Technical documentation

YOPP Data Portal concept

Versions

Version	Date	Comment	Responsible
0.3	2018-02-09	Minor modifications to clarify concept.	Øystein Godøy
0.2	2017-05-09	Integrated comments from Siri Jodha Singh Khalsa.	Øystein Godøy
0.1	2016-12-12	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 or received by e.g. email.

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, but the data is managed and hosted by contributing data centres.

The YOPP Data Portal is the interface for YOPP metadata to WMO Information System (WIS). The YOPP Data Portal will facilitate real time access to data through Internet and WMO Global Telecommunication System (GTS)[1] 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 WMO GTS WMO formats (BUFR and GRIB) are required and the YOPP Data Portal will develop functionality to transform NetCDF/CF into these formats in the dissemination process provided contributing data centres are following the required standards for documentation and interfaces to data [1] . For data consumers not connected to WMO GTS, the YOPP Data Portal will make free data from GTS available as NetCDF/CF served through OPeNDAP. This will initially be done for SYNOP, SHIP and TEMP.

1.2 Scope

This document describes the overall concept for the YOPP Data Portal. It is a very brief presentation of the distributed nature and general concept, not a full architectural design document.

1.3 Intended audience

The primary audience of this document is the system managers maintaining the data management systems at the data centres contributing to the YOPP Data Portal as well as other interested parties.

1.4 Applicable documents

[1] Guidance for data centres contributing to YOPP.

[2] Operations manual for YOPP contributing data centres.

[3] http://www.polarprediction.net/yopp/

2 System overview

The YOPP Data Portal is a metadata driven system where metadata describes the datasets and the interfaces available for the specific dataset. Metadata are collected in a central metadata

1 For datasets not routed through GTS by other agencies. Details needs to be investigated and are constrained by the available bandwidth.

catalogue while the data is served directly from the individual data centres contributing to the YOPP Data Portal. An overview of the YOPP Data Portal components and services is provided in Figure 1, and for the metadata flow in Figure 2. The latter is under continuous development and should only be considered as an illustration.

