

SOLAS Data Management Policy

Introduction

The SOLAS Science Plan and Implementation Strategy identified data and model management as critical logistical tasks for SOLAS. The implementation of SOLAS involves the collection of large quantities of environmental data by both nationally and internationally organised projects. This will include difficult, error-prone measurement of biological, physical and chemical parameters collected from process studies and experiments (field and laboratory), time series studies, and large-scale surveys (see Appendix). Similarly, SOLAS will make use of a hierarchy of modelling approaches. In most cases, the utility of the models and data involved in these projects will extend beyond the projects themselves and be of interest to other investigators. Significant benefits will result from combining data and model output collected under separate nationally-funded SOLAS projects. Further, many SOLAS data will be more useful when compiled as a global dataset, or when combined and compared with non-SOLAS data. Scientific findings and conclusions derived from SOLAS projects should be available for assessment by independent scientists; this implies that the underlying data and/or models must be readily accessible. The policy adopted by SOLAS is a trade-off between these conflicting needs. It is the intention of SOLAS to follow international best practice (e.g. www.ocean-partners.org/documents/POGO_5_data_final.pdf).

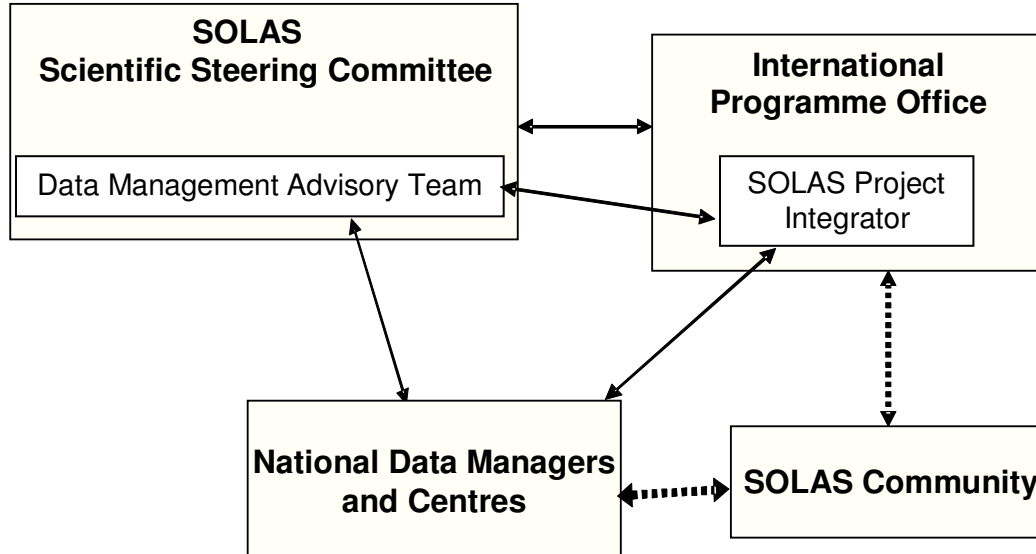
SOLAS Data Management – Guiding Principles

The overall objective of SOLAS Data Management plan is to ensure the security, accessibility and free exchange of data collected or used in the frame of SOLAS research.

Participation in SOLAS-related research requires a commitment to data management. SOLAS strongly encourages its community members to commit to the following:

1. Funding should be set aside for data management from the outset of a project to ensure that all data generated in the frame of SOLAS research is archived, documented, easily found and readily accessible.
2. Whenever possible existing national or international data centres should be used and designated data managers should be appointed.
3. Information about fieldwork activities, experiments, modelling activities, details of the data collected or generated, and name and contact of the designated data manager or data centre should be submitted to the SOLAS IPO either directly or via freely available web-based metadata catalogues (e.g. GCMD).
4. Data should be made available rapidly and submitted to a designated data centre.
5. Internationally agreed standards and protocols (e.g. those of ISO, W3C, IOC/ICES) should be used wherever possible.
6. Data should be reported together with associated metadata (i.e. relevant information about the dataset, including quality-flags).
7. Model documentation, model output and, when appropriate, the models themselves should be made available to the community on a suitable timescale.
8. Data originators should interact and collaborate with the SOLAS IPO (and thus with SOLAS Project Integration, which aims to produce global air-sea flux maps using new and existing concentration datasets).

Structural Overview



SOLAS data management will be overseen by a Data Management Advisory Team (including members from the SOLAS Scientific Steering Committee), national SOLAS data managers and the SOLAS Project Integrator, who works within the International Project Office (IPO). Effective data management within SOLAS can only be successful with full cooperation from relevant national data centres. It is clear that much of the data archiving and processing will not be carried out centrally (e.g. in the IPO), but by a distributed system involving scientists and data centres in the countries participating in SOLAS.

Responsibilities

Data Management Advisory Team

The complex nature of data management in SOLAS requires scientific planning by a multi-disciplinary team of experienced scientists and data managers. This team should, in turn, be well-connected with data management organisations in relevant WCRP and IGBP international programmes. Its role is to:

1. Oversee the compilation of a SOLAS metadata catalogue (see 1 and 2 under SOLAS IPO below) by advising the IPO and the SOLAS Project Integrator.
2. Encourage the compilation of data from individual principal investigators (PIs) and national projects into long-term integrated datasets.
3. Advise scientists without access to an effective data management infrastructure how their data might be made available to the SOLAS community in the longer-term.
4. Monitor international acceptance of, compliance with, and adoption of, SOLAS data policies.
5. Report regularly to and advise the SOLAS Scientific Steering Committee (SSC) on data matters.

SOLAS IPO

1. Create and maintain a central integrated inventory of SOLAS-endorsed projects and of their planned and completed data collection activities (cruises, aircraft campaigns, laboratory and mesocosms experiments, modelling studies, etc.).

2. Create and maintain a catalogue of actual and expected data by producing Directory Interchange Format (DIF) discovery metadata records.
3. Interact with national data centres to coordinate their activities.
4. Encourage timely delivery of metadata to the SOLAS IPO and take actions to encourage rapid submission of actual data to the designated data centres.

SOLAS Community

1. Adhere to the principles recommended within this document.
2. Provide relevant discovery metadata to the SOLAS IPO and Project Integrator for SOLAS-relevant projects.
3. Ensure timely submission of collected data to their designated data centre.

Timescales for data submission

1. In order that the SOLAS community can rapidly benefit, information on data collection activities (see item 1 under SOLAS IPO) should be submitted to the SOLAS IPO as soon as known.
2. Subsequent to project completion, short summary reports (including when, where and how each dataset was collected and the location of that data) should be submitted to the SOLAS IPO as soon as possible.
3. Finalised data should be submitted to the appropriate designated data centre as soon as possible. Nations without oceanographic data centres and/or lacking funds to build and maintain data management facilities are encouraged to submit their datasets to data centres elsewhere.

Data release

Data release can be split into two categories: (1) release to other participants within the project; and (2) public release. In general, data release will be controlled by the relevant national data centre's policy agreement. SOLAS recommends that public release normally be two years from the end of the cruise or field activity, but could be extended where, for example, analytical procedures have inherent built-in delays. SOLAS strongly supports co-authorship for any data contributions that result in publication.

Validation and quality control

Individual PIs are responsible for the quality control of their data and should provide notes to the relevant national data centre about any doubtful data values in their metadata. Questionable data should be flagged rather than discarded. Participants using the data should report questionable data back to the relevant national data centre (who will also be noting problems) and to the data originator.

Experience shows that data quality problems are often revealed once the data begin to be used scientifically, frequently by individuals other than the data collector. Comparison with other parameters or comparison between datasets at the same location can reveal errors or offsets. This is one good reason for making data available to other participants as quickly as possible.

More formally, groups of PIs expert in particular compounds or substances will be encouraged to apply further quality controls, such as intercalibration.

Archiving

The data from SOLAS must ultimately be preserved for posterity. SOLAS strongly recommends that data are archived in a relevant location with secure long term funding (e.g. a national or world data centre).

SOLAS Project Integration

SOLAS Project Integration (led by Dr Tom Bell; www.bodc.ac.uk/solas_integration/) intends to work with the scientific community to collate global datasets of SOLAS-relevant compounds and particles in order to create global flux and concentration fields. In addition to the policy above, this work has adopted the following policy:

1. ALL data providers will be co-authors in any resulting publication.
2. Data contributed to the project will ONLY be used to derive fluxes and concentration fields.

Following the collation of the global concentration datasets, a dialogue with the contributing scientific community will decide their long-term fate. However, SOLAS strongly recommends that such information be archived in a relevant location with secure long term funding (e.g. a national or world data centre).

Appendix

Special requirements for different data products (or Classes)

The implementation of SOLAS science at the international level will provide and require the following general 'Classes' of data products:

1. *Streams of geographically/temporally resolved data (Class 1)*

The research activities within various field campaign provinces and on a global scale are likely to be quasi-independent of each other (i.e. differing scientific emphases, research groups, different mix of observational approaches). Nevertheless an overall SOLAS synthesis will require comparison and integration across these studies, implying that a common data management approach is required.

Activities in these studies will produce or make use of:

- Hydrographic data collection from ships, ocean time-series and autonomous ocean platforms;
- Remote sensing data from a wide range of sensors and satellites;
- Time-series meteorological and atmospheric chemistry data collected from fixed-sites;
- Data collected from aircraft and balloons;
- Data collected from volunteer observing ships;
- Data products from operational ocean and atmosphere models.
- Relevant climatologies.

2. *Data collected from experiments and mechanistic investigations (Class 2)*

These include results from laboratory, mesocosm studies and from field-based process studies (e.g. of gas exchange). For the former, the organizational principle will frequently be the experimental treatment, conditions or approach applied in an experiment, rather than the specific geographical location and time of data collection. For the latter, large and varied datasets are projected to arise from complex campaigns conducted within a number of geographically focussed 'Field Campaign Provinces'. Such campaigns will address subsets of SOLAS questions using different approaches and mixture of observational techniques. For such process studies, it is important that relevant data from different sources can be readily accessed, combined, and archived.

The data from such studies are frequently associated with some deliberate manipulation of environmental conditions: this implies that the details of the manipulation should be associated (as metadata) with individual data sets. Another source of *Class 2* data are field-based process studies. In such studies (e.g. eddy correlation studies or studies of upper ocean mixing), very large amounts of specialized data are collected, often at one location over discontinuous periods of time. In some cases, data are collected with custom-made, specialized equipment, the characteristics of which may change over the course of the experiment or process study. This poses special challenges for data management beyond the situation for *Class 1* data.

With these more specialised studies, it is also less clear *a priori* what types and levels of data should be reported and archived. This depends on the potential significance of the individual data streams for other investigators. For many laboratory experiments, the normal practice of writing papers that include summaries of experimental data (in figures and tables), together with a detailed description of the experiment (as a methods section) will remain adequate for SOLAS needs.

Two examples of experimental approaches that would benefit from organising and managing their data in compatible and accessible forms include:

- A. Mesocosm and mesoscale patch experiments. Here there is a need for the investigators themselves to have access to different data streams from within these studies. Larger-scale synthesis will also be promoted if SOLAS investigators can readily compare the results of different experimental treatments conducted by different groups worldwide.
- B. Studies of gas exchange. In these studies, large amounts of diverse data are collected and need to be accessed by a range of investigators. In addition, it will be useful to be able to compare raw data collected in one experiment under certain conditions with data collected by other groups under different conditions.

Most importantly, *Class 2* implies a need for flexible data management to handle differing amounts of non-uniform data and complex metadata (e.g. experimental protocols/conditions) and such results should be archived in the same location(s) as the core activity data (*Classes 1 and 3*) as appropriate. A final summary report should be produced for each process study.

3. *Models, model documentation and model output (Class 3)*

Key model output used and produced by SOLAS investigators should be accessible to the wider SOLAS community. In addition, SOLAS should try to ensure that useful products of non-SOLAS modelling can be readily accessed. This includes operational forecasting / nowcasting models of the atmosphere and ocean, data assimilation models and inverse models; and reanalysis models. Ideally, the output should be accessible using the same tools as for some of the *Class 1* data products.