Bavay, M, et al. 2020. Automatic Data Standardization for the Global Cryosphere Watch Data Portal. *Data Science Journal,* 19: 6, pp. 1–8.
DOI: https://doi.org/10.5334/dsj-2020-006

PRACTICE PAPER

Automatic Data Standardization for the Global Cryosphere Watch Data Portal

Mathias Bavay[1], Joel Fiddes[2,3] and Øystein Godøy[3]

[1] Institute for Snow and Avalanche Research SLF, Davos, CH

[2] World Meteorological Organization, Geneva, CH

[3] Norwegian Meteorological Institute, Oslo, NO

Corresponding author: Mathias Bavay (bavay@slf.ch)

The Global Cryosphere Watch (GCW) was initiated by the World Meteorological Organization (WMO) as a mechanism to support the delivery of Earth System monitoring, modelling and prediction services focused on the cryosphere. GCW fosters international coordination and partnerships with the goal of providing authoritative, clear and usable data, information and analyses on the past, current and future state of the cryosphere. It fosters sustained and mutually beneficial partnerships between research and operational institutions, by linking research and operations as well as scientists and practitioners. This is important as most available cryospheric data come from the scientific community. It is generally managed by research institutes which often do not have the infrastructure, the resources, nor the mandate to enable FAIR data management, which is necessary for interoperability and discovery at data level. This implies that data do not fit into standardized systems or dataflows for broader data access and exchange (as exists at the WMO) and thus have been unavailable for operational meteorological and climate applications. This lack of standardization also impairs the reuse of data within the scientific community. GCW is bridging this gap through a data portal and software stack enabling the transformation of sparsely documented and highly variable data into standardized and well documented data suitable for downstream applications with data level interoperability. A processing engine converts raw data provided by the data producers into NetCDF-CF standard files with NetCDF Attribute Convention for Dataset Discovery (ACDD) metadata. The data portal web front end harvests the metadata necessary for its search engine through an OPeNDAP server so no manual editing of the medatadata is necessary. When a user downloads some data from the web portal, it gets the requested data through the OPeNDAP server.

Keywords: meteorological data; data processing; standardization; publication; automatic

1. Introduction

The Cryosphere is an important component of the Earth system that includes snow, ice in all its forms, permafrost and seasonally frozen ground. The cryosphere is a global phenomenon existing at all latitudes, is critical for water supply and has a large influence on global climate. However, it is sensitive to climate change and provides us with the most visible evidence of our changing climate. The Cryosphere is therefore critical to monitor, yet covers some of the most remote regions of the earth in altitude and latitude and therefore is generally undersampled. Moreover, a large amount of globally available cryosphere data come from the scientific community, particularly in polar and other remote areas and is therefore not readily usable by national meteorological and hydrological services (de Rosnay et al., 2015; Brun et al., 2013).

In 2011, The World Meteorological Organization (WMO) launched the Global Cryosphere Watch (GCW) in order to address this gap by supporting key cryospheric in-situ and remote sensing observations. GCW aims to disseminate authoritative data to WMO members as well as to the scientific community and to provide sustained monitoring of the cryosphere. As such, it bridges the gap between the scientific community and the national meteorological and hydrological services and agencies through several activities. Such activities encompass standardization (developing snow and ice measurement standards and best practices, creating

an official Glossary of cryospheric terms), strategic planning (defining the observational requirements) and data exchange (building the CryoNet GCW network of surface stations as well as by setting up a data portal with a strong focus on data interoperability).

2. Data portal requirements

The data producers in this use case are usually relatively small groups that are responsible for the entire data acquisition process, from setting up and maintaining the stations to publishing the data. A survey was conducted with 37 institutions worldwide that act as data producers, in order to identify how they operate regarding data collection (Bavay and Fiddes, 2019). The survey revealed that although there might be dedicated resources allocated for managing the stations' data (78% of the respondents), these resources are insufficient to setup complex data sharing schemes (panels a) and d) of **Figure 1**). This means that the data producers are often not familiar with WMO's complex standards, both in terms of data formats, data exchange protocols and metadata and therefore can not realistically interface with the WMO dataflows. When there are systems to further disseminate the data, the data do not generally come in a standard format that could directly be reused by other members of the scientific community (panels b) and c) of **Figure 1**). Based on the experience of setting up the data portal, these in-house defined file formats are mostly variants of csv files with little metadata (panel c). Furthermore, the metadata are often in the language of the originating country, which could be challenging. This means that on a practical level, the data producers can not provide data-level interoperability, which is a major hurdle to global scale data sharing (and often even at the local scale). This problem of interoperability of the data models and exchange protocols has received less attention for extremely heterogeneous and distributed (as opposed large volumes) 'big-data' and requires radically different strategies (Parsons et al., 2011).

