



## **Paper 4 - Frequently Asked Questions on Data Sharing**

**Draft prepared by the NCRI Informatics Coordination Unit  
March 2004**

### **1. Why should I share my final research data?**

Data sharing achieves many important goals for the scientific community, such as

- reinforcing open scientific inquiry
- encouraging diversity of analysis and opinion,
- promoting new research, testing of new or alternative hypotheses and methods of analysis
- supporting studies on data collection methods and measurement
- facilitating education of new researchers
- enabling the exploration of topics not envisioned by the initial investigators
- permitting the creation of new datasets by combining data from multiple sources.
- promoting data quality
- expediting the furthering of scientific knowledge thus benefiting contributors, those who access the data and patients

### **2. Who benefits from data sharing?**

Everyone benefits, including investigators, funding agencies, the scientific community, and, most importantly, the public. Data sharing provides more effective use of NCRI member organisations resources by avoiding unnecessary duplication of data collection. It also conserves research funds to support more investigators. The initial investigator benefits, because as the data are used and published more broadly, the initial investigator's reputation grows.

### **3. Is data sharing widely accepted as a good practice?**

Yes, Data sharing is already either required or recommended by Funders in other disciplines, e.g. British Academy, Wellcome Trust's History of Medicine Programme, Economic and Social Research Council (ESRC). Further examples are:

- In the biological sciences, protein and DNA sequences are made available to researchers through data archives, such as GenBank
- In a bid to encourage and promote the sharing and preservation of research data, the Medical Research Council (MRC) is setting up an initiative initially focussing on population-based research and clinical trials
- The microarray community is embracing data sharing through the Microarray Gene Expression Data (MGED) society - an international organisation of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments. Furthermore, many scientific journals require that authors make available the data included in their publications (e.g. Nature)

- The NIH has recently included requirements for data sharing in grant applications (for further information please see <http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>)

**4. What do you mean by final research data?**

By "final research data", we mean recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Final research data do not include laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens.

**5. Does "final research data" include data that were not originally produced under an NCRI member organisations grant or contract?**

Sometimes. For example, if support is sought from an NCRI member organisation to transform or link existing datasets (as opposed to producing new data), the investigator should include a data-sharing strategy in the application.

**6. What do you mean by unique data?**

By "unique data" we mean data that cannot be readily replicated. Examples of studies producing unique data include: large surveys that are too expensive to replicate; studies of unique populations, such as centenarians; studies conducted at unique times, such as a natural disaster; studies of rare phenomena, such as rare metabolic diseases.

**7. What kinds of data are candidates for sharing?**

Potentially all kinds of data are candidates for sharing, but unique data are especially important. Some biological sciences already have data-sharing strategies in place, such as genetic mapping. But other basic science data are also amenable to sharing. Data from human subjects (e.g., surveys, clinical studies) also can be shared if the identity and privacy of research participants can be protected.

**8. Can you give me some examples of data that have been shared?**

Here are some examples of publicly available datasets

Illinois State Cancer Registry Public Datasets  
<http://www.chas.uchicago.edu/healthdata/illinois/idphcancer>

Age-specific incidence of all cancers: by gender and age 1997: Social Trends  
<http://www.statistics.gov.uk/statbase/ssdataset.asp?vlnk=5218&B3.x=48&B3.y=8>

Examples of shared datasets from the basic sciences include a growing number of genome sequences and maps, as well as protein and nucleotide databases (see ENTREZ <http://www.ncbi.nlm.nih.gov/Database/index.html> and other resources for molecular biology at the National Center for Biotechnology Information at <http://www.ncbi.nlm.nih.gov>)

**9. Data from my studies are generated from a very small number of rats, and I publish the final data. Am I expected to provide these data to other investigators as well?**

Publishing these final data constitutes an acceptable mechanism for sharing data.

**10. How soon after data collection am I obliged to share the final data?**

Recognising that the value of data often depends on their timeliness, data sharing should occur in a timely fashion. NCRI member organisations expect the timely release and sharing of data to be as soon as possible after publication of the main findings from the final dataset or earlier than this if appropriate. This time point will be influenced by the nature of the data collected. Data from small studies can be analysed and submitted for publication relatively quickly. If data from large epidemiologic or longitudinal studies are collected over several discrete time periods or waves, data should be released in waves as data become available or main findings from waves of the data are published.

NCRI member organisations recognise that the investigators who collected the data have a legitimate interest in benefiting from their investment of time and effort. NCRI member organisations continue to expect that the initial investigators may benefit from the first and continuing use, but not from prolonged exclusive use. NCRI member organisations also understand that an institution's desire to exercise its intellectual property rights may justify a need to delay disclosure of research findings. Any anticipated delay in the publication of data, for whatever reason, should be justified in the data sharing strategy for consideration by the funding organisation.

