ICES User Handbook: Best practice for Data Management
January 2019

ICES
CIEM

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 purposes of citation:

ICES. 2019. ICES User Handbook: Best practice for Data Management. 12 pp. http://doi.org/10.17895/ices.pub.4889

Contents

1 About this document

This document is an open and collaborative development, ensuring the ICES community has access to guidance on best practices for managing data collections.

The initial document is a collaboration between ICES Data Centre and the Data and Information Group (DIG). The intention is that this will develop into a handbook used by working groups themselves; to ensure that data are managed, structured, and developed in a robust way. This will allow the best possible use of the data.

Feedback and dialogue are essential in shaping the guidance contained in this handbook. This means that both DIG and ICES Data Centre want to ensure we capture the lessons learned and experiences gathered from different exercises in working groups. Doing so effectively will benefit the ICES community on a continuous basis.

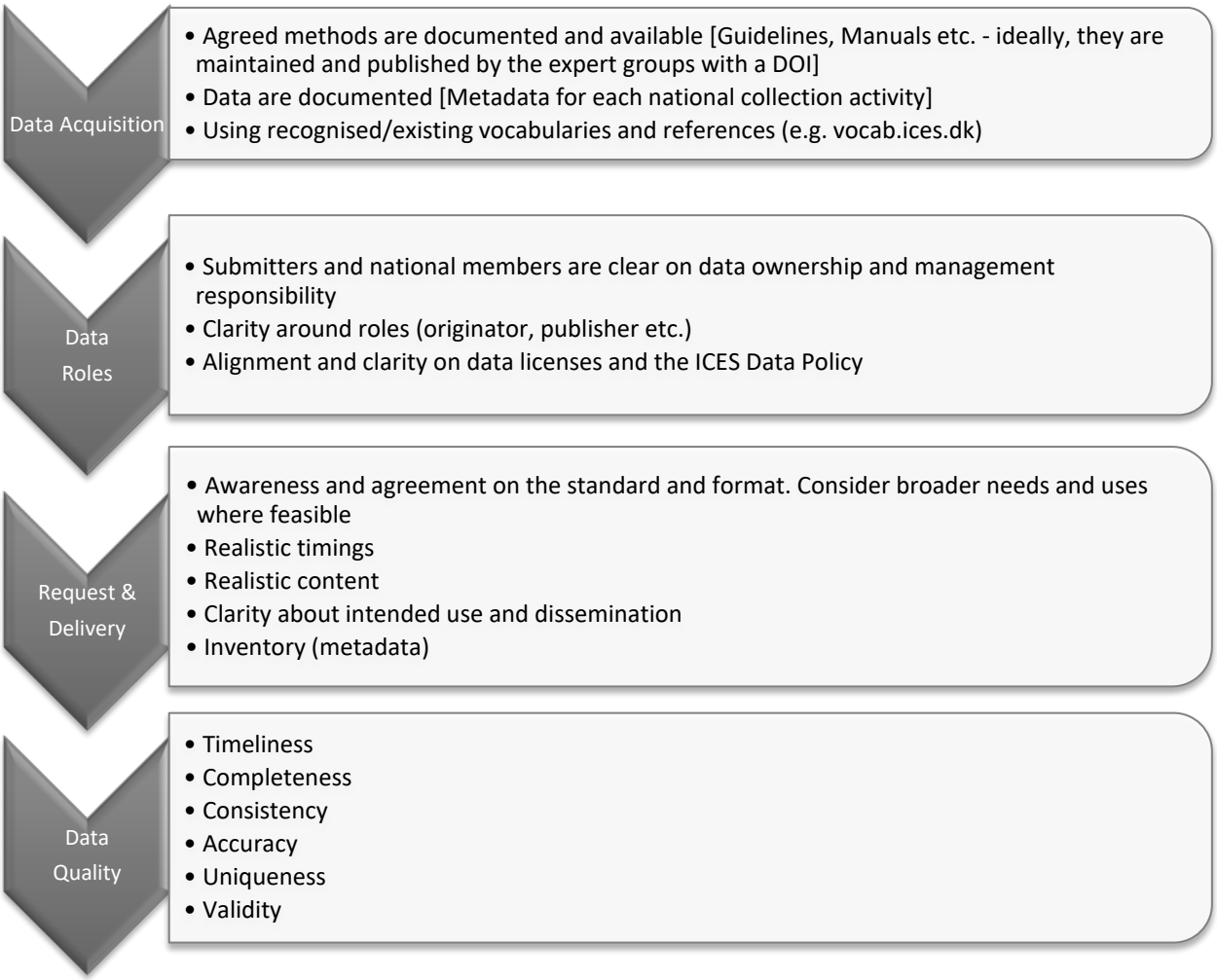
This initial document focuses on what expert groups can do when establishing collaborative data collections. There will be variation in the levels of complexity and in the processes associated with different types of data, but the overarching principles are similar for almost all types of data under consideration. We consider data a generic concept that applies anywhere we want to collect, collate, and analyse any type of observation.

Data are generally collected by the members of the ICES community; the work on methods, quality checks, and submission of data is done by the individual member countries. Throughout the data pipeline, from collection to advice products, there are good practices that expert groups can (and do) apply. These ensure the smooth progress of that journey through the steps described in the following document.

2 The Data Pipeline

We consider there to be 5 general steps in the data pipeline, from acquisition through to a quality profiled data collection that is ready for use. These steps include several points of best practice that must be considered when we approach the data pipeline from an ICES Community and Data Management perspective.

The graphic below provides an overview of these points, and of how they can be considered part of the best practice management of data. Each data collection will be different, and some considerations may be more important for some workflows than others. But these points do represent universal considerations. They should be applicable to any data collections that ICES expert groups acquire or analyse, and any that are managed within ICES Data Centre.



Each point related to these 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; rather that issues are documented thoroughly and that recommendations on appropriate use can be made clearer.

There is a short section on each of the steps, expanding slightly on most of the related points, in Appendix 1: Detailed description of Best practice points.

3 Dialogue and realistic approaches!

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

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; 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, particularly as there is a need to ensure consensus is reached. But it is equally important to recognise that the 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 or document them. This is far more efficient than passing on the data, only for another step or group to locate an error later.

4 Working with existing data

ICES is increasingly delivering data through a service-oriented approach using a variety of web services: http://ices.dk/marine-data/tools/Pages/WebServices.aspx.

This is consistent with other large-scale data centres.

The advantage of web services is that their multiple applications and analytical tools allow data to be queried and retrieved in a standardised way. This makes the availability of data more consistent.

The most common formats for retrieving data from web services are quite technically complex (e.g. XML and JSON structures). It means that training in these specific skillsets would be required if expert groups were to be most effective. Translation tools (in the R environment, for example) can, however, convert data back into more familiar formats like excel or comma separated values. The main benefit of accessing services directly is that the collection data retrieved is the most recent and is quality assured, without the user having to make separate new extracts and versions.

For individual expert groups, a gradual increase in the use of web services and associated technologies is likely. Documents and internal material will remain stored in the SharePoint portal, but this should be regarded as a "walled garden" that is not used for direct data distribution to a wider community. ICES Data Centre and many expert groups are increasingly using R scripts, as well as repositories in online libraries such as Github, for documentation. The methods and tools used to work with or transform data are also made available in this way.

The ICES Transparent Assessment Framework (TAF) offers many examples of using R and Github to control the flow of data into scientific assessment. This is an increasingly important consideration for expert groups, who over a 3 year cycle can produce multiple dataset outputs based on the same source data. The ability to identify and trace back specific data products for publishing and/or assessment purposes is a key consideration for an expert group's integrity.

A large amount of metadata already exists for data held in ICES, but current levels of visibility are not optimal. The metadata's utility in terms of searching across collections is not highly exploited either. As more services start interacting and more persistent identifiers (DOI's) are added to datasets, we can expect to see wider integration. Use of the metadata in direct location of datasets will also increase.

5 Groups and contacts that provide data oversight: Who should I speak to?

The ICES community has great resources in the form of expert groups. There are many of them, however, and it could be difficult to locate or identify all the potential users of data. Finding those who actually "decide" what changes can be made for a given database could also be a challenge. For data related questions, the main contact is and will continue to be ICES Data Centre. They receive increasing support, however, from operational and expert groups. This provides greater governance and review capabilities, and helps ensure that solutions are fit for purpose. This is how the best possible services will be provided across the community.

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

In addition to ICES Data Centre and DIG, there are a number of data-focused groups and governance groups in operation. The governance groups have the task of aligning particular existing ICES data products or applications with best practice requirements. The governance groups communicate with other expert groups that provide the data, as well as the data user groups for these collections. They have a greater focus on particular data collections or systems, and they evaluate to what extent the best practice elements are being realised within the given applications.

Other groups have a specific interest in methodologies, historical data, and specific types of data or tools.

Examples of these groups are as follows:

- PGDATA: Data Needs for Assessment and Advice
- WGHIST: History of Fish and Fisheries (incl. dark data/rescue)
- SC-RDB: Regional Database steering group
- WGZE: Zooplankton ecology (dark data rescue)
- WGDG: DATRAS Governance
- WGSMART: SmartDots application Governance

6 Future challenges and opportunities

Data acquisition, storage, analysis, access, and management are changing rapidly. This is driven by huge leaps in computing power and interconnectivity.

DIG and ICES Data Centre are maintaining a list of challenges and opportunities relating to future developments. These include cloud technology and machine learning, as well as the task of making data open and processes transparent. A list like this is far better, however, with input from the wider community. So this is an open invitation for expert groups to help identify the main challenges and opportunities in managing data for the future ICES community.

One of the key processes is to ensure that data are as open as possible and as closed as necessary. This centres on the FAIR principles, ensuring that all data held in ICES are:

- Findable (through documentation and metadata)
- Accessible (through clarity on licensing, formats and the ICES data policy)
- Interoperable (through extended use of shared reference systems and services)
- Reusable (by having known data quality and good documentation)

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!

1 Appendix 1: Detailed descriptions of Best Practice points

1.1 Data Acquisition

Data acquisition can cover many aspects of collecting data; these can range from survey work in the field to literature reviews and collations, to the rescue of historical data. The key concerns are clarity on data acquisition methods and how much is known about those methods, as well as the data origin.

1.1.1 Agreed methods

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

It is considered best practice that this material is openly available to the community in a traceable and citable format.

1.1.2 Data Documentation

When data are collected or collated in line with an agreed method, it is naturally important to document any variations. But something might not go entirely to plan; there may be technical issues, or there may be differences between the data collected. To reduce the risk of inaccurate assumptions being made, it is important to document any variations from the agreed methods as early on in the process as possible.

Metadata offer an ideal way to document these variations, as well as the basic profile information on each data set that forms part of the acquisition. Some expert groups have their own defined and documented metadata standards for this purpose. Others can use existing standards of metadata; this depends of course on the complexity and nature of that data. It should be noted that metadata can also 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 data collection will be sufficient.

1.1.3 Using existing references and vocabularies

The use of a shared reference system avoids ambiguity and can have many useful applications. Some conspicuous examples of common reference systems are Aphia ID or LSID for species. These are useful as there are common names in different languages in each ICES member country, and scientific names are sometimes reorganised.

By storing an existing reference in the most appropriate reference system (where references ideally form part of the agreed methods), commonality immediately starts to form between data from multiple collections. ICES Data Centre maintains the vocabulary server that can host reference lists, but many other reference systems also

exist. As long as they are openly available, these other systems may be very helpful in improving the structuring of data.

1.2 Data Roles

The data roles apply across the entire data life cycle, but particularly after acquisition. At this stage it is essential to be clear and open about the roles that relate to the data.

Typical roles related to data are:

- Custodians: The persons or organisations 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, because where a data set feeds into a larger collections they may co-exist. There may be custodians, for example, responsible for national datasets which are in turn submitted to a larger international dataset with its own custodians.

It is key that the responsibilities of different roles are clear within an expert group's specific data flow, as well as when and where they are transferred between roles.

1.2.1 Data ownership, responsibilities, and licenses

While ownership of data is rarely transferred within the ICES community, it is essential that expert groups are clear on their own national mandate when agreeing to share data. Before there is progress on new data sharing agreements or data calls, interested expert groups should understand and discuss these matters. The dialogue should cover the positions of national member's licensing, and their potential alignment with the ICES Data Policy. This decreases the risk of progressing to a point where data have been collated, only to find that only partial data can be obtained (or made available for a scientific output) due to differences in licensing.

This step can take time to resolve and can often require additional dialogue within member countries. If best practices have been considered thoroughly, it should be straightforward to find out exactly what needs to be shared, how, and by whom.

1.3 Data Request and delivery

It is only once the agreed responsibilities and licensing aspects of the data are known that the process should move on to the next step. This is the actual creation of data calls, and their 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. It is really only at this stage, however, that we can start considering the data process as ready in terms of establishing a data collection.

1.3.1 Realistic timings

Realistic timings translates 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. Methods, formats, or responsibilities of roles may need to be clarified, improved, or agreed at this stage.

Realistic timings also mean that expert groups should be encouraged to think about both the longer-term and the wider use of data (where this is appropriate). Rather than treating a data call as a one-off exercise, expert groups could consider how to make the process sustainable over a longer term. This could mean for next year + 1 + 1 + 1, or when 1–10 new data submitters are added to the exercise. By considering these points early, a much higher degree of consistency can be achieved with a lot less inconvenience.

1.3.2 Realistic content

While data quality and usability of data are often associated with those data being as complete as possible, we must 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, the formats and requirements set for data collection must be set at a realistic level. This balance will have to be evaluated by experts, who should consider resolution versus coverage and completeness versus volume.

The key aspect of the best practice is to be clear about what the data should be able to support. Perhaps equally important is what the data will not be able to support immediately. The best approach is often to develop data content over a period of time, taking it step by step.

It does mean that approaches must be kept open enough to accommodate future changes, and that realistic timings are kept in mind. This should help in making the process sustainable over time.

The documentation of changes to data formats or profiles through metadata are, of course, essential.

1.4 Data Quality

Many ICES Data collections are used for multiple purposes. Perceptions and requirements for data quality can vary widely between different disciplines or assessment needs.

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 analysts to determine its fitness for use.

The most common characteristics used to profile and express data quality are as follows:

1.4.1 Timeliness

The time difference between, for example, the acquisition or submission of data and the moment at which it is available to users. Most timeliness considerations in ICES are handled via data calls. Points covered elsewhere in this document about having realistic timings and a sustainable, repeatable process can have a big impact on timeliness.

1.4.2 Completeness

Either 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. Completeness can also come down to the proportion of unreported or null values in a dataset.

Completeness will be evaluated primarily by the expert group using the data. If there is a defined temporal or spatial coverage target, the data should be profiled against this requirement; completeness can then be expressed or evaluated.

There are more complex aspects of completeness, of course, such as catchability of different species. But the agreed methods in the data acquisition process should ideally state the targets for each of the data quality dimensions.

1.4.3 Consistency

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

1.4.4 Accuracy

An expression of the degree to which the data correctly describes the "real world" object or observation. The comparative "real world" observation is ideally made through primary research, or in a way that compares with third party data of a known quality.

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

1.4.5 Uniqueness

Uniqueness is ultimately about traceability. It is a question of how quickly and easily we can verify observations from an ICES database in either a member level, or an internal database used to hold the collected data. Queries, corrections, or comparisons may be sped up in the future by considering and keeping original national identifiers in an ICES data system.

Many national institutes are also looking at, or being asked to publish, data openly on a national level. The ability to recognise duplicates, such as those introduced by merging a national dataset and the ICES portal, is very important.

There may be predefined validity requirements in an agreed method for data acquisition, and this typically determines validity. Data validity should generally be applied to every single field in a data format. A latitude value greater than 90 degrees, for example, cannot be valid; a gear code not listed in the selected vocabulary would also not be valid.

Much of the data validation is performed during the data submission stage. The clearer a data format description can be about validity rules, however, the sooner issues of rejected submissions can be avoided.

2 Appendix 2: 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

3 Change log

Date	Change	Prepared by
10 January 2019	Initial version created	Jens Rasmussen (Marine Scotland), Neil Holdsworth (ICES)