Comments on the STFC Scientific Data Policy document (SDP)

Prepared by the STFC Computing Advisory Panel (CAP)

Jan 2012

Status: Pre-Final Version.

The STFC has released a document entitled "Scientific Data Policy". This document here provides some initial comments to that policy to STFC.

Throughout we use the following terms:

"activity" to generically refer to some area of STFC science. This may be an experimental or theoretical collaboration, a facilities users group or a facilities operation group.

"preservation" to refer to the simple operation of physically preserving the information represented by data so that it can be physically accessed. This includes dealing with evolution of physical media, but not the means to interpret and the data.

"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 software required to do this.

The Principles and Recommendations of the SDP policy are referred to by **P(n)** and **R(n)** respectively.

Comments

1. First and foremost, CAP welcomes the spirit embodied in the SDP document :

   "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. We agree that data which is produced on behalf of STFC represents an investment by the public and is a source of scientific and economic impact. These data are often obtained at considerable taxpayer expense and should be maximally exploited. Where such exploitation dictates, then the data should be preserved, where appropriate curated and where appropriate made available to the public.

3. **P(iii)** and **R(v)** enumerate the type of data covered by the policy, ranging from *Raw* data through various stages of *Processed* data to *Published* data. **P(x)** states that it must all be made publically available.

4. It is important to establish that in general not all types of data can be considered in the same way. In general the nearer data is to its *Raw* form, 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.

5. In respect of data at the *Published* end of the scale, 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).

6. In respect of *Raw* data, we fully agree that they should be preserved when they have been created at significant expense (this has to be qualified by it either being impossible to recreate the Raw data, or the expense of re-creating it would exceed the cost of preserving it).

7. **P(xi)** is an example of a seemingly simple statement about data associated with publications, but one which is ill defined in some contexts, and this could, if interpreted "legalistically", lead to significant problems in meeting the policy requirements. Where the meaning is clear (i.e. an obvious set of data goes with a publication and it is meaningful to make it available) then we agree and encourage that this should be done. However, for large long term experimental collaborations (e.g. LHC, FAIR) with very large raw data rates, should this be interpreted as the figures and numbers of entries in figures ?, the reduced datasets upon which the analysis was performed ? or does it mean the original *Raw* data files (as defined by **P(iii)**)?. We rather doubt that the policy was meant to intend this, but if interpreted to do so could be impractical. This issue is addressed in more detail below.

   We hope that the spirit of this principle may be interpreted as applying to those data for which it is meaningful and practical.

8. 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 meta data associated with the data needed to interpret it.
- The tacit knowledge needed to understand and interpret the data.

9. An important issue is that of capturing the "tacit knowledge". This is 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 solution.

10. The biggest concern of CAP comes from the potential interpretations of **P(x)** stating *"Data [..] should be made publically available after a limited period"*, and **P(xii)** stating *"available to anyone"* . Taken together with **P(iii)** and **R(v),** and 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 which 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 considerable. 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. Here we repeat some text from the RSE submission to the Royal Society on this matter : "*It is clear that whilst the overarching principle of open data must be embraced, this does not necessarily mean that access will be easy, immediate, or that the data will be in a readily accessible form. "*

    CAP recognises that it is not the intent of the policy to impose impractical burdens, but notes that without context (activity) specific guidance and interpretation, then the letter of the policy may impose such a burden. We recommend that such context specific guidance is drawn up and that grant proposers may use this when drawing up data management plans.

    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.

11. CAP agrees with principle (v), that data management plans should be drawn up at an early stage. We also suggest that designing data models at the outset with long term preservation in mind is likely to be useful. In fact many of the policy aspirations are only truly likely to be achieved in future experiments where open data principles are built in from the outset.

12. Proper data curation is a first class activity, and requires professionals who will not in general be the scientists creating the data. 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 (this is not to be confused with short term initiatives designed to establish a capability -these certainly do require scientists to be involved).

13. It follows that data management plans drawn up by those proposing grants are likely to be incomplete – since it is unlikely that the proposers will have the knowledge to address long term curation completely at the time of proposal. We suggest that STFC should recognise this limitation. This may be mitigated if there is any "standardised" way of meeting some of the obligations (e.g. a centrally supported data store).

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.

14. We suggest that all 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 curation –vs- the cost of deferring this until needed.

15. 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.

