Technical documentation

Guidance for data centres contributing to SDMS

2025-11-11: Consolidated version following internal review.

Versions

Version	Date	Comment	Responsible
1.2	2025-11-11	Major modification, mappings are updated.	Lara Ferrighi
1.1	2023-11-29	Minor modification, added reference to granularity perspectives document.	Øystein Godøy
1.0	2023-02-25	Major rewrite	Øystein Godøy
0.9	2022-11-30	Minor modifications to text, added section about SIOS-KC training, added some links.	Luke Marsden
0.8	2022-06-10	Minor modifications to text, fixed and removed dead links and updated visual profile.	Ilkka Matero
0.7	2022-01-22	Updates following lessons learned in harvesting especially from GeoNetwork. Added outline for product manual. Updated information on interpretation of ISO 19115.	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

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 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 developed and thus become a leading observational system in the Arctic and Polar regions. The SIOS Data Management System (SDMS) [RD-2] Data Access 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 [RD-3] data centres contributing to SIOS. These data centres manage the data on behalf of the owners and providers of the data.

A major innovative element of SIOS will be the Knowledge Centre, 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 [RD-1]. 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 datasets.

For the SDMS the term “dataset” 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, as well as designated start and end times. A dataset 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 dataset 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 provide 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. Some guidance on granularity of datasets is provided in [RD-29].

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, and in the long term to harmonize accessing data via a 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 in the document represent long term goals, requiring 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 outlined in the implementation plan of the SIOS Preparatory Phase [RD-8], see section 5.3.1[1].

1.3. Intended audience

System managers at the data centres contributing to the SDMS.

1.4. Training and support

SIOS-KC offers help and services related to preparing metadata and data in accordance to the FAIR data management principes (Wilkinson et al., 2016) and the interoperability guidelines outlined within this document. You can find information about training courses, online material, and what support SIOS-KC can offer you here.

1.5. Applicable documents

[RD-1] Svalbard Integrated Arctic Earth Observing System – Preparatory Phase (SIOS-PP) Description of Work.

[RD-2] Robert Huber and Michael Klages (lead authors), 2012. SIOS Data Management System – User Requirements Document. SIOS Deliverable D6.3.

[RD-3] Distributed SIOS Data Management System Implementation Plan, 2014, SIOS Deliverable D6.6.

[RD-4] SIOS Data Policy

[RD-5] SDMS Acronyms and Concepts Document

[RD-6] SDMS System Requirements Document

[RD-7] SDMS Architecture Design Document (not public available)

[RD-8] SDMS System implementation and integration plan

[RD-9] SDMS Operations Manual (not public available)

[RD-10] WMO Information System

[RD-11] WMO Core Profile of the ISO 19115

[RD-12] WIGOS, including the metadata standard

[RD-13] The Open Archives Initiative Protocol for Metadata Harvesting, Version 2

[RD-14] OAI-PMH tools

[RD-15] OGC CSW specification

[RD-16] GCMD DIF Writers Guide

[RD-17] GCMD Science Keywords

[RD-18] Climate and Forecast Standard Names

[RD-19] WMO Code Lists

[RD-20] NetCDF

[RD-21] Climate and Forecast Conventions

[RD-22] OPeNDAP

[RD-23] UNIDATA's Common Data Model

[RD-24] OpenSearch

[RD-25] Wilkinson et al., 2016: The FAIR Guiding Principles for scientific data management and stewardship

[RD-26] SPDX License List

[RD-27] OSGeo Link Properties Lookup Table

[RD-28] GCMD URL Types

[RD-29] SDMS Granularity Perspectives Document

2. Interoperability interfaces

2.1. Discovery metadata

2.1.1. Background

Metadata are generated by the data centres hosting the datasets. Metadata are harvested and ingested in the central catalogue for usage by the SIOS Data Access Portal user community. SIOS Data Access 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 dataset.
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 Access 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 Access Portal harvests 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 [RD-7]. 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 passes 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.

NOTE	A separate document is developed on the harvest process (release pending conversion from LibreOffice document to ASCIIDOC and addition 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.

