
# Technical documentation

*Guidance for data centres contributing to GCW*

# Versions

| Version | Date | Comment | Responsible |
|---|---|---|---|
| 0.1 | 2015-11-25 | First draft for internal discussion. | Øystein Godøy |

# Table of Contents

# 1 Introduction

## 1.1 Background

The GCW Portal is the entry point to cryospheric datasets. It offers a web interface that contains information about datasets (metadata). These metadata are harvested on a regular basis from data centres actually managing the data on behalf of the owners/providers of the data.

The GCW Portal utilises interoperability interfaces to metadata and data in order to provide a unified view on the datasets that are relevant for GCW activities. The GCW Portal is also the interface for GCW metadata to WMO Information System (WIS) and WMO Integrated Global Observing System (WIGOS)[1]. The GCW Portal will also facilitate real time access to data through Internet and WMO GTS[2] as requested by the user community. This require a certain level of interoperability at the data level in addition to at the metadata level. On GTS WMO formats (BUFR and GRIB) is required and the GTS Portal can transform into these formats in the dissemination process provided contributing data centres are following the required standards for documentation and interfaces to data.

## 1.2 Scope

This document provides an overview of the GCW Portal system and identifies aspects that has to be handled as well as key performance indicators that are required to operate the GCW Portal system in a sustainable manner.

## 1.3 Intended audience

System managers at the data centres contributing to the GCW Portal system. This includes data centres managing CryoNet data as well as other data centres managing cryospheric data. Requirements are stricter for data centres managing CryoNet stations data than for other contributing data centres. Concerning the roles that should be defined at contributing data centres, the reader is referred to [2] .

## 1.4 Applicable documents

[1] Global Cryosphere Watch (GCW) Implementation Plan, Version 1.5

[2] GCW Portal Operational Manual, Version 0.1

[3] http://globalcryospherewatch.org/

[4] WMO Information System

[5] WMO Core Profile of the ISO 19115

[6] WIGOS, including the metadata standard

[7] The Open Archives Initiative Protocol for Metadata Harvesting, Version 2

[8] OAI-PMH tools

[9] OGC CSW specification

1 Details on how to avoid duplicate information in WIS and WIGOS needs to be defined.

2 For datasets not routed through GTS by other agencies.

[10] GCMD DIF Writers Guide

[11] GCMD Science Keywords

[12] Climate and Forecast Standard Names

[13] WMO Code Lists

[14] NetCDF

[15] Climate and Forecast Conventions

[16] OPeNDAP

[17] UNIDATA's Common Data Model

# 2 Types of contributing data centres

## 2.1 CryoNet

The GCW surface observation network is comprised of a core component (CryoNet). These stations are following GCW measurement practises and have continuous temporal records of a certain length and quality.

Contributing stations are those that provide useful measurements of the cryosphere, but whose data records may be shorter or with large gaps, do not completely follow CryoNet measurement practices, or in some other way do not provide the quality and consistency of data required of CryoNet stations. These stations may be in remote, hard to access regions where cryospheric observations are scarce or in regions where they complement other cryospheric measurements.

CryoNet and contributing stations must expose metadata as well as data in GCW standardised form enabling the GCW Portal to catalogue all datasets, access archived data as well as feed real time data into real time data streams used by the WMO GCW user community.

All CryoNet and contributing stations must provide sufficient information to the minimal requirements of WIS[4] and WIGOS[6] metadata.

## 2.2 Other

In addition to the stations listed above, there is a large number of data centres managing relevant datasets. In order to identify these datasets, the GCW Portal is harvesting metadata from a number of data centres and filter the information for information about the cryosphere. Data centres wishing to contribute to this will also need to follow some minimal requirements concerning interoperability at the metadata and data level.

These requirements are provided below.

# 3 Interoperability interfaces

## 3.1 Metadata

### *3.1.1 Background*

Metadata are generated by the data centres hosting the data sets. Metadata are harvested and ingested in the central catalogue for usage by the GCW Portal user community. GCW Portal metadata are divided in 3 categories:

1. Index metadata for identifying relevant products for a specific purpose.
2. Configuration metadata for tuning of user services for a specific data set.
3. Use metadata for understanding the data accessed.
4. Site metadata for understanding the context in which a dataset has been generated.

