The International Journal of Digital Curation
Volume 8, Issue 1 | 2013

Data Management and Preservation Planning for Big Science

Juan Bicarregui,
Scientific Computing Department, STFC

Norman Gray,
School of Physics and Astronomy
University of Glasgow

Rob Henderson and Roger Jones,
Department of Physics,
University of Lancaster

Simon Lambert and Brian Matthews,
Scientific Computing Department, STFC

Abstract

'Big Science' - that is, science which involves large collaborations with dedicated facilities, and involving large data volumes and multinational investments – is often seen as different when it comes to data management and preservation planning. Big Science handles its data differently from other disciplines and has data management problems that are qualitatively different from other disciplines. In part, these differences arise from the quantities of data involved, but possibly more importantly from the cultural, organisational and technical distinctiveness of these academic cultures. Consequently, the data management systems are typically and rationally bespoke, but this means that the planning for data management and preservation (DMP) must also be bespoke.

These differences are such that 'just read and implement the OAIS specification' is reasonable Data Management and Preservation (DMP) advice, but this bald prescription can and should be usefully supported by a methodological 'toolkit', including overviews, case-studies and costing models to provide guidance on developing best practice in DMP policy and infrastructure for these projects, as well as considering OAIS validation, audit and cost modelling.

In this paper, we build on previous work with the LIGO collaboration to consider the role of DMP planning within these big science scenarios, and discuss how to apply current best practice. We discuss the result of the MaRDI-Gross project (Managing Research Data Infrastructures – Big Science), which has been developing a toolkit to provide guidelines on the application of best practice in DMP planning within big science projects. This is targeted primarily at projects' engineering managers, but intending also to help funders collaborate on DMP plans which satisfy the requirements imposed on them.

International Journal of Digital Curation (2013), *8*(1), 29–41. http://dx.doi.org/10.2218/ijdc.v8i1.247

 doi:10.2218/ijdc.v8i1.247

Introduction: The Big-Science Paradigm

There appears to be a rough consensus that many of the central concerns regarding data management and preservation (DMP) for the majority of academic disciplines relate to the management of a large collection of disparate data generated in a relatively undisciplined manner by a wide variety of independent researchers, who are mainly concerned with their own research. Thus the concerns are about the usability of repositories, the challenges of persuading researchers to deposit their data, and how best to manage the citation of data, with additional concerns about how researchers may best receive credit for the data they have collected.

However, implicit within this view appears to be a rather simple conceptual model of what it is that researcher-users actually do to create the data. Researchers: (i) obtain grants, which (ii) they use to generate data within their own research methods which (iii) they manage locally and then (iv) share either as datasets or linked to publications. Much of the interest in research information systems within this area presumes a rather simple relationship between (i), (ii) and (iv), and much of the DMP effort appears to be concerned with persuading researchers to do step (iii) better, possibly with suitable institutional assistance, cajoling or prescription.

This model is less appropriate for large-scale projects in the physical sciences, which have decades of experience with data management and sharing, at scale, incorporating a data management workflow that is different from this paradigmatic one under each of its four headings. This incompleteness suggests firstly that the DMP solutions created under this paradigm, when applied to other disciplines, may not be as generally applicable as expected; and secondly that there are data management problems outside those automatically considered by that paradigm, which are nonetheless well-understood, and for which practical solutions already exist.

For our purposes, such ‘Big Science’ projects tend to share many features which distinguish them from other research disciplines. These include the following:

1. The projects are large collaborations, involving hundreds or thousands of researchers from many institutions, typically in different countries.
2. The projects last many years, with extended planning and set up phases, and long lifetimes of experimental running, data collection and analysis.
3. The projects are funded with long term budgets, and typically from multiple sources, thus requiring complex legal agreements on resource provision and ownership.
4. The projects typically establish dedicated experimental facilities, with their own structures and dedicated technical staff, including computing support.
5. The projects typically generate large volumes of complicated and instrument-specific data (1–10PB per year, with exabyte-per-year rates anticipated in the next decade).

The key feature, from the point of view of this paper, is that this is facilities science. There is a core facility, with multinational funders, a multi-decade existence, and a conceptual and administrative separation between the elaborately-engineered resource and the research scientists.