The data users stand at the other end of the system and range from single individuals to large organizations and from relatively simple statistical analysis on the data to full numerical models such as weather forecast systems. Unfortunately, getting the data and preparing them in a way that they can be used in the user's toolchain is very time consuming (in a different field, it has been estimated that data preparation takes about 60 to 80% of the time involved in a data mining exercise (Romano, 1997)). A small survey was conducted with more than 50 cryospheric data users (contacted through various mailing lists, Bavay and Fiddes, 2019) regarding third party data usage. It revealed that there is often the need for direct contact with the data producers in order to request some missing metadata as well as to clarify some of the observed behaviors in the data (this

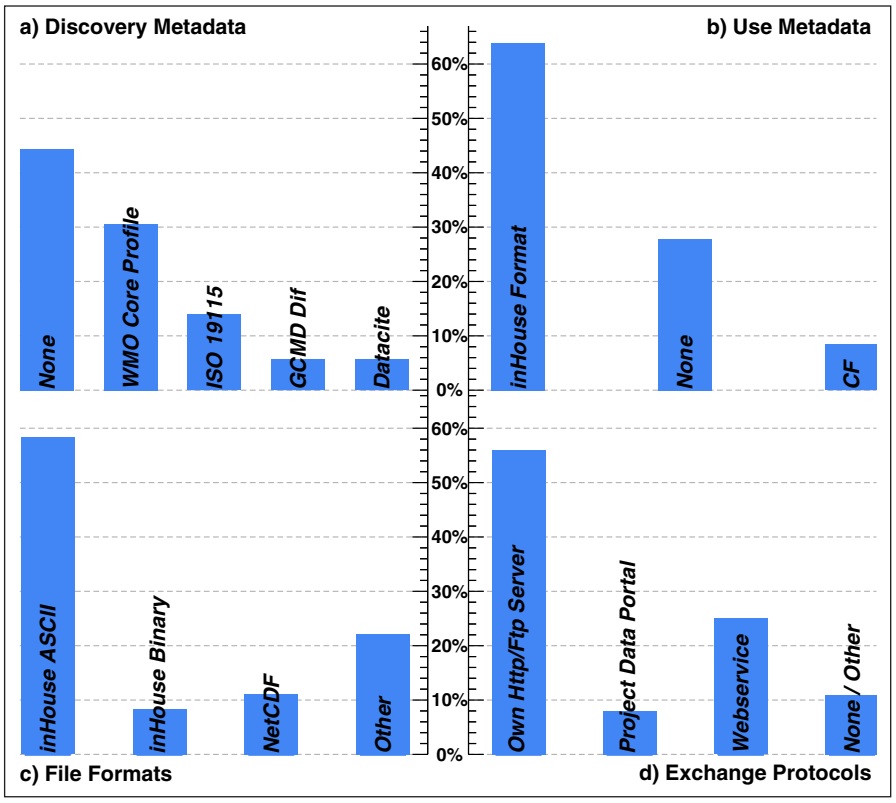

Figure 1: GCW Data Interoperability survey.