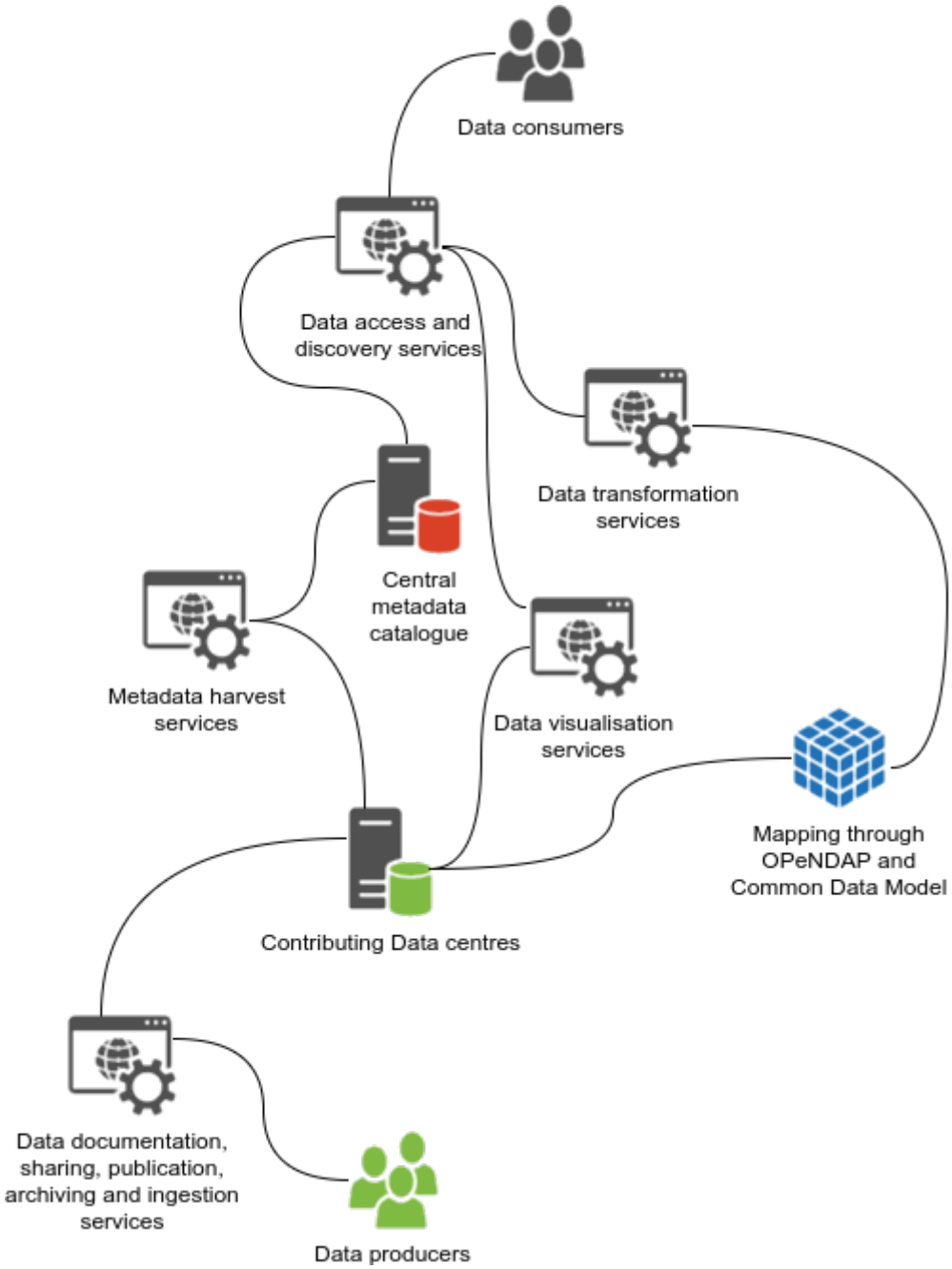

Figure 1: Conceptual overview of the YOPP Data Portal, including components and interfaces.

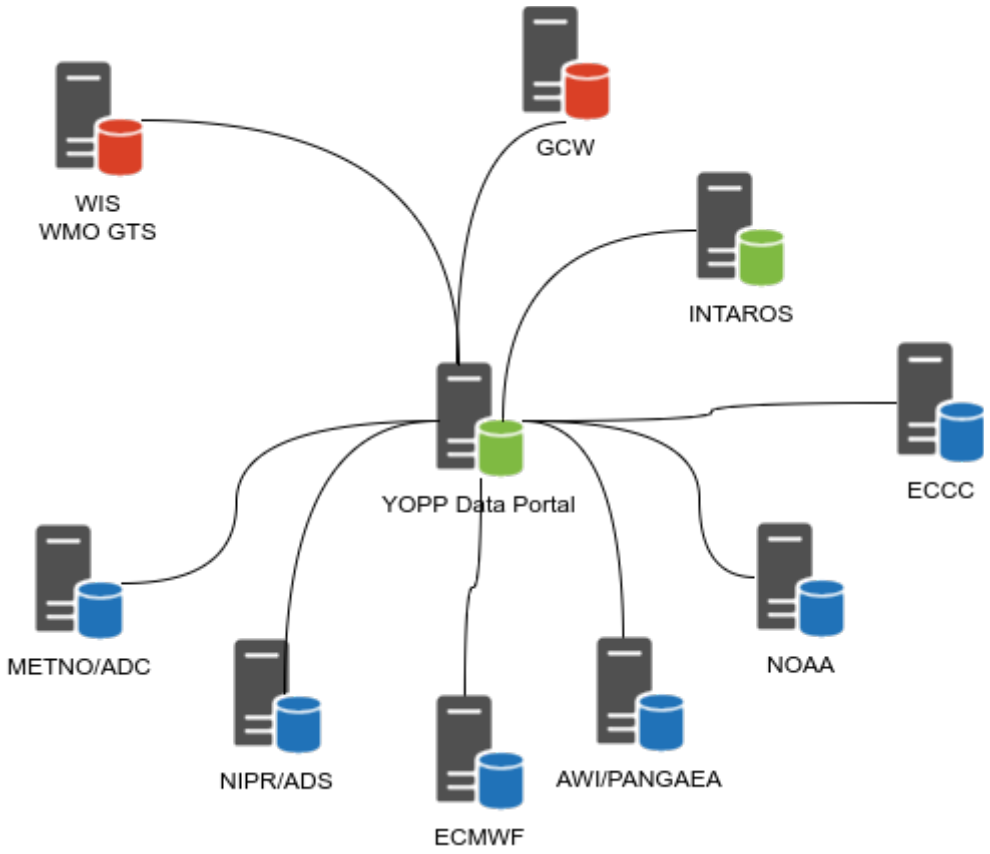

Figure 2: Schematic overview of the YOPP Data Portal nodes.

