CSK-2 Data and Information Management Plan

(Version 1, 15 May 2022)

Table of Contents

The Draft CSK-2 Data and Information Management Plan (DIMP) was prepared by the CSK-2 Task Force on Data and Information Management Plan, which was formed as per the decision of the CSK-2 International Steering Group at its first meeting (virtual, 8 December 2021).

The CSK-2 DIMP Task Force currently consists of the following members (in alphabetic order):

Ken Ando (TF Lead)	Japan Agency for Marine-Earth Science and Technology, Japan
Adi Purwandana	Research Center of Oceanography, National Research and Innovation Agency, Indonesia
Eugene Burger	Pacific Marine Environmental Laboratory, NOAA, USA
Jingsong Guo	First Institute of Oceanography, Ministry of Natural Resources, China
Joon-Soo Lee	National Institute of Fisheries Science, Korea
Marina Frederik	National Research and Innovation Agency, Indonesia
Wenxi Zhu	IOC Sub-Commission for the Western Pacific, IOC/UNESCO
Zhaohui Chen	Ocean University of China, China

1. Introduction

The CSK-2 programme aims to promote cooperation and advance understanding of the largest circulation in the Pacific Ocean- Kuroshio, as its condition changes in global warming with associated environmental and socio-economic impacts. The importance of data and information generated through CSK-2 research, monitoring and observations cannot be overemphasized. They help us understand and track the changes, improve ocean and climate predictions, and conserve and sustainably use marine resources. But this can be achieved only if those data and information are properly managed, efficiently and effectively used and distributed.

As an international programme, CSK-2 data and information management will be guided by relevant international oceanographic data policies and practices, including the IOC Oceanographic Data Exchange Policy (Appendix-1), the FAIR Data Principles (Appendix-2), and the United Nations Decade of Ocean Science for Sustainable Developments (2021-2030).

2. Objectives

The objective of the Management Plan is to ensure that CSK-2 data and information will be properly handled, documented, made accessible, and preserved for long-term use.

The purpose of the CSK-2 Data and Information Management Plan (DIMP) is to provide standardized methods and criteria to facilitate the management of data and information generated and/or collected from the CSK programme. It serves as a guide for CSK-2 Principal Investigators (PIs), CSK Data Portal Manager(s), and their institutions; relevant National Oceanographic Data Centers (NODCs) and Associate Data Units (ADUs); and any other partners who wish to play a part in and contribute to CSK-2 data and information exchange and management.

The target user groups of CSK-2 data and information include, but are not limited to, scientific communities, ocean and meteorological agencies, industry/business sectors, educational sectors and local/coastal communities. Those user groups were identified as important stakeholders of CSK-2 and have been gradually engaged in the co-development of the CSK-2 programme.

3. Scope of CSK-2 data and information

This includes the expected types of data and information to be generated during the CSK-2 programme implementation.

“CSK-2 Data” refer to the observational data, model-based data (forecasting and reanalysis), satellite data, and their associated metadata. Physical, biological, biogeochemical, fishery-related, atmospheric, and land-based hydrological parameters are among those likely to be covered by CSK-2 Data.

“CSK-2 Information” will include but not be limited to cruise reports, mooring deployment reports, scientific papers associated with the CSK-2 Programme, and any other information materials resulting from the CSK-2 Programme.

4. Data and information management and exchange architecture

The CSK-2 adopted a distributed data and information exchange network system composed of a series of CSK-2 Data Nodes and a few CSK-2 Data Portal(s) (Figure 1). The CSK-2 Data Portal sites are linked to each other and also to each Data Node.

4.1 CSK-2 Data Nodes and Data Portal(s)

Ideally, a CSK-2 Data Node is developed by one CSK-2 project or several projects (PI or his/her institution), or by accredited oceanographic data center (e.g., NODC and ADU). The CSK-2 Data Nodes are responsible for managing and maintaining the CSK-2 data and information provided by CSK-2 projects, respectively, in accordance with relevant regulations. The CSK-2 Data Node(s) shall be developed using the most common protocols (e.g., http, https or ftp) and must be easily accessible by users, including other CSK-2 Data Nodes and Portals.

The data generated from an individual CSK project are managed by its PI and shared primarily through the Data Node and broader CSK-2 data and information exchange network, as a contribution to the CSK-2 programme. Data and information in the CSK-2 Data Node include, but are not limited to, data, metadata and information generated from the CSK-2 project, specific instructions for using the data, and activities and outputs generated from the project.

In the event that any PI and/or his/her institution is unable to manage their data, it is strongly recommended that she/he submit these data to his/her country's accredited oceanographic data center (e.g., NODC and ADU) or other international oceanographic data repository.