Requirement	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 Access Portal will not specify or change a unique identifier unless the dataset is hosted by the SIOS Data Access Portal.

NOTE	The only exception to this is when SIOS is harvesting discovery metadata using THREDDS Data Catalogues (described later) and the underlying NetCDF files don't have a proper and guaranteed unique identifier.

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.

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 Access Portal. Suitable interfaces for this are OAI-PMH and OGC CSW. Other interfaces are constantly evaluated, but to ensure a cost effective solution the number of 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 are freely available.

Although OAI-PMH is recommended, OGC CSW will also be harvested. Work is in progress to enable support for other exchange mechanisms for discovery metadata.

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 Access 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 is a number of open source tools available that implements this (e.g. pycsw, GeoNetwork). Some of these are listed on [RD-14]. The central node of SDMS is using pyCSW to provide OAI-PMH (and other machine readable en points).

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.

Requirement 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 reduces the load on the SDMS, 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 [3]. More information on this feature is available in OAI-PMH specification of deleted records.

Requirement 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.

Recommendation Details on the discovery metadata specifications are provided below. In order to properly identify the metadata standards in the OAI-PMH service endpoint it is recommended to use the keywords below.

Keywords

- “dif” for GCMD DIF, preferably (dif9 or dif10 to reflect the version)
- “iso19139” for the ISO19115 minimum profile

2.1.2.3. OGC CSW

The Open Geospatial Consortium Catalogue Services for the Web (OGC CSW [RD-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[4] and is supported by the SIOS Data Access Portal.

Requirement	Do not use SOAP.
Requirement	Serve ISO19115/ISO19139. Detailed information on ISO19115 profiles is provided later in the document.
IMPORTANT	Only version 2.0.2 exchange has been thoroughly tested so far.

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.

WARNING	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 being tested, but is yet not supported for regular exchange of discovery metadata within SIOS. According to [RD-24] 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.”

Within the Polar Data Management Community, there is an ongoing effort to support the exchange of discovery metadata using schema.org. The intention is to make it easier for data repositories to announce their datasets by embedding schema.org in web pages. The work is based on ESIP’s Science on schema.org[5]. Harvesting information from data repositories serving schema.org is under testing for third party data repositories.

NOTE	More information on schema.org to be added.

2.1.3. Information structures for discovery metadata

2.1.3.1. Internal discovery model

SIOS is harvesting discovery metadata from data centres using established discovery metadata schemes and transforms this information into the internal information model used. The Table 1 below shows the elements of the information model while the mapping between the information model and the metadata standard ISO 19115 and GCMD DIF are presented in the following sections. Further mappings are under development.

Table 1. Discovery model elements

Element	Description
metadata_identifier	A unique identifier (A UUID with namespace is recommended) for the dataset.
last_metadata_update	Last date of updated metadata using the form YYYY-MM-DDTHH:MM:SSZ
title	A short title for the dataset.
abstract	Short summary describing the dataset embedded in gco:CharacterString.
temporal_extent	Temporal extent of the dataset. Currently gaps are not handled.
geographic_extent	Spatial extent of the dataset. Requires all 4 corners of the BoundingBox to be set, also for point measurements. Points are interpreted if values are identical.
iso_topic_category	ISO Topic Category, using a controlled vocabulary.
keywords	A word or phrase that describes some aspect of a resource. It 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.
personnel	This field is used to identify personnel with various roles in relation to the dataset. It should also include contact information, at least email address and name of the affiliated institution, role (see below) and name.
data_access	URL to the actual dataset accompanied with identification of the protocol supported.
related_information	URL to relevant information regarding the dataset, accompanied with identification of the protocol supported. This could be, for example, a landing page, project page or scientific publication.
use_constraint	License for the metadata using SPDX License List. The identifier (adhering to the SPDX formatting) goes into gmx:Anchor and the link to the text into the attribute of this xlink:href. This is currently a recommended field, but it is strongly recommended and suggested to become mandatory in the future.

