, LSHTMResearchDataWorkingGroup; Leon, D (2011) Raising Standards - Lowering Barriers: Documentation of, access to and preservation of research data at the London School of Hygiene & Tropical Medicine. Working Paper. UNSPECIFIED.

DOI: 10.17037/PUBS.03409897

Raising Standards - Lowering Barriers

Documentation of, access to and preservation of research data at the London School of Hygiene & Tropical Medicine

Draft report and recommendations of the LSHTM Research Data Working Group

30 June 2011 (v5)

Contents

1. Summary and recommendations

Most research funders are now requiring that data collected in projects they have financed should in principle be made available to the wider scientific community. This report provides a basis for LSHTM to develop guidance, infrastructure and expertise to meet these expectations.

Underlying the move towards data sharing and improved access is the creation of adequate documentation of datasets. We argue that there is an overwhelming scientific case to improve the standard of documentation of the data we collect. This will be of great benefit to the research teams that generate the data in the first place, as well as significantly lowering the costs of providing data to third parties. While the level of documentation that we should aspire to will vary from study to study, there is a minimal standard which we need to ensure. Beyond this, the advent of web-based documentation, in which it is possible to provide easy-to-use and powerful ways of displaying information down to the level of individual variables is now possible at relatively low cost.

Aside from issues of documentation, there are important principles of access that include the need to protect confidentiality and to ensure that appropriate consent has been obtained from participants, the importance of establishing transparent procedures for access and making decisions on requests for data.

This is a rapidly moving area, with new initiatives being launched on an almost monthly basis. The School is in a good position to be one of the leading institutions in the UK in the way we document and make accessible our research data. This will not only ensure we are compliant with the requirements of funders, but it will also mean that the quality of the science we do is improved.

Recommendations to SMT

- LSHTM develop an institutional portal/gateway for discovery of our key data assets that is visible on the internet;
- Develop an intranet web-page with resources and links on issues to do with data documentation, archiving and sharing;
- Develop LSHTM policies/guidance on issues including :
    - obtaining appropriately wide consent to permit data sharing,
    - maintenance of confidentiality and minimising risk of disclosure of identities,
    - establishment of data access processes and procedures including model data sharing agreements,
    - inclusion of adequate budget lines on new grant applications (reflected in pFACT),
    - best practice and minimal standards for data documentation;
- Review institutional incentives for developing good and accessible meta-data;
- Review career pathways/opportunities for web-programmers and information specialists
- Explore with UK Data Archive possibility of exemplar archiving of some key LSHTM studies;
- Introduce staff training, including workshops on documentation and meta-data – principles and practice in line with guidance from the Research Development Framework;[1]
- Introduce same issues into taught courses and doctoral training also with reference to the Research Development Framework;
- Identify flag ship data sets/resources that should be encouraged to develop good meta-data and access procedures;
- Investigate how far Archives Service should be strengthened to provide relevant central support and guidance on these issues, and what the resource implications of this may be.

[1] http://www.vitae.ac.uk/policy-practice/234301/Researcher-Development-Framework.html

2. Terms of reference and membership

In February 2010 the LSHTM Senior Management Team (SMT) established the Research Data Working Group with the following terms of reference:

To advise SMT on the issues relating to the documentation, preservation and access to research data. Specifically:

- Advise on how good research practice can be promoted in the School to ensure the effective documentation and management of research data
- Consider the position of major research funders in relation to the preservation and access to research data to ensure that the School is able to conform to their requirements
- Review good research practices within the School and elsewhere and to assess how these are meeting funding bodies requirements
- Review guidance and advice available from external bodies, for example, Joint Information Systems Committee
- Investigate the potential solutions to issues relating to the archiving and preservation of data, including whether purely in-house or in collaboration with other HEIs or institutions, and whether a process for prioritising which data should be archived and preserved is required
- Investigate the extent to which recommendations on good practice may entail expenditure of resources, including the establishment of data access committees
- Develop a strategy to ensure that the School is at the forefront of good practice in relation to the documentation, preservation and access to research data.

The membership of the group was as follows :

David Leon (EPH) - Chair
Taane Clark (ITD, EPH)
Victoria Cranna (Archivist & Records Manager)
Paul Fine (EPH)
Andrew Gray (Repository Manager)
Judith Green (PHP)
Caroline Lloyd (Head of Library & Archives Service)
Alfredo Taqueban (Library) - Secretary
Sheena Wakefield (NST)

3. Process

The group met a total of five times. The rationale and aims of the group were communicated to each Department (now Faculty) by a member of the group speaking at a Departmental management group meeting. A web-based proforma was used to collect information about the range of datasets collected by LSHTM researchers and about their experience of and issues around data sharing. In April 2010 David Carr (Wellcome Trust) and Peter Dukes (MRC) came and gave presentations on the current state of the data sharing strategies. Contact was made with JISC and ESRC staff to learn more about any data sharing initiatives they have been developing. The Chair of the group met with Dr Elizabeth Pisani (an honorary member of staff) who has been centrally involved in helping the Wellcome Trust and other funders develop their data sharing strategies.

4. Background

The volume and diversity of data being collected or used for research purposes is continuing to increase year on year. This is driven by parallel advances in our capacity to collect and store increasing amounts of data as well as in the technologies employed to generate the data in the first instance.

Maximising the value of research funds. Research funders are increasingly expecting scientists to be prepared to share data they have collected. This is in part driven by a legitimate desire to ensure that the maximum value of their investment is realised for the broader public good. It is also in recognition of the wasted resources needed, and ethical issues raised by, generating new primary data sets when existing datasets could address the research questions asked. However, in this report, we argue that many of the things required to move in this direction are going to improve how we do science – both within our projects – and by facilitating wider access to data we have collected through making this process far less onerous and time consuming.

Documenting data. Good research practice relies on robust methods for documenting the data generated, as well as robust methods for ensuring its quality. Typically, considerable efforts are directed at increasing the reliability and validity of research design and the measurement of variables, but rather less on documenting the data produced or collected. Good quality documentation is an essential element of research governance, with advantages for the research team that collects/generates the data and for other data users.

The creation of quantitative, analytic datasets is done using software packages which generally allow implicit documentation of variables through for example "value labels". However, most analytic packages are not optimal vehicles for documenting or even linking to comments or information on source or validity of individual data items or details of the data collection instruments and overall study design. Much of this sort of higher level information is often held in the heads of the investigators or written down in Word documents or other digital documents that may not be easily accessible or searchable.

Recent developments in IT provide powerful tools that can result in a step-change in the way researchers can document and describe their data sets. Using web-based interfaces it is now possible to create easily used and understood descriptions of data sets that link information about individual data items (variables) to instruments and protocols. In this report we refer to this sort of information as *variable-level metadata*.[2] Not all studies will need to or be able to develop this standard of documentation. However, those that do will have a tool that hugely facilitates the use of their data by their own teams, by PhD students, by collaborators and other scientists.