The first category is the metadata provided by the data centres and is e.g. GCMD DIF or ISO 19115 (i.e. WIS metadata). The second category is maintained in the central metadata repository and is used for configuration of higher order services like visualisation, transformation etc and is created internally in the GCW Portal based on information retrieved from contributing data centres. The third category is covered e.g. by utilisation of NetCDF files formatted according to the Climate and Forecast Convention where sufficient information to actually use the data is provided. The fourth category links directly to WIGOS metadata. These metadata describes the station, its surroundings, instrumentation, procedures etc. There is some overlap between these metadata and the first category[3].

The GCW Portal harvest metadata to a central repository that is used to search for relevant datasets. It does not utilise distributed search as this is a slower process compared to searching in a central repository. Metadata are harvested at regular intervals and checked for conformance according to the standards identified herein and in [2] .

Regardless of the metadata standard used and the mechanism for transport of the information the following recommendation should be implemented at the repositories.

REC. 1. All datasets should have a unique identifier. This is used to track datasets in the central repository and check for duplicates. The identifier is set by the authoritative source for the dataset.

REC. 2. REC. 1. implies that GCW Portal will not specify or change a unique identifier unless the dataset is hosted by the GCW Portal. This kind of support is currently not supported.

### *3.1.2 Exchange mechanisms for metadata*

#### 3.1.2.1 Introduction

Metadata should be exposed using a suitable interface that allows information on existing datasets as well as changes to the inventory to be conveyed to the GCW Portal. Suitable interfaces for this are e.g. OAI-PMH and OGC CSW[4]. Other interfaces may be evaluated, but

3 Details to be figured out in cooperation with WIGOS and WIS teams.

4 Not fully tested yet.

to ensure a cost effective solution the number interfaces must be limited.

OAI-PMH is the recommended interface to use due to its simplicity and cost effective nature. A number of software solutions supporting this is freely available.

#### 3.1.2.2 OAI-PMH

The Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH) is the recommended interface for exchanging metadata with the GCW Portal. It is a cost effective and robust implementation for exchange of metadata between data centres, is used extensively by WMO Information System and is under consideration for WIGOS metadata exchange. It is much cheaper to implement than most alternatives and there are a number of tools available. Some of these are listed on [8] . Some not listed but worth examining are pyOAI and MOAI.

When implementing OAI-PMH there is a number of GCW recommendation that are based on experience during the initial period of metadata exchange for GCW.

REC. 3. PAI-PMH version 2 must be used.

REC. 4. When implementing OAI-PMH for large repositories containing much more than GCW relevant data, configuration of a dedicated cryosphere or GCW set is strongly recommended as this reduce the load on the GCW Portal which else has to do filtering of all harvested metadata. The name of the set that GCW should harvest has to be communicated and names like “GCW” or “Cryosphere” is recommended. More information is available in OAI-PMH Set specification.

REC. 5. When records are deleted in the contributing data centres catalogues, information on this has to be communicated to the central catalogue. In order to achieve this OAI-PMH identifies the support for deleted records  this through the **deletedRecord** element retrieved in the Identify request. Valid responses are no, persistent and transient. GCW contributing data centres must support **transient** and must maintain transient records for at least 1 month[5]. More information on this feature is available in OAI-PMH specification of deleted records.

REC. 6. The OAI-PMH interface by default offers metadata in Dublin Core. This is insufficient for GCW purposes. Metadata has to be offered in ISO19115 and/or GCMD DIF. Details on these specifications are provided below. In order to proper identify the metadata standards it is recommended to use the following keywords: “dif” for GCMD DIF, “iso” for ISO19115 minimum profile, “wis” for the WMO Core Profile of ISO19115 and “wigos” for WIGOS metadata. The latter is yet not fully defined in XML.

REC. 7. CryoNet stations are required to support WIS and WIGOS metadata[6].

#### 3.1.2.3 OGC CSW

The Open Geospatial Consortium Catalogue Services for the Web (OGC CSW [9] ) is another standard for exposing the content of a catalogue in a standardised form. As for OAI-PMH records are exposed using XML. Compared to OAI-PMH, OGC CSW is a bit more expensive

5 This may change.

6 In the current situation details on these standards should be discussed between the GCW Portal and CryoNet data centres.

to implement from the specification although there are several tools supporting it. It is the recommended exchange mechanism for metadata within the European framework INSPIRE and will be supported by the GCW Portal although OAI-PMH is recommended from a cost benefit perspective.