Element	Description
data_center	The host data center of the dataset. This should have both a long and short name, but only specification for the long name is currently identified for ISO mapping.
project	Project where the dataset was generated or collected. Preferably both short and long names should be provided.
dataset_language	The language used in production, storage etc. of the dataset. The default for all datasets is English.
storage_information	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.
platform	The platform used to collect the data.[6]

2.1.3.2. ISO19115

ISO19115 is an 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 [RD-17]. A mapping exists between Climate and Forecast standard names [RD-18] and GCMD Science Keywords. ISO 19115 is used by e.g. INSPIRE at the European level and WMO (WMO Core Profile [RD-11]) at the global level.

Requirement ISO19115 records **must** at least state the unique id, temporal and spatial locations, scientific content, responsible data centre and PI as well as links to the actual data[7]. Datasets without such specifications are not ingested in the catalogue.

Requirement All times **must** be encoded as ISO8601.

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 [RD-25]. 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 [RD-19][8]. 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.

NOTE SDMS interprets both gmd:MD_Metadata (ISO 19115-1) and gmi:MI_Metadata (ISO 19115-2). The mdb:MD_Metadata (ISO 19115-3) is not supported.

Below is a mapping table between the internal metadata model described in Table 1 and the ISO19115 elements.

Table 2. Mapping between the discovery model elements and ISO19115

Internal metadata model element	ISO19115	Comment/Note
metadata_identifier	gmd:fileIdentifier/gco:CharacterString	
last_metadata_update	gmd:dateStamp	ISO supports both gco:Date or gco:DateTime. A default time is added when a gco:Date type is provided.
title	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/gco:CharacterString	
abstract	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:abstract	
temporal_extent	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent/gmd:extent/gml:TimePeriod	Relies on gml:beginPosition always to be present and in the date/datetime domain: The use of indeterminatePosition is not supported. If gml:endPosition is missing or of type indeterminatePosition it will not be parsed and considered an ongoing observational effort. A default time is added when a date only entry is provided.
geographic_extent	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox	
iso_topic_category	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:topicCategory/gmd:MD_TopicCategoryCode	If this element is not provided a default "Not available" will be used.
keywords	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword/gco:CharacterString	The scope for keywords has to be identified by identification of the purpose (parameter/variable definitions, projects etc) of keywords in ISO records. Details on how this is done is provided in the Section 2.1.3.2.1 section below.

Internal metadata model element	ISO19115	Comment/Note
personnel	• gmd:contact/gmd:CI_ResponsibleParty • gmd:identificationInfo/gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty	Extraction and crediting people involved relies on gmd:role/gmd:CI_RoleCode to have attribute codeListValue set according to a predefined set of values. ISO codes principalInvestigator, pointOfContact, and author are translated into roles of Principal Investigator, Technical Contact, Metadata Author respectively. Roles not listed above are translated into Technical Contact.
data_access	gmd:distributionInfo/gmd:MD_Distribution/gmd:transferOptions/gmd:MD_DigitalTransferOptions/gmd:onLine/gmd:CI_OnlineResource	This implies that elements gmd:protocol/gco:CharacterString and gmd:linkage/gmd:URL must be set. See Section 2.1.3.2.4 section below.
related_information	• gmd:distributionInfo/gmd:MD_Distribution/gmd:transferOptions/gmd:MD_DigitalTransferOptions/gmd:onLine/gmd:CI_OnlineResource • gmd:dataSetURI/gco:CharacterString	A combination of gmd:function/gmd:CI_OnLineFunctionCode/@codeListValue and gmd:name/gco:CharacterString is used to parse this information. See Section 2.1.3.2.3 section below.
use_constraint	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:resourceConstraints/gmd:MD_LegalConstraints/gmd:useLimitation or gmd:identificationInfo/gmd:MD_DataIdentification/gmd:resourceConstraints/gmd:MD_Constraints/gmd:useLimitation or gmd:identificationInfo/gmd:MD_DataIdentification/gmd:resourceConstraints/gmd:MD_LegalConstraints/gmd:useConstraints	See Section 2.1.3.2.2 section below more information.
data_center	gmd:distributionInfo/gmd:MD_Distribution/gmd:distributor/gmd:MD_Distributor with gmd:distributorContact/gmd:CI_ResponsibleParty/gmd:role/gmd:CI_RoleCode/@codeListValue = 'distributor' or 'publisher'.	The long name goes into gmd:distributorContact/gmd:CI_ResponsibleParty/gmd:organisationName/gco:CharacterString and the URL for the data center into gmd:distributorContact/gmd:CI_ResponsibleParty/gmd:contactInfo/gmd:CI_Contact/gmd:onlineResource/gmd:CI_OnlineResource/gmd:linkage/gmd:URL.