Strengthening science. Throughout biomedical science, from genomics through to epidemiological studies of environmental exposures, from trials to observational studies, there is increasing recognition of the importance of combining all available evidence from studies to produce the best estimates of effects of interest. Individual studies, and their investigators, increasingly contribute to meta or pooled analyses, which in turn requires sharing of data with groups outside of the initial investigating teams.

Many health studies, particularly cohorts, are now recognised to have a utility that goes beyond what was originally envisaged by the investigators who originally set them up. New questions can be asked of already-collected data, often by people outside of the original study team. This once again requires sharing of data.

[2] See Glossary (Annex 6) for a definition of types of metadata, together with Section 6.3.

Finally, one of the defining characteristics of science as a way of understanding and acting in the world is that it is a self-correcting activity. This often takes the form of assembling new evidence to challenge hypotheses. However, it also works through reanalysing existing data and sometimes showing that original inferences were flawed. While it must be acknowledged that some reanalyses could be mischievous or biased by commercial interests, in general, science would be strengthened if access to datasets for purposes of appropriate reanalysis was easier.

Improving research practice. This report attempts to summarise key aspects of the current, rapidly changing landscape of data sharing and documentation and to recommend the priority areas that LSHTM may wish to take forward. The highest standards of documentation, research governance and full exploitation of research data are not going to be implemented overnight. Indeed we have identified the importance of adopting a flexible and proportionate approach to moving in the direction of improved documentation and discoverability. Some studies may be too small or limited to require any significant change in how things are done. In contrast, newly funded large studies whose value is likely to be exploited over an extended period by a wide range of researchers should aim to adopt current best practice. However, what defines "best-practice" is not always clear, and this report identifies the need for LSHTM to develop information sources, guidelines and policies in this area.

5. Findings

In this section we summarise key information about the external environment and the resources both within and outside LSHTM that have a bearing on our terms of reference.

5.1 Funders

Over the past year there have been a number of major developments in this field, with funders in particular publishing strategies and position statements.

In May 2010 funders of biomedical research met in Washington DC to move towards a consensus statement, which was finally published in January 2011 followed shortly by a Lancet Comment by the head of the Wellcome Trust Mark Walport.[3] The statement is reproduced as Annex 1 at the end of this report.

Wellcome Trust.

The Wellcome Trust's statement on data management and sharing requires "that all of our funded researchers maximise the availability of their research data with as few restrictions as possible. When developing research proposals, we expect all researchers to consider their approach for managing the data they will generate." However, the studies that are a priority in this respect would have one or more of the following characteristics :

- large-scale (requiring significant resources over time)
- broad utility
- creating reference datasets
- associated with [research] community buy-in

3

1. Walport M, Brest P. Sharing research data to improve public health. Lancet 2011;377(9765):537-9.

Illustrative examples of applications that would require a data management and sharing plan would include large-scale genetic association studies of common diseases; genome-wide or large-scale functional genomic studies in a specific organism; and longitudinal studies of patient and population cohorts.

For studies generating small-scale and limited data outputs, a data management and sharing plan will not normally be required. Generally, the expected approach for projects of this type would be to make data available to other researchers on publication, and where possible to deposit data in appropriate data repositories in a timely manner. While a formal data management and sharing plan need not be submitted in such cases, applicants may find the guidance below helpful in planning their approaches for managing their data."

Medical Research Council.

The MRC has been developing a comprehensive strategy on data generated by studies funded entirely or in part by the Council. This has two aspects:

- Expectations of MRC regarding data sharing, and their incorporation into policies and criteria for future grant funding
- Development of infrastructure to enable data to be discoverable i.e. visible to researchers searching for already existing datasets that may help them address scientific issues of interest

In parallel, and likely to converge, MRC is developing a similar approach to tissue banks, although this is slightly less developed.

The MRC has been working closely with a set of its flagship cohort studies (including ALSPAC, Whitehall II and the 1946 National Survey of Health and Development) to improve use of these valuable data by researchers. This Data Support Service (DSS) initiative will enable population researchers to make the best possible use of data assets for new science without unnecessarily duplicating expensive resources. This includes creation of meta-data as well as developing policies for establishing data access procedures particularly for third party researchers. The aim is to provide within the near future a single MRC web portal or gateway through which researchers may discover these data. The details have yet to be worked out in detail, but it seems likely that the information about each study included will ultimately be maintained by the research group itself.

Economic and Social Research Council (ESRC).

The ESRC adopts the OECD principle that 'publicly-funded research data are a public good, produced in the public interest and, therefore, should be openly available to a maximum extent possible' (ESRC 2011 'Research Data Policy'). Their revised Research Data Policy (2010) has now been implemented. This advises applicants to seek advice from the ESRC Economic and Social Data Service (ESDS) on management, sharing and confidentiality, and obliges new applicants to make a detailed statement in proposals about the nature of the data sets to be generated, and the plans for their storage, management and archiving. The award holder has a responsibility to manage the data, make preparations that enables them to be shared and to offer any data set produced by the project (with appropriate metadata) to the ESDS for data sharing/archiving within three months of the award end. The ESDS manages the process for the UK Data Archive (see below), which is the repository for most datasets from ESRC projects. Applicants have to make an explicit case for not providing a data sharing plan. Data sets accepted must be prepared in line with guidance from ESDS.

5.2 UK Data Archive (UKDATA)

This is a centre of expertise in data acquisition, preservation, dissemination and acts as the UK's national centre of expertise in data management and data sharing, but with a primary focus on humanities and social sciences. It has the largest collection of digital data in the social sciences and humanities in the UK. It is funded by the ESRC and JISC[4] and the University of Essex. Anyone who is funded by the ESRC is required to offer their data to the UK Data Archive but others can also offer their datasets. The preferred formats and the metadata they require are in their guidelines. LSTHM had offered 8 datasets to UKDATA; 4 were accepted.

Elizabeth Pisani suggested that LSHTM could as an institution directly approach UKDATA to explore whether we could develop a joint project aimed at using them to archive datasets at LSHTM.

5.3 Existing LSHTM initiatives and policies

As of 2010, a repository manager and assistant are now in place, Andrew Gray and Emma Golding respectively. Their first task is to create a digital repository that will hold metadata for, and full text of, research articles written by LSHTM staff. Over the 2 years of the project this will be integrated with other relevant school systems including the publications database. A set of policies and guides will be produced to support the repository and users. It is intended that the School's repository will have the functionality to hold most classes of digital objects including datasets. However, whether LSHTM should attempt to extend this service to support datasets will require careful consideration.

The School currently has Guidelines on Good Research Practice[5] but these do not deal in any depth with the issues considered in this report.