REC. 8. OGC CSW version 2.0.2 must be used.

REC. 9. It is **not** recommended to embed OGC CSW requests in messaging frameworks like e.g. SOAP.

Details on how to interact with a OGC CSW interface has to be discussed when there is a GCW CryoNet station that wants to use this interface.

#### 3.1.2.4 Other

Other mechanisms like OpenSearch could also potentially be supported in the future, but is currently not supported. CryoNet data centres wishing to test this needs to establish a dialogue with the GCW Portal.

### *3.1.3 Structures*

#### 3.1.3.1 ISO19115[7]

The WMO Core Profile [5] is a profile of the ISO19115 metadata standard and is recommended for use within GCW for discovery metadata. However, ISO19115 is a container that can be populated with several controlled vocabularies in some of the elements. The search model for the GCW Portal is currently built around parameter descriptions using the GCMD Science Keywords [11] . A mapping exist between Climate and Forecast standard names [12] and GCMD Science Keywords.

REC. 10. ISO19115 records must at least state the unique id, temporal and spatial location, scientific content, responsible data centre and PI as well as links to the actual data[8].

REC. 11. ISO19115 records, regardless of whether being mandatory elements or the full WMO Profile must contain GCMD Science Keywords.

REC. 12. It is mandatory that CryoNet and contributing stations at least have one keyword from the WMO CategoryCode list [13] [9]. Relevant keywords for GCW are e.g. weatherObservations, meteorology, hydrology, climatology, glaciology.

REC. 13. All times must be encoded as ISO8601.

Table 1 shows elements in ISO19115 and whether these are **M**andatory, **R**ecommended or **O**ptional, as well as whether they are **U**nique (only one occurrence allowed) and require utilisation of **C**ontrolled vocabularies.

7 Needs to be further elaborated.

8 This recommendation will be revisited.

9 There is currently no way of including this information in GCMD DIF, although a mapping to ISO TopicCategories may be used.

**Table 1: ISO19115 core elements. The WMO Core Profile is more extensive.**

| ISO 19115 Core metadata elements | Description | ISO | GCW |
|---|---|---|---|
| Dataset title | | M | M |
| Dataset reference date | | M | M |
| Dataset responsible party | | O | M |
| Geographic location of the dataset | | O | M |
| Dataset language | | M | M |
| Dataset character set | | O | O |
| Dataset topic category | | | |
| Spatial resolution of the dataset | | | |
| Abstract describing the dataset | | M | M |
| Distribution format | | O | M |
| Additional extent information for the dataset (vertical and temporal) | | O | M |
| Spatial representation type | | O | RC |
| Reference system | | | |
| Lineage | | O | R |
| On-line resource | | O | MC |
| Metadata file identifier | | O | M |
| Metadata standard name | | O | RC |
| Metadata standard version | | O | RC |
| Metadata language | | O | RC |
| Metadata character set | | O | RC |
| Metadata point of contact | | M | M |
| Metadata date stamp | | M | MC |

### 3.1.3.2 GCMD DIF

The Global Change Master Directory (GCMD) Directory Interchange Format (DIF) [10] is a metadata standard that is widely used (e.g. by the Antarctic Master Directory) and that was

Table 2 shows elements in GCMD DIF and whether these are **M**andatory, **R**ecommended or **O**ptional, as well as whether they are **U**nique (only one occurrence allowed) and require utilisation of **C**ontrolled vocabularies.

