

TEMPLATE FOR DATA REVIEW (Diepenbroek and Wiebe)

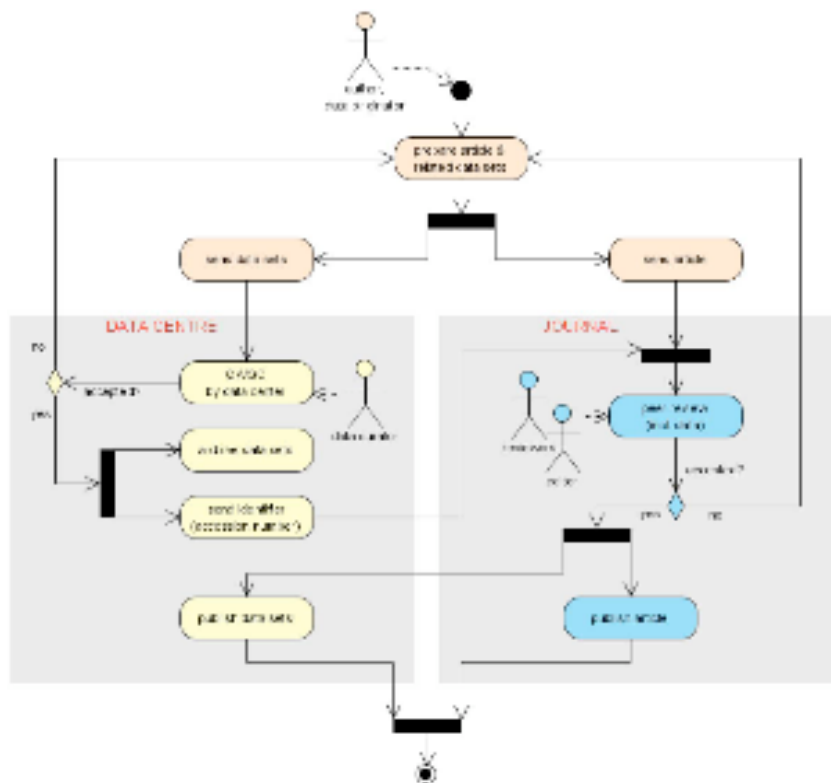
1. Scope of the use cases: supplementary data and stand-alone data publications.

There are two distinct forms of data publication considered below. One involves the publication of the background or “backbone” data that were used to prepare the figures, tables, and statistical analyses reported in an associated paper. The second involves the publication of a summary document that describes and presents the results of original research data sets as a stand-alone publication that is intended to foster knowledge about the data set(s) existence and their re-use.

a. In essence there are no real boundaries on the types of data that may be subject to publication in support of a publication. All data sets and associated metadata required to create the publication’s figures, tables, and summary statistics are included. For stand-alone data publications, the data set(s) must have relevance and re-use purposes beyond those of the originators.

b. A number of individuals will be involved in the process of data publication; included are the authors, reviewers, editors, data curators, and users. Each of their roles will be briefly described in what follows.

2. Work-flows for the different types of data publication (supplementary versus standalone)



The workflow, illustrated in the above figure (taken from the SCOR/IODE Workshop on data publishing Workshop Report No. 207), for both cases may be summarised briefly as follows: The author (data originator) sends the data (including the necessary metadata) to a certified data centre. The data centre supplies a basic level of QA/QC. Once the data are accepted, a persistent identifier (accession number) will be assigned and sent back to the data originator. The article and the persistent identifier for the data are sent to the journal. Peer review of the article will include review of related data sets. After acceptance, both article and data sets will be published and cross-referenced.

3. Writers guideline (authors role)

- a. For supplementary data. (Revised from the guidelines in first report).

Table 1: Draft guidelines for content contributors

Increasingly, journals in all areas of science are requiring that all data that are required "...to understand, assess, and extend the conclusions of the manuscript..." (Science) be made available to any reader of the paper. But journals that have such mandates, by-in-large, offer limited or generalized guidance on what the nature of the data should be, what suitable repositories exist, or what the process is to accomplish a data submission. The following is an outline of the set of steps that can be carried out to fulfill the mandate.

- 1) Background data that were used to create summary tables, statistics, and figures in the paper should be provided in suitable format (preferably in a non-proprietary form and with sufficient metadata to support accurate reuse) in order that a reader is able to reproduce a summary table, statistics, or figure.[Footnote 1].

- 2) Where substantial processing of an original data set is done to produce a resulting product, the modified data set should be provided. For example, for irregular spaced data, a common technique is to "krig" or objectively map the data to a regular spaced grid and then present the gridded data as a contour plot. The gridded data, a table of the parameter values used to do the objective mapping, an accounting of the software used, and adequate (good enough to allow reproduction of the processing if required) references to the original data should be provided. [Footnote 2].

