; the Office for National Statistics, (ONS); the Public Records Office, (PRO); the National Centre for Social Research, the UK National Digital Archive of Datasets, (NDAD); and the Data Archive. The representation of these organisations brings to the group a wealth of experience in both the collection and preservation of data from a range of sources including historical and administrative records, survey data and spatially referenced data.

The goals of the group are as follows:

- To define the extent to which materials, including questionnaires, data coding dictionaries, instructions for computations, working drafts and definitions of terms should be archived for future use.

- To establish a code of best practice for doing this.

- To suggest how data creators, custodians and users can co-operate to ensure that best practice is observed.

The paper will explore the need for such standards and will describe progress to date with a view to stimulating debate and eliciting wider opinions on some of the key issues that the group will be addressing.

Why establish a working group on the archiving of statistical material?
In July 1998, the Royal Statistical Society convened a meeting, ‘Archiving statistics: challenges and prospects’. The meeting was opened by Dr. Tim Holt, the Director of the Office for National Statistics and was well attended by over 60 data custodians and archivists, data producers and both public and academic researchers with interests in a diverse range of subject areas. In his introduction Dr. Holt recognised the importance of recording the processes by which statistics have been produced and acknowledged that the approach to preservation of such material within government has been inconsistent and varied between departments. Indeed, the Government Statistical Service had no overall policy on the archiving of the statistical material it generates. Dr. Holt also recognised the influential role of the Data Archive in demonstrating what could be achieved in the preservation of such material and drew attention to the recent establishment of the National Digital Archive of Datasets (NDAD). He welcomed the meeting and hoped that it would lead to improved procedures that would be agreed between the various sectors with an interest: data producers; data custodians and archivists; and data users.

All of the speakers recognised the importance of preserving those materials that explain the research or data collection process in order to allow fully informed use of the statistical material for future historical use and secondary analysis. Consequently, the speakers all contributed to the key aim of the conference: the stimulation of discussion about which paper and electronic materials are needed for the informed use of published statistics and how these can be preserved. There was general agreement that such material should include the contextual material associated with a data collection exercise. The list of possibly relevant material is potentially extensive and can include original questionnaires and data; coding notes; instructions for the creation of derived data; working drafts and definitions of terms. It can even be extended to include policy documents explaining why a particular set of data were collected or compiled in the form they were and at a particular time. The discussion included not only the provision of material associated with the collection of statistics through surveys but also the preservation of material produced during the collection and collation of administrative statistics such as birth and death counts or unemployment figures.

A recurrent theme of the meeting was the recognition that data producers need to ensure the implementation of good practice throughout the data collection exercise. In particular, it was recognised that it is critical that this need be met from the earliest stages of any project involving data collection. Ideally, guidelines needed to be established as a
reference tool both for the funders of data collection exercises and their project managers, to enable them to build preservation requirements into their management procedures at the development stage of any project. The application of such guidelines should then facilitate the collection and collation of all the relevant contextual material in readiness for archiving and preservation, once a project is completed.

Thus participants at the July 1998 meeting were unanimous in calling for a need for a coherent approach and defined guidelines for data preservation. Speakers in turn noted the loss of historical material, the need to preserve the contextual material relating to data collection exercises and, associated with this the need to ensure that a complete historical record is captured. There was general agreement that although these are recognised and worthy goals, the lack of a set of standards and guidance on the collation and preservation of such material is a major factor in the failure to meet the goals. In summary, there was an acknowledged need for action in this area. Thus, with the support of the President of the Royal Statistical Society, an RSS working group was proposed which has subsequently been approved by the RSS Council. This group is now well established.

The RSS Working Group.
Following an invitation to those attending the July conference, to express interest in participation in the group, the inaugural meeting took place in October 1998. Its composition reflects the breadth of interest demonstrated at the conference itself, including representatives from the spheres of custodian and archivist, data producers and data users. Thus, the committee includes data providers from the Office for National Statistics (ONS), the Home Office and the National Centre for Social Research. Data custodians are represented by the Public Record Office, (PRO), the UK National Archive for Datasets (NDAD), the Data Archive and Qualidata and users through the dissemination role played by the data custodians.

Terms of Reference.
The first meeting agreed the following terms of reference that have been subsequently agreed by the RSS.

• To define the materials, including questionnaires, data coding dictionaries, instructions for computations, working drafts and definitions of terms that should be archived for future use.
• To suggest how data creators, custodians and users can co-operate to ensure that best practice is observed.
• To establish a code of best practice for achieving this.

Existing literature
Subsequent meetings have been held in November 1998 and in February 1999. At the first of these the group established the need for a project plan which is now in place. The first task of the working group was to discover existing material that might be relevant and to review this. We have set ourselves a fairly daunting task since the breadth of statistical material under consideration is great. We are considering, amongst others, survey material, administrative records such as health records, observational data such as road traffic counts, census material and geo-coded data. The inclusion of contextual material extends the range of material significantly and we have extensive discussion about precisely what material needs to be preserved.

