Technical documentation

Guidance for data centres contributing to SDMS

Versions

Version	Date	Comment	Responsible
0.x	2022-01-21	Updates following lessons learned in harvesting especially from GeoNetwork. Added outline for product manual.	Øystein Godøy
0.6	2020-04-14	Corrections of incorrect cross references, added descriptions on file formats and cleaned recommendations	Øystein Godøy
0.5	2020-03-15	Cleaning of comments, preparation for conversion to ASCIIDOC and establishment of work flow on GitHub.	Øystein Godøy
0.4	2017-05-27	Comments from SDMS WG incorporated and cleaned.	Øystein Godøy
0.3	2017-03-17	Comments from the SDMS WG incorporated.	Stein Tronstad Torill Hamre Markus Fiebig Terry Hannant Angelo Viola Øystein Godøy
0.2	2017-01-15	Comments from Angela Schäfer and Torill Hamre are included.	Øystein Godøy Angela Schäfer Torill Hamre
0.1	2016-10-20	First draft for internal discussion in the SDMS Working Group.	Øystein Godøy
0.0	2016-04-15	First draft based on a similar document developed for the WMO Global Cryosphere Watch.	Øystein Godøy

Table of Contents

1. Introduction . . . 3
1.1. Background . . . 3
1.2. Scope . . . 4
1.3. Intended audience . . . 4
1.4. Applicable documents . . . 4
2. Interoperability interfaces . . . 6
2.1. Discovery metadata . . . 6
2.1.1. Background . . . 6
2.1.2. Exchange mechanisms for metadata . . . 7
2.1.2.1. Introduction . . . 7
2.1.2.2. OAI-PMH . . . 7
2.1.2.3. OGC CSW . . . 8
2.1.2.4. Sensor Description (Not yet supported) . . . 8
2.1.2.5. Other . . . 9
2.1.3. Information structures for discovery metadata . . . 9
2.1.3.1. ISO19115 . . . 9
2.1.3.2. GCMD DIF . . . 12
2.2. Data . . . 20
2.2.1. Background . . . 20
2.2.2. Exchange mechanisms for data . . . 20
2.2.2.1. Introduction . . . 20
2.2.2.2. HTTP/FTP . . . 20
2.2.2.3. OPeNDAP . . . 21
2.2.2.4. OGC WFS . . . 21
2.2.2.5. OGC WCS . . . 21
2.2.2.6. OGC WMS map projections . . . 22
2.2.3. File formats . . . 22
2.2.3.1. Introduction . . . 22
2.2.3.2. Darwin Core Archive . . . 23
2.2.3.3. JSON/GeoJSON/JSON-LD . . . 23
2.2.3.4. NetCDF/CF . . . 23
2.2.3.5. WMO BUFR . . . 24
2.2.3.6. WMO Grib . . . 24
2.2.3.7. XML . . . 24
Appendix A: Outline for Product Manual . . . 26
A.1. Introduction . . . 26
A.2. Outline . . . 26

1. Introduction

1.1. Background

Environmental and climate changes are currently observed at a global scale and in particular in the Arctic. In order to give better estimates of the future changes, the Arctic has to be monitored and analysed by a multi-disciplinary observation system which is suited to validate and gradually improve Earth System Models. The best chance to achieve significant results within a relatively short time frame is found in regions with a large natural climate gradient, and where processes sensitive to the expected changes are particularly important.

Svalbard and the surrounding ocean areas fulfil all these criteria: Svalbard is located in a region with a very large climate gradient, being alternately influenced by cold central Arctic or mild marine climate conditions at time scales of weeks to years. It is also located in the region with the strongest inflow and outflow processes between the Arctic and lower-latitude oceans. In addition, Svalbard is the only region in the world (and has the facilities) where one can study and quantify one of the remaining unknowns in the climate puzzle: the extraterrestrial and especially solar influence on climate.