If both the CSK-2 project and the accredited NODC/ADU are unable to manage the data, the CSK-2 Task Force will define a practical arrangement, in consultation with parties concerned.

To provide access to CSK-2 data and information, the CSK-2 Programme will establish a few needs-based CSK-2 Data Portals, managed by ocean research institutions in the region on a voluntary basis. These CSK-2 Data Portals serve as a gateway for a wide range of users, and thus include links to all CSK-2 Data Nodes, CSK-2 data and information inventories, general guidelines for using data, and additional links to other data sites relevant to CSK-2 to ensure the findability (discoverability) of data.

All CSK-2 Data Portals are also linked with each other and other relevant international data portals and catalogues to CSK-2 (e.g., OceanSITES, GEOTRACES, GEO Portal, ODIS) to ensure easy and fair access to the CSK-2 data and information exchange network.

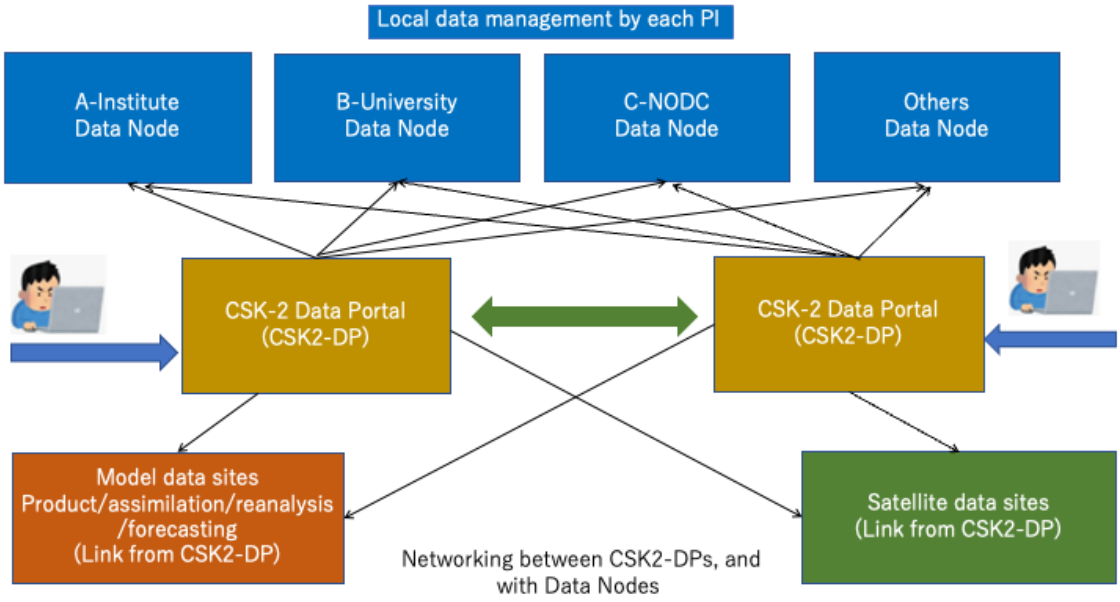

Figure 1 Scheme of the CSK-2 data and information exchange network.

4.2 Metadata

The CSK-2 programme recognized the need for careful description of data to enable data discoverability and to determine data quality. Thus, metadata shall be included, to the greatest extent possible, in the CSK-2 Data Nodes and Data Portals, following generic guidelines as provided below.

a. Metadata of in-situ observations

In principle, metadata of in-situ observations are documented in netCDF format attributes, so there is no need to make a new file. In terms of time-series observations, the OceanSITES netCDF data format is recommended (www.oceansites.org/documents/). For other in-situ observations, as a minimum requirement, relevant information about the PI's name, contact point(s), observation time(s) in UTC, location(s) (lon., lat., and depth), parameters, instruments and sensors, processed or not, and date of process shall be included as netCDF attributes. If data were processed, the QA/QC flags and documentation information shall be included with necessary explanations as a netCDF attribute. While netCDF is strongly recommended, CSV formatted files will also be accepted, especially for non-standard observation feature-type data (biological data).

If for some reason data cannot be shared and disseminated, the development of an independent metadata file, to include the minimum information as described above, is recommended. The metadata file shall be shared/disseminated via either a Data Node or a Data Portal.

b. Metadata of gridded datasets (forecast, reanalysis, satellite)

For gridded datasets generated in the CSK-2 programme, again the use of the netCDF format is recommended. The information to be documented as attributes include PI's name, contact point(s), date(s) and time(s) (start and end time in UTC, time step, grid number), location(s) by latitude, longitude, and depth or height (start and end locations, step, grid number), parameters, dimensions of parameters, and date of generation, as a minimum requirement.