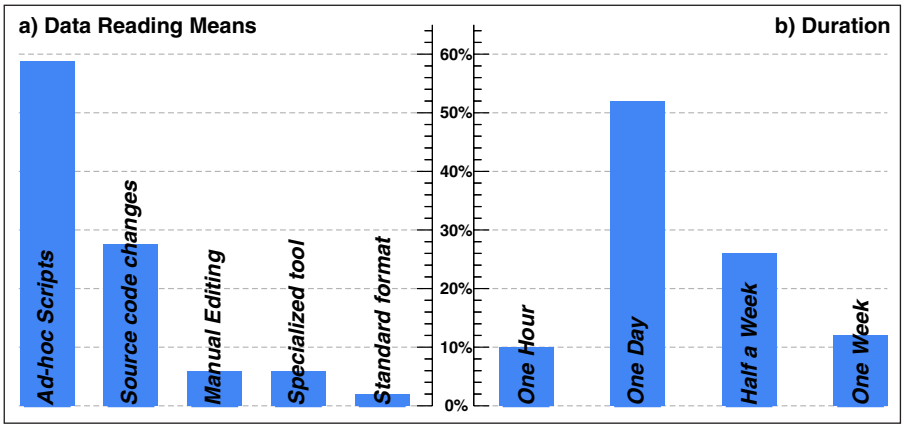

Figure 2: GCW Third Party Data Usage survey.

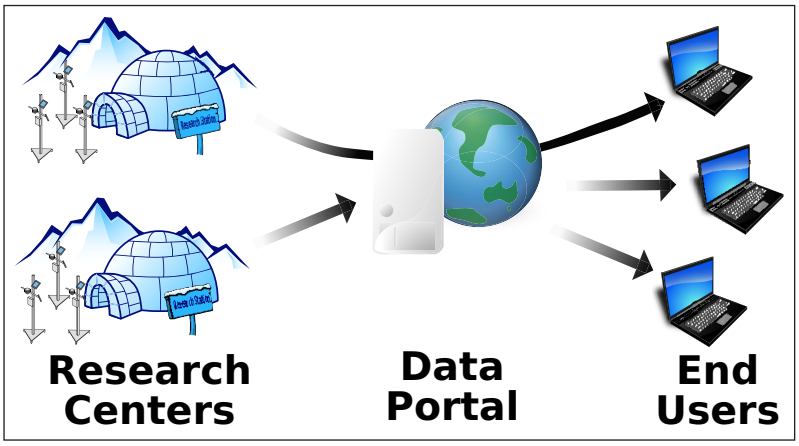

Figure 3: GCW Data portal goals: connecting the data producers to the data users both in a centralized and decentralized way.

was the case for 47% of the respondents). Since all data sources have their own file formats, the data are often read by editing the source code of the data consuming application or with an ad-hoc script that is seldom reused and error-prone (only 7% of the users can directly read third party data, see panel a) in **Figure 2**). This leads to frustrating experiences for the data users and high costs to integrate new data sources (with 90% of the users investing around one day and up to one week to deal with a new data source). A proper solution for the end users is to offer standardized data formats with standard metadata, if possible embedded in the data files. This means that regardless of the origin of the data, they can be reused in the same toolchain once it has been modified to support the standard format and metadata: this is data-level interoperability.

The link between the data producers and the data users is the GCW data portal (**Figure 3**) that should strive to accommodate the needs of both. In order to keep it manageable (with very low budget), it is necessary to establish a system where adding new data sources (i.e. new stations) comes with very low overhead and where everything runs automatically once set up. Moreover, the system should offer both a distributed operation (so the data producers retain full control over their data) and a centralized operation (offering more convenience for the smaller data producers). Although the system supports any timeseries at any sampling rates, for the GCW data portal all the data sources that have been submitted so far are Automatic Weather Stations (AWS) measurements. In the current phase, it has also been decided to exclude large measurement networks that might already have ad-hoc solutions to share their data with the national meteorological and hydrological services.

3. Design principles

Based on the requirements laid out in Section 2, it has been decided that the data portal would ingest raw data directly from data producers. This significantly lowers the requirements on their side since they only have to send the data as they have them (without any additional processing). Of course, they also have to provide the metadata associated with the stations that they contribute.