**Table 2: GCMD DIF elements.**

| GCMD DIF elements | Description | GCMD | GCW |
|---|---|---|---|
| Entry_ID | The **<Entry_ID>** is the unique document identifier of the metadata record. The **<Entry_ID>** is determined by the metadata author or data center contact personnel and may be identical to identifiers used by the data provider's data center or organization. For example, the National Snow and Ice Data Center (NSIDC) Distributed Active Archive Center (DAAC) identifies their metadata records as *NSIDC-xxxx*, where *xxxx* is a numerical designator. Also, the identifier is case insensitive meaning nsidc-xxxx and NSIDC-xxx refer to the same metadata record. | MU | MU |
| Entry_Title | The <Entry_Title> is the title of the data set described by the metadata. | MU | MU |
| Parameters (Science Keywords) | The **<Parameters>** field allows for the specification of Earth science keywords that are representative of the data set being described. These keywords are important for the precise search and retrieval of information from the GCMD. The author must select these keywords from the controlled set of science keywords. The **<Parameters>** field consists of a 7-level hierarchical classification of science keywords | MC | MC |
| ISO Topic Category | The **<ISO_Topic_Category>** field is used to identify the keywords in the ISO 19115 - Geographic Information Metadata (http://www.isotc211.org/) Topic Category Code List. It is a high-level geographic data thematic classification to assist in the grouping and search of available geographic data sets. | MC | MC |
| Data Center | The **<Data Center>** is the data center, organization, or institution responsible for distributing the data. | M | MC |
| Summary | The **<Summary>** field provides a brief description of the data set along with the purpose of the data. This allows potential users to determine if the data set is useful for their needs. | MU | M |
| Metadata Name | The ISO 19115 **<Metadata_Name>** field is used to identify the current DIF standard name. | MU | MC |
| Metadata Version | The **<Metadata_Version>** field is used to identify the current DIF metadata standard. | MU | MU |
| Data Set Citation | The **<Data_Set_Citation>** field allows the author to properly cite the data set producer. | R | R |
| Personnel | **<Personnel>** defines the point of contact for more information about the data set or the metadata. | R | R |
| Instrument | The Instrument or **<Sensor_Name>** is the name of the instrument used to acquire the data. | RC | RC |
| Platform | The Platform or **<Source_Name>** is the name of the platform used to acquire the data. | RC | RC |
| Temporal Coverage | The **<Temporal_Coverage>** field specifies the start and stop dates during which the data was collected. | R | M |
| Paleo-Temporal Coverage | For paleoclimate or geologic data, **<Paleo_Temporal_Coverage>** is the length of time represented by the data collected. | R | O |
| Spatial Coverage | The **<Spatial_Coverage>** field specifies the geographic and vertical (altitude, depth) coverage of the data. | R | M |

| GCMD DIF elements | Description | GCMD | GCW |
|---|---|---|---|
| Location | The **<Location>** field specifies the name of a place on Earth, a location within the Earth, a vertical location, or a location outside of Earth. | RC | OC |
| Data Resolution | The **<Data_Resolution>** field specifies the resolution of the data, which is the difference between two adjacent geographic, vertical, or temporal values. Controlled keywords representing horizontal, vertical and temporal data resolution ranges can be selected. Selection of data resolution ranges will assist users in refining their search for data within specific resolution ranges. | RC | OC |
| Project | The **<Project>** is the name of the scientific program, field campaign, or project from which the data were collected. | R | RC |
| Quality | The **<Quality>** field allows the author to provide information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata. | R | MC |
| Access Constraints | The **<Access_Constraints>** field allows the author to provide information about any constraints for accessing the data set. | R | MC |
| Use Constraints | The **<Use_Constraints>** field allows the author to describe how the data may or may not be used after access is granted to assure the protection of privacy or intellectual property. | R | MC |
| Distribution | The **<Distribution>** field describes media options, size, data format, and fees involved in distributing the data set. | R | RC |
| Data Set Language | **<Data_Set_Language>** describes the language used in the preparation, storage, and description of the data. | RC | RC |
| Data Set Progress | The **<Data_Set_Progress>** describes the production status of the data set regarding its completeness. | RC | RC |
| Related URL | The **<Related_URL>** field specifies links to Internet sites that contain information related to the data, as well as related Internet sites such as project home pages, related data archives/servers, metadata extensions, online software packages, web mapping services, and calibration/validation data. | RC | MC[10] |
| DIF Revision History | The **<DIF_Revision_History>** allows the author to provide a list of changes made to the DIF over time. | R | R |
| Keyword (ancillary keywords) | The **<Keyword>** field allows authors to provide any words or phrases needed to further describe the data set. | R | R |
| Originating Center | The **<Originating_Center>** is the data center or data producer who originally generated the dataset. | R | R |
| Multimedia Sample | The **<Multimedia_Sample>** field allows the author to provide information that will enable the display of a sample image, movie or sound clip within the DIF. | R | O |
| References (Publications) | The <**Reference>** field describes key bibliographic citations pertaining to the data set. | R | R |
| Parent DIF | The **<Parent_DIF>** field allows the capability to relate generalized aggregated metadata records (parents) to metadata records with highly specific information (children). Population of the **<Parent_DIF>** field should be reserved for instances where many metadata records are | R | O |

