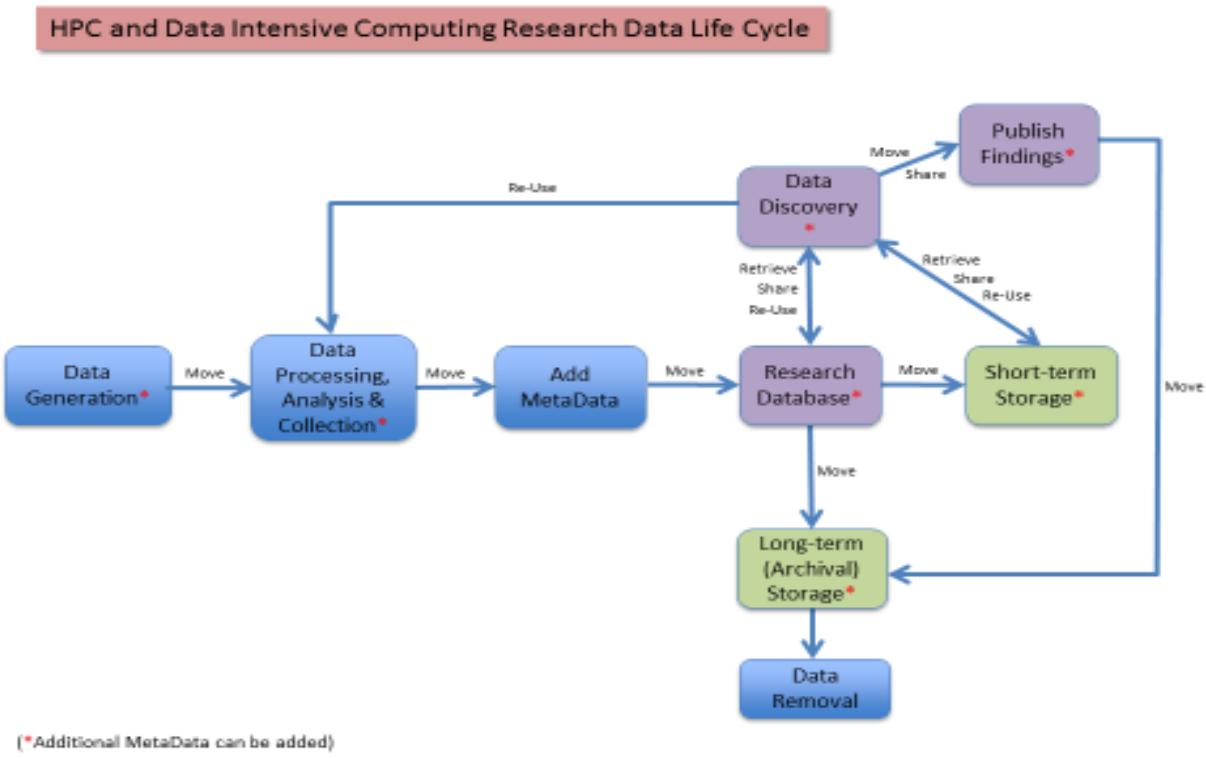
DiRAC Data Management Plan

The long form of the DiRAC Data Management Plan was approved by the STFC DiRAC Oversight Committee in May 2016. This is available upon request from j.a.yates@ucl.ac.uk



The above diagram (author: Jenner, DIRAC), taken from the recent National E-Infrastructure Project Directors Group Report, *The National Data E-Infrastructure (Colling et al)*, outlines the basic requirements of a data e-infrastructure service if it is to enable a productive life cycle for data and researchers.

Using the STFC Data Management Policy and Guidelines as our guide (the above diagram is designed to reflect these), please find below a summary of our Data Management Plan.

1. Types of Data

Broadly speaking the DiRAC-2 simulations produce four "levels" of real data.

- **Level-4: Raw data.** These are the raw data produced by the simulations performed on the DiRAC systems. For some of the research projects these data are used for a considerable time for research by the original owners, but over the years by a considerably wider community. Their useful research lifetime could be in the region of 5-10 years and sometimes longer. However much of these raw data could be regenerated and may not need to be backed up for such long periods. The user will have to decide what needs to be preserved.

