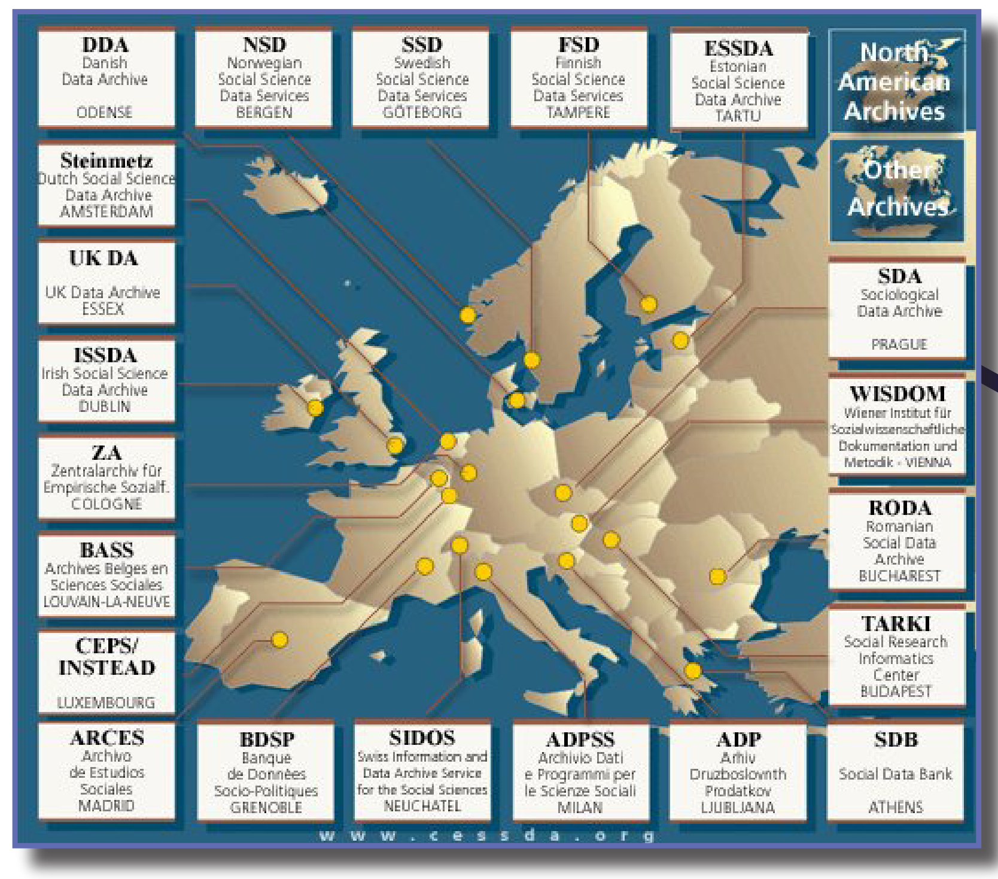
# The Social Science of Data Sharing: Distilling Past Efforts



There is an active coalition of social science data archives internationally.

Data Sharing for Demographic Research
HOME PAGE
DATA PRODUCERS
DATA USERS
Announcements - Major Studies - About the Project

The Inter-university Consortium for Political and Social Research (ICPSR), established in 1962, is an integral part of the infrastructure of social science research in the US.

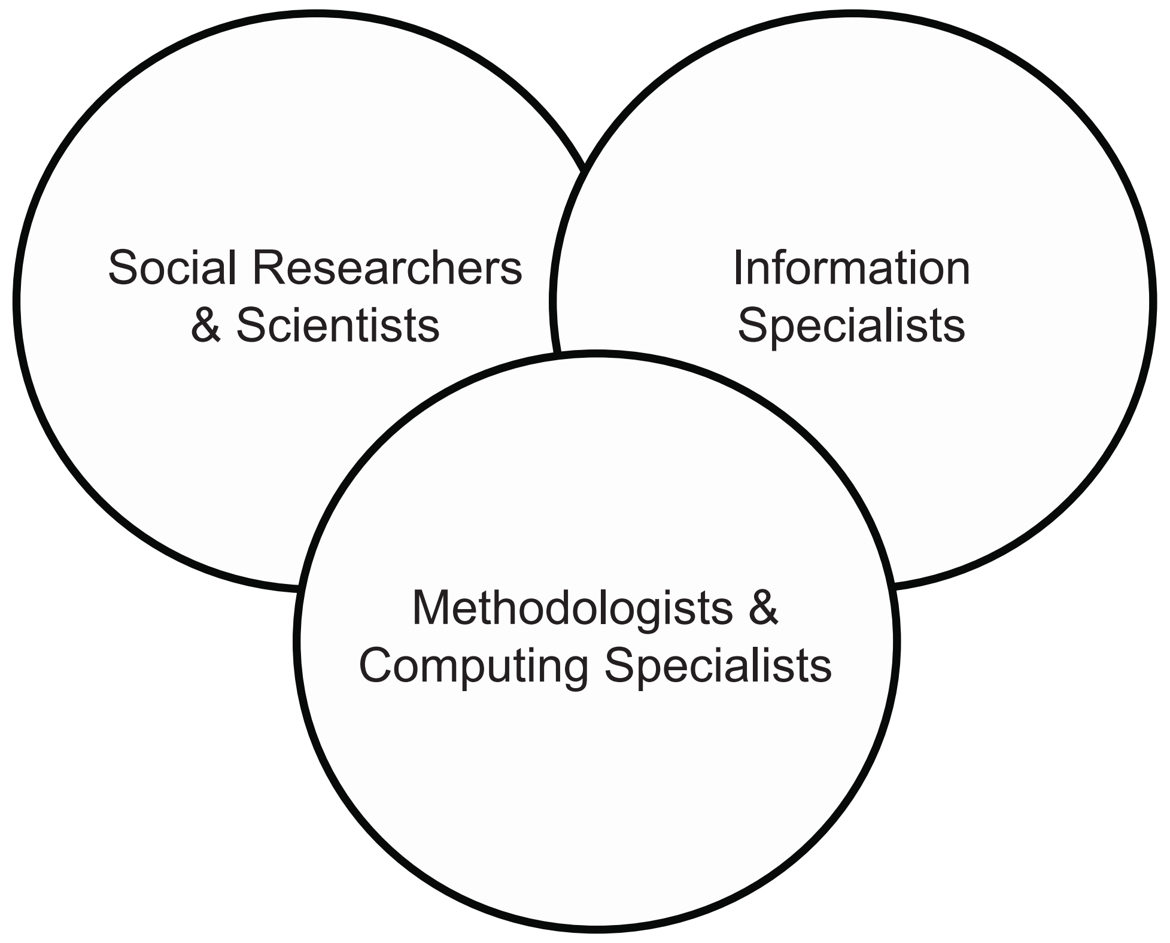
Founded in 1967, the UK Data Archive (UKDA) is curator of the largest collection of digital data in the social sciences and humanities in the UK.

## A strong culture of data sharing for secondary analysis and reuse

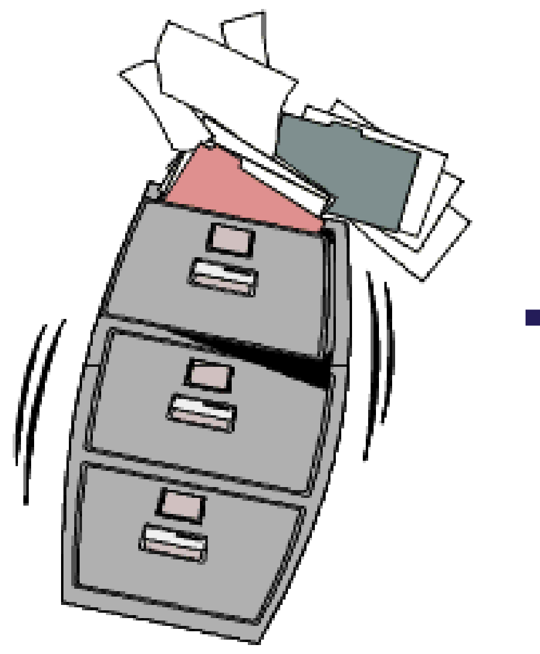
The secondary analysis of data, generated by third parties in the social & economic sciences arose in part because researchers could not command the authority and finances of government required to generate the data they needed. Funding agencies often fostered data sharing, or what we would now recognize as 'open access', as well as maximised their own financial resources by requiring that individual researchers deposit data in a trusted archive as a condition of grant award. Government data collections also became fodder for reuse as the results of large-scale surveys, administrative data, and streams of observational data were deposited and curated in economic, social and national archives thus promoting these data to be analyzed in novel ways. These digital data collections created by government or individual researchers offered research opportunity beyond the original intent of their collection including possibilities of combining data from different domains, e.g., demographic data and earth observational data to look at the human dimensions of climate change. This history of practice over the last 40 years, has not been without problems as the cost of data sharing requires considerable investment and inventiveness in creating archival and curatorial infrastructure.

## An infrastructure of data archives, data repositories, and data professionals

Since the 1960s, an infrastructure of national and domain-specific data archives and data libraries has evolved to assist secondary analysis by the international social science community. A cadre of data archivists, data librarians and social scientists, drawn from a variety of backgrounds, has grown up around this infrastructure, meeting annually over a thirty year period and otherwise operating as what is recognizable as virtual, as well as formal organization: the IASSIST (International Association for Social Science Information Service and Technology). The importance of this type of organization cannot be underestimated as it provides a collegial base for otherwise dispersed professional activities both geographically and institutionally, a means for sharing data and the best practices to manage them, and a forum for showcasing innovation and promoting technological advances. In addition, these data professionals have become a major resource in and of themselves and have a wealth of accumulated, intimate knowledge of datasets, understand their potential for appropriate re(use), and provide valuable individual support for researchers embarking on analysis.


Social Researchers & Scientists
Information Specialists
Methodologists & Computing Specialists

"IASSIST is an international organization of professionals working with information technology and data services to support research and teaching in the social sciences. Its 200 members work in a variety of settings, including data archives, statistical agencies, research centers, libraries, academic departments, government departments, and non-profit organizations." The diagram illustrates how IASSIST helps bridge the interests and concerns of three communities.



Appraisal decisions are not only needed for physical (paper) materials. Datasets must be appraised to determine level of curation activity, even if disk storage space is not an issue.

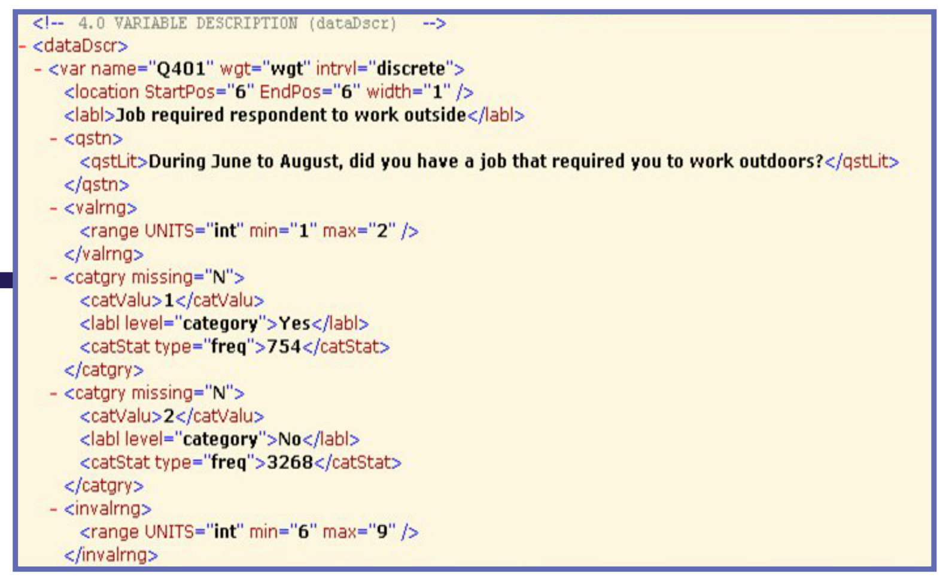
## Appraisal activities, or deciding what needs to be saved for future use

Collection development policies and procedures that include appraisal guidelines have been routinely used in traditional archive and library environments. Similarly social science data archives have developed well-established criteria for appraisal in order to prioritize the effort required to support data curation: long-term storage, creation of public use versions of data, mark-up in XML, and "publication" for online browsing and access. Data not meeting a host of defined intellectual, usability, and cost parameters may not warrant long-term preservation. For example, data may be rejected if insufficient metadata are available to establish context and relevance of a dataset including relationships among variables, data structures, software dependencies, provenance, and intellectual rights.

For additional commentary:
M. Gutmann, K. Schurer, D. Donakowski and H. Beedham. "The Selection, Appraisal, and Retention of Digital Social Science Data," Data Science Journal, 3 (30 December 2004): 209-221.

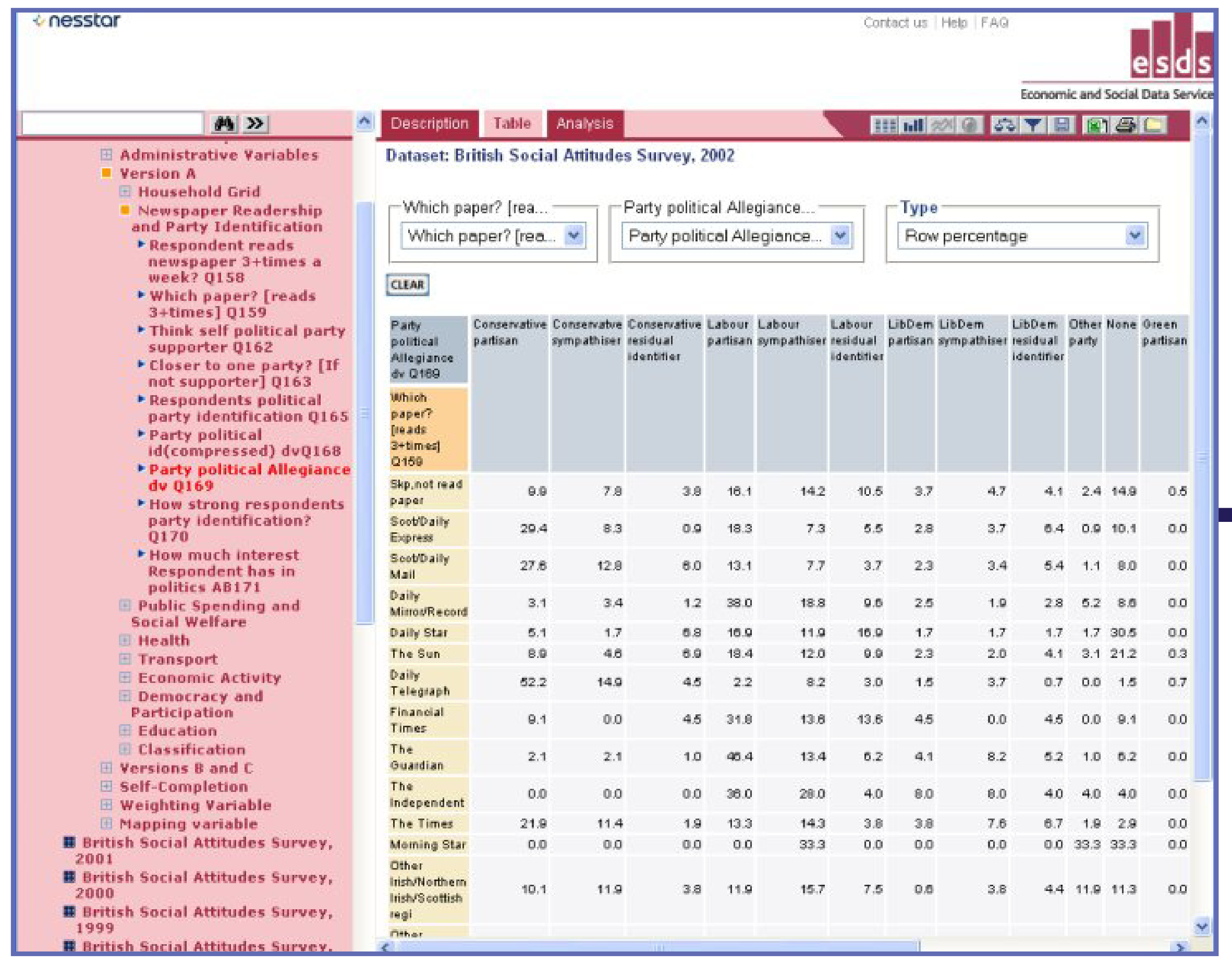
## Metadata matters, or retaining the context and relationships

The enhanced metadata standard, Data Documentation Initiative (DDI), for social science data, documents data elements and their relationships in micro (individual-level) and macro (aggregate) level social datasets in XML. Similarly, the UK National Geo-spatial Data Framework (NGDF) metadata standard, and the US Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geo-spatial Metadata (CSDGM) are needed for discovering, using, managing geospatial data. Each of these standards contains well over 300 elements. Creating metadata records of this magnitude is not trivial. A number of projects are underway to automate metadata capture at least in part and provide incentives for data producers to create their own metadata in order to reduce archival costs and deliver data to researchers in a more timely manner. The data life cycle should be considered at the time of data collection rather than post hoc so that a plan for capturing the appropriate metadata is implemented at the start of a research project. Establishing context, relevance, and relationships between data elements, assures that datasets will be reusable in the future and minimizes the risk of data loss.



```
  <!-- 4.0 VARIABLE DESCRIPTION (dataDscr)   -->
- <dataDscr>
 - <var name="Q401" wgt="wgt" intrvl="discrete">
     <location StartPos="6" EndPos="6" width="1" />
     <labl>Job required respondent to work outside</labl>
   - <qstn>
       <qstLit>During June to August, did you have a job that required you to work outdoors?</qstLit>
     </qstn>
   - <valrng>
       <range UNITS="int" min="1" max="2" />
     </valrng>
   - <catgry missing="N">
       <catValu>1</catValu>
       <labl level="category">Yes</labl>
       <catStat type="freq">754</catStat>
     </catgry>
   - <catgry missing="N">
       <catValu>2</catValu>
       <labl level="category">No</labl>
       <catStat type="freq">3268</catStat>
     </catgry>
   - <invalrng>
       <range UNITS="int" min="6" max="9" />
     </invalrng>
```

Variable-level XML markup using the DDI DTD.
(Data Resource Centre, University of Guelph)


nesstar
Contact us | Help | FAQ
esds
Economic and Social Data Service
Description
Table
Analysis
Administrative Variables
Version A
Household Grid
Newspaper Readership and Party Identification
Respondent reads newspaper 3+times a week? Q158
Which paper? [reads 3+times] Q159
Think self political party supporter Q162
Closer to one party? [If not supporter] Q163
Respondents political party identification Q165
Party political id(compressed) dvQ168
Party political Allegiance dv Q169
How strong respondents party identification? Q170
How much interest Respondent has in politics AB171
Public Spending and Social Welfare
Health
Transport
Economic Activity
Democracy and Participation
Education
Classification
Versions B and C
Self-Completion
Weighting Variable
Mapping variable
British Social Attitudes Survey, 2001
British Social Attitudes Survey, 2000
British Social Attitudes Survey, 1999
British Social Attitudes Survey,
CLEAR

Dataset: British Social Attitudes Survey, 2002

Which paper? [rea...: Which paper? [rea...
Party political Allegiance...: Party political Allegiance...
Type: Row percentage

