UK Data Archive

Opening up access to birth cohort study data:

A UK Medical Research Council pilot project

Jack Kneeshaw
Senior Data and Support Services Officer
UK Data Archive

May 17 2007

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

The context: From principle to action

"MRC policy on data sharing recognises the value of making scientific data more widely available across the research community, a recognition that is agreed to be a next necessary step by other researchers, research funders, national and international government bodies. After a period of raising awareness in the research community about the value of timely and responsible sharing of data, MRC needs now to move from principle to action."

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

The subject:
The NSHD, aka the 1946 British Birth Cohort Study

- established in 1946, the National Survey of Health and Development (NSHD) is one of the longest-running large-scale longitudinal studies in existence
- since 1962, the study has been funded continuously by the MRC
- data include a wide spectrum of risk exposures and of clinically validated measures of mental and physical health, and biological and cognitive function; survey has data on periods of the life-course that cannot be reliably accessed in retrospect or in GP records
- 12,000+ variables across the various component datasets

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

The state of play (1): Access/dissemination

- the study team receive slightly upwards of 30 data access requests p.a.: a figure increasing year-on-year
- process of request through to supply can be drawn-out and episodic: specifying and retrieving data and documentation to be sent is time consuming and involves a high level of manual intervention
- the 'ship is now creaking'

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

The state of play (2): Finding/using the data

- the data collection is not well publicised: besides the NSHD web site itself, there are few finding aids that may guide potential users to the data
- potential users of the data have no means of searching for the survey instruments, topics, questions and variables that they might be interested in
- aside from the variable names, the data files supplied by the study team do not include any metadata

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

What might be?

- Is restricted access placing a ceiling on the level of scientific output that results from the use of the NSHD data?

Cohort	Annual usage (all users)	Annual usage academic (exc. students)	Publications since start of study
1946 cohort (NSHD)	**c. 30**	**c. 30**	**354**
1958 cohort (NCDS)	172	117	950
1970 cohort (BCS70)	169	119	329

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

How do we get from here to there?

Four criteria identified in order to make the data resource widely usable in the scientific community:

- **(1)** data have to be indexed to an international standard;
- **(2)** searching content of data and metadata has to be possible via the web for both in-house and remote users;
- **(3)** once identified through a search, data have to be easily accessible, along with all the information needed for informed research use;
- **(4)** technical and procedural (governance) arrangements need to respect data subject confidentiality and take account of statutory and other regulatory requirements.

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

The recurring theme:
Wider access vs. risk of disclosure

- special and increased risk of disclosure for longitudinal studies – more data points, vast range of information collected – rightly concerns the study team and sponsor
- important to start from position that disclosure risk can never be *eliminated* but can only be *managed*
- balance between attracting wider use of the data and retaining an appropriate level of disclosure risk becomes key

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

"It should always be borne in mind that, once data have been collected, the risk of disclosure can never be eliminated entirely; and, indeed, elimination of risk cannot be the aim if there is a policy to share the data. Instead, the aim must be to *limit* or *control* the risk. That is to say, in the context of sharing data so as to increase the scientific output, the aim of a disclosure risk strategy ought to be to define a level of risk that is acceptable: a policy aimed at reducing the risk of disclosure to a point as close as possible to elimination is not likely to be optimal."

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

Solutions?: From managing to sharing data

- 2-year pilot project initiated – project board convened – membership includes presenter
- specific aims of project:
- (1) prepare a subset of NSHD data, along with data descriptions and documentation (metadata) in a digital format suitable for entry into the Nesstar software;

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

- (2) define, document and implement governance arrangements for access to NSHD data through Nesstar;
- (3) implement management tools for data security and integrity, such as logging and access controls, where appropriate;
- (4) evaluating the benefits of implementation for the perspectives of the NSHD research team and other data users;
- (5) determine the financial costs, time and effort required, and other implications for extending the approach to the whole NSHD data library;
- (6) document 'lessons learned' to inform similar activities undertaken in the future.

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

Key outcomes expected

- generate a baseline measure of current activities (NSHD research team time, cost, etc.) required for provision of data and documentation
- help define access arrangements for whole study, including digitised genetic/phenotypic data
- *widen* (inc. geographers, social scientists, inc. non-UK) and *deepen* (current users find it more accessible, Nesstar facility to share derived variables) user base > > scientific output

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

Nesstar as the data sharing tool: Why? What? How?

Supported by:

University of Essex

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

JISC

UK Data Archive

- **Why Nesstar?**
- Nesstar is primarily a tool for data discovery with a strong focus on metadata that allows users to browse study information down to the level of variable
- software allows users, via a standard web browser, to view frequencies, conduct simple tabulations, produce graphs, sub-set and weight data
- 'user defined variables' function of specific interest to study team
- Nesstar is not the only package of its type – but study team's view is that, for searching, browsing, locating and *exploratory* analysis purposes, Nesstar is almost certainly the package best suited for the project's needs

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

- **What's to do?**
- estimated 2,000 variables (of the total of 12,000+) will be described in terms of their origins, distribution and response in the cohort, derivation methods from other variables if appropriate, and code book sources
- each variable will be assigned keyword(s)
- datasets published on web via Nesstar

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

- **How to do it?**
- NSHD 'in-house' solutions (e.g. scrambling ID for issued files)
- Nesstar modifications, especially to download function, to protect against inappropriate use
- publication of new derived variables via 'user defined variables'

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

Where next?

- project proper begins in July, though data/metadata prep. work already underway
- aim to make 2,000 variables available to selected user test group by end of year 1
- testing year 2: evaluation inevitably limited in scope but user feedback very important
- findings/recommendations published at end of year 2 and a successful report may see more MRC data rolled out via modified Nesstar product

Supported by:
University of Essex
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JISC

UK Data Archive

- Further details:

Jack Kneeshaw – kneejw@essex.ac.uk

Supported by:

University of Essex

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

JISC