

**ANDS responses to
the data management challenges in the
*Australian Code for the Responsible Conduct of Research***

David Groenewegen
Deputy Director
Australian National Data Service
david.groenewegen@ands.org.au

ABSTRACT

In early 2009, selected senior staff at a number of research institutions were invited by the Australian National Data Service (ANDS) to attend Forums to discuss the issues surrounding the Australian Code for the Responsible Conduct of Research and to share their strategies around compliance. These issues included the governance arrangements, the technological and policy requirements, and the best way to reach research staff. This paper discusses the Code and its relevance to data management, the issues identified by ANDS and the strategies recommended.

Introduction

The *Australian Code for the Responsible Conduct of Research* guides institutions and researchers in responsible research practices and promotes integrity in research for researchers. Developed jointly by the National Health and Medical Research Council (NHMRC), the Australian Research Council (ARC) and Universities Australia, the *Code* has broad relevance across all research disciplines. Released in October 2007, it replaced the *Joint NHMRC/AVCC Statement and Guidelines on Research Practice (1997)*.

While both the current *Code* and its predecessor mention the need for effective data retention policies, the massive growth in the production of digital data in the last decade has posed new management, compliance and policy challenges for institutions. In the past, it was expected that data retention could be managed at a departmental level, but this proved increasingly difficult. It is now recognised in many quarters that institution wide responses are needed. This paper looks at what some of those responses might be, and advocates that particular sections of the organisation need to work together for the effective provision of services.

The Australian National Data Service (ANDS) was founded in 2008 to influence national policy in the area of data management in the Australian research community, to inform best practice for the storage and management of data and to transform the disparate collections of research data around Australia into a cohesive collection of research resources. As such, we identified a role in assisting institutions to meet their obligations to the *Code*.

About ANDS

ANDS is an initiative of the Australian Federal Government being conducted as part of the National Collaborative Research Infrastructure Strategy (NCRIS) and the Education Investment Fund (EIF). It is a collaboration between Monash University, the Australian National University (ANU) and Commonwealth Scientific and Industrial Research Organisation (CSIRO), with office locations at Monash and the ANU.

In addition to the areas listed in the previous section ANDS has been charged with the role of the establishment of the Australian Research Data Commons to support the discovery of, and access to, research data held in Australian universities, publicly funded research agencies and government organisations for the benefit of research. This investment will enable the construction of a range of Information and Computing Technologies (ICT) utilities to capitalise on and ensure greater use and re-use of data resources.

The ANDS constituency is made up of organisations such as universities, government agencies, research institutions, galleries, libraries, archives and museums, as well as people such as researchers and their support staff, repository and data centre managers and technical staff, research administrators, government policy-makers and funders.

ANDS and the Code

ANDS has no formal responsibility for the *Code*, which was developed before ANDS was founded. However, many of the aims of the *Code* coincide with the aims of ANDS. Specifically, the *Code* makes significant references to research data and responsibilities related to it, specifically in Section 2: "Management of research data and primary materials". This section describes the responsibilities of institutions and researchers in the management of research data and primary materials. These are that the institutions are to retain research data, provide secure data storage, identify ownership, and ensure security and confidentiality of research data. The researchers are to retain research data and primary materials, manage storage of research data and primary materials, and maintain confidentiality of research data and primary materials. (NHMRC, 2007, 2.1) The *Code* also encourages the idea that not only the results of the research (such as journal articles and conference papers), but the research data itself should be made available for re-use by others (NHMRC, 2007, 2.5.2).

At present this section of the *Code* is not being enforced by the funding bodies that developed them; however, all Australian universities are signatories to the *Code*, and have policies in place to encourage *Code* compliance in many areas (see <http://www.ands.org.au/resource/code.html> for examples). It is worth noting that the *Code* is not exclusively concerned with data management, and that ANDS did not attempt to address these other issues in its work.

As ANDS has an interest in encouraging good data management practices and data re-use, the service is interested in promoting awareness and supporting the uptake of the *Code* and this has been an ongoing area of activity. In addition to preparation of support materials around the data management aspects of the *Code*, ANDS has been conducting outreach efforts to better understand the issues and inform the research sector. Initially these took the form of forums with invited audiences on institutional responses to the data aspects of the *Code*, which were held in Melbourne and Sydney in the first half of 2009. These were followed by a series of roadshows, beginning in September 2009, and continuing across the major Australian cities and research hubs into 2010. This paper is informed by the discussions that have arisen from these programs.

The *Code* describes distinct responsibilities for both institutions and researchers regarding their research practices and behaviours. While ANDS has been providing advice and support in both these areas, this paper will focus on the responsibilities of institutions, while mentioning some areas of interest to researchers.

Institutions and the Code

In the area of data management, the *Code* requires a number of areas of responsibility of the institution to which the researcher belongs. These are that they:

- Retain research data and primary materials
- Provide secure research data storage and record-keeping facilities
- Identify ownership of research data and primary materials
- Ensure security and confidentiality of research data and primary materials

In order to achieve this the institution needs to provide the policies, infrastructure and facilities to ensure that the data is “owned” (i.e. it is clear who is responsible for it), retained, stored and made retrievable and accessible. If the institution does not do this, and/or does not provide sufficient support, it is highly unlikely that the majority of the researchers for whom the institution is responsible will be able to meet their obligations under the *Code*.

Key considerations

In order to provide this service and support the institution needs to understand that the work involved in complying with the *Code* is as much about cultural change as research practice. This will be a long term process, and needs effective buy-in from across the institution, including senior administrators. It is crucial that there is high level support for the work, including the identification of a senior “Champion”. This needs to be coupled with a clear steering process to coordinate work across the institution, as it is likely that various support services and resources will need to be brought into play. These will need to be identified, recruited and have clear directions set for them. Where they come from, and how much resource can be provided will vary between institutions.

One key piece of advice that ANDS has gained from its outreach activities is the necessity of effective collaboration between the various research support arms of the institution, especially the need for areas that can be broadly defined as the library, IT support and research offices to work closely together, with the policy and administrative sections providing a framework. These bodies bring together the skills, experience, knowledge and institutional breadth required to provide an efficient and wide-ranging service. Without the provision of all these skills, the potential for effectively managing data across the institution is much reduced. We are aware of this occurring at universities such as Monash and Newcastle already, due to the efforts of senior staff seeking out opportunities to work with their colleagues in other divisions of the university.

Getting started

There are a number of steps that need to be undertaken by an institution when looking at how to respond to the *Code*. These might not apply or be feasible in all cases, but they should at least be considered as part of the process.

1. Undertake an assessment of institutional commitment and priorities – where will the drive come from within the institution, and which areas might be less involved?
2. Review of all existing relevant policies and responsibilities, especially information management policies, practices, planning and responsibilities, with a view to using these as a basis for encouraging *Code* compliance.
3. Review of data storage capacity, including possible use of cloud services, to understand what other IT infrastructure might be required.
4. Assessment of capability of individual researchers to respond to the *Code*, through, for example, a survey of data management practices.
5. Audit of data currently held within the institution to understand the scale of the issue, and inform other areas such as IT.

6. Give due consideration to the financial implications, which impact across the institution.

Going through this process will help to understand what can be done, and by whom. As discussed above, institutional support areas such as the library/archives/records, IT services and the research office all have important roles to play.

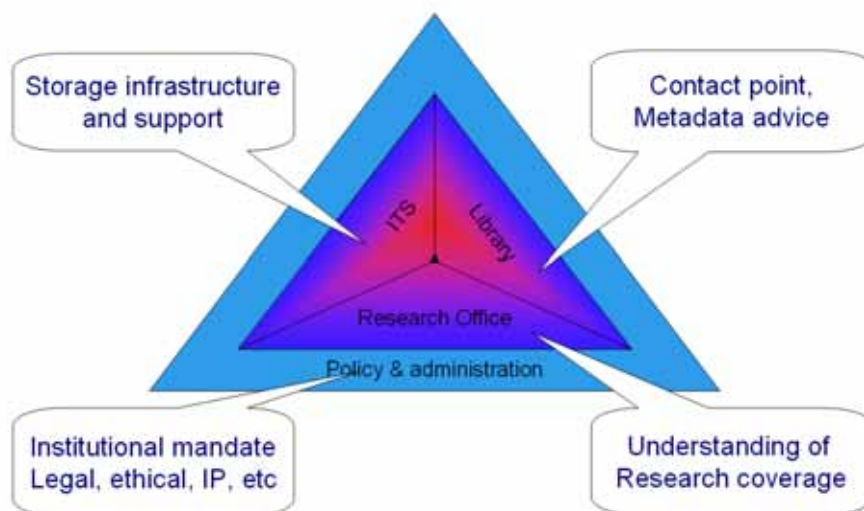


Figure 1: Institutional Support Services

Roles of institutional support areas

As seen in Figure 1, the areas should all intersect smoothly, to provide a consistent level of service to users. The nature of the roles will vary from institution to institution, based on available resources and existing practices. Similarly, the leadership of integration and coordination activities may also vary depending on the environment. The following are suggested roles, and the overlaps between roles are designed to anticipate that different organisations may assign work differently; for instance training might be provided by the library, or the research office.

Potential roles for library and/or archive and record sections

Librarians, record managers and archivists have many useful potential roles to play in assisting an institution to better provide useful data management services and support to its researchers. One key role is their many years of experience in the management of “traditional” forms of data – books, papers, records and so forth. Many of the skills learned from this (metadata, descriptions, disposal policy and sustainability for instance), can be cleanly transferred over to the data management world, and will provide essential guidance for many.

Similarly, librarians generally have extensive contacts with researchers within the institution, and extensive existing outreach initiatives. This network can be used to aid in information gathering about available data, providing advice on data management issues and options available within the institution and finding out about researcher concerns, and then providing training and support. Some universities (such as Monash (Macmillan & Clarke, 2009), or Purdue (Garritano & Carlson,

2009)) have already put significant programs in place to try to leverage this. The Monash program, for instance, has involved the creation of a cross faculty team of librarians who are seeking to advise researchers on data management policy and practice.

In Australia, many university libraries are the primary owners and managers of their institutional repository. While these were established with publications such as journal articles and conference papers in mind, they could be used for storage of research data, or of metadata about that data, with connections to the storage location. The management experience from running these repositories and encouraging access to them will also be of value in the area of broader data management and this will aid in *Code* compliance. To do this librarians need to proactively approach their existing contacts, both in the research areas they support, and in other “support” divisions of their institution. They also need to make themselves aware of the issues around data management, so that they are seen as the experts.

Potential roles for ITS or ICT departments

In an environment in which digital data is increasingly the primary source of research material, the IT department will naturally need to play a key role in the provision of effective services to researchers at their institution, specifically of storage hardware, access control, data protection and back-up services. On top of this will come an increased role as provider of software tools for data capture, metadata capture and discovery. In Australia, connections to the ARCS Data Fabric will need to be negotiated and maintained.

However, compliance with the *Code* offers the potential for an expanded role as a provider of support for the use of this storage and the other services being provided. Increasingly it is not sufficient for the hardware solutions to be in place. They must be developed in such a way as to be made easier to use, to the point where use becomes second nature, rather than a chore.

An interesting analogy might be made in the take-up of email. When first widely introduced in universities in the mid-1990s it was clearly a tool designed for computer experts by computer experts. Line editing, arcane commands and limited reliability substantially impinged on the usability of the service, and many researchers refused to use it. In the intervening years, the usability of the service and its reliability has increased to the point that research without it seems inconceivable. This is the direction in which IT data management support needs to be headed.

Potential role of Research Office

The Research Office (or some similar body within an institution that coordinates research funding, grant applications and reporting of research outputs as required) offers an important set of knowledge and skills to the process of helping an institution comply with the *Code*. First among these is identifying significant researchers and data producers, so that they can be approached and aided in their data management needs. Bringing the most important researchers in an institution into good data management practice can be expected to help attract the cooperation of others.

The Research Office also has an important leadership and advocacy role to play in the promotion of the requirements around the *Code* to help researchers understand how they need to meet it. This is especially true in situations where the funding agency may wish to require that data needs to be made available at the conclusion of the funding period. As such, the Research Office may also need to be involved in research practice training.

Policy and administration

Institutions already have policies and procedures covering research, records management, intellectual property and other research-related activities. These can be adapted for *Code* compliance purposes; for instance:

- Intellectual property, which covers areas such as copyright, moral rights or patents. The institution needs to make clear to the researcher what rights are held where: can the researcher take data with them when they leave, for instance, or in a collaboration, how are rights able to be shared between the various parties? These issues need to be made clear at the beginning of the research project, to avoid potential disputes at the end.
- Provision of data management, including:
 - Storage, which will require policies and procedures on how to provide appropriate storage, how much storage will be provided and under what conditions. There needs to be a clear and easy process to allow access to storage, or researchers will not make use of it.
 - Retention – while many institutions have policies around the retention of corporate records and data, the application of these to research is less well understood. There are many factors that impact on how long data needs to be stored, such as ethical and privacy concerns, legal requirements or the desires of funding bodies. The procedures and knowledge used for corporate data can and should be applied in the research area.
 - Disposal will become critical as research data growth outstrips storage (Gantz, 2008, p.2). Again, the skills and experience of the corporate data and archiving sections of the institution may be useful here in setting guidelines, and establishing how this is to be recorded and managed.
 - Access control is extremely important, as doing it well will create trust. In addition there are a number of other decisions that need to be made, such as how to make data available to those engaged in the research project; how data is to be “published” and made available more widely; how the institution is to keep a record of its data assets; and the institutional position on open access of both data and publications.
- Existing conflict of interest policies may need to be examined. While the *Code* recommends that data be made available, a researcher may find problems in meeting the requirement of the institution or the funder.
- Collaboration and contractual agreements are extremely important in this area, as many researchers will need to work with their peers outside the institution. So they will need help in designing and understanding agreements that include provision for data sharing after the research is complete and

agreement on who will host and store the data. This latter area is potentially very complicated, as it will need to cover who will host the data, whether there will be copies kept at each institution involved, and who is able to use the data at the end of the collaborative process.

- Ethics and privacy are another complicated areas. The *Code* makes allowances for the different data retention needs required for this. However, the institution may find that internal ethics processes make recommendations that are in conflict with the *Code* or the needs of the institution, such as the length of data storage and the ability to share.
- Last, but not least, is the issue of managing compliance. If the institution is going to comply with the *Code*, it needs to be able to demonstrate that it has done so. Therefore there need to be measures in place to track this, which ideally should not be so onerous that they deter researchers from doing any compliance at all.

Each of these areas needs to be considered in terms of both institutional and researcher responsibilities. Not all are specific to data management (and most cross over all areas of research, as does the *Code* itself), but ideally each should have a data management component and procedures behind them for translating policy into practice.

The Strategic Response to the *Code*

Once an appropriate policy framework is in place, there will be a need to develop strategies to cover:

- Publicising the policies within the community – without awareness there can be no compliance.
- Providing appropriate infrastructure to support researcher requirements and training in data management and use of infrastructure, to simplify the process for researchers.
- Establishing support services related to the infrastructure, so that assistance is available as required.
- Establishing record-keeping procedures to track compliance.
- Demonstrating compliance through review or audit frameworks.

Costs of *Code* compliance

It is undeniable that there are very real costs related to the effective implementation of policies and practices to comply with the *Code*. The process of policy review and amendment outlined above can be time consuming and costly, as data management runs across many areas of most research institutions. Development of a separate data management policy and procedures related to this offers a more cost effective method in some ways, but may have issues in being integrated into the broader research practice.

Similarly, legal review and the changes that arise may be costly, but the potential for problems if this area is not addressed seem enormous. Clearly establishing

ownership of shared data will need to be integrated in the normal contractual processes (especially when institutions are equal partners in a process).

The model proposed above, where various support services come together to direct and facilitate the processes required, can be expected to create more work for existing services, and this will need to be balanced against the other responsibilities of these areas. Some priorities will need to be reassessed and changed, and there may also be new training needed for the staff in these areas, to give them the experience and confidence to advise and support others in what is a daunting area for some.

One of the most significant initial costs will be the necessary IT infrastructure required for storage, as well as tools and software that will enable effective use of the data and the level of security and access control needed by researchers. Establishment of trust in these systems, and clear processes for their use will be critical for success.

As noted above, there will also need to be the creation of a compliance monitoring system that can overcome the potential view of *Code* compliance and the associated data management as another burden for time-poor researchers. To avoid this it is critical that the benefits of the process are made clear.

Benefits of *Code* compliance

While there are costs in complying with the *Code*, these should be offset with an understanding of the benefits. For instance, security of storage for data offers a value to the researchers that cannot be underestimated. It is likely that much valuable data is regularly being lost, and is either irreplaceable, or must be replicated at some expense.

Coupled with this storage infrastructure is the enhanced ability to share data not only with other researchers, but also with the original author or creator at some future date. By retaining and storing research data effectively and efficiently, the researcher will know that they can use it again in the long run.

For the institution, the benefits include a better understanding of what their researchers are doing, and where they might be able to best allocate resources to assist key researchers. It will also establish a corpus of past research for the institution to build upon, rather than lose material as researchers leave. This assumes that the ownership of the policy has been established in advance of it being deposited.

Finally, there is an advantage in promoting readiness if the *Code* does become mandatory, and data sharing or management becomes a condition of funding. It is worth noting that, while the Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC) do not currently require this, some other overseas funding bodies do require it (Wellcome Trust 2007), and that this may well become more prevalent in years to come.

What do researchers get?

In addition to the benefits of more secure data as outlined above, there are less tangible but equally important gains for the researcher, who is able to share and manage their data well. One of these is greater recognition of their work through increased citations and follow-on research. In a 2007 study, sharing detailed research data is associated with an increased citation rate: 48% of 85 cancer microarray clinical trial publications with publicly available microarray data received 85% of the aggregate citations (Piwowar 2007). In addition, citations remain an important indicator of the perceived value of research, as they demonstrate the value of the work by allowing continued re-use of the data. This may in turn lead to further funding.

The additional attention may also lead to other gains such as new collaborations with users of the available data or those who follow the citations to the original data. This may then lead to new research and funding opportunities, as well as a sense of satisfaction as a result of benefiting the broader research community.

Summary

There are a number of critical findings that ANDS has established from its involvement in discussions around the *Code* and compliance issues related to it. Firstly, if it is not easy and clear to manage data, then researchers will not or cannot do it. This may cause them to be unable to meet their own obligations under the *Code*. To make it easy to do effective data management there needs to be collaboration across the organisation, to provide the services and support needed. Without this organisation it is unlikely that the effective provision of services can occur.

While this may mean that there are substantial costs involved in *Code* compliance, there are also substantial benefits for both institution and researcher. This process is about change and it will be gradual, but the end result should be worthwhile.

Acknowledgement

The author would like to acknowledge the work of Ian Barnes, Margaret Henty and Frances Watson. This work was vital to this paper.

Reference list

Australian National Data Service 2009, *Responding to the Code*, <<http://www.ands.org.au/resource/code.html>>, accessed September 15, 2009

Gantz et al, 2008, *The Diverse and Exploding Digital Universe*, IDC White Paper, <<http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>>

Garritano & Carlson, 2009, A Subject Librarian's Guide to Collaborating on e-Science Projects, *Issues in Science and Technology Librarianship*, Spring 2009, <<http://www.istl.org/09-spring/refereed2.html>>

Macmillan, W & Clarke, N 2009, To Dare or Not to Dare: Collaboration in eResearch at Monash University, EDUCAUSE Australasia, 3-6 May, 2009

National Health and Medical Research Council, Australian Research Council, Universities Australia 2007, *Australian Code for the Responsible Conduct of Research*, National Health and Medical Research Council, Australian Research Council, Universities Australia, Canberra.

NHMRC - see National Health and Medical Research Council

Piwowar HA, Day RS, Fridsma DB 2007, Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.0000308

Wellcome Trust, 2007 *Policy on data management and sharing*, January 2007, <<http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm>> accessed September 15, 2009