| Party political Allegiance dv Q169<br>Which paper? [reads 3+times] Q159 | Conservative partisan | Conservative sympathiser | Conservative residual identifier | Labour partisan | Labour sympathiser | Labour residual identifier | LibDem partisan | LibDem sympathiser | LibDem residual identifier | Other party | None | Green partisan |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Skp,not read paper | 9.9 | 7.8 | 3.8 | 16.1 | 14.2 | 10.5 | 3.7 | 4.7 | 4.1 | 2.4 | 14.9 | 0.5 |
| Scot/Daily Express | 29.4 | 8.3 | 0.9 | 18.3 | 7.3 | 5.5 | 2.8 | 3.7 | 6.4 | 0.9 | 10.1 | 0.0 |
| Scot/Daily Mail | 27.6 | 12.8 | 6.0 | 13.1 | 7.7 | 3.7 | 2.3 | 3.4 | 5.4 | 1.1 | 8.0 | 0.0 |
| Daily Mirror/Record | 3.1 | 3.4 | 1.2 | 38.0 | 18.8 | 9.6 | 2.5 | 1.9 | 2.8 | 5.2 | 8.6 | 0.0 |
| Daily Star | 5.1 | 1.7 | 6.8 | 16.9 | 11.9 | 16.9 | 1.7 | 1.7 | 1.7 | 1.7 | 30.5 | 0.0 |
| The Sun | 8.9 | 4.6 | 6.9 | 18.4 | 12.0 | 9.9 | 2.3 | 2.0 | 4.1 | 3.1 | 21.2 | 0.3 |
| Daily Telegraph | 52.2 | 14.9 | 4.5 | 2.2 | 8.2 | 3.0 | 1.5 | 3.7 | 0.7 | 0.0 | 1.5 | 0.7 |
| Financial Times | 9.1 | 0.0 | 4.5 | 31.8 | 13.6 | 13.6 | 4.5 | 0.0 | 4.5 | 0.0 | 9.1 | 0.0 |
| The Guardian | 2.1 | 2.1 | 1.0 | 46.4 | 13.4 | 6.2 | 4.1 | 8.2 | 5.2 | 1.0 | 6.2 | 0.0 |
| The Independent | 0.0 | 0.0 | 0.0 | 36.0 | 28.0 | 4.0 | 8.0 | 8.0 | 4.0 | 4.0 | 4.0 | 0.0 |
| The Times | 21.9 | 11.4 | 1.9 | 13.3 | 14.3 | 3.8 | 3.8 | 7.6 | 6.7 | 1.9 | 2.9 | 0.0 |
| Morning Star | 0.0 | 0.0 | 0.0 | 0.0 | 33.3 | 0.0 | 0.0 | 0.0 | 0.0 | 33.3 | 33.3 | 0.0 |
| Other Irish/Northern Irish/Scottish regi | 10.1 | 11.9 | 3.8 | 11.9 | 15.7 | 7.5 | 0.6 | 3.8 | 4.4 | 11.9 | 11.3 | 0.0 |
| Other | | | | | | | | | | | | |

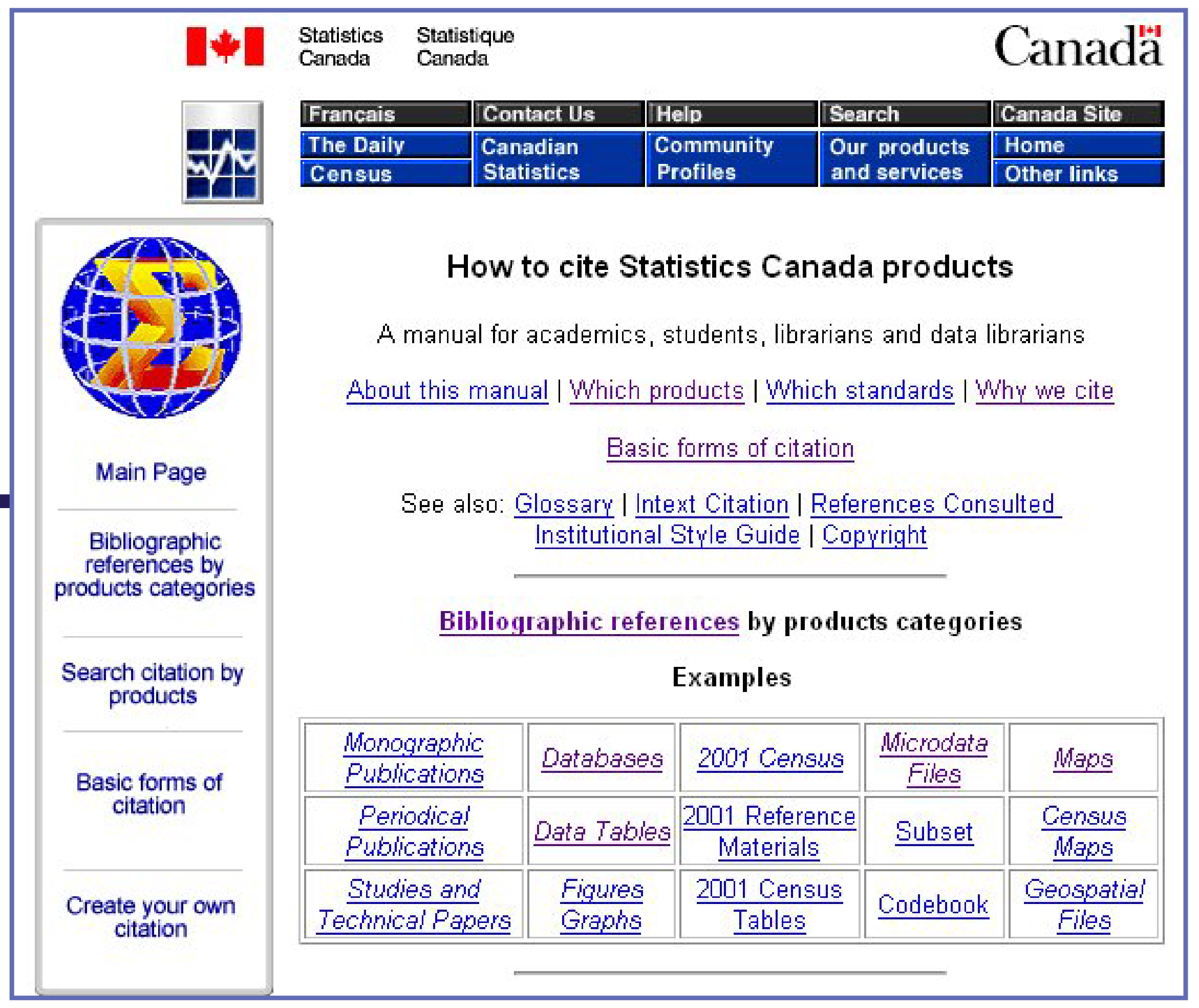
Online cross-tabular analysis of a UKDA survey using the Nesstar tool.

