15 Safeguarding and sharing ocean acidification data

Stéphane Pesant[1], Leslie Alan Hook[2], Roy Lowry[3], Gwenaëlle Moncoiffé[3], Anne-Marin Nisumaa[4,5] and Benjamin Pfeil[6]

[1]Centre for Marine Environmental Sciences (MARUM), University of Bremen, Germany
[2]National Aeronautics and Space Administration, Distributed Active Archive Center (NASA DAAC), Oak Ridge National Laboratory, USA
[3]British Oceanographic Data Centre (BODC), UK
[4]Laboratoire d'Océanographie, CNRS, France
[5]Observatoire Océanologique, Université Pierre et Marie Curie-Paris 6, France
[6]Bjerknes Centre for Climate Research, University of Bergen, Norway

15.1 Introduction

Despite the fact that most marine scientists agree with the principle that scientific data must be safeguarded and shared among the scientific community and must become public as soon as possible, there are several cultural and technological issues that minimise and often put off the application of that principle. For example:

- Many scientists do not realise that their data may be useful to other scientists; do not realise that other scientists' data may be useful to them; do not know where and how to archive their data; and do not know where and how to access other scientists' data.
- Organising and documenting data in order to meet the requirements of data archives requires time and efforts as with any other media for scientific communication (e.g. scientific papers, posters or oral presentations) but unlike these other media, archiving data is not perceived as bringing recognition of scientific work and is thus placed low on the list of priorities.
- The numerous sampling and analysis protocols used in ocean science are not described consistently in databases and are often not reported at all by scientists, which makes it difficult to harmonise masses of data and to have any confidence in the quality of meta-analyses performed on them, to the point where scientists loose interest in safeguarding and sharing data.
- Many scientists are anxious at the thought that anyone, and even colleagues, may extract data from databases and publish them without informing or acknowledging the authors. This is especially true for data that are not yet published in scientific journals or could still be reused as original contributions to science.

The objective of the present chapter is to address these issues and recommend solutions and best practices that lead to safeguarding and sharing data and metadata.

15.2 Sharing ocean acidification data

This section addresses the following issues:

"Many scientists do not realise that their data may be useful to other scientists."
"Many scientists do not realise that other scientists' data may be useful to them."
"Many scientists are anxious at the thought that anyone, and even colleagues, may extract data from databases and publish them without informing or acknowledging the authors."

Sharing data is a delicate issue because it deals with intellectual property rights, including the privilege to "be the first one to publish your own data", within a scientific community where competition for funding enforces the saying "publish or perish". The lack of confidence in the common respect of intellectual property rights is at the heart of the problem. When in doubt, most scientists will prefer to keep their data to themselves,

to share it with a few trusted ones, and to communicate data only in the form of posters, oral presentations and scientific publications. We must therefore find ways to raise trust within the scientific community and to prevent abuse of trust.

The purpose of data policies is to establish general guidelines and regulations regarding the fair exchange of data and effective collaboration between partners. We reviewed data policies of several data centres (e.g. SeaDataNet for National Data Centres, and three World Data Centres for oceanography and the marine environment) and those of major European and North American research projects relevant to ocean acidification, such as EPOCA, CARBOOCEAN, Ocean Carbon and Biogeochemistry Data Management Office (US-BCO-DMO) collections and EUR-OCEANS. We summarise here the common guidelines of these policies:

1. metadata (i.e. data about data; see section 15.6) are freely accessible without any condition;
2. data are freely accessible unless otherwise stipulated;
3. users must acknowledge/cite the original data providers;
4. all restrictions on the use and reproduction of data must be respected;
5. data must not be given to third parties without prior consent of data providers;
6. regardless of whether data are quality controlled or not, data archives and original data providers do not accept any liability for the correctness and/or appropriate interpretation of data;
7. any mistake in the data and metadata must be communicated to data archives.

The "principles and guidelines for access to research data from public funding", published by the Organisation for Economic Co-operation and Development (OECD) is also a general reference of interest (http://www.oecd.org/dataoecd/9/61/38500813.pdf). It addresses the following principles: openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability and sustainability.

Research projects generally include restrictions in their data policy about *who* has access to data and for *how long*. Access to data is generally restricted to a specific research community, but the implementation of that restriction is problematic since passwords can be shared outside the community, individual e-mail requests to data providers rapidly becomes unmanageable, and files containing datasets can be passed around. Therefore, research communities that wish to effectively implement restrictions on data access usually channel all data requests through a single or a few individuals who guard the data. In EU projects within the Sixth Framework Programme, the main regulation imposed by the European Commission was that data generated by these projects must be disseminated within two years after the completion of the project. However, data policies sometimes propose shorter moratorium periods in order to stimulate the flow of data within a restricted community and outside of it, for example *"data must be accessible to the community within three months after the time of measurement"*, or *"data is restricted to the community for a period of two years after the time of measurement"*.

Some research projects have intellectual property right panels for protection, dissemination and exploitation of knowledge to ensure that data policies are observed. These panels become involved only if conflicts between partners arise, in which case they moderate and help resolve the conflict.
The Data Management Committee of the IGBP/SCOR project for Integrated Marine Biogeochemistry and Ecosystem Research (IMBER; http://www.imber.info/DM_home.html) acknowledges that "sticks" such as policies and conflict resolution panels are effective but also recommends promoting "carrots", that call for technological developments as well as cultural changes. We highlight here some "carrots" to consider:

1. The anxiety related to misuse of data is partly due to a feeling of inequity between the many scientists who do not know how to extract data from databases and the few who are experts at it, for example modellers and bioinformaticians. A solution is to develop dissemination tools (e.g. data warehouses) that are easy to use by all scientists, i.e. not only by informatics literates, and therefore to reach a point where all scientists start using each other's data.

2. Universally Unique IDentifiers (UUIDs including DOIs, URLs, URNs or LSIDs) are now widely used by scientific journals to cite their papers. Similarly, information systems such as the Publishing Network for Geoscientific and Environmental Data (PANGAEA®) are automatically registering every dataset with persistent Digital Object Identifiers (DOIs) that are used to cite data. Like most UUIDs, DOIs are used in web browsers to get online access to metadata and data (although sometimes restricted), which considerably helps identifying intellectual property. Several data centres are now tagging their datasets with UUIDs and recently, peer-reviewed and indexed journals such as Earth System Science Data (ESSD; http://www.earth-system-science-data.net/) allow researchers to rapidly publish data prior to their full analysis, thus obtaining public recognition of their property rights on the data.
3. Another means to reduce the anxiety related to misuse of data is to request from data archives and information systems that data users register before accessing data, offering the possibility to track who downloads what data. Researchers could then receive reports on usage of their data.

In conclusion, we recommend the ocean acidification community to insist upon database developers that "carrots" must be implemented, and to progressively relax restrictions on data usage in their data policies as a mean to "give trust a chance".

15.3 Safeguarding ocean acidification data

This section addresses the following issue:

"Many scientists do not know where and how to archive their data."

Safeguarding ocean acidification data is the business of permanent data archives, which, in the field of ocean science, comprise mainly the National Oceanographic Data Centres (NODCs) and the World Data Centres (WDCs). NODCs such as BODC in the UK, SISMER in France and US-NODC in the USA are designated by the International Oceanographic Data Exchange programme (IODE) of UNESCO Intergovernmental Oceanographic Commission (IOC), while World Data Centers such as WDC-MARE in Europe, WDC-Oceanography in the USA, Russia and China are designated by the International Council for Science (ICSU). The respective role of the NODCs and WDCs is ambiguous. The IODE model is that oceanographic laboratories archive their data in the NODC that has been designated for their country, so that data contents of one NODC is in principle distinct from that of the other NODCs. There is however no authority dictating the laboratories to follow the IODE model so that a wealth of data, especially in biology, are not submitted to NODCs. In contrast to this federated approach, the WDCs were set to fill a complementary role, which was to replicate data from all NODCs into centralised data archives, and thus to act as multiple backups around the globe. Over the years however, WDCs did not replicate each other but have instead specialised and now complement each other. WDCs often offer advanced services and data products because information systems are technically easier to develop using centralised databases. These services facilitate the integration and access to large data collections, and consequently, large collaborative research projects tend to submit data to WDCs. Furthermore, in many countries, NODCs do not yet have the capacity to archive and distribute genomics, molecular, taxonomic and ecological data or to handle data objects such as images or audio files. We recommend that in such countries, data be safeguarded in a WDC that has this capacity, with the understanding that a copy of the data will be delivered to the designated NODC once it has the capacity to archive such data. Lists of all current NODCs and WDCs can be found at http://www.ngdc.noaa.gov/wdc/contact.shtml and http://www.iode.org.

In conclusion, we wish to remind the reader that, while it is the business of NODCs and WDCs to archive ocean science data, it is the responsibility of the scientists (data providers) to properly document their data, to prepare them with a minimum of structure, and to submit them to data archives. The following sections discuss the importance of documenting data with care, and provide instructions on how to prepare data.

15.4 Harmonising ocean acidification data and metadata

This section addresses the following issue:

"The numerous sampling and analysis protocols used in ocean science are not described properly in databases and are often not reported at all by scientists."

Metadata are data about data (see section 15.6). They describe *what* was measured *by whom*; *when*, *where* and *how* it was sampled and analysed; with *which instruments*, and following *which protocol*; and finally they describe the *units* and *currencies* in which measurements are expressed. The way metadata should be structured is defined by several standards. The most important ones in the geographic information domain are ISO 19115 (Kresse & Fadaie, 2004), Federal Geographic Data Committee (FGDC) (http://www.fgdc.gov/metadata), Directory Interchange Format (DIF) (http://www.gcmd.nasa.gov/User/difguide), and Dublin Core Metadata Initiative (DCMI) (http://dublincore.org/documents/dcmi-terms). They provide the structure to document not only bibliographic information, such as authors, title, date, publisher, and keywords, but also spatial and temporal coverage, methods and data quality. These "contents" standards are the backbone of metadata; the challenge for the scientific and data management communities is to develop standard vocabularies to document the data in a consistent and accurate way.

Registers are authoritative bodies that build and maintain vocabulary databases, and offer web services to encourage the implementation of standard vocabularies. Taxonomic registers and chemical substances registers are particularly relevant to ocean acidification research. The main taxonomic registers are the World Register of Marine Species (WoRMS which grew from the European Register of Marine Species (ERMS); http://www.marinespecies.org/) and the International Taxonomy Information System (ITIS; http://www.itis.gov/). To our knowledge, the most authoritative collection of disclosed chemical substance information is the Chemical Abstracts Service (CAS; http://www.cas.org/) of the American Chemical Society. All these registers are updated by editorial boards of specialists and provide unique identifiers for the entities that they describe. We recommend using these registers as much as possible to describe data.

For other vocabularies pertaining for example to mesocosm experiments or carbonate chemistry computations (see section 15.6.2), we recommend the voluntary appointment of "standard vocabulary editors" from the ocean acidification scientific community to ensure a coherence between scientific requirements and data management practices (see section 15.7), building on existing initiatives such as those undertaken by the Intergovernmental Oceanographic Commission's IODE programme (http://www.iode.org/), the NERC Data Grid programme (http://ndg.nerc.ac.uk/), the SeaDataNet programme (http://www.seadatanet.org/), and the Marine Metadata Interoperability network (http://marinemetadata.org/).

Defining standard vocabularies for analytical methods and defining standard units can prove most challenging. We are now faced with vast repositories of digital data where the same parameter is measured differently and is stored in a range of units. There is no standard vocabulary for analytical methods and these are often not reported at all by data providers, or even sometimes not requested by data centres. The present guide recommends ways to report analytical methods relevant to ocean acidification research (see section 15.6), but in the case of historical data archives, differences in analytical methods will need to be sorted out by going back to the original papers or contacting data providers. On a brighter note, differences in units may be due simply to scaling factors (e.g. per millilitre vs. per litre), the refusal of imperial measurements to die (e.g. knots vs. metres per second), or confusion among mass and amount concentrations (e.g. grams per litre vs. moles per litre and cells per litre). These differences can be sorted by simple conversions. However, the harmonisation of parameters reported per litre vs. per kilogram of seawater requires additional data (*in situ* seawater density or salinity, and temperature) that are often estimated in the absence of *in situ* measurements (see section 15.6.2 for examples with carbonate chemistry).

Centralised data centres such as BODC, SISMER, the Carbon Dioxide Information Analysis Centre (CDIAC) and World Data Centres have tried to address these problems by adopting "in house" standard vocabularies for methods and "in house" standard units using different conversion factors. However, it is not a viable option for distributed systems of data centres since different standard units have been used in different data centres.

There are two ways to deal with the units issue in a distributed scenario. One is to keep units as part of the parameter vocabulary and create the necessary additional entries in the parameter dictionary for the different units. The other approach is to maintain a totally separate metadata field for units, which can then be attributed to parameters. The former way was soon seen to be undesirable because it results in parameter dictionaries of such a size that they become unmanageable. Consequently, the approach now advocated by most data centres is that units should have their own standard vocabulary that is dissociated from the parameter vocabulary.

For the sake of archiving data in a standard way, data providers and data managers should follow the International System of Units (SI) as much as possible in their choice and scaling of units, for example using molinity (amount per mass of seawater) for molar concentrations instead of molarity (amount per volume of seawater), and scaling molinity in units of moles per kilograms instead of micromoles per kilograms. These choices may in some cases increase the number of digits (integers or decimals) and may get us away from "practical units" that are usually reported in scientific communications or from "output units" given by instruments. In cases where metadata about the methods are missing, the original units may be the only way to differentiate among methods that are not comparable and should therefore not be combined, for example primary production measured in a few millilitres sample incubated under constant irradiance over 1 hour vs. primary production measured in 1 litre sample incubated under natural light conditions over 24 hours. This is the main argument in favour of "sticking" to the original units, as a means to stress methodological differences. We recommend that data providers submit data with their original units, and provide detailed methods (i.e. sample treatment and analysis) and conversion factors to SI units (e.g. molarity to molinity) as part of metadata (see section 15.6 on how to report data and metadata). We also recommend that data be converted and archived in SI units by data managers, keeping record in the metadata of the original units used by the data provider. Major data centres should eventually be able to automatically "read out" units in their archives so that major information systems can provide users with harmonised data outputs in which units have been converted according to the user's preferences.

In section 15.3, we discussed the use of UUIDs to facilitate the citation of datasets and improve the preservation of intellectual property rights. Here we wish to recommend that UUIDs be also assigned upstream of the data archiving process, i.e. to the samples themselves (e.g. unique sample identifiers), ensuring that all data generated from a given sample can be tracked, interconnected, re-assembled and harmonised during meta-analysis. The use of open-source and commercial "sample-tracking platforms" is now common in medical, molecular and genetics research and should be envisaged as well by the ocean science community, particularly in the case of multidisciplinary projects conducted in the field and in experimental facilities where one sample leads to several analyses including metagenomics, molecular and elemental composition, taxonomy, metabolic activities and trophic interactions. Groups of experts such as the Biodiversity Information Standards, previously known as the Taxonomic Database Working Group (TDWG; http://www.tdwg.org/) are investigating the use of UUIDs to track samples in ocean science. We recommend that research projects be proactive by initiating their own "sample tracking system" and keeping an eye out for any standard practices that are being developed.

15.5 Disseminating ocean acidification data and metadata

This section addresses the following issue:

"Many scientists do not know where and how to access other people's data."

Metadata are data about data (see section 15.6). Disseminating data and metadata is the business of information systems such as Google™ and various wikis. In the geoscientific world, information systems are also known as "data portals", "data directories" or "clearinghouses". The purpose of information systems is to search for and harvest data from various archives and repositories, to repackage them into predefined or customisable collections of data/products, and to disseminate these products to the public or sometimes to a restricted group of users, as discussed in section 15.2. Information systems differ in the type of archive that they use and in the service they provide. It may consist of static collections of files containing data and metadata (e.g. FTP style portals), of a relational meta-database supporting a dynamic search of predefined datasets (e.g. Google-like

portals), or of a relational database that enables mass extraction and re-packaging of data from a large number of datasets, using advanced search functions (e.g. data warehouses).

The first product (e.g. FTP style portals) is typical of large research projects that generate finite collections of files containing datasets that need to be shared among collaborators who already know about the data. However, these datasets include a lot more than what general users are looking for, and most importantly, data and metadata are often organised and formatted in very different ways from file to file.

The second product (e.g. Google-like portals) allows users to target more specifically the type of data they are looking for, but the "granularity" of data in a relational database can be very fine (e.g. each CTD cast can constitute a dataset) so that users end up with a huge number of datasets to download. Although the organisation and formatting of these datasets may be more homogeneous, assembling them proves to be a challenge that is often difficult to manage.

The third product (e.g. data warehouses) allows users to select parameters (variables) that are of interest and to extract only the corresponding values out of the entire database, and to re-package them in a table format. These products must be used with care to ensure that users do not loose essential metadata information in the process. For instance, in the resulting data table, each value should always be accompanied by geographical and temporal references (latitude, longitude, date, time, depth), parameter names and units, some details on the methodology used for sampling and/or analysis, and a citation.

There are a few information systems related to ocean acidification knowledge; they include OCB-DMO (http://ocb.whoi.edu/), CDIAC (http://cdiac.ornl.gov/) and PANGAEA® (http://www.pangaea.de/). The first two systems offer a dynamic search of metadata that leads to an "FTP style portal", while the latter information system offers a "Google-like portal" and a "data warehouse" (beta version).

It is the responsibility of the ocean acidification community to request from NODCs and WDCs that ocean acidification data and metadata be made available to the relevant information systems, notably the three mentioned above. To ensure a wider dissemination of ocean acidification knowledge, data and metadata should also be distributed to other communities via their own relevant information systems, for example the Ocean Biogeographic Information System (OBIS; http://www.iobis.org/) for the biodiversity and Census of Marine Life community, (COPEPOD; http://www.st.nmfs.noaa.gov/plankton/) for the International Council for the Exploration of the Sea (ICES), the Mediterranean Science Commission (CIESM) and the plankton research community, and the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk/) for the molecular and genomics community. In turn, these information systems could become a useful source of complementary data to the ocean acidification community.

15.6 Reporting data and metadata

This section addresses the following issue:

"Organising and documenting data in order to meet the requirements of data archives requires time and efforts, as with any other media used by scientists to communicate their findings, e.g. scientific papers, posters or oral presentations."

The previous sections of this chapter outlined important issues that must be regularly discussed by the ocean acidification scientific community to ensure that ocean acidification knowledge is shared, safeguarded, harmonised and disseminated. The first step to address these issues is to adequately prepare ocean acidification data by providing the relevant metadata, describing *who* measured, observed or calculated *what, where*, *when* and *how*. Metadata about "*where*" and "*when*" are generally well documented, while metadata regarding the other questions are very often overlooked. Metadata about "*who*" measured, observed or calculated data are essential to ensure intellectual property rights. Metadata about "*what*" was measured and "*how*" are essential to harmonise data and gain confidence in the quality of meta-analyses performed on them. Preparing metadata does require time and efforts, but we must educate researchers to these practices and raise the value of "archiving data" to the level of other types of scientific communications, such as scientific papers, posters or

oral presentations. The Data Management Committee of the IGBP/SCOR project IMBER (http://www.imber.info/DM_home.html) recommends a few actions to reach that goal.

First, research programmes must allocate funding to data management. Each programme should hire a person to create metadata, contact scientists to prepare and submit their data, aggregate datasets that are related but come from different sources, and submit/import data into a database. This person could be a recent graduate who would get trained by a data centre. Recent research programmes on ocean acidification, such as the European projects CARBOOCEAN and EPOCA, have implemented this recommendation and are indeed contributing masses of ocean acidification knowledge to data centres and information systems such as CDIAC and WDC-MARE/PANGAEA®.

Second, the scientific community, including field and laboratory scientists, modellers and data managers, must define together best practices as well as clear instructions and templates for the preparation and submission of data and metadata. There are a few recent guides of "best practices" for preparing environmental data (Hook *et al.*, 2007) and CO_2 measurements (Dickson *et al.*, 2007). We reviewed the best practices reported in these guides and those put forward by ocean acidification-related research programmes (e.g. CARBOOCEAN, EPOCA, EUR-OCEANS and MARBEF) and ocean acidification-related data centres and initiatives (e.g. CDIAC, WDC-MARE/PANGAEA®, OCB-DMO, IODE's network of NODCs and SeaDataNet). We recommend four general best practices to report data and metadata:

1. ***be informative when you assign names, titles and descriptions***, for example with data files, parameters, sampling and analysis methods, units and formats;
2. ***be consistent in the way you assign that information and refer to well-recognised vocabulary registers when applicable***, for example with taxonomy, equipment and sensors, standard parameter names and standard units (see section 15.4);
3. ***be consistent in the way you format values***, for example decimal degrees vs. degree minutes seconds, YYYY-MM-DD (ISO 8601, recommended format) vs. DD-MM-YYY vs. MM/DD/YYYY, local date and time vs. GMT, decimal and digit group separators; and
4. ***be conscientious in performing basic quality assurance***, for example lookout for outliers, duplicates, mistakes in latitude or longitude, and properly distinguish between "missing values" and true "zero values".

Beyond these general recommendations, research programmes and data centres related to ocean acidification also provide instructions and templates to guide the preparation of data and metadata. We have reviewed the most common ones and provide here a description of the core metadata and data requirements (section 15.6.1), followed by a synthesis of the specific metadata and data requirements that were identified in chapters 1 to 15 of the present guide (section 15.6.2).

15.6.1 Core metadata requirements

It is recommended that metadata always include the following information:

- ***Dataset citation:*** title, summary, date created and last updated, authors contact details (last name, first name, e-mail, institution name, address and description), reference to related publications (authors, year, title, source, volume, pages, UUID), and reference to related research project (name and UUID).
- ***Sampling events:*** event name (code), fieldwork/experiment name, research infrastructure name (e.g. ship, mesocosm, laboratory), responsible scientist name (include contact details if different from authors given in the citation), sampling device name, sampling device method, sampling quality, sampling comment, sampling reference (text or UUID), latitude (start, end and/or nominal), longitude (start, end and/or nominal), date (start, end and/or nominal), time (start, end and/or nominal).
- ***First order parameters:*** parameter name, parameter short name (often used in data tables where values are reported), parameter type (e.g. computed/calculated individual value, measured/determined

individual value, sum of individual values, statistical value obtained from individual values), units, responsible scientist name (include contact details if different from authors given in the citation), access rights (public or restricted), analysis device name, analysis device method, analysis quality (e.g. flags, detection limits and uncertainty), analysis comment, and analysis reference (text or UUID).

Second order parameters are used to subdivide the first order parameters that are for example taxon-specific, gender-specific, pigment-specific or size-fractionated. When appropriate, the following information should be provided:

- ***Second order parameters about taxonomy:*** a list of taxonomic name used in the dataset, if possible a taxonomic reference for each taxon (name of the taxonomy register, registered taxon-ID, registered taxon name, and UUID), and reference material (ID, location, condition, last check). See section 15.4 for details on taxonomy registers.
- ***Second order parameters about life cycles (e.g. gender, age, development stage):*** name, description, reference (text or UUID).
- ***Second order parameters about chemical composition (from atoms to large molecules):*** name, description and, if possible, a chemical composition reference (name of the chemistry register, chemical component ID and UUID). See section 15.4 for details on chemistry registers.
- ***Second order parameters about metrics (e.g. size classes and wavelengths):*** metrics name, description, units, lower limit, and upper limit.
- ***Second order parameters about digital objects (e.g. maps, figures, pictures, audio or video files):*** name, description (including format, method used to generate the object, and recommended method to read/use the object).

There are several forms and templates being proposed by the different data centres and data management groups to help prepare and organise metadata and data. In many cases, data providers are asked to fill in text forms with the relevant metadata and to submit it separately from the files containing the data. Sometimes, the files containing the data are submitted long after the metadata forms are submitted, risking that datasets become updated and no longer correspond to the information provided in the metadata form. Sometimes, files containing datasets are simply not submitted, kept locally, and made available upon request. Generally, we recommend that data and metadata be prepared and submitted together and, as much as possible, in the same file.

15.6.2 Specific metadata requirements

Here we summarise and complement the recommendations of parts 1 to 4 of this guide concerning data reporting for seawater carbonate chemistry, perturbation experiments, CO_2 sensitive processes and model outputs.

Seawater carbonate chemistry and ancillary parameters (Part 1)

It is essential to provide measurements of temperature and salinity when reporting data for seawater carbonate chemistry. Moreover, we strongly recommend that the following metadata information is included:

- ***Dissolved inorganic carbon (DIC) concentration (mol kg^{-1}):*** information about sample replication, sample volume, poisoning (poison volumes), analysis method (technique description, reference), CRM information (correction magnitude, batch number, analysis log), overall precision, and accuracy.
- ***Total alkalinity (A_T) (mol kg^{-1}):*** curve fitting method, type of titration (reference), description of other titration, cell type, CRM scale, sample volume, magnitude of blank correction, overall precision and accuracy.
- ***Carbon dioxide partial pressure ($p(CO_2)$; atmosphere):*** information about sample replication, storage method, analytical method (technique description, reference), sample volume, headspace volume, *in situ* temperature, temperature during measurement, temperature normalisation, temperature correction method, variable reported, gas, standard gas concentrations, frequency of standardisation, overall precision and accuracy.

- ***pH:*** pH scale, analytical method (technique description including, when appropriate, probes, buffer names and reference), *in situ* temperature, temperature during measurement, temperature normalisation, temperature correction method, *in situ* pressure, calibration method, overall precision and accuracy.

The R package seacarb (Lavigne & Gattuso, 2010) is often used to compute a complete and consistent set of carbonate chemistry parameters, using original values of temperature and salinity, and any pair of the carbonate chemistry parameters listed above (see chapter 2). Also, if available, the concentrations of silicate and phosphate can be used in seacarb as additional ions contributing to the carbonate chemistry and thus allowing for more robust computations. When original values are expressed in mol l^{-1}, we recommend converting them first to mol kg^{-1} by using seawater density that is calculated from salinity and temperature as indicated in Dickson *et al.* (2007). Seacarb can only use pH measured on the total scale as an input term. When pH is reported on a scale other than the total scale, we recommend to first calculate DIC and A_T with CO2SYS (Lewis & Wallace, 1998) using pH (other scale) and another carbonate chemistry parameter, and subsequently use these DIC and total alkalinity values in seacarb to estimate pH on the total scale, as well as nine other carbonate chemistry parameters. Seacarb uses flags to specify which pair of seawater carbonate chemistry parameters are used for computations. We extend this list of Carbonate Chemistry Computation flags (CCC flags) to include cases where pH is not available on the total scale (Table 15.1). When archiving computations from seacarb (Lavigne & Gattuso, 2010) and/or CO2SYS (Lewis & Wallace, 1998), we strongly recommend that the method of calculation and the appropriate flag be written out fully in the metadata. For example with pH, "pH was computed on the total scale using seacarb (Lavigne & Gattuso, 2010) from DIC and total alkalinity. DIC was first calculated with CO2SYS (Lewis & Wallace, 1998) using pH (other scale) and total alkalinity as input parameters (CCC flag 29)". It is essential that a table explaining the CCC flags is provided to the data centres.

The primary goal of data reporting for climate targets is to provide a template for comparing experimental results among the atmospheric, ocean and terrestrial science communities. Towards that goal, ocean acidification studies should carefully report the $p(CO_2)$ levels (in μatm) of interest for the study, with the various parameters of the ocean carbonate chemistry (see above). The use of common currency of atmospheric carbon dioxide levels and the use of standard or key $p(CO_2)$ values for most studies will elevate the value of ocean acidification science for society.

Perturbation experiments (Part 2)

It is relatively easy to distinguish between parameters that are determined in the field (e.g. measured *in situ* or determined directly or experimentally from samples collected at sea) and those determined on samples that are not specific to any geographic location (e.g. in laboratory experiments). However, the distinction between field experiments under "natural conditions" (*in situ* or simulated) and those under "artificial conditions" can be unclear. With respect to data reporting, we propose the following distinction:

- ***Perturbation experiments under natural conditions:*** they include short-term field experiments under *in situ* conditions (e.g. *in situ* incubations for primary production), simulated natural conditions (e.g. deck incubations for primary production), modified environmental conditions within natural range (e.g. photosynthetron, chemostats and nutrient uptake), or modified assemblages within natural range (e.g. dilution method to measure grazing rate).
- ***Perturbation experiments under artificial conditions:*** they include long-term experiments (>1 day) in mesoscale enclosed systems (mesocosms) with natural or modified assemblages under modified environmental conditions, and long-term enrichment experiments (>1 day) in the field, for example iron enrichment experiments.

Perturbation experiments under artificial conditions allow for biological interactions and are considered to replicate natural conditions more accurately than laboratory experiments, but on the long term these systems drift from the initial conditions and measurements should no longer be considered as "natural" observations.

Table 15.1. List of Carbonate Chemistry Computation (CCC) flags describing which pair of carbonate chemistry parameters is used for computations in seacarb and CO2SYS in addition to temperature and salinity.

CCC flag	Computation software	Input parameters (in addition to temperature and salinity)
1	seacarb	pH (total scale) and CO_2
2	seacarb	CO_2 and HCO_3^-
3	seacarb	CO_2 and CO_3^{2-}
4	seacarb	CO_2 and A_T
5	seacarb	CO_2 and DIC
6	seacarb	pH (total scale) and HCO_3^-
7	seacarb	pH (total scale) and CO_3^{2-}
8	seacarb	pH (total scale) and A_T
9	seacarb	pH (total scale) and DIC
10	seacarb	HCO_3^- and CO_3^{2-}
11	seacarb	HCO_3^- and A_T
12	seacarb	HCO_3^- and DIC
13	seacarb	CO_3^{2-} and A_T
14	seacarb	CO_3^{2-} and DIC
15	seacarb	A_T and DIC
21	seacarb	$p(CO_2)$ and pH (total scale)
22	seacarb	$p(CO_2)$ and HCO_3^-
23	seacarb	$p(CO_2)$ and CO_3^{2-}
24	seacarb	$p(CO_2)$ and A_T
25	seacarb	$p(CO_2)$ and DIC
26	***Step 1.*** CO2SYS	pH (other scale) and $p(CO_2)$
	Step 2. seacarb	A_T and DIC (from CO2SYS)
27	***Step 1.*** CO2SYS	pH (other scale) and HCO_3^-
	Step 2. seacarb	A_T and DIC (from CO2SYS)
28	***Step 1.*** CO2SYS	pH (other scale) and CO_3^{2-}
	Step 2. seacarb	A_T and DIC (from CO2SYS)
29	***Step 1.*** CO2SYS	pH (other scale) and A_T
	Step 2. seacarb	A_T and DIC (from CO2SYS)
30	***Step 1.*** CO2SYS	pH (other scale) and DIC
	Step 2. seacarb	A_T and DIC (from CO2SYS)
31	***Step 1.*** CO2SYS	pH (other scale) and $p(CO_2)$
	Step 2. seacarb	A_T and DIC (from CO2SYS)

Apart from their time-zero measurements, data collected from an artificial environment created in the laboratory or in mesocosms must not be confounded with data measured *in situ* or from field experiments under natural conditions. Given that some metadata fields are sometimes not requested by users during mass extraction of data from databases, especially metadata fields such as "comments", "notes" or even "methods description", it is not sufficient to simply mention in these fields that data are from an artificial environment. We strongly recommend that data collected from an artificial environment be archived without any geographic coordinates (latitude or longitude). Sampling date, time and depth (e.g. in mesocosms or large scale enrichment experiments) can be archived as usual, but the geographic coordinates of the artificial/perturbed environment should be archived as an attribute of the sampling infrastructure/platform, not as an attribute of the data itself.

When reporting data for perturbation experiments, it is strongly recommended to include the following metadata information:

1. ***Initial state and quality:*** where and when samples or specimens were collected; in the case of plankton, describe the initial environmental and community conditions; in the case of specimens, describe the body size and other biometrics, information on their life cycle such as gender, reproductive state, age and developmental stage.
2. ***Relevance of experimental treatments to natural field conditions:*** the environmental conditions where samples/specimens were collected; and the natural values of experimental end points in field community/population.
3. ***Experimental environmental conditions:*** the temperature and salinity in each treatment; measurement of at least two carbonate chemistry parameters (see Table 15.1) from each treatment; accurate description of the methods used to measure carbonate parameters including pH scales and buffers where appropriate; values at the beginning and end of the experiment, and if available, values during the experiment should be provided.
4. ***Experimental treatment (carbonate chemistry – chapter 2 and 6):*** time course of CO_2 manipulation; control of carbonate chemistry during the experiment; control of $p(CO_2)$ in closed headspace vs. open headspace with ambient $p(CO_2)$; method of CO_2 manipulation; in the case of aeration with air at target $p(CO_2)$, indicate $p(CO_2)$ level and flow rate; in the case of addition of high-CO_2 seawater, indicate $p(CO_2)$ and mixing ratio; in the case of addition of strong acid as well as CO_3^{2-} and/or HCO_3^-, indicate volume and normality of acid added as well as the quantity of inorganic carbon added; in the case of addition of strong acids and bases, indicate volume and normality; in the case of manipulation of the Ca^{2+} concentration, indicate the recipe of artificial seawater used.
5. ***Experimental treatment (batch culture – chapter 5):*** basic information characterising the physiological state of the initial inoculum including cell density of the stock culture, number of cells inoculated, chlorophyll per cell, growth, irradiance, temperature, and composition of the initial culture media. F_v/F_m would give information on whether the stock culture was nutrient replete and growing in exponential phase, nutrient limited, or in stationary phase; the investigators should also report whether or not the cultures were axenic and indicate in the metadata if frozen culture is available for future examination.
6. ***Experimental treatment (mesocosms – chapter 6):*** mesocosm dimensions and duration of the experiment; experimental design, layout of treatments and replication; enclosure filling methods; initial conditions; mixing configuration and turbulence characteristics and, wherever possible, direct velocity measurements of turbulent mixing should be conducted and reported; sampling methods; unintended perturbations such as shifts in plankton community composition, aggregation of dissolved or particulate matter and wall growth.
7. ***Experimental treatments (specimens – chapter 7):*** nature and magnitude of incremental changes in specimens' acclimation; the length of time between steps or total acclimation period or whether they were immediately exposed to the full treatment levels; indication or measure of specimens' stress;

length of time that specimens were exposed to the treatment; and comparison with "control" field specimens if possible.

8. ***Experimental treatments (natural gradients and* in situ *perturbations – chapter 8):*** for *in situ* perturbation, describe the experimental design, layout of treatments and replication; for natural gradients in pH or other carbonate system parameters, describe potential limitations of the design (e.g. lack of interspersion or replication, temporal and spatial variability, etc.); whenever possible, potentially confounding factors (e.g. methane, sulphide, temperature, oxygen) should also be monitored and reported.

CO_2 -sensitive processes (Part 3)

Metabolism, pH enantiostasis (chapter 9):

- Describe the experimental procedure: exposure regime, length of exposure, water physicochemistry values (levels of e.g. pH, bicarbonate, carbonate, calcium) on physiologically relevant scales.
- Describe the sampling methods: animal acclimation and treatment, sampling procedure for tissues, dye, wavelengths of excitation and emission, time resolution, local resolution (whole tissue layer, whole cell, cellular compartment).
- Describe the analyses: tissue and cell type, investigated parameters, experimental tools (buffer systems, calibration procedures, ion gradients) and pharmacological tools (specific transport inhibitors, inhibitors of carbonic anhydrase).
- Report ancillary data: fluorescence intensities vs. wavelength (in a region of interest), variability between replicates.
- Describe metabolism and pH enantiostasis data: ratios that are linearly correlated to pH in a given range, pH, H^+ flux values (nM H^+ $time^{-1}$ $(membrane\ area)^{-1}$).
- Report time constants for pH recovery due to systemic or cellular mechanisms and relative (%) change in ratio in paired experiments.
- Report rate of pH change under pH disturbance or recovery and acid-base variables during quantitative treatments of acid-base status.

Organic and export production, elemental ratios (chapter 11):

Data normalisation is often required for the assessment of how acidification affects biogeochemical, physiological and ecological processes. Normalisation can be defined as a mathematical process that adjusts for differences among data from varying sources in order to create a common basis for comparison. For example, determining the amount of $CaCO_3$ ("calcimass") or its rate of production generally involves measuring the dissolved and particulate concentrations of an element and its uptake by organisms or communities, and normalising these quantities using the proportion of calcifiers in the community and the ratio of $CaCO_3$ to the selected element in calcifiers' biomass. In that case, it is recommended to archive values for the measured variables, values of the computed yield or production rate of $CaCO_3$, and values of the normalisation factors used in the computation.

Pelagic calcification (chapter 12):

- In studies that directly measure calcification rates in planktonic calcifiers, provide clear descriptions of the experimental design and protocols.
- Precision and accuracy in measurements of the parameters of the CO_2 system and associated factors (e.g., temperature, salinity, nutrient concentrations) should be reported.
- Report other experimental conditions that may affect calcification rates in photosynthetic or heterotrophic organisms, for example, irradiance, light/dark cycles, nutrient and trace element concentrations, food availability, feeding frequency, and grazing.
- Describe the method of collection of organisms and provide the size range (and age, if known) of the organisms used in the experiments.

- For manipulative experiments at different $p(CO_2)$ levels, describe any acclimation period and the conditions experienced by the organisms during that time.
- In isotope tracer measurements of calcification rates, the blank values and their variability over the duration of the experiment should be reported. Equations for the calculation of calcification based on isotope measurements should be described, and all relevant information included, such as whether the bicarbonate concentration is assumed constant regardless of water mass or whether an isotope discrimination factor is assumed.
- In any case, if calcification rates are normalised (e.g. to chlorophyll or shell mass), it is recommended to archive values for the measured variables, normalised values, and values of the normalisation factors used in the computation.

Benthic calcification (chapter 13):

- Describe the mode of chemical manipulation: acid only, bicarbonate followed by acid or bubbling, etc.
- Report the strength of the acid, carbonate chemistry of the treatment water (ideally A_T and DIC), scale and calibration of pH electrode, temperature at which the pH is measured, when bubbling with air, report the source of the air (i.e. outside air vs. room air).
- Report the method used for measuring calcification, temperature and salinity, irradiance in quantum units, dissolved nutrient levels, organism's feeding state.
- Rates of calcification should be reported per area of living tissue (cm^2 or m^2), per weight of protein or per mass of skeleton.
- In the case of field based measurements of calcification, report water depth, current speed and percent cover of calcifiers as well as a description of community structure and the usual physical measurements of temperature, salinity, nutrients, light and water clarity.

15.7 Avoiding pitfalls and addressing challenges

The previous sections have identified a few serious pitfalls that raise a number of challenges, notably addressing intellectual property rights and raising trust among the scientific community, and harmonising data and metadata by adopting standard vocabularies and unit conversion procedures.

To address these challenges, an emerging trend in scientific data infrastructures is to create expert groups, comprising field and laboratory scientists, modellers and data managers, for different categories of data. These categories can be defined based on the type of equipment used or analysis performed or, more generally, on research fields. Different expert groups may for example address "zooplankton net sampling", "flow cytometry" or "primary production and plankton community metabolism". The goals of these expert groups are to:

1. Promote and facilitate the submission of data into designated National Oceanographic Data Centres and/or World Data Centres. Each expert group needs to identify the information system that is most relevant to disseminate its respective data to their scientific community, including modellers (see section 15.5) and must request from NODCs and WDCs that data be made available to the information system of their choice.
2. Develop standard vocabularies describing variables, sampling protocols and analytical methods. This work should be lead by data managers from the relevant information systems and data centres, building on existing initiatives (see section 15.4). The expert groups are expected to constitute or to take part in a network of "standard vocabulary editors" that review existing vocabularies and later approve changes and additions.
3. Recommend best practices to harmonise existing data and plan the collection of new data. Harmonisation involves the cross-validation of traditional and emerging methodologies (sampling and analysis), the organisation of expert-to-expert validation, and the review of conversion factors and algorithms. Planning involves selecting and promoting preferred sampling and analytical protocols for future studies.

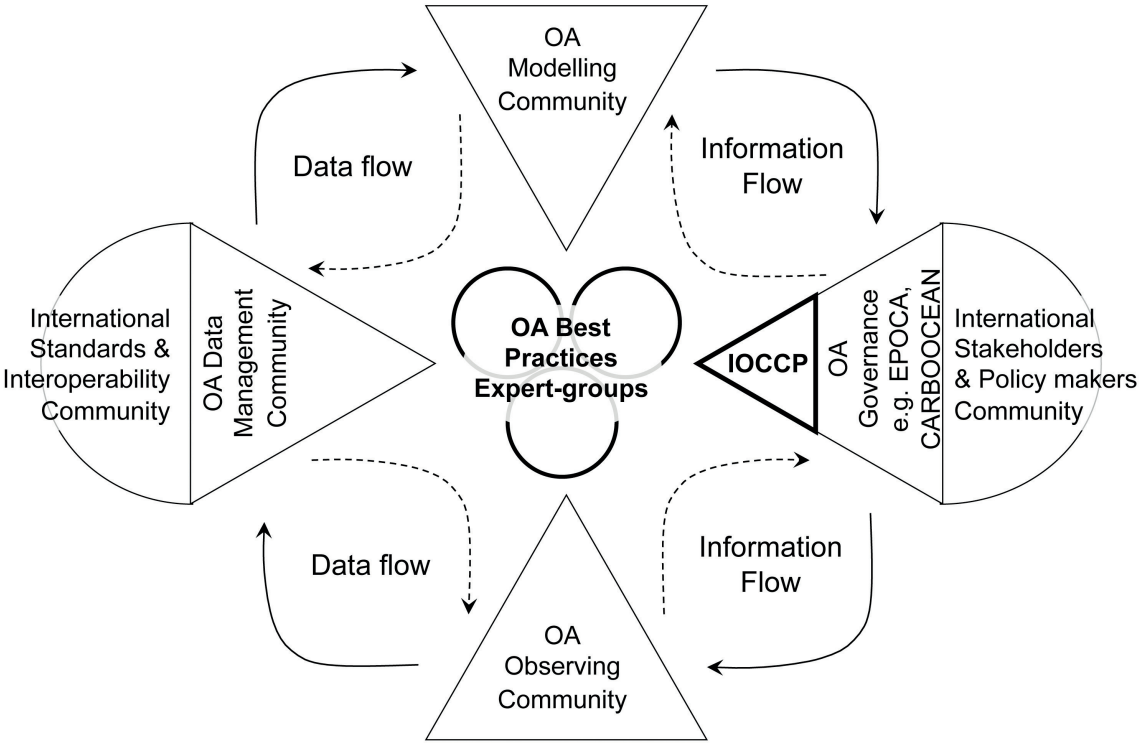

Figure 15.1 Conceptual diagram of the proposed scientific data infrastructure for ocean acidification knowledge, showing the end-to-end flow of data and information (full arrows) and including feedback loops that are often under-represented in scientific data infrastructures (dashed arrows).

IODE's Group of Experts on Biological and Chemical Data Management and Exchange Practices (GE-BICH) has initiated work in that respect on sampling instrumentation used in biological and chemical oceanography (see report in references). In the field of ocean acidification, we recommend to create three expert groups based on the structure of the present guide (seawater carbonate chemistry, perturbation experiments, and CO_2-sensitive processes). The authors of these chapters are potential candidates. Figure 15.1 illustrates how a scientific data infrastructure for ocean acidification could work in line with recommendations from the Global Earth Observation System of Systems (GEOSS; http://www.epa.gov/geoss/). The proposed scientific data infrastructure for ocean acidification basically brings together four communities with distinct roles: the observing community (field and laboratory scientists), data management community, modelling community and governance community. Representatives from all four communities work together as part of ocean acidification "OA" Best Practices Expert Groups and could be coordinated by the IOCCP and the IOC/IODE, with the overall goal to facilitate the end-to-end flow of ocean acidification data and information in the scientific community, and to stakeholders and policymakers. The proposed infrastructure imposes that data be archived and harmonised via the data management community, following the recommendations of the expert groups. In the proposed infrastructure, arrows show the flow of data and information, not the interactions among the different communities. Interactions among observing, modelling and data management communities occur in the expert groups and of course in research projects outside the infrastructure, but this is not illustrated here.

15.8 Recommendations for standards and guidelines

1. **Sharing knowledge.** In addition to using "sticks" such as data policies and conflict resolution panels, "carrots" should be used to promote sharing ocean acidification knowledge. These "carrots" call for both technological developments by the data management community and cultural changes in the scientific community. Technological developments include the use of Universally Unique Identifiers to reference data, and the development of tools to access masses of data and to track usage of data (section 15.2).

2. **Safeguarding knowledge.** Ocean acidification data and metadata should be safeguarded in National and/or World Data Centres that have the capacity to archive and distribute as needed genomics, molecular, taxonomic and ecological data, and data objects such as images and audio files (section 15.3).

3. **Harmonising data and metadata.**
   - A "standard vocabulary editor" from the ocean acidification scientific community should take part in existing initiatives such as those undertaken by the Intergovernmental Oceanographic Commission's IODE programme (http://www.iode.org/), the NERC Data Grid programme (http://ndg.nerc.ac.uk/), the SeaDataNet programme (http://www.seadatanet.org/), and the Marine Metadata Interoperability network (http://marinemetadata.org/). The objective is to develop and maintain standard vocabularies and ontologies describing *what* is measured and *how* (section 15.4).
   - Research programmes should use unique sample identifiers and use "home-made" sample tracking systems, until standard ones are available. The objective is to ensure that all data generated from a given sample are tracked, interconnected, re-assembled and harmonised during meta-analysis (section 15.4).

4. **Disseminating data and metadata.** National and World Data Centres should systematically distribute ocean acidification data and metadata to the relevant information systems, notably the Ocean Carbon and Biogeochemistry Data Management Office (OCB-DMO; http://ocb.whoi.edu/), the Carbon Dioxide Information Analysis Centre (CDIAC; http://cdiac.ornl.gov/) and the Publishing Network for Geoscientific and Environmental Data (PANGAEA®; http://www.pangaea.de/) (section 15.5).

5. **Reporting data and metadata.**
   - Data and metadata should be prepared and submitted together and as much as possible in the same file, following the detailed guidelines given in section 15.6. Metadata must describe *who* measured *what, where*, *when* and *how*. Metadata about "*where*" and "*when*" are generally well documented, while metadata regarding the other questions are often overlooked. Metadata about "*who*" measured data are essential to ensure intellectual property rights. Metadata about "*what*" was measured and "*how*" are essential to harmonise data and gain confidence in the quality of meta-analyses performed on them (section 15.6).
   - Research programmes on ocean acidification must allocate funding to data management, hiring a person (data curator) to help scientists preparing and submitting their data, to aggregate datasets that are related but come from different sources, and to submit/import data and metadata into a database (section 15.6).

6. **Avoiding pitfalls and addressing challenges.** It is recommended to create a scientific data infrastructure for ocean acidification with the overall goal to facilitate the end-to-end flow of data and information within the scientific community, and to stakeholders and policy makers. The central components of this scientific data infrastructure are a number of "Expert Groups" that include representatives of the observing community (field and laboratory scientists), data management community, and modelling community. Initially, we propose three expert groups based on the chapter structure of the present guide, i.e. seawater carbonate chemistry, perturbation experiments, and CO_2-sensitive processes. The infrastructure could be coordinated by the IOCCP and the IOC/IODE (section 15.7).

15.9 References

Dickson A. G., Sabine C. L. & Christian J. R. (Eds.), 2007. Guide to best practices for ocean CO_2 measurements. *PICES Special Publication* 3:1-191.

Hook L. A., Beaty T. W., Santhana-Vannan S., Baskaran L. & Cook R. B., 2007. Best practices for preparing environmental data sets to share and archive. Environmental Sciences Division, Oak Ridge National Laboratory. USA.

IODE Group of Experts on Biological and Chemical Data Management and Exchange (GE-BICH), 2009. Fourth Session, 27 – 30 January 2009. Reports of meetings of experts and equivalent bodies. 54 p. Paris: UNESCO.

Kresse W. & Fadaie K., 2004. *ISO standards for geographic information*. 322 p. Berlin: Springer-Verlag.

Lavigne H. & Gattuso J.-P., 2010. Seacarb: calculates parameters of the seawater carbonate system. R Package 2.3 (portions of code were contributed by J.-M. Epitalon, A. Hofmann, B. Gentili, J. Orr, A. Proye & K. Soetart). http://cran.at.r-project.org/web/packages/seacarb/index.html.

Lewis E. & Wallace D. W. R., 1998. Program developed for CO_2 system calculations. Oak Ridge, Tennessee: ORNL/CDIAC-105. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy. http://cdiac.ornl.gov/oceans/co2rprt.html