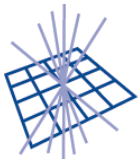

GridPP
UK Computing for Particle Physics

GridPP Project Management Board

Project Status

Document identifier :	**GridPP-PMB-158-DataPolicyComments.doc**
Date:	**30/5/2012**
Version:	**1.0**
Document status:	**Final**
Author	**PMB**

Comments on the STFC Scientific Data Policy

Prepared by David Britton on the behalf of GridPP, based on input from the STFC Computing Advisory Panel, (CAP) provided to STFC in May-2012[1]

The STFC has released a document entitled "Scientific Data Policy"[2] that outlines, at a high-level, a general policy on data preservation and curation. However, translating these high-level aspirations into a set of practical actions is likely to be a challenging and domain-specific exercise. This document examines some of these challenges from the perspective of High Energy Physics (HEP), the primary clients of GridPP. It is of note that almost all HEP data is already digitised so this document does not address an orthogonal set of issues related to the overhead in digitising and organising data, which may well apply to other disciplines.

The vast majority of data held by GridPP must be subject to the data preservation policies of the experiments that own the data and not to an independent GridPP policy. Although GridPP may be required to provide access to data, this must done be in accordance with the experiment's polices and adhere to the Grid security policies that we implement.

For clarity, we use the word "preservation" to refer to the simple operation of physically preserving data so that it can subsequently be accessed. This includes dealing with evolution of physical media, but not the means to interpret the data. We use the word "curation" to refer to the complete process of preserving, and making available on a long-term basis, the means to read and interpret the data. This specifically adds the metadata, knowledge and any necessary software.

Particle Physics Background

As early producers of large datasets, experimental HEP has evolved various strategies both for data preservation and for the exchange of data. Although much of this policy is implicit within the implementations of these strategies (the tools and procedures), the major accelerator laboratories (notably CERN) and the associated experiments are in the process of developing more explicit policies. In addition, initiatives such as DPHEP (Data Preservation in HEP) are informing this discussion (but do not lead it). Given the international nature of HEP collaborations, bound together by agreements, memorandum of understanding, and convention, it is important that the implementation of the STFC Scientific Data Policy be done in a way that is compatible with international expectations and obligations. The current experiment policies are also implicit in the Collaboration Board documents that govern the working rules and practices of the collaborations such as the authorship policy. This reflects the close association between data access and the recognition and credit systems of the collaborations.

HEP data typifies the pyramid of knowledge, with a broad base of "raw-data" delivered by the experiments and online systems, which gets progressively refined and reduced through three or four clearly identifiable steps. The tip of the pyramid is the published information progressively derived from the raw data *plus* associated metadata *plus* other knowledge embedded in software processes. As one climbs the pyramid, "data" becomes "information". There is typically a high cost for access to

[1] http://www.stfc.ac.uk/Resources/pdf/120100-CAP-SDP-Comments.pdf
[2] http://www.stfc.ac.uk/About+STFC/37459.aspx

the raw forms of data in large experiments, and individual access is not usually granted. This is effectively a cost/benefit judgment, which reflects not only the cost of access, but also the complexity of the environment required to make meaningful use of that raw data and the difficulties of ensuring the integrity (correctness) of the process.

In its most processed forms, several approaches have been employed for the curation and exchange of experimental data. The contents of figures and additional tables of information supporting the publications are stored in the HEPDATA project, as well as with journal services where available and appropriate. Further details are also provided through public notes provided through document servers. For educational and outreach purposes, datasets in simplified forms are also made available. These allow 'realistic' analysis to be performed, but are not intended to support a genuine 'publication quality' analysis.

At a lower level of the data pyramid, more extensive data formats are developed between the particle physics experiments for the comparison and combination of data. These formats are developed on a case-by-case basis, tailored to the study in hand and agreed between the experiments. This is an open but expert activity, as at this level the peculiarities, strengths and weaknesses of the different experiments become relevant, as do the tools (such as detector simulation packages and Monte Carlo generators) used to abstract the data to this level. Without a high level of dialogue and expertise, misleading conclusions can easily be drawn.