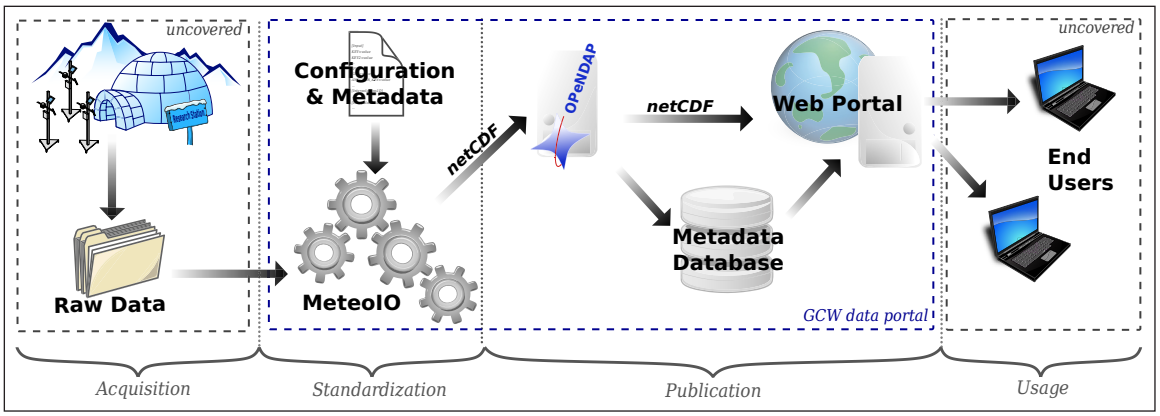

Figure 4: System overview: The raw data are delivered to the data portal where MeteoIO standardizes and OPeNDAP serves the metadata and data to the web portal as NetCDF.

In order to satisfy the requirements for the users, the NetCDF file format (Rew and Davis, 1990; Unidata, 2020) has been chosen with the Climate and Forecast Convention (CF) for the metadata. The NetCDF file format provides a standard file format that can be read by many different applications while quite compact and efficient for handling large amounts of data. The CF-1.6 convention provides standard names for the different meteorological parameters as well as the units and other metadata fields, allowing an application to read and interpret the data without any manual action (knowing that otherwise some variables names found in published dataset are also impossible to decipher for a human who has to get in contact with the data producers in order to interpret the data). Furthermore, the CF convention (starting with CF-1.6) also has a mechanism to identify the nature (e.g. timeseries, profile, trajectory) of a dataset. This is useful for services attempting to combine data from different sources.

Another design goal is to strictly avoid having to enter the same metadata more than once. But the data portal needs to offer a search feature where the end users can, for example, look for data availability of a specific variable in a given area for a given time period. These search metadata are written into the standardized files using the NetCDF Attribute Convention for Dataset Discovery (ACDD) and then harvested automatically by the data portal to populate its metadata database. The ACDD standard provides standard search metadata, describing the data origin and the spatial and temporal coverage. The data are delivered to the data portal (both for metadata harvesting and to further serve the data to the end users) through an OPeNDAP server (Pydap, 2020). Therefore no data are stored on the data portal web frontend but only requested on demand to a backend. The OPeNDAP client/server architecture allows subset queries of datasets on a temporal, spatial or by variable basis. This NetCDF-CF-OPeNDAP framework is becoming a standard in other environmental disciplines, notably climate science and oceanography (Hankin et al., 2009).

The above design considerations lead to a processing engine that can flexibly read in raw data files in various data formats and write out standardized NetCDF files. An existing software package has been used for this purpose, the MeteoIO meteorological data pre-processing library (Bavay and Egger, 2014) with some additional developments (**Figure 4**).

4. The GCW portal data processing engine

The MeteoIO library (Bavay and Egger, 2014) was initiated in 2008 for applications consuming meteorological data time series. It targets the very time consuming task of preparing the meteorological data by addressing the issues of 1) the vast diversity of data formats, 2) the necessity to correct the data for all kind of known measurement errors, 3) the variety of sampling rates and the mismatch between the measured sampling rate and the application's sampling rate of choice. It has been designed both for a research context (where flexibility is key) and operational contexts (where robustness and performance are key). As an open source C++ library, it can easily be integrated into other applications, such as numerical models, dashboards or simple data converters. It is also computationally efficient: reading, processing and writing out one year of AWS half-hourly data (in an ascii format) takes less than one second on an eight years old computer.