Internal metadata model element	ISO19115	Comment/Note
project	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword/gco:CharacterString with /gmd:type/gmd:MD_KeywordTypeCode[@codeListValue = 'project']	Project information is conveyed through keywords in ISO19115 profiles.
dataset_language	gmd:identificationInfo/gmd:MD_DataIdentification/gmd:language	
storage_information	/gmd:distributionInfo/gmd:MD_Distribution/gmd:distributionFormat/gmd:MD_Format	use gmd:name/gco:CharacterString to provide the format of the dataset
platform	/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:type/gmd:MD_KeywordTypeCode with codeListValue="platform"	Not parsed.[9]

2.1.3.2.1. Keywords and vocabulary

The following approach is currently used for keywords and vocabularies:

- parsing of keyword type in gmd:type/gmd:MD_KeywordTypeCode/@codeListValue is performed looking for the values 'theme' or 'project' to fill the keywords or project elements in the internal metadata model
- parsing of thesauri in gmd:thesaurusName/gmd:CI_Citation/gmd:title is performed looking for GEMET or CF Standard names
- parsing of gmd:keyword/gco:CharacterString or gmd:keyword/gmx:Anchor is performed for GEMET and CF Standard names
- parsing of gmd:keyword/gco:CharacterString starting with 'EARTH SCIENCE >' is used to discriminate GCMD Science keywords. The full path of the GCMD science keywords including '>' is required.

2.1.3.2.2. Use constraints

The recommended way to provide a license is:

```
<gmx:Anchor xlink:href="http://spdx.org/licenses/CC-BY-4.0">CC-BY-4.0</gmx:Anchor>
```

Where the identifier (adhering to the SPDX formatting) goes into gmx:Anchor and the link to the text into the attribute of this xlink:href. This is currently a recommended field, but it is strongly recommended and suggested to become mandatory in the future. Licenses that cannot be mapped to standard licenses will be parsed as text. Some mapping stragegy is also in place during harvesting license.[10]

2.1.3.2.3. Related information (URLs)

Links to related information can be quite difficult to parse due to the degrees of freedom in ISO. The internal model is supporting related information of different types based on the Related Information Types controlled list. Currently:

- SDMS still interprets gmd:dataSetURI as a "Dataset landing page", but this is deprecated in ISO.
- gmd:distributionInfo/gmd:MD_Distribution/gmd:transferOptions/gmd:MD_DigitalTransferOptions/gmd:onLine/gmd:CI_OnlineResource and a combination of gmd:function/gmd:CI_OnLineFunctionCode/@codeListValue='information' and gmd:name/gco:CharacterString can also be used to identify a dataset landing page. SDMS is currently interpreting phrases like "Extended human readable information about the dataset" and "Landing page" as URL of type "Dataset landing page" when found.
- Information tagged "Project on RiS" is translated into the category "Other documentation"
- "Homepage" is interpreted as a "Project home page" reference.

2.1.3.2.4. Data access

The gmd:protocol containing a predefined keyword[11] and the gmd:linkage/gmd:URL are used to parse the different data access links. This is used both to identify direct download of datasets (i.e. HTTP or FTP) as well as services on top of dataset (e.g. OPeNDAP, OGC WMS). It is important to note that direct download should not refer to a website requiring manual intervention. Direct download will be handled by the basket in the data portal and enables bundling of data for download etc.