- 3) In order to make the data discoverable and therefore useful for other purposes, additional information or metadata in standardized form may be needed. It is recognized that a standardized form may not yet exist for certain types of contributions. For example, Dublin Core could be used for publications and the Directory Interchange Format (DIF - GCMD) could be used for data. [Footnote 3].

- 4) Data and associated metadata should be submitted to a repository

conforming to the IODE technical specification prior to manuscript submission. A list of possible repositories is provided at URL. [Footnote 4].

- b. For standalone data publications (building on the Earth System Science Data. The Data Publishing Journal template). [www.earth-syst-sci-data-discuss.net]

Table 2: Template for Stand-alone Data Publication.

The following is generic listing of topics to be covered in a stand-alone data publication:

Title:

Authors:

Abstract:

Data Coverage and parameter(s) measured

1) Introduction

2) Instrumentation

3) Dataset(s)

4) Data Access

Acknowledgments

References

4. Editors role (data editor?): A journal may elect to assign a Data Editor as well as a manuscript editor to a submitted paper in the case where an elaborate data set(s) is involved in the publication. In such a case, the Data Editor would provide the author(s) with information about the procedure for submitting a data set(s) and suggestions for appropriate accredited data repositories to which the data could be submitted if the author was not aware of one. The editor would track the submission of the backbone data to a repository and provide the manuscript editor with the information that an “accession” number or d-space identifier had been assigned to the data set, where upon the manuscript editor would send the manuscript out to reviewers.

5. Reviewers role: In addition to receiving the manuscript from the editor, the reviewer will also be provided access to the backbone data. In the course of evaluating the manuscript, the reviewer may elect to use the data for verification and substantiation of publication results, or to check the validity of methods used through evaluation of the accompanying metadata. Reviewers may also review the originality, lineage, and usability of the data. Usability covers issues such as format readability and the quality and completeness of metadata, particularly metadata describing what the measurements mean.

6. Data curator (data centre) (Michael)

- a. Technical QA/QC for the data sets (formats, completeness of metadata, vocabulary, spelling, etc). Data sets will be checked for completeness

and correctness of metadata. Abnormalities (e.g. outliers) in the data will be identified and reviewed with the data originator. The degree of quality control necessary will depend upon the origin of the data set(s). If the data set(s) is a subset of an existing data set resident in a data center, modest review may be all that is required. If the data are original and not previously stored in a data center, a more thorough vetting of the data set will be required.

b. QA/QC on higher levels in particular using more sophisticated methods (as e.g. checking variances of longer time series includes the data submitted)

7. Metadata : Metadata are required for discovery of data sets, their evaluation in terms of use appropriateness, and re-use of the data. These three elements (Discovery, Evaluation, Re-Use) need increasing levels of detailed metadata descriptions. Dublin Core, ISO19115, or DIF metadata are sufficient for data discovery, but lack the information for evaluation. FGDC is OK for some kinds of evaluations, but not all. The Open Geospatial Consortium (OGC) approved version 1.0 of the Observations and Measurements Encoding specification may provide the necessary metadata for evaluation of data sets. [<http://xml.coverpages.org/ni2008-01-31-a.html>]. There are no existing agreed-upon standards in place to specify the metadata needed for re-use of data-sets. [see Footnote 3 for additional discussion]

8. Citation syntax (Michael) for Data publications:

Author(s), Date, Title of Data Publication, DOI, Supplement to: Author(s), Title, Date, Title of paper, Journal, volume, pages, DOI

Example: Hillenbrand, C-D et al. (2008): Sedimentology of various sediment cores from the West Antarctic continental margin, doi:10.1594/PANGAEA.671520, Supplement to: Hillenbrand, Claus-Dieter; Moreton, Steven Grahame; Caburlotto, Andrea; Pudsey, Carol J; Lucchi, Renata G; Smellie, John L; Benetti, Sara; Grobe, Hannes; Hunt, John B; Larter, Robert D (2008): Volcanic time-markers for marine isotopic stages 6 and 5 in Southern Ocean sediments and Antarctic ice cores: implications for tephra correlations between palaeoclimatic records, Quaternary Science Reviews, 27(5-6), 518-540, doi:10.1016/j.quascirev.2007.11.009

9. Identifier (Michael)

10. Examples (Peter, Michael)

a. Demonstration of selected data sets from PANGAEA (Michael)

- b. Description of practical scenarios (Appendix I.)
11. Metadata for data review (includes discovery metadata - may be given as links for further sources):
- a. Technical level:
 - Standard being used (ISO19115, DIF, FGDC, (DC), Darwin Core)
 - Completeness of meta-information (who, what, how, when, where)
 - b. Semantical level:
 - Validity of methods used
 - Values: precision, range (outliers)

[Footnote 1]: In what form data would be sent to a repository is an open question. Although there is a desire by data managers to receive data in a small number of “standard formats”, this was not realistic. The diversity of instruments, data acquisition methods and data types, and the difficulty of (and lack of funds to) reformat data, would cause the manuscript authors to contribute their data in a format convenient to them, not necessarily in a format convenient to or required by a data repository. It also may be difficult to contribute certain kinds of data in non-proprietary forms since they may have been acquired with commercial software packages. In addition, authors may have prepared (or processed) their data using commercial software (e.g. Golden Software’s Surfer contouring program) and that a subsequent reader might find it difficult to recreate a plot or graph if that version of the software was no longer available.

[Footnote 2]: However, it must be recognized that providing the parameters used to process data would likely be meaningless without detailed knowledge of the processing software. Even non-proprietary software, like EasyKrig, requires the commercial software Matlab to function, and is subject to the same issues of version release (D. Chu, “The GLOBEC Kriging Software Package – EasyKrig3.0”, http://globec.who.edu/software/kriging/easy_krig/easy_krig.html, July 15, 2004).

Footnote 3]: The metadata records discussed are fine for data or paper discovery, but they are not “use metadata”. They are inadequate to provide another person with sufficient information about the data to recreate a plot. For example, if several data sets were used in the creation of a summary table, use metadata would include detailed information about the meaning of the data fields, their units, how they were collected, and how they were processed, to ensure proper combining. A key question is ‘Does a data repository also store the necessary “use metadata” that would make the data stored in support of a particular contribution reusable?’ Perhaps not, but the data repository should contain sufficient information to find such use metadata, however.

Furthermore, without sufficient use metadata, data reuse can easily become data misuse, either intentionally or unintentionally. This requires sufficient effort be made to provide essential “use metadata” to minimize this possibility. The issue about metadata is thus not a simple one. The Marine Metadata Interoperability Initiative (MMI) has been attempting to develop a standard for over three years. They have developed strategies and have identified the existing standards, but fully documenting data sets is still a difficult task, with no single, agreed upon standard for accomplishing it.

[Footnote 4]: No such list of possible repositories currently exists. The issues of what is a “data repository?”, “how does it relate to the storage of original data sets (the ‘bits’)?”, and “what is its relationship to a ‘data center’?” are not resolved. The attributes of a trusted repository include persistence, stability, vetted backup procedures, and fully follow required and accepted metadata standards. Some sort of an accreditation or certification process needs to be defined as described in the SCOR/IODE Workshop on data publishing Workshop Report No. 207. A repository would hold the data that was presented or used within a particular publication. Oftentimes a publication makes use of subsets of original data sets and the subsets are what need to be archived in a repository so that others can reproduce the results presented in a publication. In spite of the fact that the subsets may be duplications of the original data sets, it is important to archive the exact data used within a publication, so that they can be “frozen” in time. This is a safe-guard to insure that the backbone data used in the paper remain unchanged even if the original data are modified given new information.

Appendix 1. Test-bed Manuscript:

Title: Data Supplement to: Acoustic properties of *Salpa thompsoni*.

Authors: Wiebe, P.H., D. Chu, S. Kaartvedt, A Hölder, W. Melle, E. Ona, and P. Batta-Lona

Abstract: Aggregations of the salp, *Salpa thompsoni*, were encountered in the vicinity of Bouvet Island (54° 26'S; 3° 24'E) during the Antarctic krill and ecosystem studies (AKES) cruise on the Norwegian research vessel G.O. SARS from 19 February to 27 March 2008. The salp's in situ target strength (TS), size, number of individuals in aggregate chains, and chain angle of orientation were determined from acoustic and optical data collected with a submersible TS-sonde instrument. Shipboard measurements were made of *Salpa thompsoni*'s material properties -sound speed contrast *h* and density contrast *g*. In addition, a model of salp acoustic backscattering was developed that takes into account the fact that aggregate salps occur in multi-individual chains. Individuals of *Salpa thompsoni* in aggregates were mostly 45.5 to 60.6 mm in mean length and the relatively rare solitaries were about 100 mm. Aggregate chains ranged from 3 to at least 121 individuals and in the upper surface waters (<20 m) showed no preferred angle of orientation. Sound speed contrast (*h*) ranged from 1.003 to 1.021; the density contrast (*g*) estimates varied

between 1.0 and 1.0039. The in situ TS-distributions peaked between -75 and 76 dB at 38 kHz, with a secondary peak of stronger targets at ~-65 dB. TS ranged between -85 and -65 dB at 120 kHz and 200 kHz, and peaked around -74 dB. The measured in situ TS of salps with TS-sonde at three frequencies match the Distorted Wave Born Approximation theoretical scattering model predictions reasonably well. The backscattering from salps when aggregates (chains made up of multiple individuals rather than from single individuals) dominate gives rise to TS values that can be similar to krill and other zooplankton with higher density and sound speed contrasts.

Backbone Data Sets:

Table 1: Position of the trawls and net tows making the salp collections.
Data are OK as is - no backbone data needed for support.

Table 2: Mixed layer (to 50 m) water properties where salps were collected for experimental work.
The data for each cast are in CTD####.txt files
C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\sarslog\CTD_Profile_Data
CTD085.txt, CTD087.txt, CTD090.txt, CTD095.txt

Table 3: Adjustment of target strength measurements, based on close to on-axis measurements of the WC38.1 calibration sphere during the salp measurements.
Backbone data for this table summarizing sphere TS measurements need to be obtained from Egil Ona (IMR).

Table 4: *Salpa thompsoni* lengths in trawl and net tow collections.
Data in:
C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\Salp_Work_Paper
salps size distribution for acoustics.xls

Table 5: *Salpa thompsoni* material properties (sound speed contrast - *h* and density contrast - *g*).
Backbone data for this table summarizing salp material properties needs to be obtained from Dezhang Chu

Table 6: Summary statistics of mean TS, confidence interval for the mean, 25 and 75% quartiles and Q75 – Q25 as a measure of spread.
Data in
C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\Salp_Work_Paper\Ona_stuff
Salp38_1-selection_inner part_2.xls, Salp38_2-selection_inner part_2.xls,
Salp120_1_selection_inner part_2.xls, Salp120_2_selection_inner part_2.xls,
Salp200-1_selection_inner part_2.xls,
Also see combined files for other work.
Salp38_total.xls, Salp120_total.xls

Table 7: Simulation parameters used in modeling the backscattering by salp aggregates.

Data are OK as is - no backbone data needed for support.

Figure 1: Salp study area and the position of tows used to collect the salps.

Position Data given in Table 1. Bathymetry from etopo2 global bathymetry data set.

Figure 2: Instrumentation used in the study. A) APOP sound speed tubes; B) APOP frame with seawater reservoir; C) the weighing vessel in the micro-balance; D) the TS-sonde being deployed from the G.O Sars on 14 March 2008.

Pictures of gear need dates and times.

Figure 3: A TS-sonde image of *Salpa thompsoni* aggregates at 20 m depth in the evening of 13 March 2008

Backbone data consist of a series of Jpeg images that were used in the analysis. Images are on a backup hard-drive (USB Western digital).

Figure 4: 120 kHz EK 60 echogram of a backscattering layer immediately after a Pelagic Trawl collection #51 was made in the layer and caught 5 tons of salps on 11 March 2008.

Acoustic image. How to create a backbone data file. Use Simrad proprietary data or table of sigma values for each auxel (acoustic boxel determined by the time increment horizontally and depth increment vertically)

Figure 5: Salp chain size and orientation were determined from the TS-sonde photo images.

Data in

C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\Salp_Work_Paper
Salp_Image_Statistics.xls

Figure 6: An echogram (120 kHz) from the TS-sonde illustrating the pattern of acoustic backscatter from salp chains.

Acoustic image. How to create a backbone data file. Use Simrad proprietary data or table of sigma values for each auxel (acoustic boxel determined by the time increment horizontally and depth increment vertically)

Figure 7: TS-distributions for targets ascribed to salps.

Data in

C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\Salp_Work_Paper\Ona_stuff
Salp200-1_selection_inner part_2.xls, Salp38_total.xls, Salp120_total.xls

Figure 8: Length distribution of aggregated salps derived based on the analysis of number of salps per chain aggregation and average length of individual salps in aggregated form

Data used for this plot in
C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\mfiles\Chu_mfiles
salp_size.txt (for station 54, 90, 90, 90).

Figure 9: Averaged TS of aggregated salps based on the DWBA backscattering
model.

Model data used in this plot can be extracted from
C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\Salp_Work_Paper\New_Figures_1
6Dec2008
TS_vs_freq_orient_ave_ona_data.fig

Observed values computed from Data in
C:\Antarctic\Antarctic\GOSars_Cruise_2007_2008\Salp_Work_Paper\Ona_stuff
Salp200-1_selection_inner part_2.xls, Salp38_total.xls, Salp120_total.xls

[Note: a draft of the manuscript with figures and tables along with the data sets listed
above and the associated metadata will be put onto a BCO-DMO server before this
report is completed. Web links to the manuscript and data sets will be provided.].