MeteoIO goes through several steps for preparing the data (**Figure 5**), aiming to offer within a single package all the tools that are required to bring raw data to an end data consumer: first, the data are read by one of the more than twenty available plugins supporting that many different formats or protocols (such as CSV files, NetCDF files, databases or web services). Then some basic data editing can be performed

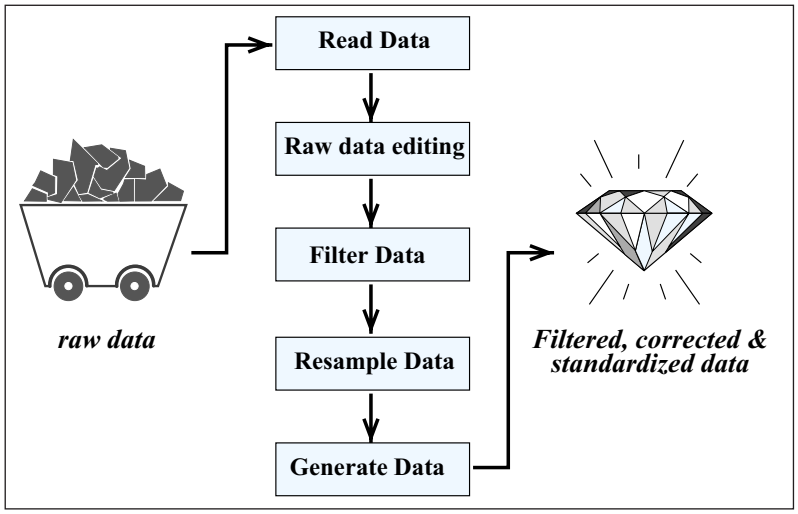

Figure 5: MeteoIO general processing work flow.

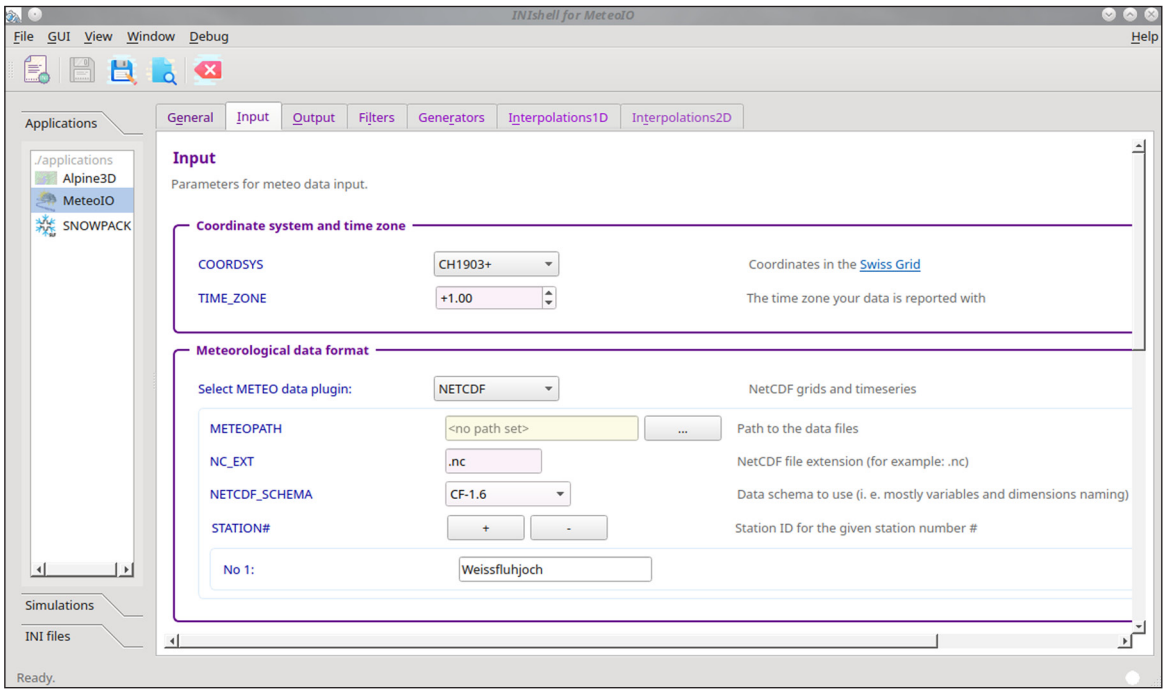

Figure 6: The Inishell graphical user interface to MeteoIO's configuration files.