Below the list of currently mapped protocols:

ISO Inspire gmd:protocol (OSGEO)	ISO Inspire gmd:protocol (GeoNetwork)	Internal model
download	WWW:DOWNLOAD-1.0-http—download	HTTP
OPeNDAP:OPeNDAP	WWW:LINK-1.0-http—opendap	OPeNDAP
OGC:WMS	OGC:WMS	OCG WMS
OGC:WFS	OGC:WFS	OCG WFS
OGC:WCS	OGC:WCS	OCG WCS
ftp	WWW:DOWNLOAD-1.0-ftp—download	FTP

2.1.3.3. GCMD DIF

The Global Change Master Directory (GCMD) Directory Interchange Format (DIF) [RD-16] is a metadata standard that is widely used (e.g. by the Antarctic Metadata Directory) and that was the foundation for data management during the International Polar Year. SDMS is supporting both DIF 10.2 (the latest version of the standard), and also older versions (DIF 9) although deprecated.

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[12].

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. See Section 2.1.3.3.2 and Section 2.1.3.3.3 below for more information.

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

Below is a mapping table between the internal metadata model described in Table 1 and the DIF elements.

Table 3. Mapping between the discovery model elements and DIF (version 9 and 10.2)

Element	GCMD DIF 9	GCMD DIF 10.2	Comment/Note
metadata_identifier	Entry_ID	Entry_ID/ShortName	The subfield ShortName for DIF 10.2
last_metadata_update	DIF_Creation_Date and Last_DIF_Revision_Date	Metadata_Dates/Metadata_Creation and Metadata_Dates/Metadata_Last_Revision	Use the subfields Metadata_Creation and Metadata_Last_Revision in DIF 10.2. Text entries such as "Not provided" are not parsed.
title	Entry_Title	Entry_Title	
abstract	Summary/Abstract	Summary/Abstract	
temporal_extent	Temporal_Coverage	Temporal_Coverage/Range_DateTime	Use Start_Date and Stop_Date for DIF 9 and use Beginning_Date_Time and Ending_Date_Time for DIF 10.2
geographic_extent	Spatial_Coverage	Spatial_Coverage	For DIF 10.2 the subfields dif:Geometry/dif:Bounding_Rectangle, dif:Geometry/dif:Point and dif:Geometry/dif:Polygon are supported.
iso_topic_category	ISO_Topic_Category	ISO_Topic_Category	
keywords	Parameters and Keyword	Science_Keywords and Ancillary_Keyword	Parameters (DIF 9) and Science_Keywords (DIF 10.2) are translated into GCMD Science Keywords. Keyword (DIF 9) and Ancillary_Keyword (DIF 10.2) are translatde into generic keywords.
personnel	Personnel and Data_Center/Personnel	Personnel and Organization/Personnel	Information about Data Center Contact is conveyed through the Personnel subfields of Data_Center (DIF 9) or Organization (DIF 10.2)

Element	GCMD DIF 9	GCMD DIF 10.2	Comment/Note
data_access	Related_URL	Related_URL	Use the appropriate combination of URL_Content_Type/Type and URL_Content_Type/Subtype to identify the type of access and URL to provide the resource. Details are provided in the Section 2.1.3.3.3 below.
related_information	Related_URL	Related_URL	Use the appropriate combination of URL_Content_Type/Type and URL_Content_Type/Subtype to identify the type of access and URL to provide the resource. Details are provided in the Section 2.1.3.3.2 below.
use_constraint	Use_Constraint	Use_Constraint	Use the subfields License_URL/Title and License_URL/URL in DIF 10.2 to provide license indentifier and URL. Information that cannot be mapped to standard licenses will be translated into license text.
data_center	Data_Center	Organization	Use the subfields Data_Center_Name/Short_Name, Data_Center_Name/Long_Name and Data_Center_URL in DIF 9.+ Use the subfields Organization/Short_Name, Organization/Long_Name and Organization_URL in DIF 10.2 with Organization_Type "ARCHIVER"
project	Project	Project	Use the subfields Short_Name and Long_Name.
dataset_language	Data_Set_Language	Dataset_Language	

Element	GCMD DIF 9	GCMD DIF 10.2	Comment/Note
storage_information	Distribution	Distribution	Use the subfield Distribution_Format.[13]
platform	Source_Name	Platform	Subfields Sensor_Name (DIF 9) and Instrument (DIF 10.2) are used to convey information about instruments.[14]

2.1.3.3.1. Location

When provided, the DIF element Location is mapped into keywords of type GCMDLOC.