16. There is an important distinction between *data* and *information* which is not addressed by the policy. The two are not the same. Of course the interpretation of data as "instrument readings" is fairly unambiguous, and in many cases this is what needs to be preserved. However rarely does data of this form represent information, which may only be extracted from data after some other inputs and other transformations. It may be that in some cases the imperative is to preserve the information – not the data. The cost differential may be significant.

17. Finally, we note that the breadth of science activities covered by SFTC is great. As such the prevailing common sense interpretation for each activity is likely to vary considerably.

We recommend that where a community wide policy (possibly a global community) exists or is being formed, then where possible this should be accepted by STFC as meeting part or all of the requirements of the SDP.

Current practice

Although the previous comments may focus on potential concerns (or perhaps better stated as unintended consequences) in the interpretation and application of the SDP, CAP emphasises that it nevertheless embraces the spirit. In fact within some activities there are already good examples of an open data policy.

We have recommended that each activity should draw up some guidelines which are appropriate and practical for that activity. To move this process along, we outline current good practice and current developments in some example activities.

A. Facilities experiments

The major STFC facilities used by UK researchers are the ISIS and Institut Laue-Langevin (neutron scattering), Diamond and the ESRF (X-ray scattering) and CLF (lasers). Of these ISIS and ILL have published data policies, and Diamond and ESRF are currently formulating.

ISIS (http://www.isis.stfc.ac.uk/user-office/data-policy11204.html) has laid out a clear set of principles and definitions, including those pertaining to free and commercial access to data and opening of data to researchers outside the original team:

- "3.1.1 All raw data and the associated metadata obtained as a result of free (non-commercial) access to ISIS, reside in the public domain, with ISIS acting as the custodian."

- "3.1.2 All raw data and the associated metadata obtained as a result of 'commercial-in-confidence' access to ISIS will be owned exclusively by the commercial user. Commercial users must agree with their relevant instruments scientists how they wish their raw data and metadata to be managed before the start of any experiment."

- "3.3.3 Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible. Any PI that wishes their data to remain 'restricted access' for a longer period will be required to make a special case to the Director of ISIS".

ILL (http://www.ill.eu/users/ill-data-policy/) operates under the following general principles "Central facilities for neutron scattering and synchrotron X-rays in Europe are working together increasingly to develop and share infrastructure for the data collected there. Such co-operation should make it easier and more efficient for users to access and process their data, and provide more secure means of storage and retrieval. It should also increase the scientific value of the data by opening it up to a wider community for further analysis and fostering new collaborations between scientific groups. Ultimately this should

improve the quality and quantity of publications from such data. However, with these developments comes a need to define how such data and any associated metadata are stored and made accessible, and for this a common data policy has been established to provide a suitable working framework". Perhaps the key clauses are those given below, and we note that 3.3.3 mirrors the policy set out by ISIS.

- "3.1.1 All raw data and the associated metadata obtained as a result of publically funded access to the research facilities are open access, with the research facility acting as the custodian".

- "3.3.3 Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of 3 years after the end of the experiment. From 3 years to 5 years after the end of the experiment people from outside the designated experimental team may request to access the data, in such situation, the facility management in coordination with PI will study the release of the access. Any PI that wishes his data to remain restricted access for a longer period will be required to make a special case to the respective facility management". Data can always be made openly accessible earlier on simple request of the PI.

B. Observational Astronomy

Observational astronomy has traditionally instinctively recognised the need to curate data properly, driven by the demonstrated value of doing this within the professional community. The value of proper curation of the data for wider access (eg. by the public) has been recognised more recently and its implementation has been focused on a few flagship projects, often partly motivated by publicity issues.

The main drivers emerging from the professional community have been:

- the enormous value of data reuse and thus effective data sharing, particularly recognising the large costs of obtaining the data in the first place. Of course as to a large extent the sky is not static, each observation is in a sense unique and not reproducible. This places limits on data reuse, but for many purposes, eg statistical studies, the variability of the sky is not an important issue;
- the emergence of large area and all-sky surveys, already extending to 10^9 to 10^{12} objects;
- the advent of the Virtual Observatory which provides a formal and advanced framework for data sharing and data reuse.