At a still deeper level, work is ongoing to preserve the data, metadata and analysis environment. This is very much an experiment-specific activity, as it depends on the event data model, experiment-specific tools and the analysis workflow for that experiment. Attempts were made to preserve data from various experiments now no longer running to varying degrees of success. In the LEP (1989—2000) case, there was planning and prior consideration, and a significant class of analyses are reproducible and new analyses possible, but still require a great deal of effort and tacit information. This indicates both the spirit of engagement of the experiments with this objective, but also points to the difficulties and limitations.

The lowest level of data that has typically been preserved and curated beyond the end of the experimental collaboration is the 'DST' (Data Summary Tape – although it is not actually tape any more), not actually the raw data recorded from the experiment. With the very long planned lifetime of the LHC, this is changing, as the experiments themselves wish to continue to reprocess from the raw data for many years. Accordingly, the LHC computing models all have planning for the long-term preservation of the raw data at CERN and at major national computing centres. Continuing this preservation after the end of the LHC exploitation phase will require continued funding. Indeed, an open question within the experiments has been at which point the data becomes an archival set and reprocessing ceases. This is not yet clear, but it is at least several years. Indeed, many analyses will not produce their first results until several years after data is taken (which is a function of the data and detector complexity and the required number of event needed for the analysis; and the fact that some analyses require prior analyses to be complete before they can become meaningful). This time-delay of years is of particular note in the perspective of developing Data Policy implementations.

One issue that is not unique to experimental particle physics, but is perhaps exposed most strongly in this field, is that of international collaboration. The views on data preservation, curation and access differ in the various nations, and the activities are undertaken under memoranda of understanding that already exist and to a large extent already imply the policy on data access in particular. Access to the data is a large incentive used to entice nations, funding agencies and institutes to engage in the design, construction and execution of the experiments. Data availability must therefore be tensioned against the prior commitments made to those joining the collaborations.

Comments on the STFC Scientific Data Policy

1. The goals embodied in the STFC document are clearly desirable and sensible: *"STFC, through the facilities it operates and subscribes to and the grants it funds, is one of the main UK producers of scientific data. This data is one of the major outputs of STFC and a major source of its economic impact. STFC, as a publicly funded organisation, has a responsibility to ensure that this data is carefully managed and optimally exploited, both in the short and the long term."*

2. The definition of 'data' in Principle-iii of the STFC document covers the full range of the data pyramid from raw data to published information. Furthermore, Recommendation-v states that data policies should cover all these types of data. In this perspective, Principle-x (that all data should be made publicly available after a limited period of time) would be extremely difficult and expensive to implement for HEP, if interpreted literally. The first full year of LHC running has generated something of the order of 100 Petabytes of stored data, which requires something like 50-million lines of code to process, together with expert knowledge distributed globally between several thousand physicists. It is an enormous challenge to make our data accessible to our own experts in a meaningful way; to extend this to a non-expert audience would be a monumental task.

3. The STFC document makes no distinction between the layers of the data pyramid. We feel that it is important to establish that not all types of data can be considered in the same way. In general the nearer data is to its raw form at the bottom of the pyramid, the more complex it is to preserve and curate, and in particular the more difficult it is to make available easily to the public (i.e. including everything needed to use the data to derive some conclusion). Conversely the more processed the data is then the easier it may be to make it available for publications, or for educational purposes - but it may then not be suitable for serious re-analysis.

   Careful consideration should be given to the applicability of the policy in respect of different types of data within each STFC activity sector. Until this is done it would be counterproductive to impose a naïve "one size fits" all expectation upon grant proposers.

4. For data at the top of the pyramid (published information) the intent of the policy is to be welcomed: Preserving datasets, where meaningful, which lead directly to publications is to be encouraged, as are moves to make additional information available in connection with publications (i.e. making sets of numbers associated with figures available, depositing of data in some repository at the time of publication).