There was a general recognition that there are a number of initiatives which may well feed into and influence the work of the group and that there are a number of organisations which have, over many years, established their own guidelines for data collectors. It would be foolish to ignore this work: there are no benefits to re-inventing the proverbial wheel. Nor have we any desire simply to reproduce any existing document that potentially provides the standards in a given area. During late December and early January, therefore, members consulted with colleagues and trawled the Internet for papers and documents. A list of relevant documents was then compiled and each member was allocated material for review. 2

We approached the review systematically, asking the following questions for each document:

• What is its purpose?
• Who is the audience?
• What type of material has been targeted?
• How detailed is the information?
• Is the document prescriptive or for guidance only?

The review confirmed that there is a lot of material available that relates either to the deposit of material for further use or to the preservation of such material. There is also a great deal of technical information available relating to file and transfer formats and a lot of information relating to areas such as respondent confidentiality and copyright. There is also a significant body of work that gives guidance on contextual material. All of this work has been carried out by experts in the particular field and cannot be ignored. For example, the ICPSR Guide to Social Science Data Preparation and Archiving, was described during the review as “so sensible and universal, and the manner of its offering so persuasive that it could be accepted as a ‘mandatory’ standard”.

Fall 1999

23
Following this review, it was clear that although much has been written about the preparation of statistical material for preservation, there is no one document which offers a complete set of guidelines for all types of material and all data creators. Whilst many, such as the ICPSR guidance, provide sound advice, each has been designed for a select community of data providers. Understandably, then, documents tend to emphasise either the particular data type with which the organisation is concerned, for example, qualitative material, or information, such as acceptable deposit formats or media, which are specific to the organisations own procedures. A further distinction was evident in the material whereby existing recommendations can be loosely divided into two types. The first is those documents provided by institutions with whom data creators have a legal or contractual remit to deposit data, and the second are those that have been written by groups or institutions, only some of which have a custodial responsibility and are acting in an advisory capacity only.

Problems and resolutions for the working group
When determining the style, structure and content of the guidelines, a number of points were agreed to be self-evident.

There is agreement amongst the group that the most efficient and beneficial use of standards is to apply them at the data creation stage but we also recognise that preparing material to agreed standards for archiving imposes a financial burden on the data provider. These costs are incurred whether or not the provision of the material is mandatory or voluntary and whether or not the provider is a public or private organisation. It is a burden that is likely to affect the quality and quantity of material that is prepared for preservation and is regularly cited as an obstacle to archiving. One of the greatest challenges that the working group will have to overcome is the need to convince data producers that they will accrue significant benefits from the preparation of their material to agreed standards.

The group expects to recommend three approaches to this problem. Firstly, we are planning to include a section in the guidelines that will give advice on the potential costs incurred by preparing data for preservation. It is hoped that this will encourage those who commission data collection exercises to build realistic costing for preservation into their budgets from the outset. If we can achieve this, data collectors should be relieved of the budgetary constraints imposed when they are expected to send data for archiving. Careful thought will be needed in the presentation of this advice. Our current thinking is that it will need to be presented in terms of man-hours, for example, since information based on currency costing will not be relevant across national boundaries and will quickly become outdated.

Secondly, the group will seek methods of promoting the known benefits and often hidden cost savings of preservation of statistical material. For example, data collection is becoming increasingly costly. It is also becoming increasingly frequent as a means of discovering more detail about social and economic phenomena and, in the case of survey data, for example, respondent resistance is said to be an increasing obstacle to effective data collection. It is only sensible then to ensure that we get the maximum benefit from the statistical material that is collected. We can do this by promoting the re-use of material, for example where time-critical data are not essential.

Thirdly, the group does not expect to place the entire cost burden onto the data commissioners and collectors. Some of the costs will have to be borne by the custodians. We expect that as long as standards can be agreed and adhered to, data custodians will take some of the responsibility for converting material to the archival format. One possible way forward with this is to capitalise on the Data Documentation Initiative by making maximum use of the data type definition. This should enable data custodians to write and share conversion routines to convert data into a preservation standard. Work of this nature is currently being done at the Data Archive, the University of Essex and as part of the ddi/ddt beta test. With this in mind, the working group is currently reviewing the dtd as a potential generic starting point for a set of guidelines.

The group has also been involved in discussion about the presentation of the standards. Our aim will be to present the standards in a way that is acceptable to a wide audience and we must avoid the danger of producing a volume that is dense and not easily navigated. Current thinking on this is that it may be appropriate to provide an overview document that contains very basic guidelines with information that is relevant to the providers of all types of data. This document might include information on providing cataloguing records and on the costs of archiving. It might also include an index to sections of a fuller document or references to a series of individual documents that relate to specific types of data or cover complex topics such as respondent confidentiality, in depth.