**11. Does data sharing pertain only to published data?**

No. Data-sharing strategies should encompass all data from funded research that can be shared without compromising individual subjects' rights and privacy, regardless of whether the data have been used in a publication. Furthermore, data sharing prior to the publication of major results is encouraged in many instances, for example, when data are collected to provide a resource for the scientific community (as in the case of many large surveys).

**12. Due to circumstances beyond my control (an earthquake!), I was unable to recontact a substantial portion of the sample in my longitudinal study. I was planning to put my data in an archive, but the resulting high rate of attrition makes the data minimally useful. Should I still archive the final dataset?**

Investigators need to find a balance between the value of the final data and the costs associated with archiving. If the data are of limited usefulness, then it is probably not worth the expense and effort of putting them in an archive. However, if the investigator has published results based on this dataset, then the dataset should be shared.

**13. I don't want to share my data, which were generated under a grant from an NCRI member organisation. Can I be forced to do so?**

Currently there are no mechanisms whereby a principal investigator can be forced to share their data. It is not the intention of the NCRI member organisations to force the scientific community to do something they oppose. Instead it is proposed that the community should embrace the principles of data sharing and be encouraged to share their data. It will be a decision for the individual funding organisations whether they will penalise applicants who do not wish to share data which would be of value to the scientific community.

**14. Will the data-sharing strategy affect the priority score of my application?**

This will be a decision for the individual funding organisations. In America the data sharing strategy does not currently affect the scoring of an application and it is proposed that a similar model is adopted by the NCRI member organisations.

**15. My research, which seeks support from both the public and private sectors, will involve proprietary data. How do I deal with the data-sharing issue in my application?**

The NCRI member organisations recognise that there may be circumstances where a co-funder has requested restrictions on data sharing as a condition of funding. These restrictions should be identified in the application and a proposal made about how data from the co-funded project will be shared. Should you believe that you are unable to share any of the data, your justification would be considered by the funding organisation programme staff who review the application.

**16. I'm a busy investigator. I don't have time to process requests for my data. What should I do?**

In addition to publishing small datasets, there are several alternatives to responding to each separate request to share data (e.g., putting data in an archive or restricted access facility, and setting up a web site for data access). Many archives and data enclaves provide technical assistance for users with questions or problems and may spare busy investigators time.

**17. Can I share data with colleagues under my own auspices?**

Yes. Your data-sharing strategies should indicate the criteria for deciding who can receive your data and whether or not you will place any conditions on their use. Data should be made as widely and freely available as possible while safeguarding the confidentiality of the data and privacy of participants. You should not place limits on the questions or methods others might pursue nor should you require co-authorship as a condition for receiving the data.

**18. Should the data source be cited or acknowledged in papers that rely on shared data?**

It is appropriate to acknowledge the source of data upon which a manuscript is based. Many investigators include this information in the methods and/or reference sections of their manuscripts. Journals generally include an acknowledgement section, in which the authors can recognise people who helped them gain access to the data. However, you should check the policies of the journal to which you plan to submit.

**19. Should I consider contributing my research data to a data archive?**

Maybe. Archives are organisations that collect and distribute data. They understand what is needed to prepare data for wider distribution and documentation for users. They provide stable, reliable, and cost-effective means for distributing data. They also provide protections for the dataset and technical assistance for requestors.

**20. Where can I find guidance on preparing data for sharing and archiving?**

The MRC is currently working to develop a Data Sharing resource for data from clinical trials and epidemiological studies. If you are applying for a grant in one of these areas you should speak to your MRC Programme Manager to discuss this.

Internationally, guidance is available from a variety of sources. For example, the Inter-University Consortium for Political and Social Research at the University of Michigan has prepared an excellent set of guidelines for preparing data for archiving. While these guidelines were written with social science data in mind, they are broadly applicable. See <http://www.icpsr.umich.edu/ACCESS/dpm.html>

For molecular biology information, the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the National Institutes

of Health, is ready to assist researchers who have genome-specific and molecular data to submit. For more information about submitting and accessing NCBI data, see the NCBI Website at <http://www.ncbi.nlm.nih.gov/Genbank/index.html>

**21. How do I pay for preparing data for sharing and archiving?**

The NCRI member organisations recognise that it takes time and money to prepare data for sharing and investigators will be able to request funds for data archiving and sharing as part of their grant application for collecting the data. If the data has already been collected, supplementary funds could be requested to support sharing data. The NCRI member organisations recommend that you consider procedures and costs for data sharing during the application process rather than after the data have been collected.