5.4 LSHTM survey of data sharing issues

In June 2010, sixteen LSHTM-coordinated studies with large data repositories were identified by the committee and faculty planning groups. The study PIs and data managers were contacted, and asked to complete an electronic questionnaire consisting of twelve questions that described: (1) their data repository or "asset", (2) whether these data are shared, and the mechanisms for the sharing process, (3) whether there has been funding to maintain documentation or meta data for the data asset, and (4) views on how the LSHTM could contribute to archiving and the sharing process. We received a response for each study (n=16).

The majority of the studies (n=15) are externally funded (MRC 5, Wellcome Trust 4, Gates foundation 2), and involve large data collections (at least 4000 individuals, up to 10 million records) predominantly from clinical trials (n=4) or cohort studies (n=8). The earliest collection dates from 1962, and more than half are ongoing studies with active data collection. Most projects have study protocols and/or questionnaires, and data are stored in databases (e.g. Oracle, Access) with specific mention of data dictionaries to electronically describe information recorded. Nearly all the studies (n=13) are part of a collaborative project involving data sharing, and the majority of those had a data sharing mechanism, including formal agreements, access

[4] JISC supports "UK colleges and universities in the innovative use of digital technologies, helping to maintain the UK's position as a global leader in education."
[5] http://www.lshtm.ac.uk/research/policies/

through PIs and committees or a website. For the remaining studies (n=3), commercial, political and military sensitivity were stated as barriers to data sharing. Nearly all studies (n=13) have had requests for access to the study data. Only nine studies reported that the documentation for the datasets was fully sufficient for someone unfamiliar with the study to use it. Only one study has received financial resources to develop documentation and/or meta-data for the study, but two were actively seeking funding. Six studies reported regulatory or funding agency requirements for access to the data, data storage / archiving or maintenance. These included restrictions imposed by data suppliers or government agencies to prevent passing data to third parties, and Wellcome Trust or MRC guidelines for access by the research community.

Several issues were highlighted concerning the role of the LSHTM in the documentation, access and preservation of research data, including:

- The study data are an important resource for LSHTM researchers, and should be maintained in a form such that other LSHTM researchers are aware they exist and may benefit from accessing them.
- Although there is a need to archive studies, there appears to be little or no provision in existing grants, especially beyond the funding period. The LSHTM seems to be performing well in the archiving and storage of questionnaires, but there were requests for "the delivery of a (subsidised) service to assist with archiving data".
- The development of a LSHTM (or access to an externally-based) data repository was requested in a number of comments, including: "it would be very helpful if LSHTM could invest in a database facility so that we would all use the same common platform", and we need "a well structured secure data repository that PIs could have access to (e.g. a protected directory dedicated to me/people who work for me) to hold all my datasets and related files". This solution could "support the generation of metadata to a proper standard", and "would be better than just putting the study on a departmental website as the latter does not guarantee permanent archiving".
- Security and access to full and meta datasets arose as an issue to most of the respondents, and recommendations included "a clear published data security policy", the provision of "a standardised form for data access and publication agreements", and "providing web space for datafiles to be available for download."

Further details about the survey and the responses are presented in Annex 3.

In summary, we concluded that with a small number of exceptions, the level of data documentation of most studies in the school is not at a level that would not be easy for others to interrogate, and falls short of our picture of web-based meta-data. Most importantly, the LSHTM website provides no means at all for systematic discovery of the data we do have. Nevertheless, there was a recognition of the importance of improving the situation, with many people seeking guidance and advice as to the best way to do this.

6. Key issues

The key issues identified by the group were as follows:

6.1 Raising internal standards for documentation

The development of improved levels of (especially web-based) data documentation and description (meta-data) is desirable for the research teams generating the data as it:

- makes analyses more robust by forcing explicit and more complete specification of strengths and weaknesses
- reduces the risk inherent in having detailed knowledge of data sets residing only in the heads of key individuals who “know the data”
- facilitates induction of new members of the research team or PhD students allowing them to start using the data more quickly and efficiently
- provides a valuable tool for use during analyses even when conducted by analysts familiar with the data, particularly for datasets that are large and complex.

6.2 Meta-data

Meta-data is a term that is used when discussing data documentation and management. The glossary in Annex 6 provides a very general definition. However, we concluded that there are several rather different types of meta-data that are quite distinct, although they all fall within the definition of being “descriptive or contextual information” that refers to (in our case) to research datasets. For our purposes there are fundamentally two types of meta-data :

> Project-description meta-data. This is key structured information describing a study such as name, purpose (statement plus key words indicating geographical location, exposures, outcomes etc.), study design, size, calendar period, types of variables, location, responsible person(s) and so on. This is usually collected and retained in the form of a structured *proforma*. For archival purposes, it will also include agreed period of data retention and information on any restrictions/mechanisms for access. This sort of meta-data provides information that can be used as the basis for data discovery and for archival management. However, it is often not of particular value for creators or users of the actual data.
>
> Variable-level meta-data. This is structured information for original and derived variables in a dataset. At a minimum this would be equivalent to a conventional coding schedule including : variable name, type, format and value labels. However, more extensive information is now often included for each variable, including such things as precise text of questions (if variable derived from a questionnaire), comments on validity or problems with variable and also descriptive statistics e.g. univariate frequency distribution for categorical responses, or mean and range for continuous variables. This sort of meta-data is of great value to creators and users of the actual data in a research context.

Web tools now enable these types of meta-data to be combined, searched and hyper-linked to relevant resources and documentation relating to the study. This sort of *variable-level meta-data* provides the most powerful and flexible approach to documenting datasets in detail. Such meta-data systems can relatively easily be extended to provide a means for potential users to select variables (using a “shopping basket”) which makes the creation of a bespoke dataset for analysis a relatively trivial issue. Web-based meta-data of this sort is becoming the standard for genetic and other –omics data, although it is only recently becoming adopted for other types of data.

The type of meta-data that is appropriate will vary from study to study. A simple tabular project-description represents a minimal level of meta-data documentation. A fully functional web-based system for meta-data that allows management of variable by variable data requests is at the other end of the spectrum. Examples of studies that have varying types of web-based meta-data are given in Annex 5.

6.3 Sharing data

There is considerable anxiety among researchers about the issue of data sharing. The main concerns are that (i) in the future data researchers will have to place their data in some sort of "public space" where they will have no effective control on access. This will in turn risk bringing the study into disrepute by allowing incorrect inferences to be drawn by people who are not sufficiently knowledgeable about the data, as well as undermining the incentive researchers may have to collect data if it is then accessible to others who have not spent time and effort collecting the data; (ii) the costs/effort of providing access to complex datasets are potentially high, simply in terms of helping third parties understand the data, and then arranging for the data to be extracted; (iii) the consents provided by subjects may not extend to sharing data (even if anonymised).