The vision for the Svalbard Integrated Arctic Earth Observing System (SIOS) is to be a regional observational system for long term acquisition and proliferation of fundamental knowledge on global environmental change (GEC) within an Earth System Science (ESS) perspective in and around Svalbard. SIOS will systematically develop and implement methods for how observational networks are to be construed and thus become a leader regarding observational systems in the Arctic and Polar regions.The SIOS Data Management System (SDMS) Data Portal is the entry point to SIOS datasets. It offers a web interface that contains information about datasets (metadata). These metadata are harvested on a regular basis from data centres contributing to SIOS. These data centres manage the data on behalf of the owners/providers of the data.

A major innovative element of SIOS will be the Knowledge Centre (KC), which will facilitate interaction between observation, modelling and process research, strategic processes, a service point to user communities and a platform for data handling and utilization [3]. The SIOS Data Management System (SDMS) will be a functionality enabling component of the Knowledge Centre, supporting data submission, discovery, access, use and preservation of SIOS relevant data sets.

For the SDMS the term “data set” is defined in line with the INSPIRE Directive as “an identifiable collection of spatial data”, i.e. a collection of data that has a reference by name or coordinates to a geographic location or area, and which in addition have a designated start and end time. A data set can contain observations (remote or in situ), derived quantities (from either of these two types of data sources) or forecasts of future states of environmental parameters. The data values can be located at a single point, along a line or transect, in a regular or irregular grid, and be captured or estimated at one or more altitudes/depths. A data set can be stored on paper, in files (one or more), or in a database, and is often accompanied by descriptions (metadata) of its content. The purpose of the SDMS is to the SIOS user community with a single entry point to relevant datasets. It is however not the purpose to centralise all data, but rather to integrate existing and future data centres contributing to SIOS using machine readable interfaces to metadata and data.

The first version of this document is based on a similar document developed for WMO Global Cryosphere Watch.

1.2. Scope

This document identifies aspects that have to be handled in establishing SDMS. The purpose is to enable transparency on metadata and in the long term also on data through limiting the interfaces to be developed and maintained. The short term goal is to enable visibility of relevant datasets through discovery metadata, in the long term to harmonize access to data via defined set of data access interfaces. Emphasis is primarily on the semantics and structure of interfaces to metadata and data, and less on underlying file formats.

The requirements of the document represents a long term goal which require funding for full implementation. Some partners have implemented parts of the interfaces mentioned while others have a longer way to go. This is reflected in the prioritised functionality to implement as outlined in the implementation plan of the SIOS Preparatory Phase [3], see section 5.3.1[1].

1.3. Intended audience

System managers at the data centres contributing to the SDMS.

1.4. Applicable documents

TODO: Add SIOS Data Policy and SPDX licenses to this list.

1. Svalbard Integrated Arctic Earth Observing System – Preparatory Phase (SIOS-PP) Description of Work.
2. Robert Huber and Michael Klages (lead authors), 2012. SIOS Data Management System – User Requirements Document. SIOS Deliverable D6.3.
3. Distributed SIOS Data Management System Implementation Plan, 2014, SIOS Deliverable D6.6.
4. SDMS Concepts and acronyms
5. SDMS System Requirements Document
6. SDMS Architecture Design Document
7. SDMS System implementation and integration plan
8. SDMS Operations Manual
9. WMO Information System
10. WMO Core Profile of the ISO 19115
11. WIGOS, including the metadata standard
12. The Open Archives Initiative Protocol for Metadata Harvesting, Version 2
13. OAI-PMH tools
14. OGC CSW specification
15. GCMD DIF Writers Guide
16. GCMD Science Keywords
17. Climate and Forecast Standard Names
18. WMO Code Lists

19. NetCDF
20. Climate and Forecast Conventions
21. OPeNDAP
22. UNIDATA's Common Data Model
23. OpenSearch
24. Wilkinson et al., 2016: The FAIR Guiding Principles for scientific data management and stewardship

[1] The official version of this document has some issues with references, an updated version will be made available within the collaboration area for the SDMS WG.