10 Further guidelines are required compared to GCMD.

| GCMD DIF elements | Description | GCMD | GCW |
|---|---|---|---|
| | basically subsets that can be better represented by one parent metadata record, which describes the entire collection. Typically, the parent metadata record will have many children metadata records, which refer to the parent through the **<Parent_DIF>** field. In some instances, a child may point to more than one parent. The **<Parent_DIF>** is populated with an **<Entry_ID>.** | | |
| IDN Node | The Internal Directory Name (IDN) Node **(<IDN_Node>)** field is used internally to identify association, responsibility and/or ownership of the dataset, service or supplemental information. | R | O |
| DIF Creation Date | The **<DIF_Creation_Date>** specifies the date the metadata record was created. | R | R |
| Last DIF Revision Date | The **<Last_DIF_Revision_Date>** specifies the date the metadata record was created. | R | R |
| Future DIF Review Date | The **<Future_DIF_Review_Date>** allows for the specification of a future date at which the DIF should be reviewed for accuracy of scientific or technical content. | R | R |
| Privacy Status | The **<Private>** field allows the author to restrict the data set description from being publicly available. | RC | RC |
| Extended Metadata | The **<Extended_Metadata>** field will allow organizations to store user defined values within the metadata record without reusing existing GCMD defined metadata fields. | O | O[11] |

REC. 14. GCMD comes with a number of predefined controlled vocabularies that should be used in specific sections of the metadata. As indicated in the table above some sections are free text in GCMD while it is suggested to use controlled vocabularies in GCW context[12].

REC. 15. GCMD do not require a controlled vocabulary for the quality element. GCW should to improve search results[13].

REC. 16. Related_URL has several subtypes. The existing list of type and subtype must be used to allow the GCW Portal to filter the purpose of the URLs provided. When types are "View Data Set Landing Page", "View Extended Metadata", "View Professional Home Page", and "View Project Home Page", no subtype is needed.

REC. 17. All times must be encoded as ISO8601.

#### 3.1.3.3 Other

This section has to be extended with further information on both WIS and WIGOS metadata. There are still some issues under consideration for the practical implementation of the latter. These issues has to be discussed within the GCW community and input provided to the Task Team on WIGOS Metadata.

11 Depends on potential requirements within GCW.

12 These vocabularies has to be developed by the GCW community through the Terminology Team.

13 This work should relate to international activities in this field in the context of e.g. GEO, ICES, WMO etc. and must be coordinated within GCW by the Terminology Team.

## 3.2 Data

### *3.2.1 Background*

While interoperability at the metadata level is important for GCW, international exchange of observations of the cryosphere is vital to success of GCW. This implies both exchange of archived data as well as exchange of real time information. In order to facilitate such exchange of information within the GCW community a certain level of standardisation is required. This standardisation is required to ensure that all users can easily understand the data that is made available and perform intercomparisons as well as use it in analyses.

In this context standardised documentation of data through use metadata as well encoding of the information is required. By use metadata is understood identification of the variables, their structure (e.g. spatiotemporal dimensions and mapping to file format), units of variables, encoding of missing values, quality/accuracy estimates, map projection and coordinate reference system etc.

Application of a common data model simplifies integration and intercomparison of datasets. Application of NetCDF[14] as the file format, utilising the Climate and Forecast[15] convention and serving data through OPeNDAP[16] simplifies the issue of integration and combination of data through the Common Data Model[17] .

REC. 18. Where possible, OPeNDAP should be supported for data access.

Several OPeNDAP implementations exist (e.g. THREDDS, Hyrax and pyDAP). pyDAP can integrate with relational databases. Utilisation of OPeNDAP simplifies handling of both archive and real time data as the real time segmentation of data is performed by the client asking for data. OPeNDAP also minimises the overhead as no files are moved, the client connects to data streams, reads the necessary data and close the connection.