The group recognised that all these concerns have a legitimate basis. However, we believe that there are some misconceptions, and moreover that some of the solutions to these concerns would in fact have substantial benefits for researchers and their own teams as has already been outlined above in section 6.1. Certainly, the funders we have spoken to distinguish between fully open access and more controlled and scrutinised data sharing, and see a place for both. For the sorts of datasets collected by LSHTM researchers, it would often not be appropriate to place it "on the web". However, we could go much further to facilitating discovery of valuable resources by other researchers, and improving the documentation and management of our data so that costs of individual requests for controlled access would be much lower than envisaged. This would be through the development of better documentation and meta-data during the data collection phase, as well as doing this retrospectively for high value datasets that continue to be of interest to the research community at large.

The question of consent is a more challenging one, and would need to be considered on a case by case basis according to the study. However, if we as an institution move forward on this issue, we will assemble expertise and experience which can be shared. This issue is discussed in more detail in section 6.5.

Finally, we concluded that no one size fits all. There are going to be many small or highly focussed studies which may cease to be of interest to other researchers once the main paper(s) have been published.[6] Nevertheless all research data will need a minimum level of *project description meta-data*, there will be some studies where *variable-level meta-data* will be developed but this will not be appropriate for all studies. However, criteria for deciding which studies do require such investment need to be further developed.

6.4 Data Archiving

The School has an archive service which currently manages mostly paper records. It is developing guidelines on managing electronic data, including research data. The service offers advice on the management of both electronic and paper records. Part of the archives remit is to preserve and to make accessible records to the wider community, and while the archives team has the skills to do this for the majority of records received, additional skills and resources would be required for making research data accessible in the way outlined by the funders.

[6] Note that all research data needs to be retained for ten years in order to comply with the School's Records Retention and Disposal Schedule.

LSHTM has an ongoing repository initiative run by the Library & Archives Service, however it is recognised that the initial project is to make the School's publications accessible, with a view that in the future the repository could possibly hold data, depending on the system used.

It should be noted that the School states that research data is retained for a minimum of ten years, as stated in the School's Guidelines on Good Research Practice (http://www.lshtm.ac.uk/research/policies/). Data can either be retained by the Faculty or can be deposited in the records management service.

Discussions with funders have indicated that they do not plan to set up a new central archive for research data, therefore it will be necessary to investigate other options for depositing data. This would involve a mix of solutions including using the UK Data Archive for data which meets their criteria (this is currently a service for humanities and social science data), other external services and internal hosting where appropriate. There are many UK wide initiatives and other institutions investigating this issue (see Annex 4) and their work should be reviewed to get a greater understanding of how other institutions and bodies are approaching the data archiving issue.

6.5 Ownership and responsibility

This was a potentially major concern of the group. At the moment LSHTM does not have any policies or recommendations about what adequate and appropriate mechanisms could be put in place to make decisions around third party requests for access to data. Moreover, there are often issues to do with ownership of data and the consent given by study participants about the extent to which access can legitimately be granted to external researchers. In addition, we do not have guidance on the scope and nature of formal agreements that owners of the data would require third parties to sign in order to protect the confidentiality of the data and to ensure its appropriate use.

The group considered that larger studies are likely to require specially constituted data access committees with independent chairs with authority to make decisions on requests for access. These would need to be transparent and timely in the way they operate. Smaller studies may not require such formal structures, but would still need clearly specified mechanisms for requesting access.

Qualitative data sets present some particular problems, in terms 'ownership' and rights to contribute to interpretations of data that have been 'generated' by individual researchers, rather than 'collected', as may be the usual model with biomedical data. It is typical for qualitative researchers to feel rather more ownership over data they have generated, such as interview transcripts or ethnographic field notes, and a common problem (exacerbated by short term contract research) is that those who generated qualitative data sets (and may have a justifiable claim to be included in further analysis) may have left the institution by the time external requests for access to use the data arrive. This reflects a wider problem of data ownership, noted by, for instance, the BSA (2001)[7] with respect to authorship and the potential marginalisation of junior staff. Further, the requirements to maintain the confidentiality of qualitative data may be more cumbersome, with time needed to adequately check that transcripts or audio files do not contain identifiers, and that original participant consent would cover the new analysis proposed. LSHTM researchers currently have a small number of requests for access to small qualitative data sets, often from MSc students wishing to re-

[7] BSA (2001) BSA http://www.britsoc.co.uk/Library/authorship_01.pdf

analyse research study data for summer projects. These are handled informally, and have presented few problems so far, but any formal system for cataloguing existing data may generate more complex issues of access, managing the rights of past researchers to exploit their own data and ensuring confidentiality.

6.6 Software

The software that is required to document datasets depends upon the type of data and the level of documentation that is desirable given the value and anticipated future use of any particular resource. Our survey revealed that, with a few exceptions, there was relatively little experience among researchers at LSHTM of using purpose-built software for documentation and meta-data. There are a few examples of bespoke web-based interfaces as described in Annex 5. However, there are now an increasing number of off the shelf solutions such as DDI (see Annex 4).

6.7 Training

Our research staff and PhD students have had limited exposure to state-of-the-art systems, and thus are perhaps unable to fully assess the benefits of moving towards using them. This requires investment in staff training in this area, and even the extension of this area to our teaching both at Masters and Doctoral level.

6.8 Developing Central Support

As noted above in (see section above on Data Archiving) the LSHTM Archives team currently offers advice on the management of datasets to staff and is planning on developing guidelines on managing datasets which will be available to staff on the Information Management and Compliance area of the website.

Strengthening the existing service would enable wider ranging support to be provided for the management of digital datasets. Some additional resources are likely to be required for this.

A central service could liaise with researchers to provide support and guidance, and focus on supporting a minimum level of project-description meta-data. Staff in the service would develop policies and act as a source of expertise regarding generic data management. They would be responsible for the following specific areas.

- Managing the system that holds the project-description metadata for datasets
- Providing advice about appropriate places to store datasets
- Managing the interface for deposit of datasets/ project-description metadata of datasets (the interface is likely to be part of the system above)
- Managing the ‘public’ view of the project-description metadata i.e. the information that is available for searching
- Providing links from project-description metadata to full dataset, including those housed beyond LSHTM
- Advising on and ensuring compliance with appropriate standards e.g. DDI
- Developing policies to manage the retention of datasets. These would include:
    - How long datasets are kept for

- Whether the retention period extended when a dataset is used
- Liaising with researchers regarding appraisal of which datasets have their retention dates extended
- Withdrawing and disposing of datasets

6.9 Resources

Developing adequate web-based meta-data (particularly variable-level meta-data) requires resources and can be costly. This is both in terms of researcher time in providing the key information relating to study design, instruments and variables, plus the time for a skilled professional who is required to set up and maintain the IT infra-structure.

It appears that major funders are now saying that they are prepared to pay for such resources, although in reality at the present time funding panels are generally still unlikely to prioritise such issues when making funding decisions. The culture may however change within funding organisations, and there are certainly indications that WT and MRC are working on specifications for more explicit specification by applicants of their data sharing and documentation plans.