5. For data at the bottom of the pyramid (raw data) we fully agree that they should be preserved when they have been created at significant expense or cannot be recreated. In other circumstances, we can use the concept of "virtual data": Data that is created on demand by a known process because it is cheaper than curating the original data. In this case the *"process"* is curated, not the data itself.

6. Principle-xi (that data should be made public within 6-months) in the STFC policy document would, if interpreted literally, lead to significant problems. As pointed out earlier, some analysis will not converge for long periods of time (years) after the raw data is collected. There are other

constraints on the time frame implicit in the agreements that bind the international collaborations. One of the ways that Particle Physics has historically ensured scientific rigour is to have at least two completely independent experiments that tackle the grand challenges. Thus in the last 20 years we have see examples of BaBar and Belle at the B-factories; Zeus and H1 at HERA; and the four LEP experiments at CERN. To make raw data publically available (and thus available to a competing experiment) in a period that is short compared to the natural life cycle of that particular data set would undermine this basic principle. The natural lifetime of much HEP data is considerably longer than 6-months.

7. In general the curation of raw data requires:
    - The physical data to be stored in perpetuity.
    - The software able to read and reconstruct such data, plus the means to evolve with media and standards.
    - The software needed to analyse such data.
    - The metadata needed to interpret the data.
    - The tacit knowledge needed to analyse, understand and interpret the data.

    All these steps have challenges but a particularly difficult issue is that of capturing the "tacit knowledge". This is a well-established phenomenon referring to the "unwritten knowledge" which exists within a collaboration of people (scientists, engineers, operations staff, etc). This tends to die out with the termination of an activity. This is not addressed in the policy and there is no simple or cheap solution.

8. Given the points made above, the potential interpretation of the combination of Principle-x stating *"Data [..] should be made publically available after a limited period"*; Principle-xi stating *"available to anyone"*; and Principle-iii and Recommendation-5 that the policy "*refers to all data including raw data*", is of great concern. Interpreted literally, this could mean that there is an expectation that all activities are required to make all *Raw* data and the means to use it easily, available to the public. The issue here is the practicality of providing and maintaining in perpetuity a platform that makes these data easily usable by a member of the public. Whilst this may be a laudable aspiration, the cost in terms of physical and human resources may in some cases be very high. Considering also the complexity, it may be that it is effectively impossible for a member for the public to use the data in any simple way – this is a fact of complexity and not any intent to hide data. If it is effectively impossible for a non-expert to make use of such data, how is the cost of making it available justifiable?

    Without context (activity) specific guidance and interpretation, then the letter of the policy as currently drafted may impose an impractical and unjustifiable burden. Context specific guidance would be needed so that grant proposers may draw up appropriate data management plans that recognise differences between the layers of the data-pyramid.

    Where (if) it is deemed (by STFC) that Raw data must be made available to the public (i.e. properly curated), then a full and realistic assessment of the resources required should be done, and this should presumably be tensioned against the opportunity cost if these resources come from fixed budget.

9. The scientists creating data will not necessarily have the knowledge and skills required for proper data curation. It will be unhelpful if this is not openly recognised, and it would be wrong to expect scientists can easily become curators at no extra cost. It follows that data management plans drawn up by those proposing grants should not be expected to be too detailed and may need to refer to external initiatives (such as a Collaboration-wide policy).

The knowledge limitation of grant proposers in respect of preservation and curation in particular should be allowed for in the expectation of data management plans required at the time of submission of proposals.

10. Planning for data preservation and curation should be subject to a cost benefit analysis. There is no point in incurring a large expenditure to curate data for which there is little likelihood that it will ever be required again. Therefore the cost benefit analysis should assess:
    a. The cost of preserving it at the point of creation –vs- the cost of re-creating it (if this is possible).
    b. The likely need to re-use the data
    c. The cost of fully curating the data at the point of creation –vs- the cost of deferring this until needed.

11. Expanding on the last point, it may be appropriate to some activities to defer the costs of curating data until the point it is required (if ever). This means doing the minimum to preserve it at the point of creation, but not fully curating it to the point of public availability until a request is made to re-use it.