- **Level-3: Reconstructed data**. These data are derived from the Raw data by applying analytics and post-processing activities. Typical content includes catalogues, statistical

models, simulated images, tables of values. Level 3 and 4 data are used primarily by physicists for research.

- **Level-2: Data to be used for outreach and education**. Several activities have been developed whereby subsets of the derived data are made available for outreach and education.

- **Level-1: Published analysis results**. These are the final results of the research and are generally published in journals and conference proceedings.

2. Data to be Preserved

- Simulation data can in principle always be regenerated provided the software and the associated transforms have been preserved (see later). **However, for many projects the compute resources required to regenerate data are large compared to the cost of safely storing it.**
- Level-4 data is fundamental and can either be preserved in compressed format OR in regenerated form, preserving the source code and compiler and linking commands or executables/virtual machine images, and the input parameters that were used to generate the data. This is because all other data may, in principle, be derived from it by re-running the reconstruction.
- Level-3 data should also be preserved. This is done for efficiency and economy since the process to re-derive it may take significant computing resources; and in order to easily facilitate re-analysis, re-use and verification of results.
- Level-2 data has no unique preservation requirement.
- Level-1 data is preserved in the journals, and additional data is made available through recognised repositories such as astroph, ADS, Virgo, PLANCK, INSPIRE-HEP and HEPDATA

Processes for Level-4 and Level-3 data preservation

The preservation of Level-3 and Level-4 data is guaranteed by the data management processes of the DiRAC-2 Facility. The DiRAC-2 Facility uses the DiRAC-2 systems, JISC Networking Services, the STFC Scientific Computing Tier 1 Data Facility and the National Service Research Data Facility to implement those processes. The process is as follows:

- The Level-4 Raw data is passed from the DiRAC-2 systems in near real time to the RAL-based STFC Scientific Computing Division Tier 1 Computing Centre where it is immediately stored onto tape.
- If the project requires a second copy of the Raw data it is made shortly afterwards to ensure redundancy. This second copy can be stored at the National Service Research Data Facility or at another site of the users' choosing. The result is resilient copies of the Raw data.
- The STFC Scientific Computing Division Tier 1 Computing Centre have custodial obligations for the Raw data and guarantee to manage them indefinitely, including migration to new technologies.
- Level-3 data is derived by running analytics and post-processing programs. Level-3 data are kept on near line disk. One or more copies of this derived data will also be stored on tape as indicated above.

In summary at least one copy of the Level 4 Raw data are maintained in physically remote locations, at sites with custodial responsibilities. This therefore ensures the data preservation requirements of the STFC policy are met.

3. Software and Metadata implications

From 1st July 2016 the DiRAC facility required all data to be tagged with Metadata that conforms to the following standards:

- The Theoretical Particle Physics and Nuclear Physics communities can make use of the International Lattice Data Grid (ILDG). It is an international organization, which provides standards, services, methods and tools that facilitate the sharing and interchange of lattice QCD gauge configurations among scientific collaborations, by uniting their regional data grids.

  The ILDG promotes a common schema to markup metadata (e.g. physics and algorithmic parameters) that describe ensembles of gauge configurations. Each regional grid catalogs the metadata for the ensembles and gauge configurations they wish to share. Through each of the ILDG web portals, a user can search any or all of the regional metadata catalogues that implement the ILDG web service interfaces.

  Gauge configurations are provided by the ILDG in a standardized file format, which is compatible with the SciDAC QIO input/output library. Member collaborations within the ILDG are developing scripting and GUI based tools to simplify downloading data files.

- Astrophysics & Cosmology, Solar System and Planetary Science, Astroparticle Physics, Gravitational Waves and Nuclear Physics can make use of:

  - The VOTable standard developed by the IVOA and the associated software;
  - A clearly defined standard, along with the metadata writing and reading software, is to be stored as a Virtual Image at the DiRAC backup sites. Examples of these could be those projects that make extensive use of database technology or those using those standards and APIs.

The data generated and used by DiRAC systems are naturally divided into the tasks listed in the Table below

Task	Level 4	Level 3
Data Analysis		Produces data from existing data products
Data Modelling	Generates raw data	Produces data by comparing models to existing data products
Data Production	Generates Raw Data	Produces data by converting raw data to a meta format
Simulation	Generates Raw Data	Produces data by converting raw data to a meta format

As part of the DiRAC-3 upgrade all products will be labelled using standard metadata. These data will be labelled by metadata containing the necessary minimum of the following:

- Project Code;
- Project PI;
- Date of Production;
- Name of System;
- Physical Location of compressed Data OR the source code, inputs and/or a VM image of the code and inputs and runtime environment;
- Research Area;
- Associated Subject metadata that describes each data format;

The Facility will have to ensure, subject to funds awarded for data management purposes, that the infrastructure exists to tag data with these metadata. These will be in the form of:

- Appropriate APIs;
- Appropriate software to write and read metadata.

Metadata Catalogue and Data Publishing

An appropriate metadata catalogue will be deployed and installed, if funds allow. This will be readable by the stakeholders needing access to these data and will fulfill the OpenAcess Obligations of our users. This will leverage existing projects, such as those listed by the Digital Curation Centre.

In the DiRAC Phase 3 context, the new systems will have to provide, subject to new data management funding, or have access to, the following:

- Hosting services for those projects that wish to publish their metadata and data products. A charge can be made for these services to individual projects;
- An Analytics and Visualisation environment in which analytics can be run on stored data and new data generated and labelled with metadata.

Software and APIs

The Software and API environment needed to manage the data will be:

- Suitable data base technologies to hold catalogue data;
- Suitable APIs to allow the creation of metadata and its interrogation;
- Suitable backup and restore software.

4. How long does data need to be preserved?

There is currently no upper time limit foreseen for the retention of DiRAC data. Naturally the ability to do so depends upon the continuation of DiRAC, the remote data centres and on available expertise and funding.

5. Which data will have value to others and should be shared?

The DiRAC Facility takes open data access very seriously. These specify:

- Which data is valuable to others: In general, raw data could not be interpreted by third parties without them having a very detailed knowledge of the experimental reconstruction software. Derived data may be more easily usable by third parties, and provision is made to make this available upon request on a case-by-case basis.

6. The proprietary period

Experiments specify a fraction of the data that will be made available after a given reserved period. This period ranges over several years reflecting the very large amount of effort expended by scientists in construction and operation of the experiments over many decades, and in part, following the running cycle that defines large coherent blocks of data.

7. How will data be shared?

Data is currently shared by the individual projects. In some cases, individual projects and systems have also taken the initiative to develop or engage with value added open data services using resources leveraged from STFC and other Facilities. Examples of this are Virgo and PLANCK, and we expect similar leveraging of resources from projects/missions such as LSST, SKA, LIGO, CTA, JUNO, JUICE, GAIA and Euclid

However, as part of the proposed DiRAC-3 upgrade the DiRAC Facility will offer data sharing facilities to all of its users.

Data will be made available in an externally intelligible format as specified by the DiRAC Project using a suitable metadata standard and file format. The software required to read the data is also available on a similar basis, along with appropriate documentation.

8. Resources required to preserve and share the data

Data preservation activities should arise as a natural result of scientific good practice. This leads to extra staff costs over and above those needed for operating DiRAC, as well as additional storage costs. These activities rely upon the continuation of DiRAC and the remote backup sites.

Data Preservation is taken care of by a Disaster Recovery Plan, which has already been approved by the STFC DiRAC Oversight Committee. The pilot project to create one offsite was completed on the 30 November 2016. This pilot project involved the DiRAC Data Centric facility at Durham, the STFC RAL Tier 1 Facility and JISC. The remaining sites are now working to create offsite backups at the RAL Tier 1 facility and this should be completed by 1st May 2017. In addition, DiRAC is making use of the Edinburgh National Research Data Facility to store data. The Facility now enforces storage quotas for all projects. This allows new projects to have access to new parallel file system resources and will make existing projects manage their data appropriately.

As part of any DiRAC Phase 3 refresh, all sites will have to specify how they will produce and store local copies of data to supplement the copies that will be backed up offsite.

The additional cost of storage for data preservation and the staff resources required will be specified on a regular basis to the DIRAC Oversight Committee and the DiRAC Project Board.

Currently, DiRAC doesn't have any specific resources for carrying out the active open data access activities described. These will be addressed in the context of a DIRAC-3 upgrade in 2017.

To implement the full metadata tagging described above would require substantial new resources and is therefore subject to funding.