From the perspective of LSHTM, we have relatively limited experience of developing levels of web-based documentation and meta-data that are of an appropriate standard. The numbers of people currently employed at LSHTM who can develop or implement the web-based tools (especially variable-level metadata), is very limited. This skill shortage needs to be addressed.

Developing and keeping up-to-date project-description meta-data also requires resources. These are most appropriately going to be located at the School level – and would fall within an extended remit of the School's Archive service (see Section 6.8 above).

Acknowledgements

We would like to thank Dr Elizabeth Pisani and Professor Basia Zaba for providing the group with important source material and their insights into the issues, and colleagues who responded to our survey.

Annex 1 - Sharing research data to improve public health: full joint statement by funders of health research (published January 2011)

This statement[8] was the output from a global funders' meeting held in Washington, DC in April 2010.

Signatories (funders of particular importance to LSHTM in bold italics)

- Agency for Healthcare Research and Quality (USA) Carolyn M Clancy, Director
- ***Bill and Melinda Gates Foundation*** Tachi Yamada, President - Global Health Program
- Canadian Institutes of Health Research Alain Beaudet, President
- Centres for Disease Control and Prevention Thomas R Frieden, Director
- Deutsche Forschungsgemeinschaft (DFG) Matthias Kleiner, President
- Doris Duke Charitable Foundation Ed Henry, President
- ***Economic and Social Research Council (UK)*** Paul Boyle, Chief Executive
- Health Research Council of New Zealand Robin Olds, Chief Executive
- Health Resources and Services Administration (USA) Mary K Wakefield, Administrator
- Hewlett Foundation Paul Brest, President
- INSERM André Syrota, Chief Executive Officer and Chairman
- ***Medical Research Council (UK)*** John Savill , Chief Executive
- National Health and Medical Research Council (Australia) Warwick Anderson, Chief Executive Officer
- National Institutes of Health (USA) Francis S Collins, Director
- Substance Abuse and Mental Health Services Administration (USA) Pamela S Hyde, Administrator
- ***Wellcome Trust*** Mark Walport, Director

Introduction

Recent advances in information technology have revolutionised science - providing new opportunities for researchers to share data and build on one another's work. Informatics and the ability to mine large datasets and combine them with information from many other sources present a huge potential to advance developments in public health. The importance of data sharing in advancing health is becoming increasingly widely recognised, and has been strongly endorsed by the H8 group of global health organisations.

In some research fields - such as genetics and physics - data sharing is well-established and has accelerated the progress of research and its application for the public good. In public health research, however, while research collaborations are growing more common, the sharing of data is not yet the norm, even within the scientific community.

Much of the data collection that could improve public health research is expensive and time-consuming. As public and charitable funders of this research, we believe that making research data sets available to investigators beyond the original research team in a timely and responsible manner, subject to appropriate safeguards, will generate three key benefits:

* faster progress in improving health

* better value for money

* higher quality science.

[8] source : http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030689.htm

Each funding institution will work within its own legal and operational framework, and we are committed to working towards these goals together. We intend to establish joint working groups where appropriate. We call on governments and other actors that generate routine health service statistics and other types of public health data to adopt a similar approach.

This Statement establishes guiding principles and desired goals. It recognises that flexibility and a variety of approaches will be needed in order to balance the rights of the individuals and communities that contribute data, the investigators that design research and collect and analyse data, and the wider scientific community that might productively use data for further research.

The joint statement of purpose

Vision

We, as funders of health research, intend to work together to increase the availability to the scientific community of the research data we fund that is collected from populations for the purpose of health research (1), and to promote the efficient use of those data to accelerate improvements in public health.

Principles

Funders agree to promote greater access to and use of data in ways that are:

Equitable: Any approach to the sharing of data should recognise and balance the needs of researchers who generate and use data, other analysts who might want to reuse those data, and communities and funders who expect health benefits to arise from research.
Ethical: All data sharing should protect the privacy of individuals and the dignity of communities, while simultaneously respecting the imperative to improve public health through the most productive use of data.
Efficient: Any approach to data sharing should improve the quality and value of research and increase its contribution to improving public health. Approaches should be proportionate and build on existing practice and reduce unnecessary duplication and competition.

Goals

While we recognise that progress may be gradual as we develop mechanisms and resources consistent with these principles, we aim to work in concert to achieve the following.

Immediate goals

Data management standards support data sharing

> Standards of data management are developed, promoted and entrenched so that research data can be shared routinely, and re-used effectively.

Data sharing is recognized as a professional achievement

> Funders and employers of researchers recognize data management and sharing of well-managed datasets as an important professional indicator of success in research.

Secondary data users respect the rights of producers and add value to the data they use

> Researchers creating data sets for secondary analysis from shared primary data are expected to share those data sets and act with integrity and in line with good practice - giving due acknowledgement to the generators of the original data.

Longer-term aspirations

Well documented data sets are available for secondary analysis

Data collected for health research are made available to the scientific community for analysis which adds value to existing knowledge and which leads to improvements in health.

Capacity to manage and analyse data is strengthened

The research community, particularly those collecting data in developing countries, develop the capacity to manage and analyse those data locally, as well as contributing to international analysis efforts.

Published work and data are linked and archived

To the extent possible, datasets underpinning research papers in peer-reviewed journals are archived and made available to other researchers in a clear and transparent manner.

Data sharing is sustainably resourced for the long term

The human and technical resources and infrastructures needed to support data management, archiving and access are developed and supported for long-term sustainability.

Annex 2 – Summary of issues arising from presentations given to the group by MRC and Wellcome Trust April 2010

Presentations given by Dr Peter Dukes (MRC) and Dr David Carr (WT) on data access policies.

- Approach being taken by WT and MRC are similar, and there is considerable contact between key players in each organisation. This is contrasted with the approach being taken by ESRC.
- Both WT and MRC are in processing of developing/reviewing their policies, with impression given by both that they are receptive to concerns of research community. Funders have already invested considerable thought in this issue.
- Key theme (articulated by Peter Dukes) was "maximising lifetime value" of any particular study/data set/biorepository
- WT policy likely to be available by summer, while MRC revision not likely to be finalised for sometime into 2011 – as it is based on their ongoing pilot work with key cohorts
- Both funders emphasised that there was an apparent lack f understanding in the research community about what they were looking for. However, both agreed that in applications what is expected is that a reasoned case is put that states governance arrangements (in particular) about access for third parties, special periods of exclusive use by research team, and the appropriate level of resource required to document, archive and make accessible data.
- Meta-data creation seen as key component of successful strategy – which would have benefit for researchers collecting data as well as creating basis for relatively pain free access to others if it ever came to this. It might also ensure that the results of others analysing data sets they did not collect themselves might be of higher quality – as they would understand the data better.
- Clear that what is expected depends entirely on the specific project/study and what is seen as its lifetime value. At one end a very targeted specific study with no proposed follow-up which once completed would be unlikely to be of much interest to others, at the other end a large scale cohort investment where the expectation is that it would in part be a resource for the wider scientific community.
- Peter Dukes helpfully gave two examples of data sets generated for immediate public release : 1. the social science "omnibus" surveys that are contracted out to social survey organisations who once data is collected deposit them for scientific use in (for example) the ESRC Data Archive. 2. the "-omics" data generators (as in Human Genome Project) where factory type production of sequence data is placed in the public domain as soon as validated. Peter's view was that the public health/epi community did not fall into either camp – and was in fact the one group in biomedical science who were most concerned about issue of open access – in many ways for reasons that were understandable.
- Neither funder was pursuing idea of funder-maintained central data repositories (ie no equivalent of ESRC data archive). Recognition, however, that each institution creating its own archive might be unrealistic and inefficient (Judy Green pushed on this issue). However, other models also available – whereby groups working in different institutions, but with common interests got together. In the lab field there are examples of this taking form of centralised infrastructure such as EBI (http://www.ebi.ac.uk/) which holds sequence data etc. However, could imagine (perhaps) the limited number of key players in Global Health in UK getting together to develop common framework, approach and maybe repository.
- Non-digital data should not be overlooked.

