

# **SOLAS Data Management: Implementation Plan**

## **Background and Needs**

The SOLAS Science Plan and Implementation Strategy identified data and model management as critical logistical tasks for SOLAS. The implementation of SOLAS will involve the collection of large quantities of environmental data by both nationally and internationally organized 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. Presently, SOLAS envisions significant field activity in roughly 9 'Field Campaign Provinces'. Similarly, SOLAS will make use of a hierarchy of modeling 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 combined with, or compared against, 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.

## **SOLAS Data Management Principles**

:

1. Do not 'reinvent the wheel'; use existing knowledge and infrastructure, wherever appropriate.
2. Use internationally agreed standards and protocols (e.g. those of ISO, W3C, IOC/ICES), wherever possible.
3. Work with other projects towards establishing an integrated data management system and policy.
4. Plan ahead for rapid data assembly.
5. Data managers should support data gatherers.
6. Data should be made available rapidly. SOLAS should reward excellence in data collection and data release. Data users should consult and collaborate with data providers.
7. Data should be reported together with any information concerning its quality (i.e. with data quality metadata and quality-flags).
8. Model documentation, model output and, when appropriate, the models themselves should be made available to the community.
9. Participation in SOLAS research requires, when appropriate, submission of data to a SOLAS-approved database or centre.

## **The Data Management Team**

The complex nature of data management in SOLAS requires planning and oversight by a multi-disciplinary team of experienced scientists and data managers. This team should, in turn, be well-connected with data management organizations in relevant WCRP and IGBP international programs.

The Data Management Team will manage the following tasks and activities (based on suggestions in the SP&IS):

- Evaluate and document, based on the Focus 1, 2 and 3 Implementation Plans, the likely data products, and requirements of SOLAS (data types, quantities and sources, metadata requirements).
- Write and establish a data reporting policy to include time limits for data reporting, incentive and enforcement procedures associated with data reporting, access rights to data, meta-data requirements and principles associated with data reporting, data quality assessment procedures, and long-term archiving issues.
- Develop and implement a practical policy for the documentation of models and model-derived products developed by, or used for, SOLAS science.
- Based on consideration of data needs and the capabilities of existing data centres: Design and implement a data reporting and retrieval structure that is efficient and compatible with the data organization and scientific issues underlying SOLAS observational and experimental campaigns.
- Conduct negotiations with national SOLAS PIs and national funding agencies to organize, fund and implement the data management structure.
- Write a detailed SOLAS data management and quality management manual. Potentially, this document to include guides to reporting procedures, quality assessment procedures, data and model access procedures.

## **Towards Implementation**

The principles of SOLAS data management listed above imply that data should be reported rapidly, archived and made available to the community. This process should be efficient: both for the data provider and the data user. The recommended implementation steps depend on the origin and nature of the data.

### ***Data Classes:***

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

1. Geographical/temporally resolved field data. Large complex data sets are likely to arise from complex campaigns conducted within a number of geographically-focussed 'Field Campaign Provinces'. Each campaign will address different subsets of SOLAS questions using different approaches and mixes of observational techniques. In addition, a more limited set of data products from global measurement systems (large-scale surveys; remote-sensing products) must be handled.
2. Data collected from diverse experimental and mechanistic investigations. These include results from laboratory and mesocosm studies but also, potentially, from field-based process studies (e.g. of gas exchange). For these data, the organizational principle will frequently be the experimental treatment, conditions or approach applied, rather than the specific geographical location and time of data collection.
3. Products relating to modelling studies including data synthesis, data inversion and theoretical approaches conducted on regional to global scales.

Components 1 and 3 imply a need for geographical and temporal (x,y,z,t) organization of diverse data types. This is a common organizational principle for several large international programs, and significant capability exists that can be utilized by SOLAS. Component 2 implies a need for flexible data management to handle differing amounts of non-uniform data and complex metadata (e.g. experimental protocols/conditions). This sort of data stream has not been so commonly reported in the past so that some significant development work may be required. Component 3 shares significant commonality with the needs of other programmes.

**Class 1: Geographically/temporally resolved data.**

A Common Approach: The research activities within the various Field Campaign Provinces, and on the 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 between, 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.

<b>Data Type</b>	<b>Data Dimensions</b>	<b>Atmos./ Ocean</b>	<b>Existing Data Centres?</b>	<b>Addl. SOLAS Requirements?</b>
Ship-based Profile	Discrete x,y,z,t	Ocean	Several (WOCE/ CLIVAR/ JGOFS legacy)	Incorporate new parameters. Expand metadata.
Autonomous Profile	Semi-continuous x,y,z,t	Ocean	ARGO/ CORIOLIS/ other?	Incorporate new parameters. Expand metadata.
Balloon Profile	Discrete x,y,z,t	Atmos.	Yes... where?	??
Underway/ VOS Data	Continuous x,y,t	Both	Yes. CDIAC/ CORIOLIS/ other?	Incorporate new parameters. Expand metadata.
Time-Series Data	Continuous z,t	Both	Yes. WOCE/JGOFS legacy plus atmosphere data sites	Incorporate new parameters. Expand metadata; easily combine Atmos-Ocean data streams.
Aircraft Data	Continuous x,y,z,t	Atmos.	Yes (give examples)	??
Operational models	Continuous x,y,z,t	Both	Yes (ECMWF/MERCATOR/FO AM/etc..)	Model output must be archived for future use ; easily combine different model products
Remote Sensing	Continuous x,y,z,t	Both	Yes	

For most of the above data streams, there are established international protocols and data centers that have been developed for prior projects and/or international monitoring activities. These should form the basis for SOLAS data management needs and only in exceptional circumstances should independent SOLAS-solutions be developed in this area. Some extension of present activities and capabilities will however be required for SOLAS-specific needs (see Table). The implementation tasks are:

1. Identify the protocols and data centers that can potentially handle SOLAS data needs:
2. Identify any current limitations of these existing protocols/centers with respect to SOLAS needs. (e.g. ability to handle the diverse range of chemical/biological parameters likely to be measured in SOLAS; Ability to provide the required level of data quality information).
3. Negotiate with these data centers concerning their present or future ability to handle/accept/serve the diverse SOLAS data streams.
4. Inform the SOLAS community about the centers and protocols that have been selected for SOLAS data management.
5. Conduct an assessment and develop practical tools so that SOLAS data can be **efficiently and easily** retrieved from these centers. Critical is that data from particular streams can be **easily and efficiently** re-combined with other SOLAS data streams for a particular Field Campaign Province (e.g. via clever use of Live-Access-Server technology, OPENdap, etc).
6. Develop and provide software tools, and **conduct international training workshops for non-specialists** in order that PIs are able to understand and use the reporting/retrieval structures. Software tools must be compatible/interfaceable with software commonly used by the individual SOLAS research groups (e.g. by the biological and chemical communities).

Despite the wide-availability of data centers, there will remain some SOLAS geographically-focussed data streams (e.g. large data sets from high-frequency measurements made in the context of process studies) that will not be readily compatible with existing structures. This may include data collected from experimental studies conducted within Focus 2 and perhaps Focus 1 of SOLAS (e.g. micrometeorological data associated with gas exchange process studies ). These types of data streams share more in common with Class 2 data.

### ***Class 2: Data from Experiments and Mechanistic Studies.***

Experiments and process studies are frequently site-specific and the resulting data are not necessarily time-continuous. As a result, these data do not necessarily match the data dimensions of Class 1 data (see Table) and are not so obviously compatible with existing data centers. Here some independently-developed solutions to meet SOLAS requirements may be required.

The data from such studies are usually associated with some deliberate manipulation of the environmental conditions: this implies that the details of the manipulation should be associated (as metadata) with individual data points. (This is fundamentally different from the Class 1 data, where it is common to have identical metadata assigned to very large amounts of data). Another source of 'special' 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, usually over short periods of time. Again with such data, detailed documentation (metadata) concerning experimental conditions is required.

With these more specialized 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. In other cases, the direct availability of raw data from one experiment may be of immediate and direct use for SOLAS investigators planning or conducting similar experiments elsewhere. In this case, a more organized and 'harmonized' data management procedure will be beneficial. Two examples of experimental approaches that would benefit from making raw data available in compatible forms include:

- 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.
- 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.

The implementation tasks required depend on the nature of the activity:

A: Experimental approaches that will be conducted by multiple groups of investigators worldwide (e.g. phytoplankton growth experiments; mesocosm experiments; gas exchange studies; etc):

1. The relevant Implementation Panels and Groups (Task Teams????) should, as part of their planning process, make written recommendations concerning the types of data that should be reported and archived.
2. These should include recommendations concerning the specific and essential metadata that should be reported with the data.
3. The recommendations can be in the context of agreed-upon and documented experimental protocols and should be widely available to SOLAS investigators (e.g. via web-access).
4. There should be consultation with the Data Management Team during the preparation of these recommendations. The SOLAS SSC should also ultimately review the recommendations.
5. The relevant Implementation Panel (Task Team????) should examine whether there is a need for central organization, across SOLAS, of the data holdings arising from such studies. (The level of organization will depend on the anticipated use of the data: it could range from simple web-based description of experiments together with contact details for PIs to an agreement to report all data to a common data centre for long-term access and archiving).

B. Individual and group experiments that are unlikely to be replicated or directly linked to present or future activities of other groups:

1. Each SOLAS process study and/or experiment should consider its data management requirements at the time of designing the experiment: consideration should be given not only to the immediate needs of the group involved, but also to the potential relevance and utility of the raw data to other investigators.
2. Advice on this can be provided by the Data Management Team.
3. In areas overlapping with areas covered by the SOLAS ???Task Teams???, consultation and information exchange with the Task Team Chair on the recommended level of data management is recommended prior to initiating a study.

### ***Class 3: Models, Model Documentation and Model Output.***

HERE I NEED A LOT OF HELP!!

As with data, the models and model output that are produced by SOLAS investigators should be accessible by the SOLAS community. In addition, SOLAS should ensure that useful products of non-SOLAS can be readily accessed.

A: Models that are capable of being run by individuals or small consortia.

1. Models should be documented and model code made available by individual PIs over the internet for use by the SOLAS community.
2. The SOLAS web-page should maintain links and documentation concerning such resources.
3. As with data, proprietary intellectual ownership issues arise and the interests of the model developer must be protected. WHO DECIDES HOW TO DO THIS? DO WE NEED A POLICY? IS THERE AN APPROPRIATE TASK TEAM? WHO ENFORCES THIS???

B: models that require significant hardware resources, and models operated outside of SOLAS.

This includes operational forecasting/ nowcasting models of the atmosphere and ocean, data assimilation models and inverse models; and reanalysis models. The major SOLAS data management issue is to assure efficient access to model output. In some cases (e.g. operational models) this may require development of a SOLAS-specific strategies for storage of model output. Ideally, the output should be accessible using the same tools as for some of the Class 1 data products.

Specific implementation tasks include:

1. Identify the SOLAS-relevant models and model products that fall under this category.
2. Critically assess the suitability of existing availability and access procedures to meet SOLAS needs: is improved access necessary?
3. Where improved access to model output is required, identify an efficient strategy, including consideration of utilizing the tools and data management procedures developed for handling Class 1 data types (see above).
4. SOLAS-relevant models, output and associated access procedures should be clearly described on the SOLAS web-page.
5. SOLAS should, where demand exists, organize training workshops concerning the use of such model products.