Technical documentation

Guidance for data centres contributing to GCW

Date: 2020-04-22

Versions

Version	Date	Comment	Responsible
0.5	2017-12-31	A major rewrite with restructuring of the content, rephrasing requirements and addition of information on both the GCW Portal and the GCW/SLF software package.	Øystein Godøy Joel Fiddes
0.4	2017-09-06	Minor edits, correction of typos and addition of links.	Joel Fiddes Øystein Godøy
0.3	2017-03-27	Minor edits on OPeNDAP and formats.	Øystein Godøy
0.2	2016-06-20	Included comments from Joel Fiddes.	Joel Fiddes, Øystein Godøy
0.1	2015-11-25	First draft for internal discussion.	Øystein Godøy

Table of Contents

1. Introduction . . . 5
1.1. Background . . . 5
1.2. Scope . . . 5
1.3. Intended audience . . . 5
1.4. Applicable documents . . . 5
2. An outline of the GCW Portal . . . 7
2.1. Background . . . 7
2.2. The concept of metadata . . . 8
2.3. Types of contributing data centres . . . 9
2.3.1. CryoNet . . . 9
2.3.2. Other . . . 10
2.4. Quick reference to recommended interfaces and standards . . . 10
3. Interoperability interfaces . . . 11
3.1. Discovery Metadata . . . 11
3.1.1. Background . . . 11
3.1.2. Exchange mechanisms for discovery metadata . . . 11
3.1.2.1. Introduction . . . 11
3.1.2.2. OAI-PMH . . . 11
3.1.2.3. OGC CSW . . . 11
3.1.2.4. Other . . . 12
3.1.3. Structures for discovery metadata . . . 12
3.1.3.1. ISO19115[1] . . . 15
3.1.3.2. GCMD DIF . . . 15
3.1.3.3. Other . . . 15
4. Data . . . 17
4.1. Background . . . 17
4.2. Exchange mechanisms for data . . . 17
4.2.1. Introduction . . . 17
4.2.2. HTTP/FTP . . . 17
4.2.3. OPeNDAP . . . 18
4.2.4. OGC WFS . . . 18
4.2.5. OGC WCS . . . 18
4.2.6. OGC WMS map projections . . . 18
4.3. File formats . . . 19
4.3.1. Introduction . . . 19
4.3.2. WMO BUFR . . . 19
4.3.3. WMO Grib . . . 19
4.3.4. NetCDF/CF . . . 19

4.3.5. JSON/GeoJSON 20
4.3.6. XML 20
4.4. The GCW/SLF Open Source Software Package 20
4.4.1. Background 20
4.4.2. Overview 20
5. Requirements 23
5.1. Background 23
5.2. Listing 23

1. Introduction

1.1. Background

The intention of the GCW Portal is to be the entry point to datasets describing the cryosphere and form the information basis for the assessment activities of the Global Cryosphere Watch. It offers a web interface that contains information about datasets through discovery metadata provided by the data providers (or host data centre). These discovery 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)[2]. The GCW Portal will also facilitate real time access to data through Internet and WMO GTS[3] 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) are required and the GCW Portal can transform data 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 interoperability tools that simplifies integration of data from a number of sources to a unified virtual data management system.

1.3. Intended audience

System managers at the data centres contributing to the GCW Portal system are the intended audience of this manual. 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

RD 1

Global Cryosphere Watch (GCW) Implementation Plan, Version 1.5 [http://www.wmo.int/pages/prog/www/OSY/Meetings/GCW_AM1/GCW_IP_v1.5(1Nov2015).docx]

RD 2

GCW Portal Operational Manual, Version 0.2.1 (To be published)

RD 3

http://globalcryospherewatch.org/

RD 4

WMO Information System [http://www.wmo.int/pages/prog/www/WIS/]

RD 5

WMO Core Profile of the ISO 19115 [http://www.wmo.int/pages/prog/www/WIS/metadata_en.html]

RD 6

WIGOS [https://www.wmo.int/pages/prog/www/wigos/index_en.html], including the metadata standard