- Both funders accepted that proper investment in access, archiving etc. was costly – but that they would pay (note : unclear however that funding boards/panels take this line when deciding on what to award – an issue of funders to sort out internally).
- Peer review could form basis for deciding on when data could be disposed of (not everything needs to be kept forever).
- Many studies already have strong collaborations that involve sharing data. Useful point – in that funders recognise this – that there is already quite a lot of providing access to others going on.
- Analogy of freezer chaos versus having proper and well archived sample storage with modern LIMS (lab information and management system).

Annex 3 – LSHTM survey of data sharing issues

The aim of the survey was to document the current practice of data sharing, archiving, and documentation in LSHTM-based studies with large data repositories, as well as potentially define any future needs of researchers that wish to make their data available.

Sixteen studies were identified by committee members and faculty planning groups, and their respective PIs and data managers were sent an electronic survey. All sixteen studies identified responded with a complete survey. The survey consisted of the *following information*, (12 questions) and we include a summary of responses where appropriate:

1. Name of the asset (e.g. database or data source, study or sample repository);

2. Name of the funder(s);

AngloAmerican (1), British Heart Foundation (1), Cancer Research UK (1), Department of Health (2), DFID (1), EU, Gates foundation (2), Medical Research Council (5), Sigrid Rausing Trust (1), Wellcome Trust (4), No funder (1).

3. The name of the person completing the form;

4. Their role with respect to this asset (e.g. principal investigator, data manager);

PI (n=11), other (n=5)

5. A brief description of nature of the asset/study (e.g. cohort study of 12 thousand children born in Aberdeen 1950-55 with follow up data to date);

6. A description of how the data/asset is documented. (e.g. availability of information about design, data code books, variable names etc.);

Study protocols (n=3), databases (n=7), data dictionaries (9), questionnaires (8), web-access to information (2), technical reports (1).

7. Whether the asset is part of internal or external collaborations involving data or analysis sharing,

Yes 13, No 3

If yes, what procedures or mechanisms there are in place for dealing with requests for access to the data by individuals who are not part of the core study team?

- "Request referred to the Principal Investigator in the first instance"

- "The anonymised data are kept on the School's network (N drive) and the access to these data is restricted to the Group. However, other users (e.g. collaborators, Research Degree students, MSc students, etc.) can have access to these data: they have to sign a Confidentiality Ageement Form and must analyse the data within the School premises. To share these with external collaborators, approval from ONS, Ethics and Confidentiality Committee (ECC) and Medical Research Ethics Committee (MREC) is needed."

-" Anyone can download the data after registering their name, affiliation and contact details and providing a brief description of the intended use of the data."

- "The database is still being created; no mechanisms are yet in place. We anticipate that persons such as summer project students will sign agreements not to share the data or publish without permission of the core team."

- "Template form to be filled in for each specific study by prospective collaborator. Form reviewed by co-PIs."

- "Not yet established formal mechanisms as collaborations are limited at present."

- "The website is accessible to the public so no requests for access required."

- "Data sharing agreement"

- "The study is in a phase where at least half of the requests for use of the data come from people outside the core study team. From the publically accessible study website there is information about how to request access to the data. Such requests are dealt with as they arise by an Oversight committee."

- "Request sent to all CO-PIs who decide how to respond. So far we have not refused data to anyone."

- Need ethics approval (n=3).

If no, whether there are any ethical, practical or legal barriers or reasons for these data not being shared?

Yes (n=3): commercial, political and military sensitivity.

8. Whether there have been any requests for access to the study data/asset from individuals who are not part of the core study team

Yes 13, No 3

9. Whether the study has received funding specifically to develop documentation and/or meta-data for the study/asset? If yes, which funder?

No 15, Yes 1 (Rockefeller Foundation)

10. Whether the documentation is sufficient in its own right for someone unfamiliar with the study to be able to start using the data if it were provided to term?

Yes 9, Partial 5, No 2

11. Whether there are any relevant regulatory or funding agency requirements for sharing access to the asset, data storage / archiving or maintenance associated with the asset.

No 10, Yes 6

- *"Data suppliers specify that we cannot pass data to third parties"*
- *"Data "owned" by the government"*
- *"ONS, ECC and the NHS body MREC"*
- *"Wellcome Trust require data to be available to collaborators"*
- *"Those that generally apply to MRC funded projects"*
- *"The whole agreed purpose of the project was to make publicly available these documents. Funders provided support for this purpose."*

12. How the LSHTM could assist with the upkeep or possible archiving of the asset?

- *"A clear published data security policy. A well structured secure data repository that PIs could have access to (e.g. a protected directory dedicated to me/people who work for me) to hold all my datasets and related files."*
- *"Developing data documentation and providing data extracts for sharing would require a further full-time senior programmer."*