2. Interoperability interfaces

2.1. Discovery metadata

2.1.1. Background

Metadata are generated by the data centres hosting the data sets. Metadata are harvested and ingested in the central catalogue for usage by the SIOS Data Portal user community. SIOS 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 in a standard format, e.g. GCMD DIF or ISO 19115. 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 SIOS 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 Convention where sufficient information to actually use the data is provided. The fourth category links directly to WIGOS metadata [2]. These metadata describes the station, its surroundings, instrumentation, procedures etc. There is some overlap between these metadata and the first category.

The SIOS 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 [8]. The discovery metadata are harvested from partner data repositories every night, then converted to the data model used in the data catalogue and ingested if it fulfils the quality check. The nightly harvest is an incremental harvest of changes since the last time the harvest was run. Every 3 months a full harvest is done to clean up potentially stale datasets.

TODO: Add a separate document on the harvest process. I.e. convert the old LibreOffice document to ASCIIDOC and add to GitHub.

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

Recommendation: All datasets should have a unique identifier issued by the host data centre. 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.

This implies that SIOS Data Portal will not specify or change a unique identifier unless the dataset is hosted by the SIOS Data Portal.

Recommendation

Always include a license in the discovery metadata. Following the SIOS Data policy, SIOS recommends usage of the Creative Commons Attribution License and use identifiers and URL to license text from https://spdx.org/licenses/. It is recommended to include both a standardised identifier and a link to the full license text.

TODO: Add link to SIOS Data Policy.

2.1.2. Exchange mechanisms for metadata

2.1.2.1. 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 SIOS Data Portal. Suitable interfaces for this are OAI-PMH and OGC CSW. Other interfaces may be evaluated, but to ensure a cost effective solution the number interfaces must be limited.

Recommendation

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.

Although OAI-PMH is recommended, OGC CSW will also be consumed. Work is in progress with other exchange mechanisms..

2.1.2.2. OAI-PMH

The Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH) is the recommended interface for providing discovery metadata to the SIOS Data Portal. It is a cost effective and robust implementation for exchange of metadata between data centres. It is much cheaper to implement than most alternatives and there are a number of tools available. Some of these are listed on [13]. The central node of SDMS is using pyCSW to provide OAI-PMH (and other protocols).

When implementing OAI-PMH there is a number of SDMS recommendations that are based on experience during the International Polar Year, SIOS Preparatory Phase and the operation of SDMS.

Recommendation: OAI-PMH version 2 must be used.

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

Recommendation:	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. SDMS contributing data centres must support **transient** and must maintain transient records for at least 1 month [4]. More information on this feature is available in OAI-PMH specification of deleted records.
Recommendation:	The OAI-PMH interface by default offers metadata in Dublin Core. This is insufficient for SDMS 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 it is recommended to use the following keywords: “dif” for GCMD DIF, “iso19139” for the ISO19115 minimum profile.

2.1.2.3. OGC CSW

The Open Geospatial Consortium Catalogue Services for the Web (OGC CSW [14]) is another standard for exposing the content of a catalogue in a standardised form. Similar to OAI-PMH records are exposed using XML. Compared to OAI-PMH, OGC CSW is 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[5] and is supported by the SIOS Data Portal. The portal only consumes ISO 19115/ISO 19139 through OGC CSW.

Recommendation:	Do not use SOAP.
Recommendation	Serve ISO19115/ISO 19139.

2.1.2.4. Sensor Description (Not yet supported)

The implementation of Sensor Web Enablement (SWE) as a useful suite of standards of the OGC for building the Sensor Web is recommended to allow for interoperability and to reduce the integration efforts of new data sources, although not supported by the SIOS Data Management System central node yet.

