MaRDI-Gross - Data Management and Preservation Planning for Large Projects

RWL Jones¹, JC Bicarregui², N Gray³, RCW Henderson¹, SC Lambert² and BM Matthews²

¹Department of Physics, University of Lancaster, UK
²E-Science Centre, STFC, UK
³School of Physics and Astronomy, University of Glasgow, UK
E-mail: Roger.Jones@lancaster.ac.uk

Abstract. The MaRDI-Gross project aims to provide high-level guidance for the strategic and engineering development of Data Management and Preservation plans for Big Science data. The main outcome will be a document, the nominal audience for which is therefore rather narrow. However, the intention is that the document be of use to other planners and data architects who wish to implement good practice in this area.

The goal of the document is not to provide mechanically applicable recipes, but to allow the user to develop and lead a high-level plan which is appropriate to their organisation. Throughout, the document is informed where appropriate by the OAIS reference model, which generally provides a useful guide to HEP data management planning.

1. Introduction
There is an increasing call requiring that scientific data be both preserved for the future and made available to the wider public. Funding agencies are increasingly requiring detailed data management plans for existing and upcoming projects. These plans should expose the resource requirements for preserving and exposing different representations of the data and the potential uses for these representations, allowing an appropriate cost-benefit analysis to be undertaken for the eventual strategy.

MaRDI-Gross aims to produce a document that will reassure someone charged with developing a Data Management Plan (DMP) plan that a reasonable framework for addressing the problem already exists in OAIS [1], which is concrete enough to be useful; further, the project aims to provide pointers to short summaries needed to become a local expert, a writer of a proposal or an assessor of such proposals. The draft report [2] is available for comment, and feedback is welcome.

2. Big Science and Data Management Planning
Big science projects, such as at the Tevatron and LHC, have many unique features. They are characterised by very large datasets (typically multi-petabyte, and growing rapidly). These are analysed in a complex environment, and the data is inherently complex in terms of schemas and the associated metadata (both explicit and tacit) required to turn it into information. This metadata becomes particularly vital if the data is to be reprocessed and reused. Accordingly,
there tends to be a large attendant community, experienced in data management to service these datasets while the collaborations are in their active phase. Analyses make take years to accumulate sufficient data and realise an adequate understanding of the detector and data. The collaborations that collect and analyse the datasets are typically enabled by multiple, multi-national funders, each with their own view on policies. The policies are usually reconciled by binding Memoranda of Understanding by which the collaboration is formed. The collaborations themselves have hundreds or thousands of members, governed by binding rules. Crucially, data access is traditionally based on contributions made in cash or in kind to the construction and operation of the experiment. During the lifetime of the collaboration, there is usually a collective editorial process for analyses. The members are sophisticated users and provide mutual support systems. All of these features mean that data planning and the opening of access to the data have aspects that well beyond mere technical challenges.

3. OAIS
The OAIS reference model [1] is a conceptual model of the functions and responsibilities of an archive of (typically) digital objects, where the archive is viewed as an organisation or other entity, in principle distinct from the data producer, which exists to preserve those objects into the Long Term. The OAIS standard does not describe how to achieve this, but it does clearly articulate the various steps of the process (for example that data goes through phases of “submission” to an archive, “preservation” there, and “dissemination” to users), the various roles involved (for example data “producer” versus “consumer”), and what, at a high level, has to be done to let all this happen (for example the creation and management of documentation about “representation information”). Integral to its development, the OAIS standard defines a fairly extensive vocabulary for digital preservation (each of the capitalised terms in the preceding paragraph has a precisely defined meaning, and these terms have become the standard terms), and most work in this area is framed, directly or indirectly, by the OAIS concept set. Thus, although the OAIS model is not the only model for a digital archive, it is both plausible and conventional, and so makes a good starting point, and a useful shared understanding, for any discussion of digital preservation. In addition, it is worth pointing out that the model was developed by the Consultative Committee for Space Data Systems, and so has a heritage which makes it a natural fit for non-space science data.

However, while OAIS provides a terminology and model for data archival, bespoke implementations are usually required. Beyond OAIS, each collaboration must assess and balance the requirements of DMP with those of the collaboration.

4. Observations from Use Cases
The experimental cases considered included:

- ISIS, the world’s leading pulsed spallation neutron source;
- the LIGO/GEO/Gravitational Waves experiments;
- LHC Experiments, in particular the ATLAS experiment.

Funding agency requirements were also considered. As an example, the following two UK-originated documents were considered:

- Research Councils UK Data Principles [3],
- Science and Technology Facilities Council Data Principles [4].

After considering the use cases, we concluded that while OAIS provides a useful language for Data Management and Preservation (DMP), it is not in itself sufficient for DMP planning in large
Figure 1. The lifecycle of a digital archive as proposed by the Digital Curation Centre.

experiments. However, it can be coupled with the findings of the CASPAR project [5] to provide a semi-standard procedure for developing a DMP plan, with template design documents. Formally, one of the biggest challenges is that the HEP cases often have incomplete “representation information” (including the tacit information held in the heads of the many collaborators) that is unfeasible to complete. (This does not, of course, imply that efforts should not be made to capture these where practically and economically possible.) In addition, the Digital Curation Centre provides a useful lifecycle model for the long-term preservation of Big Science data [6] (see figure 1).

In terms of data openness, there are usually many representations and formats of the data. It is not feasible to store and/or make open all data. In particular, some level of raw data is usually discarded in data acquisition and triggering. The full recorded raw data samples are usually too large and the required reconstruction environment too complex to be feasibly and usefully made open to all; indeed, in many collaborations, such access is not granted by default to collaboration members, to prevent wasteful overload of the storage, retrieval and processing systems. Such general access would have very huge costs in terms of hardware and support effort.

Other formats can be more usefully made generally available. Most experiments make very simplified formats available for outreach and education purposes. They also provide supplementary data and digital records corresponding to the figures and tables in completed papers. It is possible to consider making more complex analysis formats available for external
analysis, although the costs should not be underestimated and the usefulness may be limited. Other techniques that capture completed analyses and allow them to be reinterpreted under new models may be a more effective means of preserving the information from the experiment.

An important point that emerges is that discarding data is acceptable if done in a considered manner; losing data due to thoughtlessness or lack of planning is not. In general, funders should be informed of the cost/benefit of various levels of preservation and openness. They may then make a judgement as to what they require of future projects. It is noted that data storage costs for beyond a decade are very uncertain. Another important consideration is that the validation of the digital archive is essential, but has large up-front costs.

5. Principles for Funding Agencies and Assessors
We recommend that discussions with projects use the terminology of OAIS as a basis. OAIS can provide a framework for negotiating the archiving aspects of project costs/support. The funders must recognize that there are costs, both financial and in terms of effort, associated with validating repository designs, and subsequently the depositories themselves. Funders must interact with projects at an early stage in order to prioritize the preservation goals. The final decision on what to preserve may often have to wait until costs are clearer, later in the project; it may be infeasible to make robust estimates of the costs of preservation, before a project has gained experience with the final form of the gathered data. What appears to be the best-justified long-term preservation model appears to require a large up-front payment in the form of an endowment.

6. Principles for DMP practitioners/researchers
When constructing a data management plan, one should:

- Identify the different levels of data representation and their users
- Identify the different cases of data reuse, including the likely client
- Identify the requirements for reuse (including the full environment) and the associated ongoing costs
- Identify the use cases that can be realistically supported

These things having been established, one should then prepare a resource plan, including both the collaboration and the external data repository requirements. For those formats to be made open to external access, one should establish a reasonable embargo period on the data consistent with contributions required to perform the experiment; this embargo period will generally be different for different representations of the data (publication data, outreach and education simplified formats, analysis formats and full reconstruction outputs for example). The collaboration must establish fair use rules for open data, including acknowledgement and potential authorship. The use of a Digital Object Identifier for each dataset is recommended. Equally importantly, the collaboration should establish limitations to collaboration liability and support in the case of external analysis of open data; a commitment to access to data does not imply a commitment to support external analysis or to critique externally produced papers, nor does it imply endorsement of external results by the collaboration; this should be made explicit in the licensing documentation for the data. Again, the OAIS vocabulary provides a coherent, principled and shared vocabulary for archive planning and for dialogue with the funders. It is not concrete enough to support detailed planning by itself, but other projects such as CASPAR can provide the missing pieces.

7. Conclusion
Long term data management planning is becoming a more onerous requirement on Big Science projects. At the same time, there are additional requirements to make the data open to external
users. Various projects have provided tools that seem applicable to Big Science, and provide a basis for negotiations between projects and their funders. However, a realistic assessment of the resource requirements for different levels of preservation and access should be made, and the funders should then perform a cost-benefit analysis.

The MaRDI-Gross project has provided a document as a guide to both funders and projects, which is available in draft form at the project website [7]. Comments from potential users are welcome.

Acknowledgments
MaRDI-Gross is funded by the UK JISC [8].

References
[1] Consultative Committee for Space Data Systems 2002 Reference model for an open archival information system (OAIS) – CCSDS 650.0-B-1 CCSDS Recommendation URL http://public.ccsds.org/publications/archive/650x0b1.pdf
[2] Bicarregui J, Gray N, Henderson R, Jones R, Lambert S and Matthews B 2012 Dmp planning for large projects Tech. rep. Glasgow University and Lancaster University and STFC E-Science Centre URL https://http://www.astro.gla.ac.uk/users/norman/projects/mardi-gross/manual.pdf
[3] Research Councils UK 2011 RCUK common principles on data policy URL http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
[4] Science and Technology Facilities Council 2011 STFC scientific data policy URL http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf
[5] The CASPAR Project 2009 Addressing digital preservation Project deliverable URL http://developers.casparpreserves.eu/docs/CASPAR-Handbook-4.pdf
[6] Digital Curation Centre 2010 DCC curation lifecycle model URL http://www.dcc.ac.uk/resources/curation-lifecycle-model
[7] MaRDI-Gross URL http://purl.org/nxg/projects/mardi-gross
[8] UK JISC URL http://www.jisc.ac.uk/