2.1.3.3.2. Related information (URLs)

Below a mapping between the internal model and Related information types and DIF Related_URL (Type and Subtype)

Internal model	DIF 9 (Type/Subtype)	DIF 10.2 (Type/Subtype)
Project home page	VIEW PROJECT HOME PAGE/	PROJECT HOME PAGE/
Users guide	VIEW RELATED INFORMATION/USER'S GUIDE	VIEW RELATED INFORMATION/USER'S GUIDE
Dataset landing page	VIEW DATA SET LANDING PAGE/	DATA SET LANDING PAGE/
Data server landing page	GET DATA/THREDDS DATA	USE SERVICE API/THREDDS DATA
Scientific publication	VIEW RELATED INFORMATION/PUBLICATIONS	VIEW RELATED INFORMATION/PUBLICATIONS
Data paper	VIEW RELATED INFORMATION/PUBLICATIONS	VIEW RELATED INFORMATION/PUBLICATIONS
Other documentation	VIEW RELATED INFORMATION/GENERAL DOCUMENTATION	VIEW RELATED INFORMATION/GENERAL DOCUMENTATION
Extended metadata	VIEW EXTENDED METADATA/	EXTENDED METADATA/

2.1.3.3.3. Data access

Below a mapping between the internal model and Data Access types and DIF Related_URL (Type and Subtype)

Internal model	DIF 9 (Type/Subtype)	DIF 10.2 (Type/Subtype)
HTTP	GET DATA/	GET DATA/DIRECT DOWNLOAD
OPeNDAP	GET DATA/OPENDAP DATA (DODS)	USE SERVICE API/OPENDAP DATA
OCG WMS	GET SERVICE/GET WEB MAP SERVICE (WMS)	USE SERVICE API/WEB MAP SERVICE (WMS)
OGC WFS	GET SERVICE/GET WEB FEATURE SERVICE (WFS)	USE SERVICE API/WEB FEATURE SERVICE (WFS)
OGC WCS	GET SERVICE/GET WEB COVERAGE SERVICE (WCS)	USE SERVICE API/WEB COVERAGE SERVICE (WCS)
FTP	GET DATA/	GET DATA/DIRECT DOWNLOAD

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. Use metadata helps one to identify 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 [RD-20] as the file format, utilising the Climate and Forecast [RD-18] convention and serving data through OPeNDAP [RD-22] simplifies the issue of integration and combination of data through the Common Data Model [RD-23]. Other solutions may be added, but this is an easy win and addresses requirements both within numerical simulation communities, satellite CalVal communities and WMO programmes.

2.2.2. Exchange mechanisms for data

2.2.2.1. Introduction

Traditionally, data have 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 and 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 describing 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 machine-actionable data format as guided by the FAIR principles, see Section 2.2.3 using standardised encoding structures and vocabularies.

2.2.2.3. OPeNDAP

The Data Access Protocol simplifies integration of data from various data centres by utilising the Common Data Model. This is possible when input data are encoded according to Climate and Forecast conventions. The use of a data stream removes the need to download files and manually keep track of them during analysis. It also allows segmentation of data in variable space and time and it is RESTful[15]. OPenDAP can relate to files or relational database systems and is extensively used by e.g. Copernicus services, Earth System Grid Federation and others.

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

Recommendation	Where possible, OPeNDAP should be supported for data access (combining multiple physical files to a single virtual dataset or visualisation of e.g. timeseries).

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.*

IMPORTANT	To enable automatic visualisation of timeseries data at stations, data has to be encoded as NetCDF-CF with the featureType global attribute set. It is furthermore important to avoid mixing stations in a file, but rather have one file per station.

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 process 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, in particular OGC Environmental Data Retrieval (EDR). No such activity is undertaken for the old OGC WFS.

IMPORTANT	The data portal is not able to process 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”,

NOTE	Same comment here as for OGC WFS and the support for new OGC API's like OGC EDR.

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 URL Types [RD-28] and OSGEO [RD-27] 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 service 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[16] (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) and the Ocean Biodiversity Information System (OBIS). In essence it is a set of comma separated files (CSV) bundled with a

metadata file (meta.xml) and using controlled vocabularies to describe the content. A second file (eml.xml) describes the content and strucuture of the CSVs and links the column headers to the Darwin Core terms. These XML files can be created quite easily using GBIF's Integrated Publishing Toolkit software. This not only provides a user interface for creating Darwin Core Archives - the software also hosts the published Darwin Core Archives which can be registered to be made available via GBIF. SDMS cannot do much on top of Darwin Core Archives (the diversity of types of data is too large), but the format is more or less machine-actionable and is recommended for use within SDMS. Further information is available at https://dwc.tdwg.org/terms/.

Recommendation	Darwin Core Archive is recommended for biodiversity and associated data (e.g. experimental, measurements/traits, data derived from metabarcoding) within SDMS.
Recommendation	Data centres with suitable data should consider hosting an IPT. GBIF may be able to help with the installation. This provides a way to make biodiversity data available via both GBIF and SIOS since metadata are not currently harvested from GBIF.

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 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 not yet mature enough.

IMPORTANT	While SDMS can make JSON files available for download, it is **not** able to support services (e.g. visualisation, aggregation, subsetting) based on them, as the format is not sufficiently standardised across providers.

NOTE	SDMS should work to enable ACDD and CF elements in GeoJSON files when the new OGC API's emerge.

2.2.3.4. NetCDF/CF

NetCDF is a container like JSON and XML, and as 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. CF-NetCDF is thus a machine-actionable data 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 as the go-to 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).
Recommendation	You can check your NetCDF file against the CF and ACDD conventions at https://sios-svalbard.org/dataset_validation/form

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.

IMPORTANT	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 CF-NetCDF 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.

IMPORTANT	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.

NOTE	XML is more or less fully replaced by various flavours of JSON now. The new OGC API's also

include CF-NetCDF as a information container.

IMPORTANT	While SDMS can make XML files available for download, it is **not** able to support services (e.g. visualisation, aggregation, subsetting) based on them. XML is **not** recommended as exchange format even though standardised representations (e.g. WaterML) exist.

Appendix A: Outline for Product Manual

A.1. Introduction

When datasets cannot be presented in a machine-actionable data format as guided by the FAIR principles (i.e. self describing), a product manual describing the dataset and how to interpret it **should** accompany the dataset.

IMPORTANT	This product manual has to be served as a PDF document and have the sections identified below.
IMPORTANT	The product manual must be linked to the data through the discovery metadata at least, preferably it should be available in the same bundle as well.

A.2. Outline

Outline of sections that should be covered by a product manual for data that cannot be described in a machine-actionable data format.

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

1. Introduction
   a. General overview of the data production process
   b. Information of personnel involved and their contact details
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.

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

[3] This may change.

[4] Not required for scientific data.

[5] https://github.com/ESIPFed/science-on-schema.org

[6] This is a complex field, with several subfields. It currently support a limited vocabulary.

[7] This recommendation will be revisited.

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

[9] This field is currently not parsed and mapping information might be modified in the future.

[10] An internal mapping is done upon harvesting to parse license title or exact maching urls. If provided, a gco:CharacterString can also parsed.

[11] OSGEO or GeoNetwork keywords are required for proper interpretation. These keywords are translated in the harvesting routine.

[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 is not parsed for the time being.

[14] This is not parsed for the time being.

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

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