Recommended data formats are Observations and Measurements (O&M) and SensorML. For describing devices and sensors in terms of metadata to NRT-data streams or archived and disseminated data basic SensorML should be applicable. These are Standard for modelling/encoding sensor Metadata based on XML (latest version: SensorML 2.0). Hence Editors are needed to facilitate the provision of these metadata to the SDMS with focus on ISO 19115 metadata. As controlled vocabulary for device description the NERC-SeaDataNet vocabulary could be a useful recommendation if agreed on.

IMPORTANT	SWE is not supported for data discovery purposes, but is considered for other purposes within SIOS.

2.1.2.5. Other

Harvesting of discovery information using OpenSearch is under testing, but not supported for regular exchange of discovery metadata. According to [23] OpenSearch is a “collection of simple formats for the sharing of search results. The OpenSearch description document format can be used to describe a search engine so that it can be used by search client applications.”

Work is in progress to enable support for schema.org.

NOTE	More information on schema.org to be added.

2.1.3. Information structures for discovery metadata

2.1.3.1. ISO19115

ISO19115 is a information container that can be populated with several controlled vocabularies in some of the elements. The search model for the SIOS Data Portal is currently built around parameter descriptions using the GCMD Science Keywords [16]. A mapping exist between Climate and Forecast standard names [17] and GCMD Science Keywords. ISO 19115 is used by e.g. INSPIRE at the European level and WMO (WMO Core Profile [10]) at the global level.

Recommendation:	ISO19115 records **must** at least state the unique id, temporal and spatial location, scientific content, responsible data centre and PI as well as links to the actual data[8]. Datasets without such specifications are not ingested in the catalogue.
Recommendation:	ISO19115 records **should** contain GCMD Science Keywords (with full hierarchy from Earth Science and using > as separator) or some other machine readable vocabulary following the FAIR guiding principles [24]. The vocabulary used must be identified.
Recommendation:	For stations where it is relevant (e.g. stations contributing to WMO GCW through CryoNet) it is mandatory to have one keyword from the WMO CategoryCode list [18][9]. Relevant keywords for SDMS are e.g. weatherObservations, meteorology, hydrology, climatology, glaciology. In the context of SDMS alone, this is not a requirement.
Recommendation	Metadata records **should** include information on the host data centre for the dataset.
Recommendation:	All times must be encoded as ISO8601.

NOTE	More details to be added on ISO19115 guidance. In particular on how to embed references to the vocabularies used and information on the data centre hosting the data.

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

vocabularies.

IMPORTANT The table below is under review.

Table 1. ISO19115 core elements.

Element name	Description	ISO element	ISO	SDMS
Dataset title	A short title for the dataset.		M	M
Dataset reference date	Temporal extent of the dataset.		M	M
Dataset responsible party	Principal investigator and affiliated institution as well as contact details.		O	M
Geographic location of the dataset	Spatial extent of the dataset.		O	M
Keyword	A word or phrase that describes some aspect of a resource. Can be one of several types. It is used to describe the parameters in a dataset, the project affiliation etc. Proper identification of the purpose of the keywords and the vocabularies used is required. Project names are used to tag datasets in the SDMS system, e.g. as SIOS Core Data, SESS 2020 etc.		O	M
Dataset language	Should be English.		M	M
Dataset character set	Should be UTF-8.		O	O
Dataset topic category	ISO Topic Category.		O	O

Element name	Description	ISO element	ISO	SDMS
Spatial resolution of the dataset		O	O	Abstract describing the dataset
Short summary describing the dataset.		M	M	Distribution format
Should be NetCDF/CF or Darwin Core Archive in SDMS. Other standardised formats may be supported later. Non standard formats should have a detailed product manual.		O	M	Additional extent information for the dataset (vertical and temporal)
		O	M	Spatial representation type
		O	RC	Reference system
		O	O	Lineage
What is done with the data since collection.		O	R	On-line resource
URL to the actual dataset accompanied with identification of the protocol supported. OSGEO or GCMD keywords are required for proper interpretation.		O	MC	Metadata file identifier
UUID or similar used to remove duplicates and interpret data in context.		O	M	Metadata standard name
		O	RC	Metadata standard version
		O	RC	Metadata language
		O	RC	Metadata character set

Element name	Description	ISO element	ISO	SDMS
		O	RC	Metadata point of contact
		M	M	Metadata date stamp

2.1.3.2. GCMD DIF

The Global Change Master Directory (GCMD) Directory Interchange Format (DIF) [15] is a metadata standard that is widely used (e.g. by the Antarctic Master Directory) and that was

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

Table 2. GCMD DIF elements.

Element	Description	GCMD	SDMS
Entry_ID	The <**Entry_ID**> is the unique document identifier of the metadata record. The <**Entry_ID**> is determined by the metadata author or data center contact personnel and may be identical to identifiers used by the data provider's data center or organization. For example, the National Snow and Ice Data Center (NSIDC) Distributed Active Archive Center (DAAC) identifies their metadata records as *NSIDC-xxxx*, where *xxxx* is a numerical designator. Also, the identifier is case insensitive meaning nsidc-xxxx and NSIDC-xxx refer to the same metadata record.	MU	MU
Entry_Title	The <Entry_Title> is the title of the data set described by the metadata.	MU	MU

Element	Description	GCMD	SDMS
Parameters (Science Keywords)	The <**Parameters**> field allows for the specification of Earth science keywords that are representative of the data set being described. These keywords are important for the precise search and retrieval of information from the GCMD. The author must select these keywords from the controlled set of science keywords. The <**Parameters**> field consists of a 7-level hierarchical classification of science keywords	MC	MC
ISO Topic Category	The <**ISO_Topic_Category**> field is used to identify the keywords in the ISO 19115 - Geographic Information Metadata (http://www.isotc211.org/) Topic Category Code List. It is a high-level geographic data thematic classification to assist in the grouping and search of available geographic data sets.	MC	MC
Data Center	The <**Data Center**> is the data center, organization, or institution responsible for distributing the data.	M	MC
Summary	The <**Summary**> field provides a brief description of the data set along with the purpose of the data. This allows potential users to determine if the data set is useful for their needs.	MU	M

Element	Description	GCMD	SDMS
Metadata Name	The ISO 19115 <**Metadata_Name**> field is used to identify the current DIF standard name.	MU	MC
Metadata Version	The <**Metadata_Version**> field is used to identify the current DIF metadata standard.	MU	MU
Data Set Citation	The <**Data_Set_Citation**> field allows the author to properly cite the data set producer.	R	R
Personnel	<**Personnel**> defines the point of contact for more information about the data set or the metadata.	R	R
Instrument	The Instrument or <**Sensor_Name**> is the name of the instrument used to acquire the data.	RC	RC
Platform	The Platform or <**Source_Name**> is the name of the platform used to acquire the data.	RC	RC
Temporal Coverage	The <**Temporal_Coverage**> field specifies the start and stop dates during which the data was collected.	R	M
Paleo-Temporal Coverage	For paleoclimate or geologic data, <**Paleo_Temporal_Coverage**> is the length of time represented by the data collected.	R	O
Spatial Coverage	The <**Spatial_Coverage**> field specifies the geographic and vertical (altitude, depth) coverage of the data.	R	M

Element	Description	GCMD	SDMS
Location	The <**Location**> field specifies the name of a place on Earth, a location within the Earth, a vertical location, or a location outside of Earth.	RC	OC
Data Resolution	The <**Data_Resolution**> field specifies the resolution of the data, which is the difference between two adjacent geographic, vertical, or temporal values. Controlled keywords representing horizontal, vertical and temporal data resolution ranges can be selected. Selection of data resolution ranges will assist users in refining their search for data within specific resolution ranges.	RC	OC
Project	The <**Project**> is the name of the scientific program, field campaign, or project from which the data were collected.	R	RC
Quality	The <**Quality**> field allows the author to provide information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata.	R	MC
Access Constraints	The <**Access_Constraints**> field allows the author to provide information about any constraints for accessing the data set.	R	MC

Element	Description	GCMD	SDMS
Use Constraints	The <**Use_Constraints**> field allows the author to describe how the data may or may not be used after access is granted to assure the protection of privacy or intellectual property.	R	MC
Distribution	The <**Distribution**> field describes media options, size, data format, and fees involved in distributing the data set.	R	RC
Data Set Language	<**Data_Set_Language**> describes the language used in the preparation, storage, and description of the data.	RC	RC
Data Set Progress	The <**Data_Set_Progress**> describes the production status of the data set regarding its completeness.	RC	RC
Related URL	The <**Related_URL**> field specifies links to Internet sites that contain information related to the data, as well as related Internet sites such as project home pages, related data archives/servers, metadata extensions, online software packages, web mapping services, and calibration/validation data.	RC	MC[10]

Element	Description	GCMD	SDMS
DIF Revision History	The <**DIF_Revision_History**> allows the author to provide a list of changes made to the DIF over time.	R	R
Keyword (ancillary keywords)	The <**Keyword**> field allows authors to provide any words or phrases needed to further describe the data set.	R	R
Originating Center	The <**Originating_Center**> is the data center or data producer who originally generated the dataset.	R	R
Multimedia Sample	The <**Multimedia_Sample**> field allows the author to provide information that will enable the display of a sample image, movie or sound clip within the DIF.	R	O
References (Publications)	The <**Reference**> field describes key bibliographic citations pertaining to the data set.	R	R

Element	Description	GCMD	SDMS
Parent DIF	The <**Parent_DIF**> field allows the capability to relate generalized aggregated metadata records (parents) to metadata records with highly specific information (children). Population of the <**Parent_DIF**> field should be reserved for instances where many metadata records are basically subsets that can be better represented by one parent metadata record, which describes the entire collection. Typically, the parent metadata record will have many children metadata records, which refer to the parent through the <**Parent_DIF**> field. In some instances, a child may point to more than one parent. The <**Parent_DIF**> is populated with an <**Entry_ID**>**.**	R	O
IDN Node	The Internal Directory Name (IDN) Node **(<IDN_Node**>) field is used internally to identify association, responsibility and/or ownership of the dataset, service or supplemental information.	R	O
DIF Creation Date	The <**DIF_Creation_Date**> specifies the date the metadata record was created.	R	R

Element	Description	GCMD	SDMS
Last DIF Revision Date	The <**Last_DIF_Revision_Date**> specifies the date the metadata record was created.	R	R
Future DIF Review Date	The <**Future_DIF_Review_Date**> allows for the specification of a future date at which the DIF should be reviewed for accuracy of scientific or technical content.	R	R
Privacy Status	The <**Private**> field allows the author to restrict the data set description from being publicly available.	RC	RC
Extended Metadata	The <**Extended_Metadata**> field will allow organizations to store user defined values within the metadata record without reusing existing GCMD defined metadata fields.	O	O[11]

Recommendation 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 SDMS context.

Recommendation: GCMD do not require a controlled vocabulary for the quality element. SDMS should to improve search results[13].

Recommendation: Related_URL has several subtypes. The existing list of type and subtype must be used to allow the SIOS 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.

Recommendation: All times must be encoded as ISO8601 either as YYYY-MM-DD or YYYY-MM-DDTHH:MM:SSZ.

2.2. Data

2.2.1. Background

While interoperability at the metadata level is important for SDMS, exchange of observations and data products is vital to the success of SDMS. 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 SDMS community a certain level of standardisation is required, especially in order to be able to automatically combine multiple datasets into a new dataset fulfilling user requirements (e.g. for CalVal activities). This standardisation is also 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 documentation of data through standardised use metadata is required. By use metadata is understood identification of the variables, their structure (e.g. spatiotemporal dimensions and mapping to file format), units of variables, encoding of missing values, quality/accuracy estimates, map projection and coordinate reference system etc.

Application of a common data model simplifies integration and intercomparison of datasets. Application of NetCDF[19] as the file format, utilising the Climate and Forecast[20] convention and serving data through OPeNDAP[21] simplifies the issue of integration and combination of data through the Common Data Model[22]. Other solutions may be added, but this is an easy win and address requirements both within numerical simulation communities, satellite CalVal communities and WMO programmes.

Recommendation: Where possible, OPeNDAP should be supported for data access (combining multiple physical files to a single virtual dataset).

Several OPeNDAP implementations exist (e.g. THREDDS, ERDDAP, Hyrax 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 also minimises the overhead as no files are moved, the client connects to data streams, reads the necessary data and close the connection.*

2.2.2. Exchange mechanisms for data

2.2.2.1. 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.

2.2.2.2. 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 differs 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.

Recommendation: Whenever data are served as direct download through HTTP or FTP, data should be served in a FAIR compliant data format, see [file_formats] using standardised encoding structures and vocabularies.

2.2.2.3. 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[14]. OPenDAP can relate to files or relational data base systems and is extensively used by e.g. Copernicus services, Earth System Grid Federation and others.

SDMS is currently able to consume OPeNDAP enabled datasets, but the level of support depends on the structure and standards (CF) conformance of the data served.

IMPORTANT To enable automatic visualisation of timeseries data at stations, data has to be encoded as NetCDF-CF with the featureType global attribute set.

NOTE Further information to be added on visualisation requirements.

2.2.2.4. OGC WFS

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

SDMS is currently **not** able to consume OGC WFS and implementation is not recommended until the OGC API is more mature. The new OGC API will be able to serve data as e.g. NetCDF/CF and GeoJSON in addition to the other formats.

NOTE Work is in progress to enable support for the upcoming OGC API approaches.

IMPORTANT The data portal is not able to consume data through OGC WFS.

2.2.2.5. OGC WCS

OGC Web Coverage Service (WCS) is similar to OGC WFS but focuses 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 “https://en.wikipedia.org/wiki/Web_Coverage_Service[trimming]” or “slicing”,

SDMS is currently **not** able to consume OGC WCS and implementation is not recommended until the OGC API is more mature.

NOTE	Work is in progress to enable support for the upcoming OGC API approaches.

IMPORTANT	SDMS is not able to consume data through OGC WCS.

2.2.2.6. 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.

Recommendation: 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
5. EPSG:32633: UTM Zone 33x

SDMS is only able to consume WMS services that provide a GetCapability document per dataset and where the WMS URL is clearly identified using standardised vocabularies (currently GCMD and OSGEO are supported).

Recommendation:

OGC WMS services should present a Getcapabilities document per dataset, not a common document for all datasets served.

IMPORTANT	SDMS is not capable of parsing WMS layers for specific datasets from site specific GetCapabilities documents.

2.2.3. File formats

2.2.3.1. 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 (e.g. NetCDF/CF and Darwin Core Archive). 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.

When it is not possible to encode data as NetCDF-CF or Darwin Core Archive, data can be uploaded in a

non-proprietary file format that is easy to consume for users (without specific software) accompanied by a detailed product manual[15] (in PDF format). *This approach cannot be used for SIOS Core Data.*

2.2.3.2. Darwin Core Archive

Darwin Core Archive is a file format much used within the biological community and in particular within biodiversity. It is the backbone of the Global Biodiversity Information facility (GBIF). In essence it is a set of comma separated files (CSV) bundled with a metadata file (meta.xml) and using controlled vaocabularies to describe the content. SDMS cannot do much on top of Darwin Core Archives (the diversity of types if data is too large), but the format is more or less FAIR compliant and is recommended for use within SDMS. Further information is available at https://dwc.tdwg.org/terms/.