Review of the Data Documentation Initiative (DDI)
Having identified the DDI as a potential, generic starting point for a set of guidelines, the working group is now reviewing the associated DTD for it’s suitability for this purpose.

The review is at an early stage but the DTD does have a number of acknowledged strengths and the group felt that it might provide the core for a set of guidelines that could be applied across data types. Its greatest strength is that it is intended that the DTD should be accepted as a standard. Combined with the composition of its’ committee and the inclusion therein of representatives from several continents,
it is realistic to think that the standard can be agreed internationally. The committee also comprises recognised experts in the field and the Initiative is being led by ICPSR, which the working group has already identified as providing excellent material in the field.

It needs to be noted, however, that at this stage the DTD does have weaknesses as a potential standard for the purposes of the working group. In particular, it has been designed as an exchange mechanism and at this stage it is not clear whether it can yet be used as an archival format. Nevertheless, as part of its current beta testing exercise, the DDI committee has invited comments on its potential use as an archival format. There is also ongoing discussion about how well the DTD accommodates aggregate data files and hierarchical files. This is also of concern to the working group but the DDI committee is actively considering it and the Data Archive is directly involved in the development of the DTD in these areas. The links between the working group and the Data Archive will enable the working group to keep up to date on progress and developments in these areas.

The group has three advantages that we anticipate will work in our favour and allow us to contribute to the future development of the DTD to accommodate a wider range of statistical material than it does at present. Firstly, the status of the group, with RSS support and a highly professional and respected membership, will allow us to speak with authority and make informed and respected representation to the DDI committee where we consider the DTD might be developed to meet the required standard. Secondly, the group is fortunate in having members whose interests cover a broad range of data types and statistical interests. So, for example, we have one member with an interest in Geographical Information Systems who is reviewing the DTD for its appropriateness to GIS material. Another member has an interest in textual material and open coded questions whilst a third is interested in individual level data where respondent confidentiality is a particular issue. Finally, the Data Archive is represented on the DDI committee, which has welcomed a dialogue with the working group and is keen to draw upon its expertise.

Possible ways forward.

We are not yet in a position to make definitive statements about the final model for the standards although we are clear on some issues. We do want to capitalise on the significant amount of high quality material that already exists. We also want to take account of the budgetary constraints of data producers and we want to offer standards that can be realistically adopted and maintained.

Nevertheless, it is possible to make some suggestions as to how the standards recommendations are likely to develop. It is most likely that we will adopt a position that there is already a great deal of material that could, with agreement from interested parties, be adopted as part of a formal set of standards. The group might then produce a document that directs producers, custodians and users of different types of material to organisations that have established an appropriate and agreed standard.

A second approach might be to encourage the expansion of an existing standard, such as the DTD, to include areas that it does not yet support.

In practice it is most likely that a combination of these two options will be adopted.

For more information on the RSS working group or if you would like to discuss the work of the group, please contact the author by email on beedh@essex.ac.uk.

1 The Data Archive is housed at the University of Essex, Wivenhoe Park, Colchester, England, CO4 3SQ. http://dawww.essex.ac.uk

2 A list of the documents covered can be obtained from the author at the University of Essex or email beedh@essex.ac.uk

3 DDI – co-ordinated by the International Consortium for Political & Social Research at the University of Michigan.

DATA IN THE DIGITAL LIBRARY:
Charting the Future for Social, Spatial and Government Data

June 7–10, 2000
Northwestern University

The Twenty-Sixth (26) Annual Conference of the International Association for Social Science Information Services and Technology (IASSIST) will be held on the campus of Northwestern University in Evanston, Illinois on June 7–10, 2000. This year’s conference Data in the Digital Library: Charting the Future of Social, Spatial and Government Data emphasizes the strengthening relationships between archives and libraries in managing, preserving and providing access to "digital collections".

IASSIST conferences bring together data professionals, data producers, and data analysts from around the world who are engaged in the creation, acquisition, processing, maintenance, distribution, preservation, and use of numeric social science data for research and instruction.

http://www.spc.uchicago.edu/DATALIB/ia2000/
The International Association for Social Science Information Services and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:
- Regular Membership: $40.00 per calendar year.
- Student Membership: $20.00 per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription: $70.00 per calendar year (includes one volume of the Quarterly)

I would like to become a member of IASSIST. Please see my choice below:
Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details:
http://datalib.library.ualberta.ca/iassist/mbrship2.html
- $40 (US) Regular Member
- $20 Student Member
- $70 Subscription (payment must be made in USS)
- List me in the membership directory
- Add me to the IASSIST listserv

Please make checks payable, in US funds, to IASSIST and Mail to:
IASSIST,
Assistant Treasurer
JoAnn Dionne
50360 Warren Road
Canton, MI 48187
USA

Name:
Job Title:
Organization:
Address:
City: State/Province:
Postal Code: Country:
Phone: FAX:
E-mail: URL: