# MaRDI-Gross: Requirements analysis

Roger Jones (University of Lancaster)
Norman Gray (University of Glasgow)
Brian Matthews, Juan Bicarregui and Simon Lambert (STFC)

January 2012

## 1 Introduction

The project *Managing Research Data Infrastructure: Big Science* (MaRDI-Gross) is funded by JISC in 2011–12, to support big-science projects in developing suitable Data Management and Preservation (DMP) plans for the data they generate.

For our purposes, 'Big Science' projects tend to share many features which distinguish them from the way that experimental science has worked in the past. These features include being large collaborations, with large volumes of complicated and instrument-specific data (1–10 PB/year, with exabyte/year rates anticipated in the next decade), and elaborate internal organisations. The key feature, from the point of view of this project, is that this is **facilities** science – there is a core facility, with multinational funders, a multi-decadal existence, and a conceptual and administrative separation between the elaborately-engineered resource and the research scientists.

This scale of working produces some simplifications: (i) It is well resourced – data management and preservation is not the responsibility of quarter-time junior researchers, but a key concern of the projects engineering management. (ii) There is a collaborative ethos, which has data sharing (though initially only within the collaboration) at the core of it. 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 will be multiple funders in multiple countries, imposing various, and sometimes conflicting, requirements on data management and dissemination.
- The multiplicity of funders often means that no one funder can reasonably dictate terms.
- Experiments and their datasets are governed by networks of MoUs and SLAs, and in-collaboration decision-making processes which, however intricate the process, are fundamentally consensus-based.

- The IP on the data is often complex.

The complexity of the funding landscape, combined with the fact that the data management systems will (because of the data volume) usually be bespoke, mean that it is essentially infeasible to produce any reusable repository, or to produce useful step-by-step guidance or training.

The MRD-GW project studied the data-management culture of science at this scale [1]. That reports recommendations, bearing in mind the scale and available technical expertise within big-science projects, included (Sect. 2.6, p25):

- 2. Funders should simply require that a project develop a high-level DMP as a suitable profile of the OAIS specification.
- 3. Funders should support projects in creating per-project OAIS profiles which are appropriate to the project and meet funders strategic priorities and responsibilities.

That is, the 'elevator pitch' version of these recommendations is this: funders of such projects can most effectively and appropriately discharge their data-preservation responsibilities by saying to large projects "here's a copy of the OAIS spec; get on with it!"

In many disciplines, this would be dreadful advice, but facilities-scale science projects have the financial and engineering resources, and technical expertise, to produce bespoke DMP plans for bespoke data-management sytems. What must be avoided, however, 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.

## 2 The MaRDI-Gross project

The MaRDI-Gross project will deliver an intellectual 'toolkit' which will be supporting infrastructure for project-specific DMP planning.

As emphasised above, the data-management systems for projects of this scale are essentially always bespoke, so that there is no useful way in which software or 'tick-list' components can be provided for DMP planning in this space. The 'kit' must instead be a set of documentary resources targeted at technical managers and engineers. See Sect. 4 for the requirements on the final document.

The goals will be to equip technical managers and engineers with the materials (i) to make the case to the relevant wider project community for a principled DMP strategy; (ii) to use the principles of the OAIS approach, adapted or profiled as appropriate, to develop a project-specific DMP plan; and (iii) to be in a position to make useful estimates of the development resource required to implement the DMP plan.

The materials will have an auxiliary audience in funders, who need to develop the expertise to criticise DMP plans. Such critiques might cover for example conformance with RCUK goals, the 'evaluability' of the DMP plan, or various wider social or political obligations (access to environmental data is the obvious example here, though this specific area is unlikely to come within the scope of the MaRDI-Gross project).

| Users | Main task goals |
|---|---|
| Project DMP planners | Development of principled and funder-compliant plans |
| Funding bodies | Help projects design auditable preservation systems which implement funder goals |

Table 1: User needs analysis

| Stakeholders | Interest |
|---|---|
| Data users | Research tasks must be easy to achieve |
| Archive implementers | Detailed functional requirements derivable from, or usefully comparable with, high-level requirements |
| Project management boards | Concrete object of negotiation with funders |
| International project partners | Should lead, or at least be compatible with other countries prescriptions and requirements |
| Funders | Meet politically imposed obligations |

Table 2: Stakeholder analysis

# 3 Users and Stakeholders

See Table 1 for the user needs analysis, and Table 2 for the stakeholder analysis. The obligations on funders come, effectively, from government, and are most concretely represented in the "RCUK Common Principles on Data Policy" `http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx` and the RIN framework `http://www.rin.ac.uk/data-principles`.

# 4 Document requirements

We believe that the community targeted by the MaRDI-Gross project deliverables needs a compact source of information on the following topics.

## 4.1 General and policy background

This includes basic glossary-style definitions on consensus terms such as 'data management' and 'big science', but more importantly contains material about the larger-scale, 'softer', policy context. The practical motivation for its inclusion here is that it can provide the rationale for some of the aspirations and prescriptions in the more concrete parts later.

This section will include a discussion of the meaning and context of the RCUK Data Principles, and on the forces bearing on the question of open or shared data, and data citation.

### 4.2 Technical background

The MaRDI-Gross manual will be based on the framework represented by the OAIS standard [2], though with mention of other standards where appropriate. This section will describe the conceptual structure of the OAIS framework, along with criticisms of it, and discussion of the ways in which a DMP plan can be validated as being conformant.

The resources will therefore include the following.

1. Overviews of the OAIS methodology and goals, and its strengths and weaknesses with regard to validation and specificity.
2. Indications of the state of the art in OAIS application, validation and auditing.
3. (References to) Backup documentation, including the OAIS specification and selected developments or applications of it.

In some contexts, 'discoverability' means that a DMP plan may have to include references to the OAI-PMH standard.

We are conscious that there is sometimes confusion over the terminology which has developed, within disciplines, covering relevant aspects of this area. Some of this will we hope be addressed by the terminological aspects of our OAIS discussion, but we will include discussion of those sources of potential confusion that we are aware of.

### 4.3 DMP planning specifics

In addition to the general technical background of the previous section, we will include detailed information about some aspects of detailed DMP planning which are portable to, and generic across, big-science DMP problems.

1. Case-studies of existing DMP efforts in existing projects. We will aim to include input from the relevant project personnel. The MaRDI-Gross project team will build on existing close relationships with the HEP and gravitational wave communities.
2. Costing models for DMP planning, to the extent that this is feasible and useful. These models are poorly developed at present, and it is not our goal to produce new rigorous ones, but instead to gather what information is available.

## References


[1] Norman Gray, Tobia D Carozzi, and Graham Woan. Managing research data – gravitational waves: Final report. Project report P1000188, University of Glasgow, 2010. Deliverable for the MRD-GW project. Available from: `http://purl.org/nxg/projects/mrd-gw/report`.

[2] Reference model for an open archival information system (OAIS) – CCSDS 650.0-B-1. CCSDS Recommendation, 2002. Identical to ISO 14721:2003. Available from: `http://public.ccsds.org/publications/archive/650x0b1.pdf`.