Most observational astronomy uses established facilities (telescopes, spacecraft …) which are operated by professional organisations or agencies which typically take the responsibility for a significant fraction of the data curation, data access and related activities. Space missions often provide the best practice in the area, primarily for historical reasons. A very large fraction of observational astronomy

data is made freely available in the public domain, typically on timescales ~1 year after acquisition. Private data do exist however.

Typical disposition of responsibilities is as follows:

- Raw data
    - stored and archived by the facility
    - access and reduction tools provided and documented by the facility
    - advanced or specialised reduction tools may be provided by external science teams. Long-term maintenance of these tools is rarely assured.
    - access to raw data by those outside of the professional community is rarely, if ever, catered for.
- Reduced data/derived data products
    - Many facilities produce derived data products from the raw data using standardised processing pipelines. The quality and usefulness of these products varies significantly from facility to facility, as does the related documentation and meta-data.
    - These derived data products, where they exist, are typically also stored and archived by the relevant facility who will also provide relevant documentation
    - External derived data product archives also commonly exist provided by external science teams. Long-term maintenance of these archives is rarely assured.
    - Access to derived data by those outside of the professional community is provided by some large astronomy facilities/projects, for example in the form of graphical images
- Access mechanisms
    - Access to astronomy data is currently provided by a wide range of online tools which provide the required search and retrieval mechanisms and often also include many more advanced features. Access is typically facility or project-specific, although more generic access to multiple data sources is provided by some organisations/projects. The Virtual Observatory aims to provide (amongst other things) a solution to accessing data from multiple data holdings in a one-stop-shop approach. It has been very largely successful in achieving this although long-term support for the VO is far from assured. None of these access routes is restricted to the professional community, but very little of it would in practice be useful to non-specialists
    - Access to astronomy data by non-specialists (e.g. the general public) is specifically catered for by some projects, such as the Sloan Digital Sky Survey (SDSS). This is achieved by offering simplified use interfaces and use of non-specialised data formats.
    - There are also some astronomy projects that are specifically based on non-specialist participation, like GalaxyZoo and other citizen-scientist initiatives. Interestingly the results of some of these projects are really only accessible to the professional community as they require.

Overall the curation approach for derived data could probably be described as good, and benefits significantly from the existence of widely adopted data standards within astronomy (e.g. the FITS standard and its extensions) which originated in the early 80s. The FITS standard incorporates an internal data description and provision for extensive meta-data. In general use of these standards does actually ensure that a data set obtained 30 years ago will in fact still be readable and usable today! Long-term preservation of derived data product archives is nevertheless patchy and far from secure.

C. Computational astrophysics

Computers are used ubiquitously in theoretical astrophysics to model physical processes such as the action of gravity, fluid dynamics and radiative transport. The time evolution of complex physically motivated models are simulated using computers to numerically solve sets of coupled partial differential equations, starting from some assumed initial conditions. These simulations can generate huge, very rich datasets. The analysis of these datasets is often complex with the data appearing in scientific papers being highly processed and reduced forms of the raw simulation data.

In the short term these computer generated datasets can be recreated exactly, although the cost of doing so may be prohibitive. However in practice the precise recreation of the raw data is rarely possible beyond the life-time of the hardware on which it was created. On a longer time-scale the data can often be reproduced almost exactly, to a level that would support the conclusions from typical analyses, but this often requires the tacit knowledge of a small number of people to recreate the initial conditions and to rerun the simulations and in many cases the resource cost of doing so is high. Generally it is the practice in the field to preserve the raw data itself over its useful life-time which is typically 5-10 years. Because of Moore's law the cost of reproducing data (or producing new but similar datasets) after a lag of 5-10 years is sufficiently small that preserving data more than 10 years old becomes unattractive compared to generating new data.

The size of many data sets, e.g. the Millennium Simulation, make the sharing of the raw data impossible given the resources typically available. The only way to access the data is to have access to the machine(s) that serve the data. The main value of the Millennium Simulation, however lies not in the raw simulation data (consisting of time slices each giving the positions and velocities of particles representing dark matter), but in highly processed data derived from the raw data.

The Millennium database (based at MPA Garching and Durham) serves the astronomical community by making available halo and mock galaxy catalogues based on the raw Millennium simulation data. These mock catalogues have a huge variety of applications and over 450 papers have been published using this data. The data is served as an SQL database, and recently the same interface has

been used by other groups to publish data from newer simulations e.g. the Bolshoi simulations.

In this field the preservation of the raw is carried out by scientists themselves and falls short of curation, in that tacit knowledge of a few individuals is needed to locate and fully interpret the data.

Examples are:

- URL of Durham Millennium database http://galaxy-catalogue.dur.ac.uk:8080/Millennium/
- a recent initiative making simulation data public and using essential the same interface as the Millenium Simulation http://www.multidark.org/MultiDark/nova.phyast.dur.ac.uk

D. Particle Physics Experiment

As early producers of large datasets, experimental HEP has evolved several strategies both for data preservation and for the exchange of data.

The major accelerator laboratories (notably CERN) and the associated experiments are in the process of developing more explicit policy in this area. STFC should adopt this where possible. There are, in addition, initiatives such as DPHEP (Data Preservation in HEP), which are informing this discussion (but does not lead it).

Given the history in this area, much of the relevant policy on existing collaborations exists not in the Memoranda Of Understanding that form the collaborations, but in the Collaboration Board documents that govern the working rules and practices. Notably, data management and access policies are often implicit in documents such as the authorship policy. This reflects the close association between the data access and the recognition and credit systems of the collaborations.

As in the STFC document, the effective model recognises several different levels of data in a pyramid. At the top of the pyramid are highly processed summary data formats. At the bottom is raw data (as delivered from the experiment online systems). There is typically a high cost for access to the raw forms of data in large experiments, and access is not usually granted to it in bulk on an individual basis. This is effectively a cost/benefit judgement, which reflects not only the cost of access, but the complexity of the environment required to make meaningful use of that raw data.

In the most processed forms, several approaches have been employed for the curation and exchange of experimental data (which at this level should more properly be denoted information). 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 data abstraction/processing, 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 this 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 various experiments now no longer running (e.g. the LEP experiments), to varying degrees of success. In the LEP 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.

It is of note that there are currently projects such as RECAST that will allow the testing of new theoretical models against existing data, without the distribution of the data and the required environment. This is essentially at a level of abstraction between the previous two cases. The model is brought to the data, not the data to the model. This has running costs, but is potentially a better route to allow the open reuse of data.

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 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 and 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.)

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 STFC document correctly recognises the international collaborative nature of the

activities. 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. We therefore recommend that, for existing collaborations, STFC participants engage with their collaborators to honour the spirit of the STFC policy, which recognises these limitations. Where data can reasonably and meaningfully be made available without prejudice to the collaboration's own analysis programme, it should be. This may, however, be several years after data collection, and in reduced formats to improve the tractability of the data in terms of curation and understanding by third parties. Collaborations should made efforts to make data available to the public in a form that is informative and meaningful, both for the wider public and for educational purposes.

For newly-forming collaborations, these issues should be discussed in the formation of the Memoranda of Understanding, and the policies to be pursued should be articulated in outline therein.

Recommended policy for High Energy Physics

Where proposed work is to be conducted in an experiment at a host laboratory where there is a policy already established, that should be followed as far as is reasonable, and exceptions agreed with that laboratory.

For existing collaborations, the first action should be to determine the existing de facto policy, which may reside in various collaboration documents, including the Memoranda of Understanding, the authorship policies, the physics publication policies and in some cases the Treaties and Experimental terms and conditions of the host laboratory. If a general policy is already in place, is should be followed. If that policy is against the spirit of the STFC policy, the collaboration should be asked to reconsider its position on the matter, but without the presumption that it will change.

The intended uses of any data proposed for preservation should be identified (the use cases). A cost-benefit study should then be done for each use case, and a judgement made on whether the cost is commensurate with the likely gain. In this process, it is easy to underestimate the costs, as to meet the intended objectives, it is usually required to preserve a software base, metadata, possibly the hardware resources to support analysis, and also provide extensive documentation. If the curation is for the future rather than current exchange of data, the documentation must be complete in the absence of the original authors.

Finally, consideration should be given to any restrictions to be placed on the use of the data. For example, if a collaboration has a reward system for those building the experiment, acquiring and processing the data, it may not be unreasonable to require that those people have the right to sign any subsequent publication arising from that data. A weaker condition would be that they are

acknowledged, and that a paper or a DoI for the dataset always be cited.

Given the collaborative nature of most HEP experiments, the policy must be agreed by all collaborators. This usually means that the policy should be agreed by the Collaboration Board.