### *3.2.2 Exchange mechanisms for data*

#### 3.2.2.1 Introduction

Traditionally data has been exchanged using FTP in various file formats. Modern technology opens up for other mechanisms for transporting data. Many technologies share some features, but there are differences in complexity and cost of implementation.

#### 3.2.2.2 HTTP/FTP

This is the easiest manner to support data exchange, but it has limitations for large datasets as well as there is no common data model or standardisation of file formats. Often data are served in various ASCII formats that differences from data centre to data centre without any standardised metadata simplifying the process of understanding and using the data. Integration of data from various data centres usually takes much human effort. This is simplified if standardised formats like WMO BUFR or WMO Grib are used, but also for these additional information is required to fully understand the content. Data in NetCDF following the Climate and Forecast Convention is self explainable and connects to the Common Data Model.

Segmentation of real time data has to be supported by the contributing data centre.

#### 3.2.2.3 OPeNDAP

The Data Access Protocol simplifies integration of data from various data centres as it is utilising the Common Data Model, provided input data are encoded according to Climate and Forecast conventions use metadata follows the data and the application of a data stream removes the step of downloading a file and keeping track of this while working on the data. It also allows segmentation of data in variable space and space and time.

#### 3.2.2.4 OGC WFS

OGC Web Feature Service (WFS) is a mechanism allowing subsetting of information, but relies on transferring files in Geography Markup Language (GML). There is no standardised form for use metadata in GML. GML behaves like NetCDF without the Climate and Forecast convention. It is a container that can hold anything.

TBW

#### 3.2.2.5 OGC WCS

OGC Web Coverage Service (WCS) is similar to OGC WFS but focus on information representing phenomena that varies in time and space. Like WFS it transfers files, but the number of file formats may be extended and support e.g. GML, GeoTIFF, HDF-EOS, NetCDF.

TBW

#### 3.2.2.6 OGC WMS map projections

OGC Web Mapping Service (WMS) is useful for visualising maps etc. It provides a graphical representation of data but no access to data in itself.

Each WMS server must support the following map projections:

1. EPSG:32661: WGS 84 / UPS North
2. EPSG:4326: WGS 84
3. EPSG:3408: NSIDC EASE-Grid North
4. EPSG:3410: NSIDC EASE-Grid Global

### *3.2.3 File formats*

#### 3.2.3.1 Introduction

Most of the exchange mechanisms mentioned above transfer files. In order to properly understand the content of a file some use metadata is usually necessary. File formats that embed use metadata (and also discovery metadata) are preferred. NetCDF in itself is not self describing, but NetCDF following the Climate and Forecast Convention is self describing. Adding the NetCDF Attribute Convention for Dataset Discovery embeds full discovery metadata (e.g. originator/PI, constraints etc.) in the file.

#### 3.2.3.2 WMO BUFR

Binary Universal Form for the Representation of meteorological data (BUFR) is a binary data

format maintained by WMO. Its main purpose is operational exchange of real time data and it is adapted for robust transfer on varying bandwidth connections. Data that are supposed to be exchanged using WMO Global Telecommunication System (GTS) should be encoded in WMO BUFR. BUFR is a table driven file format, implying that the format is not self explaining and the user has to have the correct table to understand the content.

TBW

#### 3.2.3.3 WMO Grib

GRIdded Binary (GRIB) is a binary format maintained by WMO. As BUFR, this format is best suited for real time exchange over WMO GTS. It is also a table driven format like BUFR, having the same limitations.

TBW

#### 3.2.3.4 NetCDF/CF

REC. 19. NetCDF following the Climate and Forecast Convention with NetCDF Attribute Convention for Dataset Discovery is recommended for file format.

TBW

#### 3.2.3.5 JSON/GeoJSON

JavaScript Object Notation (JSON) and the geographical extension of this is similar to NetCDF in that it is a container lacking standardised metadata.

TBW

#### 3.2.3.6 XML

Extensible Markup Language (XML) is similar to NetCDF in that it is a container lacking standardised metadata describing its contents. There are many variants of XML and the overhead is large.

TBW