UNIVERSITY of TASMANIA | IMAS INSTITUTE FOR MARINE & ANTARCTIC STUDIES

Data Management Guidelines

Contents

Introduction

This document is the Data Management Guidelines for the Institute for Marine and Antarctic Studies (IMAS) and should be read in conjunction with the IMAS Data Management Policy. It is the responsibility of all staff and students at IMAS to ensure they are familiar with this manual and the associated IMAS Data Management Policy.

This document provides a comprehensive overview of management, storage and archival procedures for Research Data. A reduced set of practical guidelines is available as a single page document in Appendix E and provides a quick reference for *Managing your data through the project lifecycle*.

Administration

Data Management Plan

All IMAS research projects must include a Data Management Plan (see template in Appendix B of this document). Contact the IMAS Data Manager if you require assistance.

Cost Recovery

Data Management work carried out at IMAS is not currently core funded. All project proposals for externally funded projects must therefore include a budget item for the recoupment of internal costs associated with Data Management.

UTAS has formulated an online Budget Preparation Costing Tool to assist with this process. Contact the Data Manager if you require further assistance estimating data management costs associated with a project proposal.

Destruction of Data

Unless required by overriding legislation or policy, or by contractual agreement, Data should be kept indefinitely. Destruction of IMAS Data must be approved by the relevant IMAS Centre Director.

Roles and Responsibilities

Researchers

All Researchers (which includes both Staff and RHD Students) must:

- provide a Data Management Plan for all projects (Appendix B)
- ensure Data is accurate, and recorded and documented in a format that is adequate for verification of research results (see Data Storage)
- ensure all research contracts or agreements adequately address ownership and licensing of Data and, where applicable, meet usage conditions set by third parties
- not use or access Restricted Data without authorisation or, where applicable, allow another person to use or access Data without authorisation

- unless otherwise prohibited by contractual obligations of IMAS, submit Data and associated Metadata at agreed project milestones and/or upon project completion (see Data Archival and Metadata)
- comply with all aspects of the IMAS Data Management Policy and other University policies where applicable

Contact the Data Manger if you require advice or assistance to meet any of these responsibilities.

Intellectual Property

All data collected at IMAS will be managed in accordance with the University of Tasmania Intellectual Property Policy.

Staff

All Intellectual Property created by Staff in the course of their employment duties is owned by the University (except in circumstances where the University has assigned the Intellectual Property to another party, e.g. a funding body or for the purposes of commercialisation).

Students

Intellectual Property created by Students is owned by the Student (unless they choose to assign it to another party). In most cases, Students undertaking University research are requested to enter into a Deed of Assignment in accordance with the University of Tasmania Intellectual Property Policy. In the case of PhD research not supported with Commonwealth or industry funding, students may choose not to enter into a Deed of Assignment and will retain sole ownership of their Intellectual Property.

Students required to enter into a Deed of Assignment must be given the opportunity to seek their own legal advice.

Data Storage

For the purposes of this manual, Data storage is divided into two categories:

- **Files** are data stored in disk files (e.g. text files, images, PDFs, Microsoft Excel spreadsheets, Microsoft Access databases, NetCDFs)
- **Databases** are data stored in structured storage, often with multiple pages or tables and a relational structure (e.g. Microsoft SQL Servers or Oracle RDBMS)

Files (stored on disk) containing Data collected at IMAS must be stored on the R: drive in either folders assigned to Sections or the Data folder. The R: drive should be considered a general working area for all IMAS data. Data collected on an ongoing basis may also be stored and managed using a Database Server developed in consultation with the Database Manager (see Database Technologies).

Large sets of empirical Data (e.g. >1 TB) should be stored in RDSI. Upon completion of a project all IMAS Data will be archived in RDSI storage.

Data Storage Technologies

Database Servers will be limited to the following technologies:

- Microsoft SQL Server
- Oracle RDBMS

See Appendix C for guidelines on naming and structuring relational databases.

Non server-based technologies for managing Data will be limited to the following:

- Microsoft Access
- Microsoft Excel
- NetCDF

See Appendix D for guidelines on structuring and formatting tabular Data.

Data Archival

Data Archival

All Data collected by IMAS researchers must be released into the public domain following completion of a project (with the exception of Restricted Data[1]). A project is deemed to be complete when final report(s) are submitted to the relevant funding body. Where further publications are planned, Data may be retained for an additional period of up to 24 months following project completion before it is released into the public domain.

[1] For an accurate definition of Restricted Data, please refer to the IMAS Data Management policy or ask the Data Manager. Most data collected at IMAS is *not* Restricted Data and will be adequately protected by a Creative Commons License (http://creativecommons.org.au/learn/licences).

Supported Archive File Formats

Archived Data must be in one of the following non-versioned file formats:

- **Documents:** PDF
- **Data:** CSV**,** NetCDF
- **Images:** TIFF, JPEG, PNG

Metadata

Metadata for all IMAS Data must conform to the Marine Community Profile of ISO 19115. Metadata must include the following minimum requirement to describe database structures and measured data parameters:

- Schema diagrams depicting all tables, table relationships and field names; descriptions of all tables, primary and foreign keys and indexes [Databases only]
- Description of all measured parameters including units of measure and interpretation where required [all Data]

IMAS Data Submission Tool

The IMAS Data Submission Tool (DaST) is an online tool designed to assist researchers with the creation of Metadata and submission of research Data to the IMAS Data Portal. The accompanying DaST User Guide provides additional guidance for use of the tool.

Appendix A: Useful Links

Policies and Supporting Information

- IMAS Data Management Policy (link)
- UTAS Management of Research Data Policy (link)
- UTAS Intellectual Property Policy (link)
- additional UTAS Intellectual Property information (link)
- Australian Code for the Responsible Conduct of Research (link)

Budgeting and Cost Recovery Guidelines

- UTAS online Budget Preparation Costing Tool (link)
- IMAS IT, Data Management and Information Systems Cost Recovery Guidelines [coming soon]

Discovering and Submitting Data

- IMAS Data Portal (link)
- IMAS Data Submission Tool (including User Guide - link)

Appendix B: Data Management Plan Template

IMAS Data Management Plan

Project ID

Existing Data
Outline existing data that will be used in this project.

New Data
Outline new data for this project, including a brief outline of the proposed collection and processing methods used.

IP Owner

Data Organisation, Formats and Storage

Outline organisation storage methods and technology to be used (including anticipated file/database formats).

Access and Security

Outline methods used to ensure adequate access control and backup security.

Archival

Outline data archiving measures to be used, including anticipated file formats, when the data will be archived, and any access restrictions required.

Appendix C: Database Design and Naming Conventions

The naming conventions described here are designed to meet the naming limits of databases used at IMAS. While some technologies allow you to exceed the limits given, you should remain within those limits to ensure data can be easily transferred to another technology if required.

However, for Excel spreadsheets, these conventions may be used as a loose guide and users are free to vary as needed (Excel spreadsheets are normally used for small datasets where conversion to a technology requiring adherence to these naming conventions is generally trivial).

Table Names

Table names must conform to the following standard:

- Must only use letters, numbers and the underscore character.
- Must be prefixed with a two character database identifier (see the Data Manager to obtain an identifier for a new database).
- Table names should attempt to fully describe the data. Don't use meaningless names or abbreviations.
- Table must include a full description of their use and content.

Table Design

Table design must always incorporate the following characteristics:

- Must have a primary key of data type integer, which is not null and uses a sequential number system, named "ID"
- All foreign keys (fields with an established relationship to another table) must be named with the name of the related table and suffixed with "_ID". Eg. A foreign key field for a table named DM_WIDGET will be called DM_WIDGET_ID.
- Fields must include a full description and include the units of measure and interpretation of codes if required.
- For databases that don't support Boolean (yes/no, on/off) values, use a small integer field type with value 0 meaning No/Off and value 1 meaning Yes/On

Field Names

Field name must conform to the following standard:

- Must only use letters, numbers and the underscore character.
- Must use a letter (A-Z) as the first character.
- Must not exceed 30 characters in length.
- Field names should attempt to fully describe the data. Don't use meaningless names or abbreviations.

Stored Procedures and Functions

- Stored procedures and functions related to a single table must be named with the name of the related table followed by a suffix indicating its function. Eg. "_SEL", "_MAINT", "_IMPORT".
- Stored procedures and functions *not* related to a single table must be named with the two character database prefix (see "Table Names" above) followed by a suffix indicating its function.

Appendix D: Tips on Structuring Your Data

The primary consideration when structuring your Data for public archival is ensuring it can be clearly understood when read in combination with the associated Metadata record. Data should be as complete as possible, and any notes related to the Data should be contained within the Metadata record (or associated readme file) rather than the Data itself. The following tips for data structure have been developed specifically for CSV data (typically complied as an Excel spreadsheet and exported to CSV format), but can be used as a general set of guidelines for any tabular-style data.

Content

- Use standard SI units wherever possible.
- Avoid specifying units of measurement in the data itself.
    - e.g. lobster length in mm should be specified in numerical format only, with units of measurement (mm) clearly described in the Metadata
    - note that column headings need not contain units of measurement if this is adequately described in the Metadata
- Column headings must consist of a single row only.
- All columns must be named with a clear description of the data contained in that column. Avoid giving units of measurement in the column heading itself. Use only standard characters (letters, numbers, spaces, and the underscore character). Non-standard characters (\ / : * ? , " < > () . |) are *not* appropriate for CSV archival.

Time/date formats and spatial information

- Data with a spatial context should include columns specifying latitude and longitude so that each row corresponds with a spatial coordinate.
- Latitude and longitudes should be expressed in decimal degrees format. Degrees minutes or degrees minutes seconds formats should be converted before archival.
    - Southern latitudes and Western longitudes should be expressed in a negative format; e.g. 37°17’57.3” S, 63°58’25.1” W should be converted to the format -37.29926, -63.97365
- Times and dates should be separated into different columns and expressed in a standard format. We suggest dd/mm/yyyy or dd-mm-yyyy and hh:mm:ss (24-hour format is preferable for times).

General formatting

- Each column should consist of the same number of filled rows. No cells should be left blank – use “0” to indicate zero data, or a symbol such as “.” or “NA” to indicate missing values.
- Cell values must *not* contain commas (,) or semi-colons (;), as CSV file format uses these characters to separate columns.

The table below provides an example of a data excerpt containing some common formatting mistakes (**Table 1**), and the same data set re-structured to the desired standardised format (**Table 2**).

Table 1. Sample data excerpt containing common formatting errors.

				Benthic invertebrates			Algae	
	Start longitude	Start latitude	Depth (inshore)	Centro **cryptic**	slug things (ask Jim)	RL	Eck	Phyll
21st Mar, 2008	148 56 55.7E	41 09 14.0S	11.0	0	0	0	30%	70%
			11.0	0			20%	80%
			11.1	0	2	2	40%	60%
			11.0	0	1	0	60%	40%
			10.6	3	0	0	10%	90%
			10.2	5	4	1	60%	40%
			9.8	3	5	0	80%	20%
			10.1	1		0	80%	20%
			9.7	2	0	0	90%	10%
			9.4	3	1	0	100%	0%

	PROBLEM	**SOLUTION**
	= empty cells	Fill blanks with 0 or character to indicate no data
	= unnecessary header row	Restrict header to single row
	= non-standard format for date and latitude/longitude	Use standard format for date and decimal degrees for latitude/longitude
	= non-standard characters and non-descriptive column headings	Use only standard characters in column headings, keep titles informative and descriptive.
	= units of measurement expressed in data cells	Remove % from data cells and specify “% cover” as unit of measurement in Metadata

Table 2. Sample data excerpt (same as above) with correct structure and formatting.

Date	Start longitude	Start latitude	Depth inshore	Centro cryptic	Abalone	Rock lobster	Ecklonia	Phyllospora
21/03/2008	148.9488	-41.1539	11.0	0	0	0	30	70
21/03/2008	148.9488	-41.1539	11.0	0	NA	NA	20	80
21/03/2008	148.9488	-41.1539	11.1	0	2	2	40	60
21/03/2008	148.9488	-41.1539	11.0	0	1	0	60	40
21/03/2008	148.9488	-41.1539	10.6	3	0	0	10	90
21/03/2008	148.9488	-41.1539	10.2	5	4	1	60	40
21/03/2008	148.9488	-41.1539	9.8	3	5	0	80	20
21/03/2008	148.9488	-41.1539	10.1	1	NA	0	80	20
21/03/2008	148.9488	-41.1539	9.7	2	0	0	90	10
21/03/2008	148.9488	-41.1539	9.4	3	1	0	100	0

Appendix E: Managing Your Data Through The Project Lifecycle