(such as merging stations that are next to each other or renaming sensors). The data can then be filtered, by applying a stack of user-selected generic filters. These filters can either remove invalid data (such as despiking or low and high pass filters) or correct the data (such as precipitation undercatch correction, debiasing, Kalman filtering). Once this is done, the data are resampled to the requested time steps by various temporal interpolations methods. It is important to keep in mind that during this whole process, MeteoIO works with any sampling rates, including variable sampling rate and can resample to any point in time. If there are still missing data points at the requested time steps, it is possible to rely on *data generators* to produce some data out of either parametrizations (such as converting a specific humidity into a relative humidity) or very basic strategies (such as generating null precipitation to fill gaps). Finally, either the data are forwarded to the data-consuming application or written back by a user-selected plugin.

All the steps described above are fully controlled by a single configuration file so nothing is hard-coded into the library. Although it is an ascii file, it can be edited graphically with the Inishell GUI[1] (see **Figure 6**). This configuration file can be backed up to later reproduce the processing (for example to deploy a data

[1] https://models.slf.ch/p/inishell/

preparation workflow to an operational system) or even to re-implement the processing steps in another toolchain (reproducibility). Moreover, a human-readable log of all the operations that took place on each data point can be generated but is currently distinct from the output dataset. Since at this stage, it has been decided not to correct the datasets to keep the data unchanged, this is no limitation. It could later be integrated into the dataset themselves as metadata.

5. The GCW data portal search interface

The GCW Data Portal search interface (see **Figure 7**) harvests every night the discovery metadata that describe the datasets from contributing data centers. The discovery metadata describe who measured, what, where and when as well as how data can be used and how to access them. The harvested metadata are translated to a GCMD DIF and ISO19115 compliant internal representation through an extensive setup of Extensible Stylesheets which also includes the translation of keywords describing the variables or URLs. The latter is a major challenge (and therefore not always even possible) because of the lack of use of controlled vocabularies for the description of variables and URLs, especially when ISO19115 is used.

The harvesting process supports standard discovery metadata exchange protocols (such as OAI-PMH and OGC CSW) and discovery metadata according to the specification of Global Change Master Directory (GCMD) Directory Interchange Format (DIF) and ISO19115. For GCW, support for fully self describing NetCDF/CF served through OPeNDAP was added to the system. This relies on either a file named catalog.txt containing links to OPeNDAP Data Attribute Structure (DAS) for the datasets or THREDDS[2] Catalogs for identification of relevant datasets and generation of their discovery metadata. A list of the required ACDD elements is provided in the Data Portal support section. When no unique identifier is generated by the data provider, the harvested metadata are given a unique identifier by the GCW Data Portal.