**22. Should I address data sharing in my NCRI member organisations application?**

Yes. NCRI member organisations have agreed that in the long term it will be desirable for all applications for funding to include a data sharing strategy. Initially, NCRI member organisations will be introducing this requirement in a flexible manner over the coming academic sessions. You should check with the individual funding organisation regarding their particular requirements at this time.

**23. What do I need to include in my application and where do I put the information about data sharing?**

It is proposed that the data-sharing strategy, or the justification for the absence of such a strategy, should follow after the description of the research to be undertaken. It is proposed that the data-sharing strategy would not count toward the application page limit. Additional information on data sharing might be included in other sections of the application, as appropriate. For example, if you are producing a large dataset that will become an important resource for the scientific community, you probably want to mention this in the significance section. If you are requesting funds to prepare, document, and archive the data, you would want to include relevant information in the budget and budget justification sections.

**24. The informed consent form for my recently completed study states explicitly that only my research team will see the data provided and that we will not share the data. Am I now expected to share it?**

No, but if you plan to collect additional data from those subjects under a grant with a data-sharing strategy, you should revise the consent procedure to be consistent with the data-sharing strategy. In preparing and submitting a data-sharing strategy during the application process, investigators should avoid developing or relying on consent processes that promise research participants not to share data with other researchers. Such promises should not be made routinely or without adequate justification described in the data-sharing strategy.

**25. How can I protect the privacy of my subjects?**

It is the responsibility of the investigators, and their institution to protect the rights of participants and the confidentiality of their data. Data should be redacted to strip all individual identifiers, and effective strategies should be adopted to minimise risk of disclosing a participant's identity. Options to protect privacy include: withholding part of the data, statistically altering the data in ways that will not compromise secondary analyses, requiring researchers who seek data to commit to protect privacy and confidentiality, and providing data access in a controlled site, sometimes referred to as a data enclave. Some investigators use hybrid methods, releasing a redacted dataset for general use but providing access to more sensitive data through a user contract or data enclave. In most instances, sharing data is possible without compromising participant confidentiality and privacy.

**26. I collect data of a sensitive medical nature, is this data is too sensitive to be shared?**

Not necessarily. Sensitive data can be shared so long as appropriate privacy safeguards are in place. Investigators must determine if and how the rights and privacy of the subjects can be protected.

**27. Can data from a clinical trial be shared?**

It depends. Participants' privacy must be protected in accord with all applicable laws and regulations. Clinical trial datasets are frequently rich in items that could potentially identify individual subjects. For example, many early phase trials use small samples, which make it difficult to protect the privacy of the participants. Researchers who are planning clinical trials and intend to share the resulting data should think carefully about the study design, the informed consent documents, and the structure of the resulting data prior to the initiation of the study.

Clinical trial data can be validated by aggregation with other data sets (see commentary by Brenton and Caldas ([http://dx.doi.org/10.1016/S0140-6736\(03\)14053-6](http://dx.doi.org/10.1016/S0140-6736(03)14053-6))). There are many precedents for sharing of clinical trial data. For example, data from a number of clinical trials supported by the National Heart, Lung, and Blood Institute (NHLBI) are available for research use (See <http://www.nhlbi.nih.gov/resources/deca/directry.htm>). The National Institute of Allergy and Infectious Diseases (NIAID) also lists their clinical trials datasets that they have made available through the National Technical Information Service (NTIS) for public use (See <http://www.niaid.nih.gov/research/aidsdata.htm>).

**28. Is data on DNA and protein sequences archived?**

Yes. For example, GenBank (<http://www.ncbi.nih.gov/Genbank/>) and Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>) archive gene sequencing data. The sharing of materials, data, and software in a timely manner has been an essential element in the rapid progress that has been made in the genetic analysis of mammalian genomes.

**29. I am working on a select pathogen and cannot share the data for reasons of national security. Is this an acceptable reason for not sharing?**

Yes.

**30. If I am required to submit a revised data-sharing strategy, what do I need to do?**

As is the proposed case with Principal Investigators who submit any additional or revised application material, your revised data-sharing strategy must be signed by all bodies who have ownership or IPR associated with that data (e.g. your institutional official)

**31. I want to request a dataset from a recent publication. How do I do this?**

You should check the publication to see if reference is made to an archive, an enclave, or a Website where the data might be available. If no such information is provided, you may wish to send a letter to the Principal Investigator to see if the data are available for sharing, and where you might be able to get the data and associated documentation.