

## STFC scientific data policy

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.

### Scope

This policy applies to all scientific data produced as a result of STFC funding:

- Through grants to universities in particle physics, astronomy and nuclear physics.
- Through access to beam time at STFC supported facilities, e.g. ISIS, CLF, Diamond, ILL, ESRF.
- Through STFC subscriptions to other organisations, e.g. CERN, ESO.

This includes data produced as a result of past funding from STFC or its predecessor organisations (e.g. PPARC, CCLRC) which has already been curated.

This policy does not apply to:

- Data resulting from or relating to work carried out by STFC under contract/service level agreement with other organisations, e.g. data resulting from work on HECTOR (EPSRC) or the National Grid Service (JISC, EPSRC), data curated at the BADC (NERC), or data arising from commercial use of facility beamtime. Policy regarding such data is the responsibility of the contracting organisation.
- Research outputs arising from STFC funding, such as publications.
- Software as a form of data in its own right (as distinct from software required to make use of data).
- Physical collections of items, which may be considered as a form of database. However, it would be expected that similar considerations concerning curation and exploitation would also apply in such cases.
- Data which are purely administrative in nature.

### General principles

- i. STFC policy incorporates the joint RCUK principles on data management and sharing (see appendix). Those principles are not therefore repeated here.

- ii. Both policy and practice must be consistent with relevant UK and international legislation.
- iii. For the purposes of this policy, the term 'data' refers to (a) 'raw' scientific data directly arising as a result of experiment/measurement/observation; (b) 'derived' data which has been subject to some form of standard or automated data reduction procedure, e.g. to reduce the data volume or to transform to a physically meaningful coordinate system; (c) 'published' data, i.e. that data which is displayed or otherwise referred to in a publication and based on which the scientific conclusions are derived.
- iv. STFC is not responsible for the use made of data, except that made by its own employees.
- v. Data management plans should exist for all data within the scope of the policy. These should be prepared in consultation with relevant stakeholders and should aim to streamline activities utilising existing skills and capabilities, in particular for smaller projects.
- vi. Proposals for grant funding, for those projects which result in the production or collection of scientific data, should include a data management plan. This should be considered and approved within the normal assessment procedure.
- vii. Each STFC operated facility should have an ongoing data management plan. This should be approved by the relevant facility board and, as far as possible, be consistent with the data management plans of the other facilities.
- viii. Where STFC is a subscribing partner to an external organisation, e.g. as a member of CERN, STFC will seek to ensure that the organisation has a data management policy and that it is compatible with the STFC policy.
- ix. Data management plans should follow relevant national and international recommendations for best practice.
- x. Data resulting from publicly funded research should be made publicly available after a limited period, unless there are specific reasons (e.g. legislation, ethical, privacy, security) why this should not happen. The length of any proprietary period should be specified in the data management plan and justified, for example, by the reasonable needs of the research team to have a first opportunity to exploit the results of their research, including any IP arising. Where there are accepted norms within a scientific field or for a specific archive (e.g. the one year norm of ESO) they should generally be followed.
- xi. 'Published' data should generally be made available within six months of the date of the relevant publication.
- xii. 'Publicly available' means available to anyone. However, there may a requirement for registration to enable tracking of data use and to provide notification of terms and conditions of use where they apply.
- xiii. STFC will seek to ensure the integrity of any data and related metadata that it manages. Any deliberate attempt to compromise that integrity, e.g. by the

modification of data or the provision of incorrect metadata, will be considered as a serious breach of this policy.

#### Recommendations for good practice

- i. STFC recommends that data management plans be formulated following the guidance provided by the Digital Curation Centre <http://www.dcc.ac.uk/resources/data-management-plans>. STFC (e-Science department) can provide advice upon request.
- ii. STFC would normally expect data to be managed through an institutional repository, e.g. as operated by a research organisation (such as STFC), a university, a laboratory or an independently managed subject specific database. The repository(ies) should be chosen so as to maximise the scientific value obtained from aggregation of related data. It may be appropriate to use different repositories for data from different stages of a study, e.g. raw data from a crystallographic study might be deposited in a facility repository while the resulting published crystal structure might be deposited in an International Union of Crystallography database.
- iii. Plans should provide suitable quality assurance concerning the extent to which data can be or have been modified. Where 'raw' data are not to be retained, the processes for obtaining 'derived' data should be specified and conform to the standard accepted procedures within the scientific field at that time.
- iv. Plans may reference the general policy(ies) for the chosen repository(ies) and only include further details related to the specific project. It is the responsibility of the person preparing the data management plan to ensure that the repository policy is appropriate. Where data are not to be managed through an established repository, the data management plan will need to be more extensive and to provide reassurance on the likely stability and longevity of any repository proposed.
- v. Plans should cover all data expected to be produced as a result of a project or activity, from 'raw' to 'published'.
- vi. Plans should specify which data are to be deposited in a repository, where and for how long, with appropriate justification. The good practice criteria assume that this data is accompanied by sufficient metadata to enable re-use. It is recognised that a balance may be required between the cost of data curation (e.g. for very large data sets) and the potential long term value of that data. Wherever possible STFC would expect the original data (i.e. from which other related data can in principle be derived) to be retained for the longest possible period, with ten years after the end of the project being a reasonable minimum. For data that by their nature cannot be re-measured

(e.g. earth observations), effort should be made to retain them 'in perpetuity'.

## RCUK principles on data management and sharing

<http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.
- RCUK recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these are considered at all stages in the research process.
- To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to publish the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.
- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.
- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.