- *“The School should provide a secure network which really complies with the data security rules established by British Standards 7799. So far, we have been forced to use our own "secure computer" (with no external connection) within a "secure room" in order to comply with the data security requirements of the various statutory bodies.”*
- *“Access to a proper repository which could support the generation of metadata to a proper standard would be better than just putting the study on a departmental website as the latter does not guarantee permanent archiving. However, I rather doubt that LSHTM needs its own repository.”*
- *“Person time to develop data dictionary”*
- *“Providing a standardised form for data access and publication agreements; maybe (if we did want to make the entire dataset publicly available) by providing web space for the datafile to be available for download.”*
- *“The ACCESS database was established in Dept of Social Medicine, University of Bristol and its development and maintenance is still operated from Bristol. It would be very helpful if LSHTM could invest in an database facility so that we would all us the same common platform. This is what Bristol does, it works and LSHTM could do the same or even better!”*
- *“We would benefit from a unified database”*
- *“This dataset is, sadly, not available for sharing. However the LSHTM has been very helpful in providing archive storage space for paper copies of questionnaires.”*
- *“The School has been helpful in finding archive storage space for the paper questionnaires form this study (over 100,000 questionnaires). Given the acute shortage of space, I am very grateful for this facility.”*
- *“Provide storage for database on server and maintenance costs for the resource beyond the funding period of the ACT Consortium”*
- *“I employed one of my research fellows to develop the webbased metadata about 8 years ago. At the time there was no one else at LSHTM doing such things (that we found at least). To my knowledge the skills required to put the sort of metadata together on the web are still uncommon here. We need to build up a pool of expertise in this.”*
- *“The data: Would be great if it could be archived properly by LSHTM. Ideally might require some additional funding for someone to document it more fully, but this may not be essential.“*
- *“This work is done most efficiently by skilled and experienced staff dedicated to this ... would be great if LSHTM could provide a (subsidised) service to assist with this ... upkeep and archiving takes time and resources, usually at a point in the research project when time is particularly scarce and resources are absent. Our funders are not likely to allocate resources directly for this.”*
- *“Person-time from an "archivist"”.*

A complete summary of responses in an excel spreadsheet is available upon request.

Annex 4 – Links to useful resources and initiatives (as of May 2011)

Research funders

Medical Research Council's data sharing pages :
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/index.htm

Wellcome Trust's data sharing pages :

http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/index.htm

Economic and Social Research Council's research data policy pages :
http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx

Biomedical funders joint statement of purpose on sharing research data to improve public health launched January 2011.

http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030689.htm

plus Lancet Comment by Mark Walport

http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_peda/documents/web_document/wtvm049648.pdf

Report on Sharing Data from Large-scale Biological Research Projects from the WT-funded 2003 Fort Lauderdale meeting. This meeting was driven principally by the issue of sharing genomic sequence data.

http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf

In 2009 the Toronto International Data Release Workshop concluded that rapid release of prepublication data has served the field of genomics well, and recommend extending the practice to other biological data sets.

http://www.nature.com/nature/journal/v461/n7261/pdf/461168a.pdf

Data Documentation Initiative (DDI)

This is a free meta-data system developed and so far mainly used by the social science community.

"The Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving."

http://www.ddialliance.org/

Some epidemiological/public health/biomedical studies are starting to use it. These include the Global Fund for AIDS, TB and Malaria (http://www.theglobalfund.org/html/5YEdata/?page=catalog)

UK Data Archive

The UK Data Archive is curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom. In May 2011 it published the 3rd edition of its 'Managing and Sharing Data - best practice for researchers' guide. This can be downloaded as a pdf from :

http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

The Data Archive runs the the Data Management Planning for ESRC Research Data-rich Investments project (DMP-ESRC) is funded under the JISC Managing Research Data programme. It aims to increase the data management and sharing capability within the social sciences community.

http://www.data-archive.ac.uk/create-manage/projects/jisc-dmp

In May 2011 this project has produced a first draft guide to helping estimate the relative costs of data documentation and archiving, although this is very general and does not help with estimating costs for more elaborate meta-data projects :

http://www.data-archive.ac.uk/create-manage/planning-for-sharing/costing

The Economic and Social Data Service (ESDS)

http://www.esds.ac.uk/

ESDS is a national data service providing access and support for an extensive range of key economic and social data, both quantitative and qualitative, spanning many disciplines and themes. ESDS provides an integrated service offering enhanced support for the secondary use of data across the research, learning and teaching communities.

Managing Research Data (JISCMRD)

http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

Current JISC programme to manage research data with a number of project now finishing

- piloting essential research data management infrastructures within institutions and for distributed research groups
- improving practice in research data management planning
- developing tools to help institutions plan their research data management practice
- encouraging the publication of research data and demonstrating the benefits of improved methods for citing, linking and integrating research data
- stimulating the acquisition of appropriate skills among academics and research support staff in Universities

Digital Curation Centre

http://www.dcc.ac.uk/

A national body funded by JISC to provide assistance, advice and support to institutions wanting to curate their digital data. It provides toolkits, plans, policies, guides, manuals, case studies and workshops covering all areas of managing digital research data.

DATUM for Health: Research data management training for health studies

http://www.northumbria.ac.uk/sd/academic/ceis/re/isrc/themes/rmarea/datum/

JISC funded project to complete in July 2011 to promote research data management skills of postgraduate research students in the health studies discipline through a specially-developed training programme which focuses on qualitative, unstructured research data.

Data Asset Framework

http://www.data-audit.eu/index.html

To provide organisations with the means to identify, locate, describe and assess how they are managing their research data assets. DAF combines a set of survey methods with an online tool to enable data auditors to gather this information. DAF help with planning a strategy to ensure research data produced in UK Higher Education Institutions are preserved and remains accessible in the long term.

Glasgow Data Management Support

http://www.gla.ac.uk/services/datamanagement/

University of Glasgow provide a set of service support for researchers wanting to manage their research data. The resource covers creation, access, intellectual property rights, ethics and file structures.

Annex 5 - Examples of studies involving LSHTM researchers with an explicit data sharing element

- **Aberdeen Children of the 1950s study (ACONF)**

http://www.abdn.ac.uk/aconf/pages/navigationHome.html

The web-site documenting this study is now hosted in Aberdeen – but it was originally developed and set up at LSHTM. The most valuable part of the website is the Database Search facility (http://www.abdn.ac.uk/aconf/pages/navigationHome.html) which researchers who have used the data say is enormously helpful to have running at the same time as they are analysing the data in STATA (or whatever package they are using). Illustrative screen-shots are reproduced on the following page.

- **ALPHA network**

http://www.lshtm.ac.uk/eph/psd/alpha/

- **NATSAL study National Survey of Sexual Attitudes and Lifestyles**

This illustrates the UKDA model, with data deposited as per funders' requirements with open access after a specified period (to allow research team to exploit the data for publications). Datasets consist of three cross sectional surveys from 1990-1, 2000-1, and a current tranche, clinical data on sample results and some qualitative data, codebooks. Researchers report low burden for data preparation (as high quality documentation needed anyway for multi-institutional team) and UKDA manage almost all access requests, and incentives to publish in a timely manner from obligation to deposit.

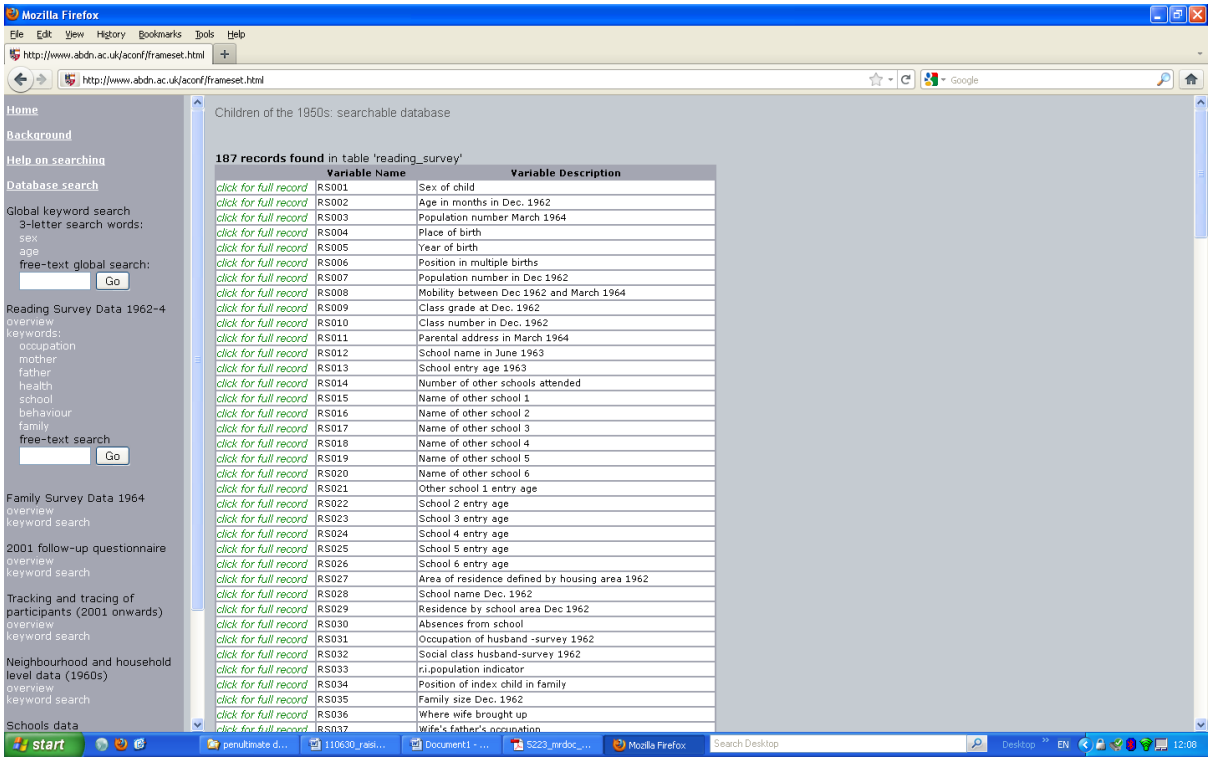
http://www.esds.ac.uk/findingData/snDescription.asp?sn=5223

- ***Mycobacterium tuberculosis* drug resistance study**

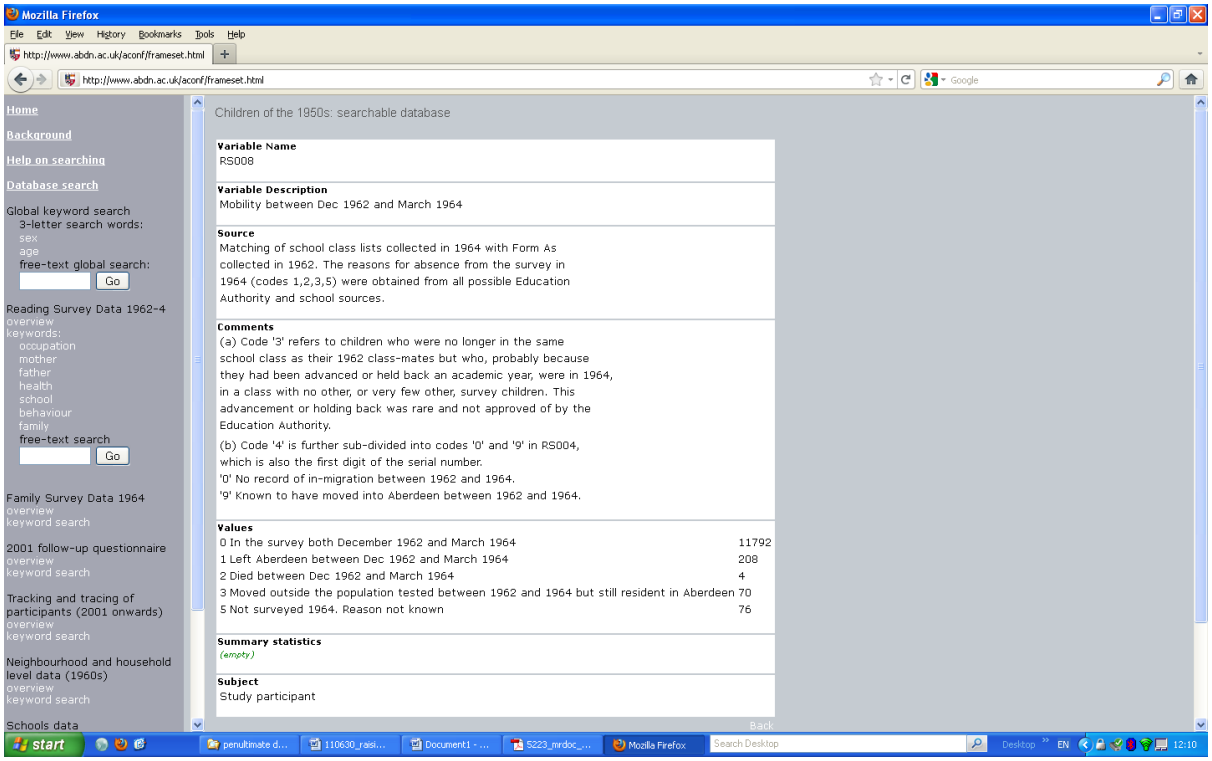
A school-led project making hundreds of whole genome tuberculosis sequences, in raw and summarised form, available to global partners and the research community. Web-based tools for data sharing, interrogation and analysis are under-development.

Screen shots from Aberdeen Children of the 1950s variable-level web-based metadata

1. Variable list



2. Details of individual variable including frequency distribution



Annex 6 – Glossary

Metadata: In the context of this report this term is used to refer to the data that supports the discovery, understanding and management of quantitative and qualitative scientific data. Further details on types of metadata are provided in Section 6.2.

Web-based metadata : Meta-data that is placed on the web, and as such is usually discoverable using search engines such as Google.

Institutional Repository: An online area for collecting, managing, disseminating and preserving in a digital form the intellectual output of an institution on a long term basis. A repository can hold a wide variety of materials including research articles, datasets and learning objects.

Depository: A term that has now become synonymous with repository, originally intended to look after material on a short term basis.

Data Discovery: A service that is provided by different systems and tools in order to locate sets of data that may have previously either been hidden or unavailable.

Data Archiving: The process of moving data that is no longer actively used to a separate data storage device for long-term retention. Data archives consist of older data that is still important and necessary for future reference, as well as data that must be retained for regulatory compliance.