RD 7

The Open Archives Initiative Protocol for Metadata Harvesting, Version 2 [http://www.openarchives.org/OAI/openarchivesprotocol.html]

RD 8

OAI-PMH tools [https://www.openarchives.org/pmh/tools/tools.php]

RD 9

OGC CSW specification [http://www.opengeospatial.org/standards/cat]

RD 10

GCMD DIF Writers Guide [http://gcmd.gsfc.nasa.gov/add/difguide/index.html]

RD 11

GCMD Science Keywords [http://gcmd.nasa.gov/learn/keyword_list.html]

RD 12

Climate and Forecast Standard Names [http://cfconventions.org/standard-names.html]

RD 13

WMO Code Lists [http://wis.wmo.int/2013/metadata/version_1-3-0/WMO_Core_Metadata_Profile_v1.3_Part_2.pdf]

RD 14

NetCDF [http://www.unidata.ucar.edu/software/netcdf/]

RD 15

Climate and Forecast Conventions [http://cfconventions.org/]

RD 16

OPeNDAP [http://opendap.org/]

RD17

UNIDATA's Common Data Model [http://www.unidata.ucar.edu/software/thredds/current/netcdf-java/CDM/]

RD 18

Attribute Convention for Dataset Discovery [http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery_1-3]

[1] Needs to be further elaborated.

[2] Details on how to avoid duplicate information in WIS and WIGOS needs to be defined.

[3] For datasets not routed through GTS by other agencies.

2. An outline of the GCW Portal

2.1. Background

The GCW Data Portal [https://gcw.met.no], or catalogue, is dedicated to data management and to providing specific information on datasets . The data management component is an enabling service in the sense that it identifies relevant datasets and their locations and provides an interface that can be used in the evaluation of GCW data and products. The portal will support simple visualization (generation of maps or diagrams like time series) and transformations such as reformatting and re-projection of data, *if the data are served through the appropriate interfaces and forms*.

GCW data management shall integrate datasets and provides access to data and information on past, present, and future cryospheric conditions. To achieve these results, the data portal must be attached to real-time and near-real-time data management systems and to data archives. While interfacing with existing data management systems, GCW respects partnership and ownership. GCW itself will rely on distributed data management technologies and partners (e.g. CryoNet stations) to establish the GCW catalogue. This process will create a unified interface to datasets in an otherwise fragmented terrain. No information on data (discovery metadata) will be kept in the GCW catalogue without an agreement with the data producer/data owner.

GCW data management follows a metadata driven approach in which datasets are described through discovery metadata exchanged between contributing data centers and the GCW catalogue.

In the GCW context, at least two types of metadata are relevant (see Section 2.2 for more information on metadata). One is “discovery” or index metadata identifying general characteristics of a dataset, including what was measured where and when, potential restrictions on data use, data custodians, and the available interfaces to the actual dataset. This is the type of metadata that will be exchanged within GCW. Another type, “use” metadata, is required when a user has accessed a dataset and begins to use it. Such metadata typically include a specification of variables, units used, how missing values are encoded, and other details on the contents of the dataset. The third type of metadata is interpretation or context metadata for observational data (e.g., data quality, instrumentation used, processing performed, and environmental conditions), which allow data to be interpreted in context. The ingested discovery metadata will be harvested from project specific, national, and international catalogues. Some examples are given in Figure 1. In addition to harvesting existing catalogues, the data management part of the GCW portal will facilitate forms for submission of metadata on datasets not handled by existing catalogues. Successful exchange of metadata will involve some degree of adaptation of systems on either side. However, in order to establish a sustainable system, the number of standards the GCW portal has to support cannot be too many. Furthermore, the actual data also has to be standardised to support integration of data among data providers. Concerning the search model used for the GCW portal, search for scientific parameters is currently based on the GCMD Science Keywords. *All datasets must be documented in the English language.*