In relation to the YOPP Data Portal a number of roles are usually defined. These are required to operate a data centre, to integrate across data centres and to fill the data centres with actual content.

Data manager – The data manager is responsible for the content of the data centre. The data manager guides data providers in the process of documenting and formatting data as well on submission. The data manager is responsible for the life cycle management of data (both real time dissemination and long term preservation). The data manager is responsible for interaction with other data centres and with the YOPP Data Portal[2].

System manager – The system manager is responsible for the technical framework used to manage and communicate data. This includes the web based tools made available to the data manager, data provider and data consumer.

Data provider – The data provider is the Principal Investigator responsible for the dataset being made available to YOPP. A data provider may be responsible for one or more datasets and relates to one or more data managers depending on the routing (real time and long term preservation) of datasets.

Data consumer – The data consumer is a user of datasets through the YOPP Data Portal (whether archived or real time). The data consumer depends on the work performed by

2 Data managers of contributing data centres will be part of the YOPP Data Task Team.

the data provider, data manager and system manager in order to find (through discovery metadata), access (through interfaces to data) and use (through use metadata fully describing the content of a dataset) relevant datasets.

3 Functionality

In order of priority, the key functionality for data consumers will be:

1. Unified data discovery, understood as the process of finding relevant datasets across the distributed data repositories contributing to YOPP.
2. Retrieval of data in the raw form offered by the YOPP contributing data centres, understood as the process of downloading data identified in the previous step. If direct access to data is not possible/allowed, provision of information on how to access the data.
3. Visualisation of data, understood as the process of generating a graphical interpretation of a dataset (either as a map, a time series or appropriate) for data identified.
4. Transformation of data, understood as the process of reformatting, reprojecting, subsetting and combining different datasets into a new dataset.

The prioritisation provided above is both from a user and implementation perspective. It starts with the easy wins and continues with more advanced functionality. However, the fallback will always be unified discovery. Often interoperability interfaces to data are missing and there is not resources to develop support for all types of interoperability interfaces.

It is anticipated that the contributing data centres forming already have procedures and software for receiving and managing metadata and data. They may however lack

- Translations of internal metadata to international standards or the standards supported by the YOPP Data Portal.
- Interfaces to metadata that scale and are sustainable from a data management perspective[3].
- Interfaces to data[4].
- Monitoring systems for data usage and system availability

4 Questions and answers raised during discussions in YOPP preparation

Some questions and issues that should be discussed further. These questions will be moved into a FAQ and an issue tracker at https://yopp.met.no/ but are left here for now.

- Will it be possible to order data in advance for a specific area and time?

3 E.g. automatic identification of datasets that are deleted/superseeded.

4 The YOPP Data Portal will rely on online access to data from the catalogue interface. Experience indicates however that standardised documentation often is missing and data access interfaces are very simple.

- This depends on the contributing data centres and the data requested (whether a contributing data centre actually offers this). It is not supported currently, but could be put as a feature request.

- How do users provide feedback on data quality?
  - This should be directed towards the responsible PI for a dataset.
- Is it possible for the YOPP Data Portal to host data (e.g. verficiation data)?
  - Potentially, the portal or contributing data centres could probably host data, but it would probably require some cost recovery for storage space.
- Will the YOPP Data Portal support Single Sign On (for easier access to restricted information)?
  - This depends on the support from the contributing data centres as well. It is not supported currently, but might be added if contributing data centres also want to support this.
- Can the YOPP Data Portal guide users with “homeless data” to a data centre willing to host the data?
  - If contributing data centres are willing to support this, a guidance document identifying archives (YOPP contributing data centres or commercial) and how to contact these could be created. The fee (if applicable) would then be negotiated between the data centre and the PI/responsible institution. Whether it is possible to integrate these data in the portal depends on the interoperability interfaces offered.
- Can real time data be identified in the portal (e.g. through keywords like “new”, “up-to-date”)?
  - Not currently, this will be investigated, but will require access to the actual data and not only discovery level metadata. In the short term this is only feasible for data through OpeNDAP.