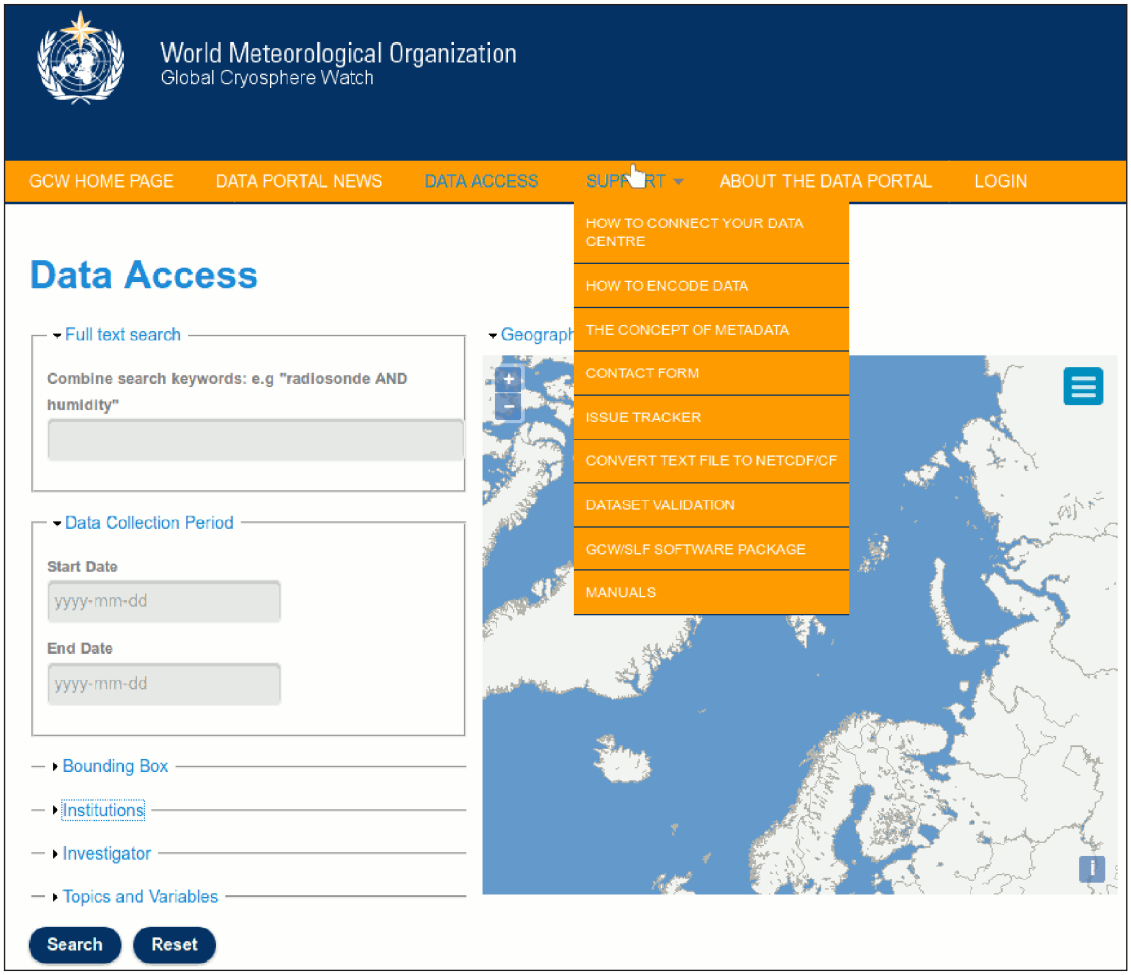

Figure 7: GCW Data Portal search interface.

[2] https://www.unidata.ucar.edu/software/tds/current/catalog/index.html

Then, the GCW Data Portal offers a standard inclusive search interface where variable, institution, temporal, spatial and full text (logical operators AND and OR are supported) search elements are merged into a search statement (search is powered by SolR). The search interface is embedded in a Drupal Content Management System to ensure that editorial information can easily be combined with the search interface and results. Currently on the fly visualisation and transformation of timeseries and gridded products are supported, while support for visualisation and transformation for profiles are under development.

6. Discussion and conclusion

Although there are many data sharing portals, few try to tackle data level interoperability which is key for easy reuse and downstream services. On the other hand, properly implementing data and metadata exchange standards has proven a great challenge to global data sharing (Tanhua et al., 2019). Such standards cover very broad fields and therefore bring complexity while simultaneously sometimes lacking some features required in a specific field. This is likely the reason why these standards are often badly implemented or not even implemented at all, at least in the cryospheric sciences. For the data producer, a custom designed inhouse format contains all the necessary features and is much quicker to implement. For the data user, the global lack of datasets complying to well recognized standards make them appear as yet another format alongside all the inhouse varieties. Moreover two datasets claiming to follow a given standard but not actually being compliant lead to distrust in the standards since they seem to offer no gain at the costs of more implementation complexity.

One way to break this vicious circle is to offer data producers toolchains that are already compliant with the standards (so there is no additional costs to using well recognized standards) as well as easy to use validators (so inhouse implementations can be validated) and to heavily promote these standards to large, well known data producers in order to ensure that the data users will encounter these formats during their normal work. Data converters are also useful tools to help both data producers and data users start to use the standards. This way, once the data users have acquired the capability to read data in a standard compliant format, they gain access to a large pool of datasets and will later request smaller data producers to support the same standards.

The system that has been setup and deployed for the Global Cryosphere Watch data portal fits into this approach: first by providing valuable data to the data users in a verified, standard compliant format, making it attractive. Then, it provides components that can be reused by data producers into their own toolchains. Since the current toolchain is highly automated and requires very little manual care while delivering data level interoperability, it shows that producing standard compliant data can be done in a cost effective way.

