Best Practice for Data Management

Version 1.2 I January 2022

ICES USER HANDBOOKS

ICES INTERNATIONAL COUNCIL FOR THE EXPLORATION OF THE SEA
CIEM CONSEIL INTERNATIONAL POUR L'EXPLORATION DE LA MER

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H. C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

Recommended format for purpose of citation:
ICES. 2022. Best practice for data management. Version 1.2. ICES User Handbooks. 15 pp.
https://doi.org/10.17895/ices.pub.8884

Contents

i Background

This document is intended as a general guidance for creating and maintaining data collections in ICES – although a number of the general principles can be applied anywhere. It will particularly be a useful resource for expert groups that seek to establish new data collections and systems within ICES, and aims to act as a guide for considerations and work that should be worked out by the contributors ahead of creating formal data calls or building systems to host the data.

The document has been initiated in a collaboration between ICES Data Centre and the Data and Information Group (DIG) to ensure that data are managed, structured, and developed in a robust way that will allow best possible use of the data. The focus of the document is for data contributors, possibly already in an ICES working group, seeking to prepare and establish data collections, or to contribute data to existing systems. Data contributors may or may not be aware of how ICES data submissions work, so some sections of the document cover who to speak to, and how to work with the existing system to ensure the information is available for new members of the ICES community.

The document is structured along the general stages that data needs to pass through from collection to use in formal advice, with a focus on what data submitters/providers needs to consider. This is followed on by the overarching principles of data sharing (either openly or more constrained), and putting the data into the context of how it is managed once it arrives in ICES. Next to that, the document outlines routes of communication – which groups and bodies within ICES would be first contact points for requests for establishing new data collections.

Last, but not least, feedback and dialogue is essential to shape the guidance in this handbook, and both DIG and the ICES Data Centre want to ensure we can capture lessons learned and experiences from different exercises in working groups for the benefit of the ICES community on a continuous basis. Thus, there is a strong encouragement of users of the document to provide feedback or ask questions to either the Data and Information Group or the ICES Data Centre.

Changes since the last version

A complete version history can be seen in Annex 1.

Location	Change description
Section 1	Replaces chapter 2 of the 2019 Best practices document - updates in the process taken into account.
Section 2	Replaces chapter 3 and 5 of the 2019 Best practices document - specification of the people/groups to contact, also taking into account the governance groups
Section 3	Replaces previous Annex 1 of the 2019 Best practices document, and has been aligned with the diagram in section 1. All topics mentioned in section 1 have now a place in this chapter.

Other relevant information

The User Handbook falls under the responsibility of the ICES Data and Information Group (DIG). It is reviewed triennially, but can also be updated more frequently when required.

1 The Data Journey

In general, data is collected by the members of the ICES community, and the work on methods, quality checks, and submission of data are done by the individual Member Countries or institutes. Throughout the pipeline of data from collection to advice products, there are good practices that expert groups can (and do) apply to ensure the following steps in that journey happen more smoothly. There will be variations in the level of complexity or processes associated with different types of data, but the overarching principles are similar. We consider 'data' as a discipline-agnostic concept that applies anywhere we want to collect, collate, and analyse any type of observation.

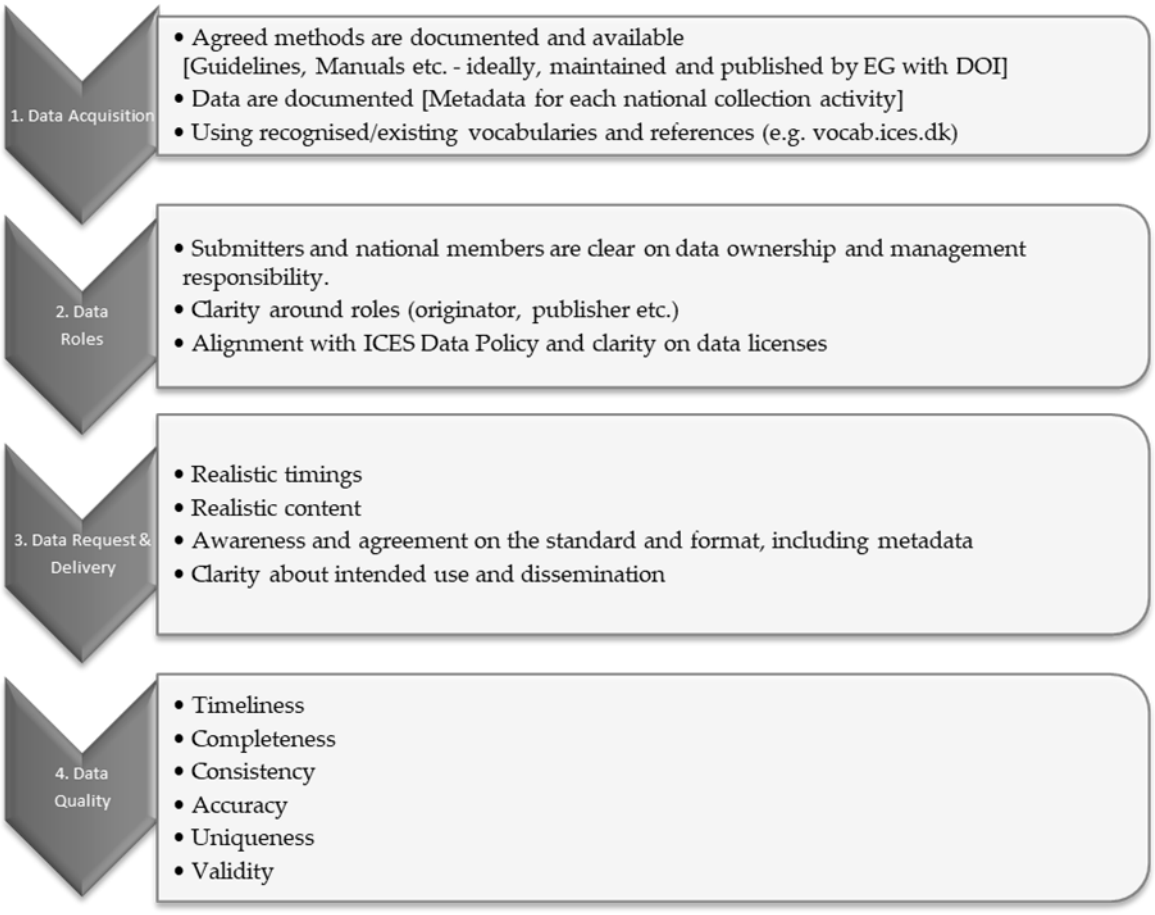

Figure 1.1. The data pipeline

We consider four general steps in the data pipeline from acquisition through to a quality profiled data collection ready for use. These steps include different considerations of best practice from an ICES Community and Data Management perspective. The output of data into different uses would follow on from the four steps described here. However, all uses of data – be it for additional research or production of advice – will benefit from having completed and profiled the data following these steps.

Figure 1.1 provides an overview and bullet points of things to consider as part of best practice management of data. Each data collection will be different, and some considerations may be more important for some workflows than others. The steps do however represent universal considerations that should be applicable to any data collection that ICES Expert groups collect, analyse and are managed within the ICES Data Centre.

Each bullet point under the steps should be considered in order to achieve a best practice. It does not mean that every single item will be achieved in the first attempt, but rather that issues are documented and recommendations on appropriate use can be clearer.

Section 3 contains a more detailed description for each of the steps and bullet points mentioned in Figure 1.1.

2 Who should I talk to and what should I consider?

While the considerations above may seem overwhelming, the most important aspect is to engage in dialogue early. This is dialogue within and across expert groups, and with the ICES Data Centre or DIG.

The best way to contact DIG, expert groups, or steering group chairs to discuss any shared consideration related to data is to look up the group on the ICES website and make contact with the listed chair of the relevant group. This information is kept up to date by the ICES Secretariat, and the chairs of the groups will be able to facilitate further dialogue. To contact the ICES Data Centre, the general contact address is accessions@ices.dk.

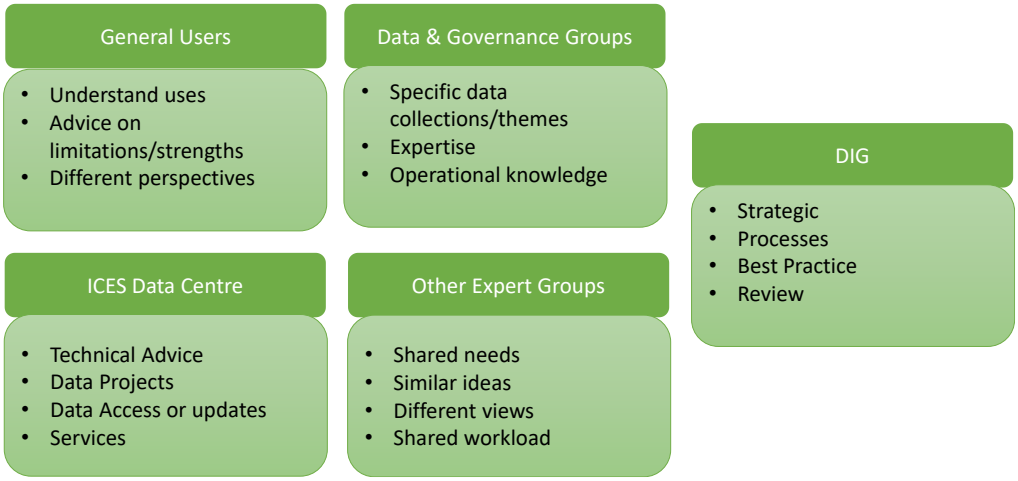

Figure 2.1. Options to retrieve more data guidance

2.1 Data submission

The sooner and the more openly the discussions around the nature, formats, and requirements for data can take place, the sooner both data practitioners and experts can evaluate data quality in an open and transparent way. The expert groups are where the greatest understanding of the data are centred, the ICES Data Centre, DIG, or other entities can provide specific technical advice and support that can supplement this.

Be realistic about time frames. Working through many of these details takes a long time, especially to ensure there is good consensus. But equally important is to recognise that support requirements for development or data management can be hugely reduced if the groundwork has been done well.

It is also important for expert groups to foster an environment where mistakes or errors can be reported openly so that a collaborative effort can be made to correct them or document them, which is far more efficient than passing on the data only for another step or group to locate an error.

2.2 Data governance

The ICES community has great resources in the form of exert groups, but there are a lot of them, and it may sometimes be difficult to locate or identify all the potential users of data, or who actually "decides" what changes can be made for a given database. For data related questions, the main contact is and will continue to be the ICES Data Centre. They are however increasingly being supported by governance groups: expert groups that provide aspects of governance and review to help ensure solutions are fit for purpose and provide the best possible services across the community.

DIG is an operational group and works closely with the ICES Data Centre, SCICOM, and steering groups to ensure alignment and review of data policies, processes, and emerging issues that may have a strategic impact on ICES Data Operations. Dialogue with other expert groups is always welcome, and the main support provided by DIG is in the form of recommendations for best practice, updates or reviews to processes, and interactions with other national and international data centres.

In addition to the ICES Data Centre and DIG, a number of data focussed groups and governance groups are in operation, all under the Data Science and Technology Steering Group. The governance groups are tasked with bringing together the state and requirements for particular ICES data products or applications. The governance groups communicate with other expert groups providing the data as well as the data user groups for these collections. They are more focussed on particular data collections or systems, and evaluate to what extent the best practice elements are being realised within the given applications.

Finally, the entire list of best practice principles and goals for ICES will only come to fruition through close collaboration and open communication. Dialogue is essential in every aspect, and so the closing chapter of this paper will be a repeated encouragement to all expert groups to engage early and communicate often.

3 Detailed description of working towards best practice points

3.1 Data Acquisition

Data acquisition can cover many aspects of collecting data, be it from survey work in the field to literature reviews and collations to rescue historical data. The key aspect is to be open and clear about how much is known about the methods of acquisition and origin of data.

3.1.1 Agreed methods

Where the data acquisition is guided directly in the ICES community, many expert groups already have manuals, instructions, or recommendations for methods that should be employed during data acquisition. Where the data acquisition is being pre-defined, the agreed methods should aim to highlight targets for each of the data quality dimensions (see Section 3.4) to the extent possible. This will allow a clearer and faster evaluation of data quality against a set of known and agreed targets.

Best practice is for this material to be openly available to the community in a traceable and citable format. Ideally the material is assigned a persistent identifier or submitted to a best practise collection such as Ocean Best Practises[1] to allow long term referencing.

3.1.2 Data acquisition documentation

When data are collected or collated in line with an agreed method, it is important to document any variations. Making notes on variations form an agreed standard as early as possible ensures it is fresh in mind, and can be described accurately. Metadata offer an ideal way to document variations as well as basic profile information about each data set that forms part of the acquisition. Some expert groups have defined and documented metadata standards for these purposes, while others can simply utilise existing standards of metadata, subject to the complexity and nature of the data. Note that metadata can simply be used as subject lines to help keep the documentation aligned – it is not always necessary to develop complex xml tools or systems when a simple structured table describing the collection will do.

3.1.3 Using existing references and vocabularies

Using a shared reference system (such as species lists, area definitions, or lists of agreed terminology) avoids ambiguity, and can be applied for many purposes. One example of common reference systems are Aphia ID or LSID for species[2]. In each member country there are common names in different languages, and scientific names are sometimes reorganised. By storing an

[1] https://www.oceanbestpractices.org/ (accession date 08 December 2021)

[2] http://www.marinespecies.org/about.php (accession date 08 December 2021)

existing reference in the most appropriate reference system (ideally, references form part of the agreed methods) data from multiple collections immediately start having commonality. ICES Data Centre maintains the vocabulary server[3] that can host reference lists, but many other reference systems already exist as well, and as long as they are openly available, they may be very helpful in structuring the data better.

3.2 Data Roles

3.2.1 Data roles, ownership and management responsibilities

The data roles apply across the entire life cycle of the data, but in particular once data are acquired, it becomes essential to be clear and open about the roles that relate to the data.

Typical roles related to data are:

- Custodians: Persons or organisation that is responsible for maintaining and ensuring access to the data.
- Originators: The persons or organisations that acquired the data. Often custodians and originators are the same, but not always.
- Publishers: The persons or organisation that is responsible for publishing the data

These roles can get a little confusing as they may co-exist where a data set feeds into a larger collection. Thus, there may be custodians for national datasets which are submitted to a larger international dataset with its own custodians.

The key importance is clarity about responsibilities within the specific data flow the expert group is working with, and when and where they transfer between different roles.

Agreement on data management roles (see above) and data management responsibilities (e.g. data quality assurance and maintenance) is crucial before any data exchange. While ownership of data is rarely transferred within the ICES community, it is essential for participants in expert groups to be clear on their own national mandate in agreeing to share data.

3.2.2 Alignment with ICES Data Policy and clarity on licenses

Before new data sharing agreements or data calls are progressed, it is important for the interested expert groups to understand and have a dialogue that examines the positions of national member's licensing and their potential alignment with the ICES Data Policy. This prevents progressing to a point where data have been collated together only to find that only partial data can be obtained (or made available for a scientific output) because of differences in licensing. This aspect can take time and often require dialogue within member countries as well, and the more the best practices considerations have been made – the more straightforward it should be to make clear exactly what needs to be shared, how, and by whom.

[3] https://vocab.ices.dk

3.3 Data Request and delivery

Only once the agreed responsibilities and licensing aspects of the data are known, should the process move on to making actual calls for data and delivery into either existing or new systems. It should be stressed that the earlier in the process the dialogue with ICES Data Centre, DIG, or governance groups can be initiated, the easier this process should be. However, it is really only at this stage that we can start considering the data process as ready to establish a data collection.

3.3.1 Realistic timings

The realistic timings translate to "it will take longer than you think". Bringing together data from multiple sources often shines a light on any consistency issues, and it might be necessary to revisit some of the earlier steps to clarify and improve agreed methods, formats, or responsibilities of roles at this stage.

The other aspect of realistic timings is to encourage expert groups to think longer term and to a wider use of data (where this is appropriate). Rather than treating a data call as a one-off exercise, expert groups are encouraged to consider how to make the process sustainable over a longer term – e.g. For next year + 1 + 1 + 1, or when 1–10 new data submitters are added to the exercise. By making these considerations early, a much higher degree of consistency can be achieved with a lot less pain.

3.3.2 Realistic content

While data quality and usability of data are often associated with having as complete data as possible, it is important to consider what is realistic across the ICES community as well as perhaps other existing data sources. While some member countries may have very high quality or high-resolution data, it is important to ensure that the data formats and requirements are set at a realistic level. Experts will have to evaluate the balance here, it will be a balance perhaps of resolution versus coverage, or completeness versus volume.

The key aspect of the best practice is to be clear on what the data should be able to support, and perhaps equally important what the data will not be able to support immediately. The best approach is often to develop data content over a period, taking it step by step. It does require considerations of keeping approaches open enough to accommodate the future changes, and relate to the realistic timings, and making the process sustainable over time.

3.3.3 Data standards and format

For data providers as well as data users, it helps when a description of the data format and the used standards (e.g. vocabulary, units) is available. Combining data is a lot easier when everyone "talks the same language", i.e. uses the same standards. This could also include metadata.

3.3.4 Intended data use

Whenever data is requested, the requestor should be clear about the intended use, further dissemination, data storage at the receiver's end, and data access. Only then a data provider can decide if data submission is according to the local (institute's/national) data policy.

3.4 Data Quality

Data quality is not a singular state of good or bad quality. Rather it is a composite of a number of known characteristics about the data that allow users and analysis to determine its fitness for use. Many ICES Data collections are used for multiple purposes, and perceptions and requirements for data quality can vary widely between different disciplines or assessment needs. The most common characteristics used to profile and express data quality are described below.

3.4.1 Timeliness

The time difference between e.g. acquisition or submission of data and until it is available to users. Most timeliness aspects in ICES are handled via data calls, and thus previous comments about realistic timings, and thinking of a sustainable, repeatable process can have a big impact on timeliness.

3.4.2 Completeness

The proportion of stored data that meets a requirement or a target, or the proportion of a requirement or target that can be met with stored data. It can also include the proportion on unreported or null values in a dataset. Most of the evaluation of completeness will be up to the expert group that use the data. If there is a defined temporal or spatial coverage target, the data should be profiled against this requirement, and completeness can then be expressed or evaluated.

Obviously, there are more complex aspects of completeness such as catchability of different species, but the agreed methods in the data acquisition should ideally state the targets for each of the data quality dimensions.

3.4.3 Consistency

The degree of consistency is perhaps best expressed as the absence of any difference when comparing parameters across a data set. E.g. there are no major consistency issues between different source datasets. If there are differences, it is very important to document these in metadata.

3.4.4 Accuracy

An expression of the degree to which the data correctly describes the “real world” object or observation. Ideally the comparative “real world” observation is made through primary research, or to compare with third party data of a known quality.

Accuracy is quite closely related to consistency as variations in accuracy will potentially impact on consistency as well. If expert groups can make comparison with independent data and verify the accuracy, or quantify the error, dataset become much more valuable.

3.4.5 Uniqueness

Uniqueness is ultimately about traceability. How quickly and easily can we verify observations from an ICES database in a member level or internal database used to hold the collected data? Considering and keeping original national identifiers in an ICES data system may speed up any queries, corrections, or comparisons greatly in the future.

In addition, many national institutes are also looking at, or being asked to publish data openly on a national level. It would be really important to be able to recognise duplicates introduced by merging datasets from a national dataset and the ICES portal for example.

Sometimes uniqueness can be achieved through composite data (e.g. a country, ship, date, and station number for example) but this method is more vulnerable to any issues (e.g. a date was changed or transformed wrongly, and the uniqueness can no longer be determined) – so a single unique identifier from the data originator being retained throughout the data collection is best practice.

3.4.6 Validity

Typically, there may be predefined validity requirements in an agreed method for data acquisition that determines validity. Data validity should typically be applied to every single field in a data format. For example, a latitude value greater than 90 degrees cannot be valid, or a gear code not listed in the selected vocabulary is not valid. Typically, much of the data validation is performed during the data submission stage, but the clearer a data format description can be about validity rules, the sooner issues of rejected submissions can be avoided.

Annex 1: Change log

Date	Change	Prepared by
10 January 2018	Initial version created	Jens Rasmussen, Marine Scotland, Neil Holdsworth, ICES
9 April 2020	Text edits, comments	Lena Szymanek, Wim Allegaert, Ingeborg de Boois
02 July 2020	Additional discussion	DIG 2020 subgroup on Action 11
15 October 2021	Updated version	Jens Rasmussen
9 November 2021	Finalised version 1.2	Ingeborg de Boois

Annex 2: Author contact information

Jens Rasmussen j.rasmussen@marlab.ac.uk

Ingeborg de Boois ingeborg.deboois@wur.nl

Annex 3: Useful links

ICES Data Collections and pages:

http://ices.dk/marine-data/Pages/default.aspx

ICES Data Policy:

http://ices.dk/marine-data/guidelines-and-policy/Pages/ICES-data-policy.aspx

DIG page:

http://ices.dk/community/groups/Pages/DIG.aspx

Guidelines:

http://ices.dk/marine-data/guidelines-and-policy/Pages/default.aspx

Metadata :

http://gis.ices.dk/geonetwork/srv/eng/catalog.search#/search?facet.q=type%2Fdataset%26orgName%2FICES&resultType=details&fast=index&_content_type=json&from=1&to=20&sortBy=relevance

Transparent Assessment Framework (TAF):

https://taf.ices.dk

Vocabularies (reference lists):

https://vocab.ices.dk/

Web Services:

http://ices.dk/marine-data/tools/Pages/WebServices.aspx