## Investing in discovery and analysis tools

With the rise of mainframe-based statistical software packages such as SPSS in the late 1970s, analysis of social science data became more prevalent and standardized input and output formats made long-term preservation of datasets simpler. But it wasn't until the 1990s when the power of microcomputers increased sufficiently to handle larger statistical analyses, web-based data extractors proliferated, and an XML-based social science metadata standard, the Data Documentation Initiative (DDI), was developed, that data curation activities became more sophisticated in the social sciences. Building on a tradition of data promotion and discovery for re-use, online systems that facilitate quick and easy access were developed both within the academy and commercially. With the DDI as the foundation, these systems allow researchers to locate, browse, subset, and download data. Some systems link datasets to the research publications using them; others allow a preliminary exploration and visualization of variables. Overall they increase interoperability among systems by converting legacy formats to DDI/XML for use by a variety of applications with enhanced portability overtime.

## Inciting citation practices

Researchers should cite computer files for the same reasons they cite traditional resources. A proper citation gives credit to the original author and also provides the information that others require to simply locate a dataset or enable replication studies. There are a range of practices in the social sciences from data archives and government agencies providing bibliographic citations for the data they disseminate all the way to journal publishers that require authors to cite their data in their bibliography. Still a search of one of ISI's citation indexes shows not only the absence of consistent citation practices, but also the dearth of primary data resources being cited at all. The emergence of trusted digital repositories using persistent identifiers for their content may provide a means for scholars to cite and retrieve data in a new way. There are certainly challenges for authors who wish to cite their data source especially if an online extraction has been run on the web, but in the spirit of research integrity, good citation practices remain a hallmark of good science.



How to cite Statistics Canada products

A manual for academics, students, librarians and data librarians

About this manual | Which products | Which standards | Why we cite

Basic forms of citation

See also: Glossary | Intext Citation | References Consulted
Institutional Style Guide | Copyright

Bibliographic references by products categories

Examples

| | | | | |
|---|---|---|---|---|
| Monographic Publications | Databases | 2001 Census | Microdata Files | Maps |
| Periodical Publications | Data Tables | 2001 Reference Materials | Subset | Census Maps |
| Studies and Technical Papers | Figures Graphs | 2001 Census Tables | Codebook | Geospatial Files |

Statistics Canada guidelines for citing both bibliographic and data products. Suggested citation form for a database: Statistics Canada. Year of publication. Title of Database [database, Beyond 20/20 or Excel]. Using Name of Distributor or Extraction System (distributor). Version. URL (accessed date).

Authors: Peter Burnhill, University of Edinburgh, UK, Diane Geraci, University of Michigan, US, Robin Rice, University of Edinburgh, UK