Recommendation: Darwin Core Archive is recommended for use for biodiversity data within SDMS.

2.2.3.3. JSON/GeoJSON/JSON-LD

JavaScript Object Notation (JSON) and the geographical extension GeoJSON of this is similar to NetCDF in that it is a container lacking standardised metadata. JSON-LD (JavaScript Object Notation for Linked Data,) enables encoding of Linked Data using JSON.

There is currently no standardised FAIR compliant implementation of JSON for the types of data SDMS is handling. The CF convention could be implemented in JSON and there is work internationally pushing in this direction, but yet not mature enough.

IMPORTANT SDMS is currently **not** able to consume JSON files.

NOTE SDMS should work to enable ACDD and CF elements in GeoJSON files.

2.2.3.4. NetCDF/CF

NetCDF is a container like JSON and XML and such not a recommended file format for data within SDMS. However, the Climate and Forecast convention constrains the degrees of freedom within NetCDF and enforces structures and application of controlled vocabularies to describe the content of the data. NetCDF/CF is thus a FAIR compliant file format and recommended for use within SDMS. However, even NetCDF/CF have too many degrees of freedom to allow higher orders services to be established for datasets. Thus some further constraints on granularity and structures are recommended. NetCDF/CF is the backbone of the Earth System Grid Federation serving IPCC data, Copernicus Marine Environmental Monitoring Service (CMEMS), SeaDataNet and several other services. The file format is recommended for meteorological, oceanographic, hydrological and glaciological data (although exceptions exist). Work is in progress within WMO to identify specific CF profiles for NetCDF for use within WMO.

Recommendation: 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.

Recommendation:	NetCDF/CF files should be encoded according to CF-1.8 or higher.
Recommendation:	NetCDF/CF files should also include global attributes according to the Attribute Convention for Dataset Discovery.
Recommendation:	For datasets representing time series or profiles it is required to add the global attribute featureType with the appropriate content. If no featureType is found in the data it is assumed that the data are gridded in nature.
Recommendation:	It is **not** recommended to combine information from several stations in a single NetCDF/CF file (granularity issue).

2.2.3.5. 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.

BUFR is, although being a standardised format, not recommended for data sharing within SDMS.

NOTE	SDMS will extract information in BUFR format and convert this to NetCDF-CF and make this available in the SIOS Data Portal.

2.2.3.6. 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.

GRIB is, although being a standardised format, not recommended for data sharing within SDMS.

2.2.3.7. 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, as the format is text-based.

SDMS is currently **not** able to consume XML files nor is XML recommended as exchange format although standardised representations (e.g. WaterML) exist.

[2] WIGOS Metadata are based on the OGC Observations and Measurements Schema.

[3] This may change.

[4] This may change.

[5] Not required for scientific data.

[6] This recommendation will be revisited.

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

[8] This recommendation will be revisited.

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

[10] Further guidelines are required compared to GCMD.

[11] Depends on potential requirements within SDMS.

[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 SDMS by the Terminology Team.

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

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

[15] There is currently no template for product manuals available. This is to be developed.

Appendix A: Outline for Product Manual

A.1. Introduction

When datasets cannot be presented in a FAIR compliant file format (i.e. self describing), a product manual describing the dataset and how to interpret it **should** accompany the dataset. This product manual has to be served as a PDF document and have the sections identified below.

A.2. Outline

Outline of sections that should be covered by a product manual for data that cannot be described in a FAIR compliant file format.

IMPORTANT	The file formats used should always be free and open, not commercial.

1. Introduction
    a. General overview of the data production process
2. Description of the data
    a. Description of which variables that are measured, their characteristics including units on variables etc.
3. Description of the file format
    a. Description of structure and organisation of the data within this file format. Should include all necessary information to decode the file.
4. References to software
    a. Description of and links to software that can read and interpret the data. Should focus on open software. Alternatively code snippets for Python, R etc could be embedded in the document.