Particle physics has the longest experience with this model of doing science, most famously in the Large Hadron Collider (LHC) collaboration centred at CERN[1], but gravitational wave physics (e.g. LIGO[2]) and radio astronomy (e.g. SKA[3]) have or will have similar or larger collaboration sizes and data volumes. Other areas of astronomy have long experience with internationally shared telescope facilities, though working at a different scale. Structural sciences (i.e. studies into the micro- and nano-scale structure of matter) are moving towards this model of working, where large-scale facilities, such as neutron and synchrotron sources, support many individual scientists working within the traditional DMP paradigm. However, the facility itself – with its continuity of funding over a long period, dedicated infrastructure and specialist staff – has the characteristics of "big science" and can leverage those characteristics to provide more systematic data management for its user community (Flannery et al., 2009).

Preservation policy and practice in big science deals with large volumes of data in large (100s to 1000s) collaborations, with technically sophisticated users and computing support. The data volume is the least significant feature in the present context, since it is 'only' a technical problem; the other two features change the game.

This scale of working produces some simplifications:

- It is well resourced: DMP is not the responsibility of quarter-time junior researchers, but a key concern of the project's engineering management.
- There is a collaborative ethos, which has data sharing at its core. Data, once acquired, goes directly into the archive and is retrieved from there for processing by researchers.

However the scale also produces a variety of complications:

- There are multiple funders in multiple countries with various, sometimes conflicting, requirements relating to data management and dissemination.
- The multiplicity of funders often means that no one can dictate terms.
- Experiments and their datasets are governed by networks of Memoranda of Understanding and Service Level Agreements and in-collaboration decision-making processes which, however intricate the process, are fundamentally consensus-based.
- The intellectual property of the data is often complex.

Thus the nature of big science determines that it cannot benefit from the considerable effort going into providing technical and software support for DMP

[1] CERN - the European Organization for Nuclear Research: www.cern.ch

[2] Laser Interferometer Gravitational Wave Observatory: www.ligo.caltech.edu

[3] The Square Kilometre Array: http://www.skatelescope.org/

planning. It is in this context that the advice of a recent JISC-funded study of data at this scale to "just read and implement OAIS" (Gray et al., 2012) is more practical than it appears. Facilities-scale science projects have the financial and engineering resources, and technical expertise to produce bespoke DMP plans and data management systems. However, what must be avoided is pointless reinvention, and so there is an outstanding need for a fast-track to an optimal solution. This is where funder support can be helpful in supporting the relevant technical personnel by connecting them to high-level DMP best practice.

The MaRDI-Gross project[4] is building on previous work by developing practical advice for large-scale DMP planning. It is based on the insights of the OAIS reference model, and includes discussions on cost modelling, with a target audience of big science practitioners and funders. The emphasis has largely been on the UK community associated with the Science and Technology Facilities Council[5] (STFC), the major funder of big science in the UK. The guidelines apply more widely, as much of the work of STFC is in collaboration with similar bodies in other countries and with cross-national institutions, and thus the guidelines can also apply in those cases. The goal of the project is to bring big science practitioners up to speed with the current best practice, and to equip funders with the means to critically engage with DMP planners, giving both groups a rapid boost towards relevant disciplinary best practice. In the rest of this paper we consider the factors considered in these guidelines, and discuss the implications of relevance for the wider DMP community. The full guidelines can be found in Bicarregui et al. (2012).

Policy Drivers

We first consider the various high-level policy drivers for DMP planning. Big science projects are high-profile, and need to take special account of the high level interests concerning the longer term goals of their discipline and of society at large.

Policy Requirements of Funders

Policy drivers within big science are largely defined by governmental and inter-governmental policy frameworks, which determine the agreements within which international big science is undertaken. Thus in 2011, Research Councils UK (RCUK) published a set of 'Common Principles on Data Policy'(RCUK, 2011). The RCUK principles are informed by the earlier OECD 'Principles and Guidelines for Access to Research Data from Public Funding' (OECD, 2007) and in turn inform the discipline-specific policies of the UK research councils. In particular, the STFC's policies are tailored to the needs of big science projects (STFC, 2011). These include specifying that big science projects and facilities should have a DMP plan under the control of their management and science communities; noting that these policies and restrictions are typically subject to international agreement and to legislation in different countries; and recognising that resources need to be released for data management, as appropriate for the long term needs of its designated community.

[4] The MaRDI-Gross project: http://mardigross.jiscinvolve.org/wp/

[5] Science and Technology Facilities Council (STFC): www.stfc.ac.uk

Open Data

Further policy considerations include the approach to open data. STFC's data sharing principles endorse the international push towards such data sharing in the more general context of scholarly research. In the US, the NSF's GC-1 document states in Section 41 that:

> "[NSF] expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work" (NSF, 2010).

Thus funders are setting policy so that scientific data should generally be universally available, partly because it is usually publicly paid for, but also because the public display of evidence is a necessary part of science. Of course, in practice it is not as simple, and a host of technical, political, social and personal issues complicate the social, evidential and moral arguments for general data release.

The arguments against general data releases are practical ones: data releases are not free, and may have significant costs. Many of these costs come from data preservation, since it is archived data products that are the most naturally releasable objects: releasing raw or low-level data may be cheap, but may also have little value, requiring specialist knowledge. Such data releases may even have a negative value, if they foster misunderstandings which are time-consuming to counter. In consequence, the 'open data question' overlaps with the question of data preservation; if the costs of data preservation are satisfactorily handled, then a significant subset of the practical problems with open data release will disappear.

The above arguments focus on the costs to the data owners, and the benefits to society, of data release. The costs are reasonably objective, and while it might be difficult to estimate them to much better than an order of magnitude, they can be grounded in money. In contrast, the latter balancing benefits are often diffuse, and involve educational and outreach benefits which are real, but which can only be translated into numbers through a formal cost-benefit analysis, as used in environmental economics, for example. We believe this is an interesting topic, which it would be instructive to explore further.

Some questions of data sharing can be discussed using the OAIS notion of the Designated Community. Higher level data products contain less detail than lower level or raw datasets; they are also intended to serve broader communities, and are more expensive to generate in terms of processing. When a scientist chooses between a project's data products, the choice represents a trade-off involving the amount of time they can invest in understanding the data product, the degree of support they receive from colleagues and the data owners, and the subtlety of the question they wish to address (more subtle distinctions might be erased by higher level products, but might be spuriously detected in poorly-understood raw data). On the other side of the exchange, a project will have a formal or informal model of whom it is serving by the provision of data, and will design data products, and allocate costs, accordingly.

 doi:10.2218/ijdc.v8i1.247

Data Preservation Objectives

A crucial question for DMP policy is: "what precisely are the preservation goals?" One should not assume that everything should be preserved, indefinitely, because this would be far too expensive. Data preservation objectives are influenced by the needs of the discipline. As an observational science, astronomy data is generally repeatable, but some of the most precious astronomical data records unpredictable, transient events or long-timescale secular changes. Astronomical data is potentially useful almost indefinitely and, because its object of study is in some sense fundamentally simple (there is only one sky), it is also broadly intelligible almost indefinitely, so it is reasonable for its target preservation time to be greater.

High Energy Physics (HEP) data is somewhat different, as the discipline's community is generally very much in control of what it observes through the successive generations of experiments. A consequence of this is that, firstly, HEP experiments tend to become obsolete with each technological generation, and secondly the complexity of the apparatus makes it hard to communicate into the future the understanding sufficient to reuse the data. Experimental apparatus is generally understood better as it is used, so data gathered early in an experiment will be periodically reanalysed with increased accuracy. However, this understanding is generally not formally recorded, but is communicated through wikis, workshops and other informal means. Even if all records were preserved with complete fidelity, an archive would still be missing the word-of-mouth information which a new postgraduate (for example) needs to acquire. In OAIS terms, the Representation Network for HEP data is particularly intricate, goes beyond format description (the most commonly considered form of representation information in many preservation scenarios) to include semantic descriptions, software and tacit knowledge, and it may be infeasible to gather all the Representation Information necessary to let a naive researcher make sense of it. Nevertheless, South (2011) describes scenarios in which HEP data should be reanalysed some decades after an experiment has finished, and describes ongoing work to preserve data for long enough to enable post-experiment exploitation. These scenarios have in common a commitment of a few full time equivalent (FTE)s of staff to actively conserve and exploit the data, acting as living "Representation Information."

Amongst open data advocates there is often an automatic expectation that 'everything should be preserved' so that an experiment can be redone, results reanalysed, or an analysis repeated later. Is this actually true? Or if it is desirable, how much effort should be expended, in the face of the inevitable costs? In fact, it is not always the case that an experiment can be redone, because it is not always feasible to document it in enough detail that the measurements can be remade. For similar reasons, if the data analysis is particularly complicated, or requires a subtle understanding of the behaviour of a particular instrument, it may not be feasible to document that analysis in enough detail. There is therefore a case that at least some details of the experimental environment – digital as well as physical – are not reasonably preservable, and that as a result little effort should be expended on preserving them, if well-documented higher level data products are available and intelligible. We should stress that we are not advocating deliberately deleting raw data – it might be useful, and it might be usable – but simply noting that one should not

overstate its value. Archivists are familiar with such choices in their retention and disposal policies, and these choices are particularly acute for big science.

Technical Frameworks

There are key technologies that someone producing a big science project DMP plan should be aware of, and thus we discuss the technical frameworks relevant to the good management of data. In particular, the OAIS model (CCSDS, 2002) is seen to be of particular relevance as a major international standard. We do not describe the OAIS model here, but note that it provides a generally applicable framework and vocabulary that identifies the key aspects of data management and preservation within open archives, and describes core functional components and workflows. However, there is no general recipe to implement OAIS, and big science systems are large or unusual enough that no one recipe is likely to be applicable. As a consequence, there is a need for profiles of OAIS and exemplars for its use, so that big science projects can rapidly adapt the standard to their needs. The CASPAR project[6] concretised the model by developing a set of methods and tools for several stages of the digital preservation lifecycle, as defined by OAIS. There were three testbeds providing validation; the science data testbed was provided by STFC and the European Space Agency. The outputs of the project are collected in Giaretta (2011). Two aspects of this project are particularly applicable for big science: preservation analysis, and the preservation toolkit.

The preservation analysis methodology (Conway et al., 2011) is designed to ensure that the archived science data is a truly reusable asset. The methodology defines a number of analysis steps to produce an actionable preservation plan, satisfying a well-defined preservation objective. To carry out preservation that will allow future users to use the data within new contexts, the information to reuse the data is captured in a formal model of the information dependencies and the alternative information required if the environment changes. This can then be used to design data packages and preservation actions. Although these analyses may at first seem burdensome, we expect that since big science projects need to develop highly functional data management systems, many of the questions in the analysis will already have answers.

The CASPAR implementation provided integrated tools to support the preservation process as described in OAIS functional model, with a particular emphasis on managing representation information to support the contextual use of data. These tools were prototypes at the end of CASPAR and their development is being continued in the SCIDIP-ES project[7] as a web service infrastructure.

The OAIS model can be criticised for being so high-level that "almost any system capable of storing and retrieving data can make a plausible case that it satisfies the OAIS conformance requirements" (Rosenthal et al., 2005), so it is important to be able to provide assurances that the project has achieved more than simply producing the statement "we promise not to lose the data." This involves both defining more detailed requirements and devising more stringent and auditable assessments of an archive's actual ability to be appropriately responsive to technology change (for example, see

[6] The CASPAR project: http://www.casparpreserves.eu/

[7] The SCIDIP-ES project: http://www.scidip-es.eu/

Giaretta, 2011). Funders are also likely to require reassurance that the DMP plan that a project has proposed is achievable. The standard 'audit and certification of trustworthy digital repositories' (CCSDS, 2012) offers a detailed specification of criteria for auditing digital repositories. It is possible to imagine levels of certification where different criteria may be appropriate for projects of different scale, or where the funder has different requirements for the resulting data. High profile big science projects might reasonably be expected to achieve the highest certification level.

The Practicalities of DMP Planning for Big Science

We add some observations on some other practical aspects of DMP for big science.

Data Release Planning

DMP plans need to consider how to manage the release of data. When large 'source' facilities service the work proposals of individual scientists or small groups, they typically release data by making it public in their facility archive, after an embargo period. Large collaborations instead typically release data in large blocks.

The LIGO collaboration has agreed an algorithm to release data triggered by a range of occurrences, including the publication of papers quoting the data, when the collaboration has probed a given volume of space-time, or when a certain time has elapsed after the start of the current phase of the experiment (Anderson & Williams, 2013). The goal was to balance the collaboration members' need for privileged access to the data, with the funder's desire to see the data made public as soon as possible. Large astronomical surveys tend to release higher level data products either after an observing season is over, or after each complete pass over a survey area. The release is not immediate, but takes place after data reduction and quality assurance checks. The ATLAS experiment at the Large Hadron Collider[8] is experimenting with a service called 'Recast' (Cranmer & Yavin, 2010), which will take a phenomenological model as input from a user, and analyse the data in the light of that model. This system means that searches can be performed on the data by physicists not connected to the collaboration, without requiring them to become familiar with the detailed structure of the data. This is effectively a type of high-level data product, which maintains control of the data without the obligation of documenting a data product, and without exposing them to the costs of handling external analysis based on misunderstandings of the data.[9]

Software Preservation

There is often substantial important information encoded in ways which are only effectively documented in software. There is therefore an obvious case for preserving this software. However, software typically includes the libraries it depends upon, the operating system (OS) it uses, and the configuration and start-up instructions. The OS may require particular hardware, the software may be qualified for a very small range of OSs and library versions, and it may be hard to gather all of the configuration information (Matthews et al., 2010). Thus software preservation is complex.

[8] ATLAS: http://atlas.ch/

[9] In OAIS terms, the Recast system itself should be considered as part of the representation information.

However, it is not certain that software preservation is necessary. If the data products are well enough described, then re-running the analysis may be unnecessary, or not worth the investment required for the software preservation. This may be both a cheaper and more reliable way of carrying the experiment's information content into the future, and this trade-off is more in favour of data preservation as we consider the longer term. These two options are not exclusive: one can preserve data and software. However, solutions for software preservation generally focus on active curation in the sense of preserving software through continuing use. This can be successful. However, the sustainability of software then depends on the continuing vitality of a community, so it is brittle in the face of significant funding gaps. A suitably open process can be encouraged, but while this may need fewer resources it probably needs more commitment, and is even less predictable than a funded solution. Further, an early software version, though later deprecated, may still be needed to regenerate or validate a historical release of a data product. Despite these qualifications, assuming continued support is still a reasonable software preservation strategy.

Costs and Cost Models

There is a good deal of detailed information and some modelling of the costs of digital preservation, e.g. the KRDS2 study (Beagrie et al., 2010), the LIFE3 project (Hole et al., 2010), and the PLANETS[10] project. However, this has not created a strong consensus and the variation in preservation contexts may mean that no simple consensus is possible. Note that these projects considered what one might call 'live' archives, where the data has an active community of individuals with expertise in using the data. The situation changes when considering long term preservation, where data is not used for extended periods and there are no living sources of advice about the data. In the case of 'unaccessed' data, there is even less in the way of robust cost modelling, although it seems likely that the cost model for this would be dominated by the costs of byte storage, rather than staff.

There is probably little actual experience of digital archives working entirely without advice from human curators. Information from two astronomy archives considered in Gray et al. (2012) was found to be consistent. The two archives held in the order of 100TB each; one spent 25–30 staff years on development and both spend in the range of 3–6 staff years per year on maintenance and support; each spends between a quarter and a third of its budget on hardware. The HEP community is now constructing more detailed plans for data preservation, and the associated costs. South (2011) estimates that a long term archive would cost 2–3 FTEs for 2–3 years after the end of the experiment, followed by 0.5–1.0 FTE per year, per experiment spent on the archive's preservation. They compare this to the hundreds of FTEs spent on the running of the experiment, and on this basis claim an archival staff investment of 1% of the peak staff investment, to obtain a 5–10% increase in output.

Ingest and Acquisition

In astronomical, HEP and gravitational wave contexts, archive ingest is generally tightly integrated with the system for day-to-day data management, as data goes directly to the archive on acquisition and is retrieved from that archive as part of

[10] The Planets Project: http://www.planets-project.eu/

 doi:10.2218/ijdc.v8i1.247

normal operations. On the other side of the archive, projects will generate and disseminate data products in their normal business, without regarding these as archival objects. Thus the submissions to the archive may consist of both raw and analysed data, and the objects disseminated will include either raw or processed data, or both. The long term planning in the LIGO DMP plan, for example, is therefore less concerned with setting up an archive than with the adjustments required to make an existing data management system robust and more accessible for the long term. This means, in turn, that some fraction of the archive's ingest costs (associated with quality control and metadata, for example) will be covered by normal operations, so the marginal costs of the additional long term archival ingest and dissemination are probably both rather low and typically borne by infrastructure. Thus if the associated activities can be contrived to overlap with normal operations, then the costs directly associated with the archive may be significantly decreased.

Discussion

The most obviously feature of 'big science' in our definition is, of course, the 'big data' aspect. It is characteristic of such projects that they are generally willing to deal with data volumes at the upper end of what is feasible, if necessary by designing instruments to produce data volumes at levels predicted to be manageable by the time the instrument comes online. Without discounting the technical challenges of such data rates, the key implication is that day-to-day data management is a core concern of the project, which is designed and funded accordingly. As a consequence of this:

- Data preservation is straightforwardly identified as an extension of data management. The former is not trivial, but some aspects of data preservation are handled "for free" by the necessary existence of a data management infrastructure, without which the experimental apparatus will be unusable.
- In particular, the problems of data ingest, which loom so large in much of the DMP literature, are reduced to the problem of documenting and adjusting archival metadata, and the identification of suitable representation information.
- Big science projects are inevitably also large-scale engineering projects, familiar with the management of cost estimates, so that the costing of DMP can be built in to the relationship between funders and funded.

So, although at first glance the development of a DMP plan appears to be a burdensome addition to the engineering of a big science project, there may not be a huge amount to do. Much big science is in the happy position of the DMP problem *being already solved to first order*, and thus a DMP planning exercise becomes a question of formalising existing practice.

It can be seen that big science projects are often very different from other forms of research, and although some disciplines have similar characteristics (e.g. population studies in social science and medicine have similar long time scales and specialist teams; clinical trials in medicine require a certified data management process), the situation in the majority of research disciplines is very different and requires a different approach. This naturally leads to questions of:

1. At what point does a repository cross the boundary between a 'small' repository and a 'big science' one?
2. How can institutions benefit from the presence of a big science repository?
3. Are there any lessons for smaller scale repositories?

Big science repositories are solving research and practice problems in DMP; institutional repositories, by contrast, are solving organisational, cultural, educational, usability and scalability problems which the big science DMP planners are able to ignore in part. The educational and usability problems are secondary ones, since the constituency is sufficiently motivated to educate itself, by the need to gain access to the data, which is only available through the DMP system. This is the result of the prior organisational decisions to route all data through a single preservation-ready archive. For institutional repositories, scalability problems should be addressed by looking at (the technologies underpinning) the big science solutions. Something that works at the largest scales will also work at the merely large scale.

In consequence, to address question 1 above, the boundary between a 'simple' and a 'big science' repository is perhaps not simply a matter of scale, but instead lies in the presence of significant technical challenges that are somehow intrinsic to the nature of the data in question, so that dedicated infrastructure is needed to make handling the data tractable *at all*. Having a large data volume represents one type of technical challenge, but others might include having to deal with particularly heterogeneous data, or rapidly changing or otherwise non-traditional data (how would one preserve the internet's connectivity graph, for example, or a service-oriented website?), or dealing with particular assurance requirements (e.g. clinical trials data). In this sense, we can perhaps regard 'big science' problems as one of a category of 'next generation data management problems', which are characterised by having significant unsolved technical preservation problems. This does not imply that the 'simple' repositories are easy, or fully solved: 'simple' repositories might be more complex overall, if producing a 'simple' repository requires simultaneously solving a number of organisational, cultural, educational and usability issues. Perhaps, therefore, a 'simple' repository is one where the required underlying techniques are well understood, so that attention can turn to the applied problem of implementation within a particular institutional or disciplinary context. Thus - and touching on both questions 2 and 3 - 'next generation repositories' can feed technical solutions into repositories at institutional or disciplinary scales.

This is where we believe the earlier report's 'light touch' advice ("*here's OAIS; get on with it*") is valuable. For a 'simple' repository, similar repositories have solved similar problems in similar ways, but a 'next generation' repository has few or no precedents, and cannot be evaluated in the same way. In this situation, OAIS represents a principled and broadly validated approach to archive design, which asks very pertinent, but still very general, questions of any repository, and which supports a natural route to internal and external criticism, up to and including an audit.

Acknowledgements

This work was supported by the RDMP strand of the JISC Managing Research Data programme. We also thank those who provided comments on drafts of the guidelines, including Rob Baxter, Peter Clarke, Catherine Jones and David Shotton.

References

Anderson, S. & Williams, R. (2013). LIGO data management plan. LIGO Technical Report. Retrieved from https://dcc.ligo.org/public/0009/M1000066/018/LIGO-M1000066-v18.pdf

Beagrie, N., Lavoie, B. & Woollard, W. (2010). Keeping research data safe 2. JISC Project Report. Retrieved from http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Hole, B., Lin, L., McCann, P. & Wheatley, P. (2010). LIFE3: A predictive costing tool for digital collections. In *iPRES 2010: 7th International Conference on Preservation of Digital Objects*. Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hole-64.pdf

Bicarregui, J.C., Gray, N., Henderson, R., Jones, R., Lambert, S.C. & Matthews, B.M. (2012). DMP planning for big science projects. MaRDI-Gross project final report, v1.0. Retrieved from http://arxiv.org/abs/1208.3754

Consultative Committee for Space Data Systems. (2002). Reference model for an Open Archival Information System (OAIS) (Blue Book CCSDS 650.0-B-1). Retrieved from http://www.ccsds.org/publications/archive/650x0b1.pdf

Consultative Committee for Space Data Systems. (2012). Audit and certification of trustworthy digital repositories – CCSDS 652.0-M-1.CCSDS Recommended Practice. Identical to ISO 16363:2012. Retrieved from http://public.ccsds.org/publications/archive/652x0m1.pdf

Conway, E., Giaretta, D. & Lambert, S.C. (2011). Curating scientific research data for the long term: a preservation analysis method in context. *International Journal of Digital Curation, 6*(2). doi:10.2218/ijdc.v6i2.204

Cranmer, K. & Yavin, I. (2010). RECAST: Extending the impact of existing analyses. Technical report, New York University. Retrieved from http://arxiv.org/abs/1010.2506

Giaretta, D. (2011). *Advanced digital preservation.* Springer-Verlag. doi:10.1007/978-3-642-16809-3

Gray, N., Carozzi, T. D. & Woan, G. (2012). Managing research data in big science. LIGO Project Report P1000188. University of Glasgow. Retrieved from http://arxiv.org/abs/1207.3923

Flannery, D., Matthews, B.M., Griffin, T., Bicarregui, J.C., Gleaves, M., Lerusse, L., Downing, R., Ashton, A., Sufi, S., Drinkwater, G. & Kleese, K. (2009). ICAT: Integrating data infrastructure for facilities based science. In *Proceedings of the 5th IEEE International Conference on e-Science.* Oxford, UK.

Matthews, B.M., Shaon, A., Bicarregui, J.C. & Jones, C.M. (2010). A framework for software preservation. *International Journal of Digital Curation, 5*(1). doi:10.2218/ijdc.v5i1.145

National Science Foundation. (2010). Grant general conditions (GC-1). Technical Report gc1010, National Science Foundation. Retrieved from http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gc1010

Organisation for Economic Co-operation and Development. (2007). OECD principles and guidelines for access to research data from public funding. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Research Councils UK. (2011). RCUK common principles on data policy. Retrieved from http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

Rosenthal, D.S.H., Robertson, T., Lipkis, T., Reich, V. & Morabito S. (2005). Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine, 11*(11). doi:10.1045/november2005-rosenthal

Science and Technology Facilities Council. (2011). STFC scientific data policy. Retrieved from http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf

South, D.M. (2011). Data preservation in high energy physics. Paper presented at the 18th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2010). Retrieved from http://arxiv.org/abs/1101.3186