4.3 Data QA/QC

It is recommended that standard methodologies be applied to quality control the data, and to indicate the results of the verification process. All CSK-2 PIs shall conform to the protocols recommended/required by his/her institutions, or other protocols assured by peer-reviewed research. In the absence of the protocols described above, it is recommended that the PIs consider using one of the various protocols recommended below (https://repository.oceanbestpractices.org/handle/11329/336).

QA/QC protocols used shall be written as an attribute in the netCDF files or the independent metadata file.

4.4 Data dissemination

The sharing of data and information will greatly benefit and foster cooperation among institutions, Member States and a wide range of ocean stakeholders. The CSK-2 programme aims to, among other things, promote the sharing of data and information in accordance with the IOC Oceanographic Data Exchange Policy (Appendix-1) and adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles, as described in Appendix-2.

CSK-2 PIs, participating institutions and countries shall endeavor to provide timely, free, and unrestricted access, via CSK-2 Data Nodes, Portals or other publicly accessible platforms, to oceanographic data, associated metadata, and information generated from the CSK-2 Programme. The timeline for data dissemination is subject to relevant practices or regulations of participating institutions or countries. It is recommended that data accessibility should occur no later than two years after the data are collected and verified. For operational purposes, such as weather forecast, we encourage the real-time data to be shared within 24 hours.

In principle, data collected from the open ocean, including associated metadata, shall be shared on a free and unrestricted basis. The sharing of data generated from a coastal nation's jurisdictional waters shall be subject to the relevant regulations of that coastal nation. The CSK-2 programme acknowledges the right of Member States and data originators to determine the terms of such exchange in a manner consistent with international conventions and good practices, where applicable. In any case, the metadata files of all CSK-2 observations shall be provided on a free, timely and unrestricted basis.

In line with the FAIR data principles, CSK-2 Data Nodes are advised to use File Transfer Protocol (FTP), Hyper-Text Transfer Protocol (HTTP), or HTTP Secure (HTTPS) (Accessible) for simplicity and ease of use, to ensure complete metadata for data documentation, and to declare Creative Commons (e.g., CC-BY) for expression of copyright (Reusable). To ensure proper attribution, it is recommended that data originators assign a Digital Object Identifier (DOI) to their data from their associated institution (Findable).

The function of a CSK-2 Data Portal is twofold: a gateway for a wide range of users, and a hub of CSK-2 data and information with links to all Data Nodes and associated forecasting, reanalysis, and satellite datasets. It is recommended that all CSK-2 Data Portals sites follow the ISO 19115 metadata standard (https://www.iso.org/standard/53798.html) and use common and widely used applications, for example, GeoNetwork (https://geonetwork-opensource.org/) and ERDDAP (https://www.ncei.noaa.gov/erddap/index.html) for interoperable data delivery. Data Portals will also serve as providers of information on CSK-2, including its progress, highlighted scientific achievements and societal applications, capacity development and literacy, etc.

4.5 Archival and Preservation

The CSK-2 Data Nodes, Portal(s), participating institutions and Member States should ensure all data generated from CSK-2 are ultimately secure and retrievable for long-term use.

Data obtained or generated from the CSK-2 project(s) could be archived by the PI's institution. If needed, delayed quality assurance/control (QA/QC) shall be conducted in respective CSK-2 Data Node and/or NODC/ADU, with relevant update to the Metadata. Raw data, and all processed data shall be also archived properly by each PI's institutions in accordance with regulations of the institution.

It is strongly recommended that PIs provide a copy of data to the NODC and/or ADU in their home country during the project period, as well as a final copy of data upon the termination of CSK-2 project. If there is neither NODC nor ADU in his/her home country, the PI may seek the assistance of the CSK-2 ISG to register his/her data to any other NODC/ADU or internationally accredited data repository .

5. Implementation timelines

The CSK-2 programme development and implementation will continue to evolve as issues emerge. Although the CSK-2 Data and Information Management Plan cannot be, and should not be, prescriptive in terms of exact timelines, it is essential to outline a reasonable timeframe for all CSK-2 partners to guide the planning, construction and operation of the data and information exchange and management network.

It is recommended that: i. any CSK-2 Data Node be developed and become operational within one year from the start of the CSK-2 project; ii. the zero-order web-based interface of the CSK Data Portal(s) be launched in Oct-Dec 2022, and iii. the whole portal system be in operation in 2023.

6. Data users outreach

The CSK-2 participating institutions, countries, CSK-2-ISG, and its DIMP Task Force shall reach out to potential data users, in whatever way possible, to maximize the value of the data and information to a wide range of stakeholder groups, including researchers, marine and meteorological agencies, and marine resource conservation and management communities.

Appendix-1: IOC Oceanographic Data Exchange Policy (IOC resolution XXII-6)
Appendix-2: FAIR Data Principles

Appendix-1: IOC Oceanographic Data Exchange Policy
(IOC resolution XXII-6: IOC Oceanographic Data Exchange Policy)

Preamble
The timely, free, and unrestricted international exchange of oceanographic data is essential for efficient acquisition, integration, and use of ocean observations gathered by the countries of the world for a wide variety of purposes that include the prediction of weather and climate, operational forecasting of the marine environment, preservation of life, mitigation of human-induced changes in the marine and coastal environment, and the advancement of scientific understanding that makes this possible.
Recognizing both the vital importance of these purposes to all humankind and the role of the IOC and its programmes in this regard, the Member States of the IOC agree that the following clauses shall frame the IOC policy for the international exchange of oceanographic data and its associated metadata.

Clause 1
Member States shall provide timely, free, and unrestricted access to all data, associated metadata, and products generated under the auspices of IOC programmes.

Clause 2
Member States are encouraged to provide timely, free, and unrestricted access to relevant data and associated metadata from non-IOC programmes that are essential for application to the preservation of life, beneficial public use, and protection of the ocean environment, the forecasting of weather, the operational forecasting of the marine environment, the monitoring and modelling of climate, and sustainable development in the marine environment.

Clause 3
Member States are encouraged to provide timely, free, and unrestricted access to oceanographic data and associated metadata, as referred to in Clauses 1 and 2 above, for non-commercial use by the research and education communities, provided that any products or results of such use shall be published in the open literature without delay or restriction.

Clause 4
With the objective of encouraging the participation of governmental and non-governmental marine data-gathering bodies in international oceanographic data exchange and maximizing the contribution of oceanographic data from all sources, this Policy acknowledges the right of Member States and data originators to determine the terms of such exchange, in a manner consistent with international conventions, where applicable.

Clause 5
Use of IODE system: Member States shall, to the best practicable degree, use data centres linked to the World Data System and IODE's National Oceanic Data Centres, such as the World Ocean Database and the Ocean Biogeographic Information System, as long-term repositories for oceanographic data and associated metadata.

Clause 6
Member States shall enhance the capacity in developing countries to obtain and manage oceanographic data and information, and assist them to benefit fully from the exchange of oceanographic data, associated metadata, and products. This shall be achieved through the non-discriminatory transfer of technology and knowledge using appropriate means, including the IOC's Training Education and Mutual Assistance programme and through other relevant IOC programmes.

Definitions

“Free and unrestricted” – non-discriminatory and without charge.
“Without charge” – at no more than the cost of reproduction and delivery, without charge for the data and products themselves.
“Data” – oceanographic observation data, derived data, and gridded fields.
“Metadata” – information about data describing the content, quality, condition, and other characteristics of the data.
“Non-commercial” – not conducted for profit, cost-recovery, or re-sale.
“Timely” – the distribution of data and/or products sufficiently rapidly to be of value for a given application.
“Product” – a value-added enhancement of data applied to a particular application.

Clause 5 was revised in 2019 by Decision IOC-XXX/7.2.1(II) of the Assembly at its 30th session, Paris, 26 June–4 July 2019

Appendix-2: FAIR Data Principles

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Scientific Data*. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

A practical "how to" guidance to go FAIR can be found in the Three-point FAIRification Framework.

Findable
The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.
F1 (Meta)data are assigned a globally unique and persistent identifier
F2 Data are described with rich metadata (defined by R1 below)
F3 Metadata clearly and explicitly include the identifier of the data they describe
F4 (Meta)data are registered or indexed in a searchable resource

Accessible
Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorization.
A1 (Meta)data are retrievable by their identifier using a standardised communications protocol
A1.1 The protocol is open, free, and universally implementable
A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
A2 Metadata are accessible, even when the data are no longer available

Interoperable
The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2 (Meta)data use vocabularies that follow FAIR principles
I3 (Meta)data include qualified references to other (meta)data

Reusable
The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.
R1 (Meta)data are richly described with a plurality of accurate and relevant attributes
R1.1 (Meta)data are released with a clear and accessible data usage license
R1.2 (Meta)data are associated with detailed provenance
R1.3 (Meta)data meet domain-relevant community standards

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).