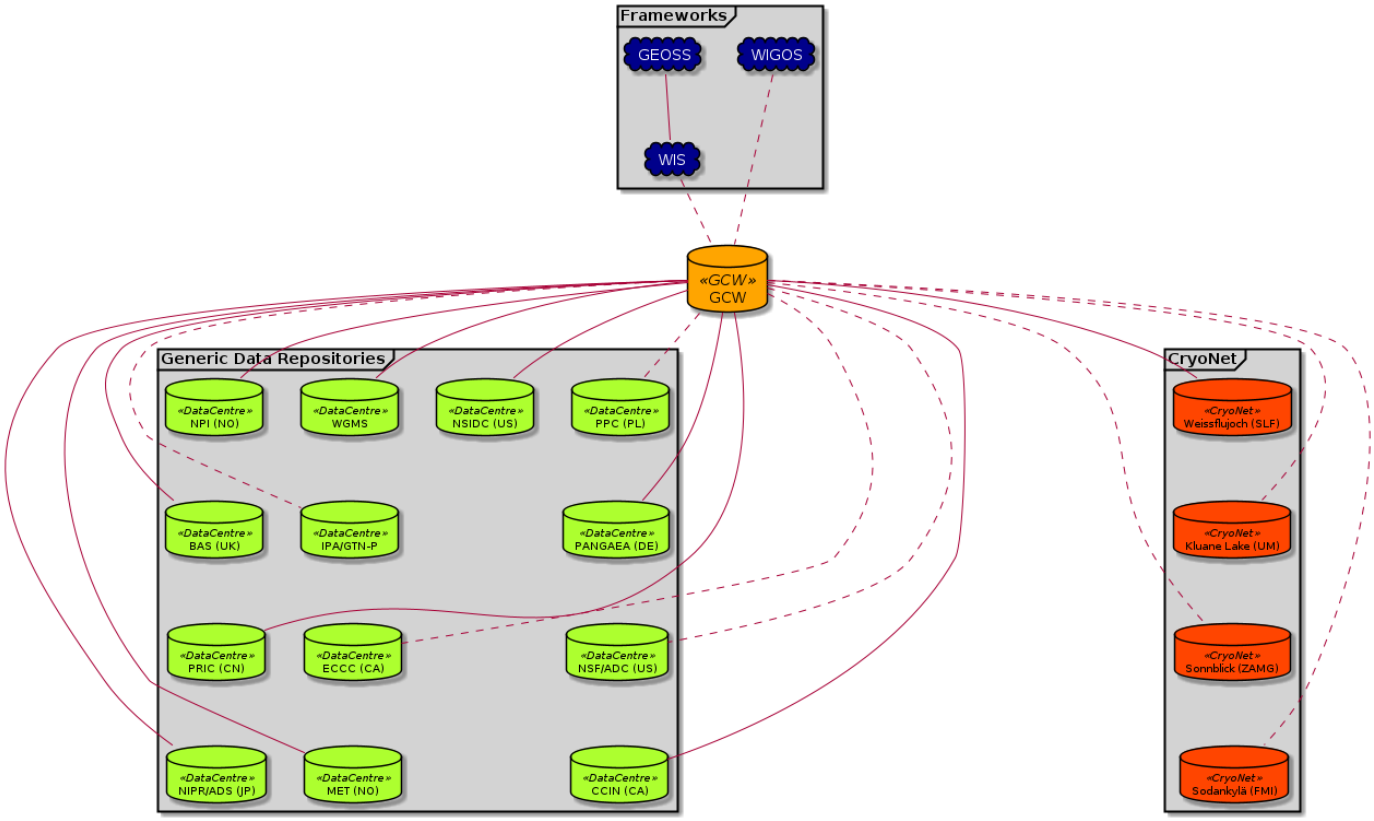


Figure 1. Data centres which the GCW Portal have discussions with.

2.2. The concept of metadata

GCW Portal metadata are divided in 4 categories which are briefly described in Table 1. Among these 4 categories, data providers or supporting data centres are supposed to provide 3 of them. Further below in this document, there are recommendations on how to provide these.

Table 1. Brief introduction to different types of metadata.

Type	Purpose	Description	Examples
Discovery metadata	Used to find relevant data	Discovery metadata are also called index metadata and are a digital version of the library index card. It describes who did what, where and when, how to access data and potential constraints on the data. It shall also link to further information on the data like site metadata. GCW is required to expose this information through WMO Information System as well. Discovery metadata are thus WIS metadata, although the GCW portal can translate to WIS for those not using WMO standards directly.	ISO19115 ISO19115 (WIS) GCMD DIF

Type	Purpose	Description	Examples
Use metadata	Used to understand data found	Use metadata describe the actual content of a dataset and how it is encoded. The purpose is to enable the user to understand the data without any further communication. It describes content of variables using standardised vocabularies, units of variable, encoding of missing values, map projections etc.	Climate and Forecast Convention BUFR GRIB
Configuration metadata	Used to tune portal services for datasets for users.	Configuration metadata are used to improve the services offered through a portal to the user community. This can be e.g. how to best visualise a product. This information is maintained by the GCW portal and is not covered by discovery or use metadata standards.	
Site metadata	Use to understand data found	Site metadata are used to describe the context of observational data. It describes the location of an observation, the instrumentation, procedures etc. To a certain extent it overlaps with discovery metadata, but also extends it. Site metadata can be used for observation network design.	WIGOS OGC O&M

2.3. Types of contributing data centres

2.3.1. 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 [http://globalcryospherewatch.org/about/cryosphere.html].

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 [http://globalcryospherewatch.org/cryonet/site_types.html] must provide sufficient information to the minimal requirements of WIS [4] and WIGOS[6] metadata.

2.3.2. 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.

2.4. Quick reference to recommended interfaces and standards

The text below provides a brief introduction to relevant interfaces for GCW data management. However, in order to establish a manageable system, the number of interfaces has to be a minimum, otherwise the development and maintenance costs of the Portal will be too high and not sustainable. The recommended interfaces for data centres integrating with the GCW Portal are:

1. Discovery metadata
    a. Provide discovery metadata enabling users to search for relevant information using ISO19115 with GCMD Science Keywords for variables or GCMD DIF records through OAI-PMH.
2. Data
    a. Provide data through OPeNDAP using the CF-1.6 convention or higher. One time series per stream, i.e. do not combine many stations into one structure. This allows streaming of data and handling of both real time and archived data.

Further details are provided below.

3. Interoperability interfaces

3.1. Discovery Metadata

3.1.1. Background

Discovery 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.

The GCW Portal harvest discovery metadata to a central repository that facilitates the search process for users. The GCW search is not based on distributed search (forwarding search criteria to supporting catalogues) as this is a slower process compared to searching in a central repository. Discovery 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 data centre repositories.

3.1.2. Exchange mechanisms for discovery metadata

3.1.2.1. Introduction

Discovery 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 OAI-PMH and OGC CSW[4]. Other interfaces may be evaluated, but to ensure a cost effective solution the number of 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 are freely available and a free and open software package (see [3.3]) will be provided.

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 [https://pypi.python.org/pypi/pyoai] and MOAI [http://pypi.python.org/pypi/MOAI].

When implementing OAI-PMH there is a number of GCW recommendations (see [3.3.2]) that are based on experience during the initial period of metadata exchange for GCW.

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 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. If OGC CSW is used it must provide ISO19139 records through GET. The keyword element has to be populated with relevant GCMD Science Keywords.

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 are under implementation, 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 for discovery metadata

The GCW Portal is consuming discovery metadata that are describing the data. In order to make the portal work properly and to ensure integration with WMO systems like WIS, the following discovery metadata elements are required ().

Table 2: GCW discovery metadata elements, purpose and mapping to DIF and ISO19115. Status is one of Mandatory (M), Optional (O) or Recommended ®.

Dataset Identifier	A unique ID for the dataset issued by the responsible data centre.	M	Entry_ID	MD_Metadata> MD_Reference> MD_Identifier
Dataset Title	A brief descriptive title of the dataset suitable for listing purposes.	M	Entry_Title	CI_Citation
Dataset Abstract	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.	M	Summary	MD_Metadata> MD_Identification

Dataset Parameters	Specification of keywords from a controlled vocabulary describing the content of the dataset and that consumers can use to identify the dataset.	M	Parameters	MD_Identification> MD_Keywords
Dataset Temporal Coverage	Specification of the start and stop dates of the dataset. If currently operating, the stop date is empty.	M	Temporal Coverage	EX_Extent> EX_TemporalExtent
Dataset Spatial Coverage	A bounding box for the data specifying the location of the dataset using latitudes and longitudes. Latitudes are positive northwards and longitudes eastwards.	M	Spatial Coverage	EX_Extent> EX_GeographicBoundingBox
Dataset Use Constraints	A description of what a consumer can do with the data after accessing them. In order to protect intellectual property rights (e.g. non commercial use).	M	Use Constraints	MD_Constraints> MD_LegalConstraints
Dataset Creator	Details on the institution and/or people responsible for generation of the dataset.	M	Personnel	CI_Citation> CI_ResponsiblePartyInfo

Dataset Progress	A specification of whether the data production is ongoing, complete or planned.	R	Data Set Progress	MD_Identification
Dataset Operational Status		O		
Dataset Access	Internet links to the data. The type of service behind a link need to be identified by using proper keywords. GCMD content type keywords are required.	M	Related URL	CI_Citation> CI_OnlineResource
Dataset Related Information	Internet link to project or site specific websites providing context information for the dataset.	R	Related URL	CI_Citation> CI_OnlineResource
Dataset Quality	A freetext formulation on the quality of the data. E.g. whether data has been quality controlled or not.	R	Quality	
Data Centre	The Data Center, organisation or institution responsible for maintaining and publishing the data. This is not to be confused with the Dataset Creator. The information required covers relevant contact information as well as URL to the website.	M	Data Center	CI_ResponsiblePartyInfo

Discovery Metadata Last Revision	Specification of the creation date for the discovery metadata or the last revision. The form YYYY-MM-DD must be used.	M	Last DIF Revision Date	MD_Metadata> CI_Date > CI_Date

3.1.3.1. ISO19115[1]

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[5], although this not actively maintained currently. GCW is working with WIGOS to establish a dedicated controlled vocabulary for GCW purpose.

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 used to establish the International Polar Year Data and Information Service (IPYDIS), hosted by the National Snow and Ice Data Center (NSIDC).

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.

[4] Not fully tested yet.

[5] http://cfconventions.org/faq.html#stdnames_mappings

4. Data

4.1. 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. Use metadata can be defined as 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 (cf. Section 3.1.1.).

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].

Several OPeNDAP implementations exist (e.g. THREDDS [http://www.unidata.ucar.edu/software/thredds/current/tds/], Hyrax [http://docs.opendap.org/index.php/Hyrax], ERDDAP [https://coastwatch.pfeg.noaa.gov/erddap/index.html] and pyDAP [http://www.pydap.org/]). 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.

4.2. Exchange mechanisms for data

4.2.1. 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.

4.2.2. 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 differ 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. FTP is not recommended for GCW purposes, HTTP can be the fallback mechanism, but in order to properly identify the physical files in a time series, more discovery metadata is required than if OPeNDAP is used. Anyway, data must be presented in a self explaining format, site or organisation specific formats are not useful.

4.2.3. 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 and it is RESTful[6].

OPeNDAP is the recommended way to exchange data within GCW. It allows access to both archived and real time data. If OPeNDAP is supproted, the GCW portal can extract the most recent observations from a time series and ingest these in WMO GTS if required by the user community. The data served through OPeNDAP has to be mapped to the Climate and Forecast convention 1.6 or higher. Further details on how to encode data in according to this convention is provided in the NetCDF description (see [3.2.3.4]).

4.2.4. OGC WFS

OGC Web Feature Service (WFS) is a mechanism allowing subsetting of information, but it relies on transferring data as 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. Thus, it is not recommended to use OGC WFS in the context of communication with the GCW Portal and it is not supported by the Portal.

4.2.5. 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. However, OPeNDAP can serve the purpose of both WFS and WCS at a much lower cost. The GCW Portal does not support OGC WCS.

4.2.6. 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. The WMS service must provide a GetCapabilities document per dataset, not combining multiple datasets into the same GetCapabilities document for the GCW Portal to be able to consume the OGC WMS service.

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

4.3. File formats

4.3.1. 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 (CF) is self describing. Adding the NetCDF Attribute Convention for Dataset Discovery embeds full discovery metadata (e.g. originator/PI, constraints etc.) in the file. *The NetCDF/CF format is strongly recommended in most cases due to widespread use in the scientific community and ease of implementation.* Some data may not fit into the model of NetCDF, but with the CF 1.8 convention there is now support for standardised descriptions of gridded data (including satellite swath data), time series at stations, time series of profiles at stations, timeseries at moving stations (trajectories) and geometries which formerly was kept in KML, ESRI Shapefiles or similar. The NetCDF container is thus providing a standardised encoding for most of the data GCW is concerned with as long as the CF convention version 1.8 or higher is used.

4.3.2. 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.

If CryoNet sites are providing observational data according to the NetCDF/CF convention over OPeNDAP, the GCW Portal will, if requested, create BUFR files and submit these onto GTS if required by the GCW Community. Furthermore, currently the GCW Portal converts GTS circulated BUFR to NetCDF/CF, this functionality of moving between NetCDF and BUFR will be extended and incorporated as a transformation service in the GCW Portal as well.

4.3.3. 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. The GCW Portal has developed functionality to map between GRIB and NetCDF/CF, although this is yet not fully included in the GCW Portal yet.

4.3.4. NetCDF/CF

This ensures a self explaining dataset where structure and content are encoded using an accepted standard that has impact beyond the original community. It can be used to handle gridded data, time series, profiles and trajectories in standardised manner if encoded according to Climate and Forecast conventions [15]. Furthermore, it includes semantics in a manner which can be used to cross walk content with other structured data descriptions.

Usage of NetCDF/CF version 1.6 or higher is recommended for gridded and observational data. If outline

data are handled, version 1.8 or higher is recommended. It is also recommended to add ACDD[18] global attributes to the NetCDF/CF files in order to incorporate the discovery metadata in the actual data.

4.3.5. 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. The consequence is that combination of data from various sources is not straightforward.

JSON/GeoJSON is not supported by the GCW Portal.

4.3.6. 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. The consequence is that combination of data from various sources is not straightforward.

XML is not supported by the GCW Portal.

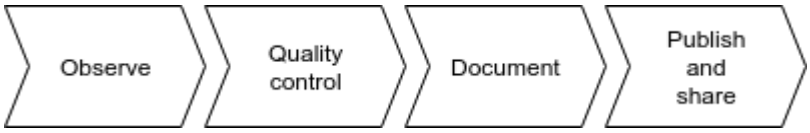
4.4. The GCW/SLF Open Source Software Package

4.4.1. Background

WMO Global Cryosphere Watch (GCW) is depending on a number of observing stations (CryoNet stations) for feeding the GCW value chain with observations. GCW has a requirement for both real time and archived data. In the period 2015-2017, GCW has been working with WSL/SLF to set up interoperability between the WSL/SLF data centre being responsible for one of the CryoNet stations. WSL/SLF has kindly agreed to make the software stack they have developed available for a wider community.

The following is a brief description of several software tool used at the WSL Institute for Snow and Avalanche Research (SLF) [https://www.wsl.ch/en/about-wsl/organization/locations/slf-davos.html] to processes and manage data at various stages of the “datacycle” from sensors to published dataset. All projects are available under open source licenses.

The software package is designed to support the value chain provided in .



4.4.2. Overview

The GCW/SLF software package consists of several modules. The purpose of these modules is listed below and how the components work together to support the purpose of the GCW portal is illustrated in .

The core element in the software package is the data preprocessor MeteoIO that takes data from the sensor, through a quality control procedure into standardised NetCDF/CF files which can be published.

MeteoIO was originally developed to provide robust meteorological forcing data to an operational model that forms part of the avalanche forecast at the SLF. However, it also happens to be very good at reading diverse data sources and producing a standardised output. It has a modular architecture which makes it flexible and fast to develop new use cases. It can handle both gridded and time series data and has various functions for cleaning/ processing data to various quality standards and produces QA reports. MeteoIO is C++ library.

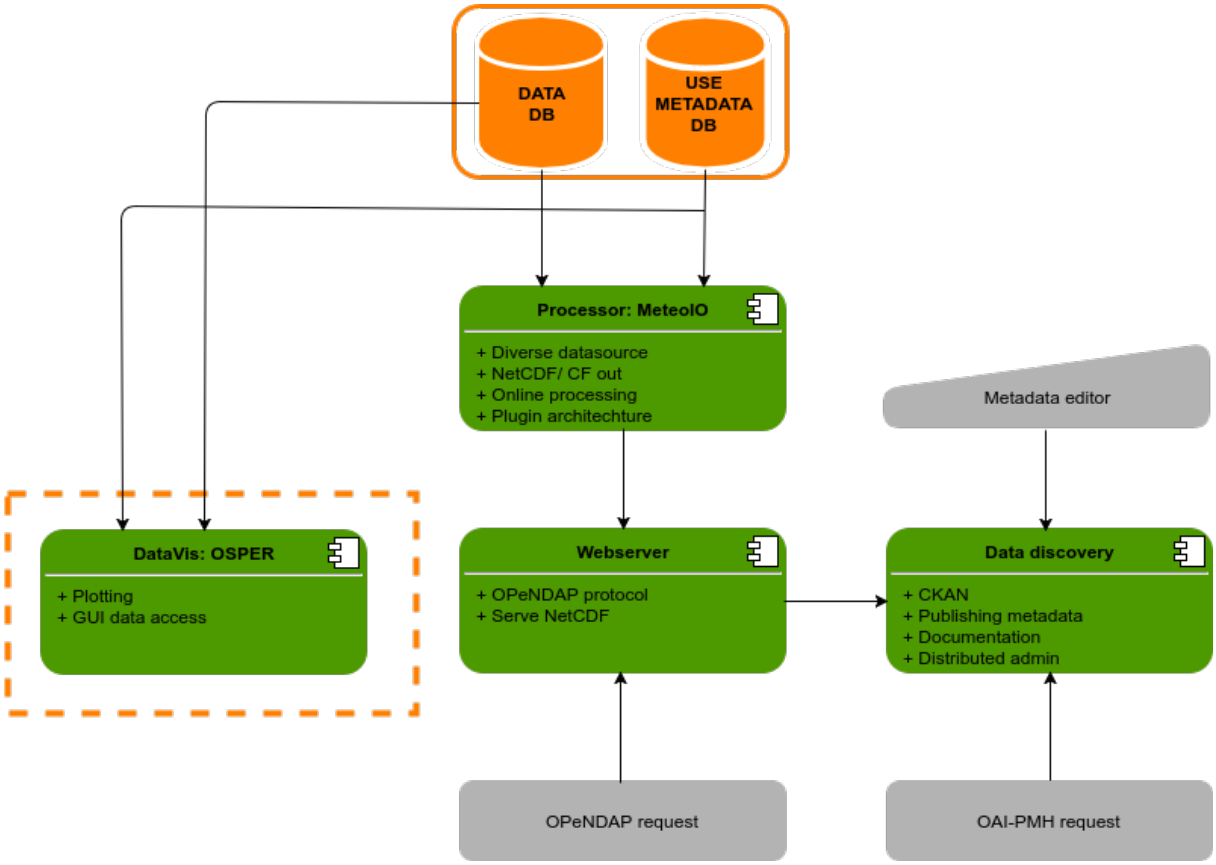
MeteoIO git: https://models.slf.ch/p/meteoio/

In order to publish discovery metadata for the data prepared through MeteoIO, software developed through the ENVIDAT project is used. EnviDat is the WSL/SLF main CKAN based dataportal and metadata repository. Core CKAN has been extended to cover specific requirements of research data management. These include an OAI-PMH server, DOI publishing and supporting metadata standards. The advantage of CKAN is that it provides a robust and intuitive UI for structured metadata submission. This enables large parts of the data management process to be decentralised to the submitter.

CKAN project: https://ckan.org/

EnviDat extensions are here: https://github.com/EnviDat

The OPeNDAP component providing external access to the data preprocessed through MeteoIO and announced through ENVIDAT is currently under development.



[6] http://apievangelist.com/2014/12/05/history-of-apis-noaa-apis-have-been-restful-for-over-20-years/

5. Requirements

5.1. Background

The following chapter is listing a number of requirements related to the information provided above. In the long term this information may be extracted into a separate document, but in order to simplify the process it is listed herein for now.

5.2. Listing

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.
2. [REC. 1.] implies that GCW Portal **must** 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. OAI-PMH **should** be used for exchange of discovery metadata.
4. OAI-PMH version 2 **must** be used if OAI-PMH is used for exchange of discovery metadata..
5. When implementing OAI-PMH for large repositories containing much more than GCW relevant data, configuration of a dedicated cryosphere or GCW set **should** be supported as this reduce the load on the GCW Portal which otherwise has to perform 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 [http://www.openarchives.org/OAI/openarchivesprotocol.html#Set].
6. 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[7]. More information on this feature is available in OAI-PMH specification of deleted records [http://www.openarchives.org/OAI/openarchivesprotocol.html#DeletedRecords].
7. The OAI-PMH interface by default offers discovery metadata in Dublin Core. This is insufficient for GCW purposes. Discovery metadata **must** be offered in ISO19115 and/or GCMD DIF. Details on these specifications are provided below. In order to properly identify the metadata standards in the responses provided by the OAI-PMH end point, 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 in the “ListMetadataFormats” response. The latter is yet not fully defined in XML.
8. CryoNet stations **must** provide WIS and WIGOS metadata[8].
9. Discovery metadata **must** be available in the English language.
10. OGC CSW version 2.0.2 **must** be used if OGC CSW is used for exchange of discovery metadata.
11. If OGC CSW is used for exchange of discovery metadata, the implementation **must** support HTTP GET (key, value in URL).
12. OGC CSW requests **must** not be embedded in messaging frameworks like SOAP. This will not be

supported by the GCW Portal.

13. 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[9].
14. ISO19115 records, regardless of whether being mandatory elements or the full WMO Profile **must** contain GCMD Science Keywords describing the parameters provided in the data.
15. CryoNet and contributing stations **must** have at least have one keyword from the WMO CategoryCode list [13][10]. Relevant keywords for GCW are e.g. weatherObservations, meteorology, hydrology, climatology, glaciology.
16. All times **must** be encoded as ISO8601 in the form YYYY-MM-DDTHH:MM:SS and in UTC.
17. GCMD comes with a number of predefined controlled vocabularies that should be used in specific sections of the metadata. As indicated above some sections are free text in GCMD while it is suggested to use controlled vocabularies in GCW context[11].
18. GCMD do not require a controlled vocabulary for the quality element. GCW records **should** to improve search results[12].
19. Within GCMD DIF Related_URL has several subtypes. The existing list of type and subtype [http://gcmdservices.gsfc.nasa.gov/static/kms/rucontenttype/rucontenttype.csv] **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.
20. OPeNDAP **should** be supported for data access. Data **should** be encoded according to the Climate and Forecast Convention version 1.6 or higher including the featureType attribute to identify gridded data versus time series, profiles or trajectories.
21. OGC WFS and OGC WCS **should** not be used for data exchange.
22. NetCDF following the Climate and Forecast Convention with NetCDF Attribute Convention for Dataset Discovery **should** be used as file format.

[7] This may change.

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

[9] This recommendation will be revisited.

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

[11] These vocabularies has to be developed by the GCW community through the Terminology Team.

[12] 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.