Technical documentation

Guidance for data centres contributing to YOPP

Versions

Version	Date	Comment	Responsible
0.4	2018-01-28	Updated following developments in the GCW interoperability guidelines.	Amélie Neuville Siri Jodha Singh Khalsa Øystein Godøy
0.3	2017-03-09	Updated following YOPP data management meeting and discussions.	Øystein Godøy Eivind Støylen
0.2	2016-06-20	Included comments from Joel Fiddes.	Øystein Godøy
0.1	2015-11-25	First draft for internal discussion.	Øystein Godøy

Table of Contents

1 Introduction

1.1 Background

The YOPP Data Portal is the entry point for YOPP datasets. It offers a web interface that contains information about datasets (through discovery 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 YOPP Data Portal utilises interoperability interfaces to metadata and data in order to provide a unified view on the datasets that are relevant for YOPP activities. The YOPP Data Portal is also the interface for YOPP metadata to WMO Information System (WIS) and WMO Integrated Global Observing System (WIGOS)[1]. The YOPP Data 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 YOPP Data 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.

This document is based upon a similar document developed for the Global Cryosphere Watch.

1.2 1.2Scope

This document provides an overview of the YOPP Data Portal interoperability interfaces that simplifies integration of data from a number of sources to a unified virtual data management system. It list more interfaces/standards than what is supported today, and initially the support will be limited.

1.3 Intended audience

System managers at the data centres contributing to the YOPP Data Portal. Concerning the roles that should be defined at contributing data centres, the reader is referred to [2] .

1.4 Applicable documents

[1] YOPP Data Portal concept, Version 0.1
[2] YOPP Data Portal Operational Manual, Version 0.3
[3] http://www.polarprediction.net/yopp/
[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

[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. Details needs to be investigated and are constrained by the available bandwidth.

[8] OAI-PMH tools
[9] OGC CSW specification
[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

Contributing data centres support the activities of YOPP and manage datasets of relevance to YOPP. In order to identify these datasets, the YOPP Data Portal is harvesting metadata from the data centres. Data centres contributing to YOPP data management will need to follow minimal requirements concerning interoperability at the metadata and data level.

Options for interoperability interfaces and standards are described and discussed below. The emphasis has been to establish a cost efficient system and reduce the housekeeping needed.

3 Interoperability interfaces

3.1 Metadata

3.1.1 Background

Metadata are generated by the data centres hosting the data sets. YOPP Data Portal metadata are divided in 4 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 YOPP Data 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 (CF) Convention where sufficient information to actually use the data is provided. The YOPP Data Portal requires CF version 1.6 or higher. 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 YOPP Data 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 the YOPP Data Portal will not specify or change a unique identifier unless the dataset is hosted by the YOPP Data Portal[4].

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 YOPP Data Portal. Suitable interfaces for this are e.g. OAI-PMH and OGC CSW[5]. 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 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 YOPP Data Portal. It is a cost effective and robust implementation for exchange of metadata between data centres, is used extensively by WMO Information System. It is much cheaper to implement than most alternatives, there are a number of tools available and it is reasonably standardised. Some of these are listed on [8] . Some not listed but worth examining are pyOAI and MOAI.

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

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

REC. 4. When implementing OAI-PMH for large repositories containing much more than YOPP relevant data, configuration of a dedicated YOPP set is strongly recommended

---

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

[4] This kind of support is currently not supported..

[5] Not fully tested yet.

as this reduces the load on the YOPP Data Portal which otherwise has to perform filtering of all harvested metadata. The name of the set that YOPP should harvest has to be communicated and a name like "YOPP" 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 through the **deletedRecord** element retrieved in the Identify request. Valid responses are no, persistent and transient. YOPP contributing data centres must support **transient** and must maintain transient records for at least 1 month[6]. 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 YOPP purposes. Metadata has to 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.

REC. 7. YOPP observational data should have attached both WIS and WIGOS metadata[7].

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 YOPP Data 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 YOPP contributing centre that wants to use this interface.

3.1.2.4 Other

Other mechanisms like OpenSearch could also potentially be supported[8]. YOPP data centres wishing to test this needs to establish a dialogue with the YOPP Data Portal.

---

[6] This may change.

[7] In the current situation details on these standards should be discussed between the YOPP Data Portal and YOPP data centres.

[8] The software supports this, but details needs to be tested.

3.1.3 Structures

3.1.3.1 Discovery metadata elements

The table below shows the discovery metadata elements that the YOPP data portal rely on, including references to ISO19115 and GCMD DIF which are the discovery metadata standards supported by most data centres within the domain. The sections below provides specific recommendations for the discovery metadata standards mentioned. The table describes the purpose for a specific field as well as where to put it in various standards. In order to support higher order functionality on datasets (e.g. visualisation, transformation/subsetting), specific information on the access mechanisms is required using controlled vocabularies.

Table 1: YOPP discovery metadata elements, purpose and mapping to DIF and ISO19115. Status is one of Mandatory (M), Optional (O) or Recommended (R).

Element	Purpose	Status YOPP	GCMD DIF	ISO19115
Dataset Identifier	A unique ID for the dataset issued by the responsible data centre. 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.	M	Entry_ID	MD_Metadata> 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
ISO Topic Category	Identification of the keywords in the ISO 19115 - Geographic Information Metadata (http://www.isotc211.org/) Topic	M	ISO_Topic_Category	MD_DataIdentification MD_TopicCategoryCode

	Category Code List. High-level geographic data thematic classification.			
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.	R	Personnel	CI_Citation CI_Responsibility
Metadata point of contact	Details on the institution and/or people responsible for generation of the metadata.	M	Personnel	CI_Citation CI_Responsibility
Dataset Progress	A specification of whether the data production is ongoing, complete or planned.	R	Data_Set_Progress	MD_Identification
Dataset Operational Status	A specification of the operational status of the product/dataset. E.g. whether it is scientific, experimental, pre-operational or operational.	O	Quality	MD_Metadata dataQualityInfo
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 MD_Distribution
Dataset Related Information	Internet link to project or site specific websites providing context information for the dataset.	M[9]	Related_URL	CI_Citation> CI_OnlineResource
Data Set Citation	Citation of the dataset producer.	R	Data_Set_Citation	CI_Citation

[9] Further guidelines are required compared to GCMD.

Project	Name of the scientific program, field campaign, or project from which the data were collected.	R	Project	MD_Identification> MD_Keywords
Dataset Quality	A free text formulation on the quality of the data. E.g. whether data has been quality controlled or not.	M	Quality	MD_Metadata> DQ_DataQuality
Dataset responsible party	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_Responsibility
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 DIF_Creation_ Date	MD_Metadata> CI_Date > CI_Date

3.1.3.2 ISO19115

The WMO Core Profile [5] is a profile of the ISO19115 metadata standard and is recommended for use within YOPP 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 YOPP Data 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. Usage of ISO19115-3 is recommended.

REC. 11. 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.

REC. 12. ISO19115 records, regardless of whether being mandatory elements or the full WMO Profile must contain GCMD Science Keywords.

REC. 13. It is mandatory that datasets at least have one keyword from the WMO CategoryCode list [13] [10]. Relevant keywords for YOPP are e.g. weatherObservations, meteorology, hydrology, climatology, glaciology.

REC. 14. All times must be encoded as ISO8601.

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

3.1.3.4 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).

REC. 15. 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 YOPP context[11].

REC. 16. GCMD do not require a controlled vocabulary for the quality element. YOPP should to improve search results[12].

REC. 17. Related_URL has several subtypes. The existing list of type and subtype must be used to allow the YOPP Data 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. 18. All times must be encoded as ISO8601.

3.1.3.5 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 YOPP community and input provided to the Task Team on WIGOS Metadata.

3.2 Data

3.2.1 Background

While interoperability at the metadata level is important to achieve an overview of the relevant data, exchange of simulations and observations are vital to the success of YOPP. This implies both exchange of archived data as well as exchange of real time information. In order to facilitate such exchange of information 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. Interoperability at the data level relies on standardised documentation and encoding of data through use metadata. Use metadata identifies 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.

---

[11] These vocabularies has to be developed by the YOPP community.

[12] This work should relate to international activities in this field in the context of e.g. GEO, ICES, WMO etc..

Application of a common data model simplifies integration and intercomparison of datasets. Application of NetCDF [14] as the primary 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. 19. Where possible, OPeNDAP should be supported for data access.

Several OPeNDAP implementations exist (e.g. THREDDS, Hyrax, ERDDAP and pyDAP). 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 minimises the overhead as no files are moved, the client connects to data streams, reads the necessary data and close the connection, removing the need for housekeeping transient files.

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 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.

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 time and it is RESTful[13]. The YOPP Data Portal supports operations on top of OPeNDAP for gridded datasets. These operations are under development for other types of data as well.

---

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

3.2.2.4 3.2.2.4OGC WFS

OGC Web Feature Service (WFS) is a mechanism allowing subsetting of information, but relies by default 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.

OGC WFS is considered a non-RESTful web service and is currently not supported by the YOPP Data Portal.

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. Like WMS, WCS can also transform a set of files to a common map projection and extract a specific area of interest in space and time by "trimming" or "slicing". The YOPP Data Portal does not support operations on top of OGC WCS today. These operations are however supported for OPeNDAP.

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:3409: NSIDC EASE-Grid South
5. 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 3.2.3.2WMO 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) must 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. The YOPP Data Portal is working on software for dumping data fra BUFR to NetCDF/CF. This has been setup for dumping SYNOP and TEMP data from WMO GTS for the Arctic and making them available through the Data Portal (in progress).

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. To a certain extent the YOPP data portal can support conversion between GRIB and NetCDF/CF provided relevant tables are available. However, as there are many tables circulating, all conversions have not been tested.

3.2.3.4 NetCDF/CF

REC. 20. NetCDF following the Climate and Forecast Convention with NetCDF Attribute Convention for Dataset Discovery is recommended for file format where possible as it is a dynamic standard with a semantic framework and it maps directly to the generic Common Data Model.

REC. 21. It is recommended to add the featureType global attribute to datasets that are not of gridded type. Without this attribute, datasets will be assumed to be gridded.

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. Furthermore, it includes semantics in a manner which can be used to cross walk content with other structured data descriptions.

When data are encoded using NetCDF, the YOPP Data Portal require data to be encoded according to the Climate and Forecast Convention (CF-1.6 or higher). Preferably datasets should include the Attribute Convention for Dataset Discovery as well to support discovery metadata. For non gridded datasets it is important to use the global attribute featureType and to set this to timeseries, profile or trajectory if the data served are not gridded (e.g. remote sensing or numerical simulations).

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. The consequence is that combination of data from various sources is not straightforward. It is not recommended to use this for dataset publication.

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. The consequence is that combination of data from various sources is not straightforward. It is not recommended to use this for dataset publication.