Data for Secondary Analysis: the Experience of the UK Data Archive

Presentation to the RESPECT Conference, Budapest, June 2003

Hilary Beedham
UK Data Archive[1]
University of Essex, England

Introduction to the UK Data Archive

Funded by the ESRC, the University of Essex and the JISC, UK Data Archive has been in existence for over 30 years. Over these years we have developed and are now the umbrella organisation for a number of data services. ESDS (Economic and Social Data Service) and HDS (History Data Service) and have gained a reputation for high quality research and development in the areas of resource discovery (data management) and data dissemination.

UK Data Archive has traditionally been funded to serve the research community in higher education but has recently been given a mandate to serve further and secondary education.

We offer a number of services: resource discovery, data dissemination (on-screen data browsing, a www download service and delivery on computer media such as CD-Rom), and the preservation of data for future use, including support for data depositors.

In addition we provide an on-line searchable catalogue for socio-economic and humanities data and the opportunity to simultaneously search catalogues of other European Social Science Data Archives.

The latter development has been made possible by participation in a series of FP5 projects, two of which are current. One of these, MADIERA[2], will further develop the on-line, cross-organisational search and browse service, to include a much-improved multilingual thesaurus for native language searching. As with other academic data archives, UK Data Archive only makes available anonymised data, with the onus for anonymisation lying with the depositor or data creator.

Our holdings originate from a number of sources: central government, research councils, charitable organisations and foundations; as well as private organisations. The data can be at individual or organisational level.

---

[1] http://www.data-archive.ac.uk/

[2] IHP-KA1-2001-1

2003 © Hilary Beedham

Intellectual Property and Data Protection

Intellectual Property and Data Protection are the two most important factors for consideration when making individual data available for secondary analysis. This applies regardless of the means by which data will be delivered. In the UK Data Archive, as in many other data archives and data libraries, the archive does not own the data it disseminates. It holds the data, under licence, for secondary use and users are required to agree a set of contractual requirements before accessing data. These include acknowledgement of the data originators, agreement not to identify respondents and to return a copy of any papers using the material to the UK Data Archive. Whilst there have been subtle changes over the years, the fundamentals of the licensing process have remained consistent: the protection of intellectual property rights and the protection of respondent confidentiality.

Data producers, whether individuals or organisations, have intellectual property in their data and, regardless of the method of dissemination, ownership must always be acknowledged. Data Archives, as explained, operate on the basis of licences which require secondary users to acknowledge data ownership in any work based on the material supplied.

Respondent confidentiality is an important issue for data archivists and is often cited as a reason for not making individual data available for secondary analysis. However, the UK Data Archive has been making such data available, under licence, for over thirty years and, latterly, using web techniques.

Managing Intellectual Property Rights

The key to the effective management of Intellectual Property Rights is the maintenance of structured catalogue records that meet commonly agreed standards. Thus the UK Data Archive, along with its European partner archives, apply the Dublin Core[3] to their catalogue records and use the DDI Metadata Standard[4] for records relating to datasets. This has allowed us to develop software that generates automatic, on-screen copyright statements when users browse datasets via the web. Similarly, the appropriate citation is created to accompany datasets that are downloaded. An additional benefit of this process is that it serves to re-assure researchers that the data they are using is from reliable sources.

As users are also required to return copies of published work that use data from the archive, these works can then be included in the catalogue record as publications associated with the original dataset and the secondary analyst is recognised for their value-added work on the data.

Protecting Respondent Confidentiality

There are three means by which the UK Data Archive ensures respondent confidentiality. The first is that data producers anonymise data before sending them to

---

[3] http://www.dublincore.org/

[4] http://www.icpsr.umich.edu/DDI/

the Data Archive. Nevertheless, as part of our standard processing for preservation and further distribution, the UK Data Archive applies procedures to ensure that there is no disclosive information has been inadvertently left in the files. On very rare occasions information has been identified that gave possible cause for concern that individuals might, with considerable effort, be identified. In these cases (three in thirty years), the depositors were contacted and their advice sought on further action before the data were released.

The legal protection offered by the licence agreement therefore represents the third means of ensuring confidentiality so that, even if identifiable data were somehow to be released, all users sign a legally binding agreement in which they agree not to identify individuals, either within the datasets or in combination with other material. Thus, there is ultimately, the potential for legal redress against any researcher who misuses data. In addition, such a user would immediately be excluded from all further access to data with the consequent effect on their reputation and career.

Technological Solutions

For many years this system worked in a pen and paper based environment with licences and user agreements passing between data owner, data archive and user by post. The process was time-consuming and administratively burdensome and often resulted in unacceptable delays to users.

As a result of technological developments, especially the benefits offered by the World Wide Web, users can now register, print an access agreement, sign and fax a copy to the archive and have their access agreed within twenty-four hours.

For our web-based dissemination service we have developed a sophisticated access control unit based on our records about users, datasets and the uses to which data are put. The system will block access to unregistered users; it can permit a user to browse the catalogue records but not the data; it can permit a user to browse the data and create ad hoc tables on screen; and it can restrict the downloading of data. In addition, it has the potential to allow differential access to variables within a file although we have not implemented this feature.

Statistical Disclosure Control

In collaboration with the other European archives and CBS Netherlands, and with Framework Programme 5 IST funding[5], UKDA investigated the possibility of real-time access control for ad hoc tables generated on the web. The work was successful insofar as it was possible to create a demonstration interface between the SDC software and the Nesstar publishing software[6] but more work and resources would be necessary to overcome a number of problems before any real-time service could be implemented.

---

[5] Faster IST-1999-11791

[6] See http://www.nesstar.com/

Conclusion

The UK Data Archive has long experience of making individual level data available for secondary analysis without any known breaches of respondent confidentiality, and in an environment where users understand the importance of ensuring that intellectual property rights are recognised and data protection laws are adhered to. This has been achieved through close co-operation with data producers, by adherence to standards and procedures, and by the maintenance of careful records about datasets and users. The systematic approach to the organisation of these records has allowed the UKDA, along with partner archives, to move relatively easily from a pen and paper based system to a web-based system with significant benefits to users in terms of reduced administration, easier access to information and speedier access to data.

Copies of deposit forms and user agreements are available from the UKDA website:

http://www.data-archive.ac.uk/

2003 © Hilary Beedham