The MeteoIO library is available for many platforms under the GNU Lesser General Public License v3.0 (LGPL v3) on https://models.slf.ch/p/meteoio/. The Global Cryosphere Watch data portal can be found at https://gcw.met.no/.

Acknowledgements

The authors would like to thank Dr. Charles Fierz and Ms. Rodica Nitu for their very valuable and long standing support. This work has also been supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 730203.

Competing Interests

The authors have no competing interests to declare.

References

Bavay, M and **Egger, T.** 2014. Meteoio 2.4.2: A preprocessing library for meteorological data. *Geoscientific Model Development*, 7: 3135–3151. DOI: https://doi.org/10.5194/gmd-7-3135-2014

Bavay, M and **Fiddes, J.** 2019. Global cryosphere watch data survey. URL: https://www.envidat.ch/dataset/global-cryosphere-watch-data-survey. DOI: https://doi.org/10.16904/envidat.133

Brun, E, Lawrimore, J, de Rosnay, P and **Friddell, J.** 2013. Proposal for a GCW action aiming at improving operational snow depth observation for snow depth. *Technical Report*. Global Cryosphere Watch. URL: https://globalcryospherewatch.org/projects/snowreporting.html.

Hankin, S, Blower, JD, Carval, T, Casey, KS, Donlon, C, Lauret, O, Loubrieu, T, Srinivasan, A, Trinanes, J, Godoy, O, et al. 2009. Netcdf-cf-opendap: Standards for ocean data interoperability and object lessons for community data standards processes. In: *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society*. URL: http://www.oceanobs09.net/proceedings/cwp/Hankin-OceanObs09.cwp.41.pdf. DOI: https://doi.org/10.5270/OceanObs09.cwp.41

Parsons, MA, Godøy, Ø, LeDrew, E, De Bruin, TF, Danis, B, Tomlinson, S and **Carlson, D.** 2011. A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, 37: 555–569. DOI: https://doi.org/10.1177/0165551511412705

Pydap. 2020. Pydap server documentation. URL: http://www.pydap.org/en/latest/server.html.

Rew, R and **Davis, G.** 1990. Netcdf: An interface for scientific data access. *IEEE computer graphics and applications*, 10: 76–82. URL: https://ieeexplore.ieee.org/iel1/38/2037/00056302.pdf. DOI: https://doi.org/10.1109/38.56302

Romano, D. 1997. Data mining leading edge: Insurance & banking. *Proceedings of Knowledge Discovery and Data Mining*, Unicom, Brunel University.

de Rosnay, P, Isaksen, L and **Mohamed, D.** 2015. Snow data assimilation at ecmwf. *ECMWF Newsletter*, 26–31. URL: https://www.ecmwf.int/en/elibrary/17328-snow-data-assimilation-ecmwf. DOI: https://doi.org/10.21957/lkpxq6x5

Tanhua, T, Pouliquen, S, Hausman, J, O'Brien, K, Bricher, P, de Bruin, T, Buck, JJH, Burger, EF, Carval, T, Casey, KS, Diggs, S, Giorgetti, A, Glaves, H, Harscoat, V, Kinkade, D, Muelbert, JH, Novellino, A, Pfeil, B, Pulsifer, PL, Van de Putte, A, Robinson, E, Schaap, D, Smirnov, A, Smith, N, Snowden, D, Spears, T, Stall, S, Tacoma, M, Thijsse, P, Tronstad, S, Vandenberghe, T, Wengren, M, Wyborn, L and **Zhao, Z.** 2019. Ocean fair data services. *Frontiers in Marine Science*, 6: 440. DOI: https://doi.org/10.3389/fmars.2019.00440

Unidata. 2020. Network common data form (netcdf). URL: https://www.unidata.ucar.edu/software/netcdf/.

How to cite this article: Bavay, M, Fiddes, J and Godøy, Ø. 2020. Automatic Data Standardization for the Global Cryosphere Watch Data Portal. *Data Science Journal,* 19: 6, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2020-006

Submitted: 02 December 2019 **Accepted:** 29 January 2020 **Published:** 24 February 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS