POLDER best practice guide to implementing schema.org for data discovery

2023 V 1.0

Authors

Pip Bricher[1,16] (https://orcid.org/0000-0001-7975-5307), Chantelle Verhey[2] (https://orcid.org/0000-0002-0047-7875), Ruth Duerr[3] (https://orcid.org/0000-0003-4808-4736), Rebekah Ingram[4], Peter Pulsifer[4] (https://orcid.org/0000-0001-9561-3640), Julia Collins[5], Matthew B. Jones[6,10] (https://orcid.org/0000-0003-0077-4738), Taco de Bruin[7] (https://orcid.org/0000-0001-9149-2095), William Manley[8] (https://orcid.org/0000-0001-7922-9753), Shannon Christoffersen[9], Amber Budden[6,10], Anton Van de Putte[11, 11X] (https://orcid.org/0000-0003-1336-5554), Allison Gaylord[12](https://orcid.org/0000-0002-3254-6791), Marten Tacoma[7] (https://orcid.org/0000-0002-3733-7278), Alice Fremand[13] (https://orcid.org/0000-0001-8272-0981) , Giuseppe Aulicino[15] (https://orcid.org/0000-0001-6406-8715), Alex Gao[4], Melinda Minch[2] (https://orcid.org/0000-0003-3878-7147), Stein Tronstad[14] (https://orcid.org/0000-0002-9743-7211), Amos Hayes[4] (https://orcid.org/0000-0003-0878-0868), Øystein Godøy[17] (https://orcid.org/0000-0001-6410-3488), Olaf Schneider[14]

[1] Southern Ocean Observing System, University of Tasmania

[2] World Data System-International Technology Office

[3] Ronin Institute
[4] Geomatics and Cartographic Research Centre, Carleton University
[5] NSIDC/CIRES/CU
[6] DataONE
[7] NIOZ Royal Netherlands Institute for Sea Research
[8] INSTAAR, University of Colorado at Boulder
[9] Arctic Institute of North America, University of Calgary
[10] Arctic Data Center
[11] biodiversity.aq, Royal Belgian institute for Natural Sciences
[11X] Université LIbre de Bruxelles
[12] Nuna Technologies
[13] UK PDC/BAS
[14] Norwegian Polar Institute
[15] University of Napoli Parthenope
[16] Australian Hydrographic Office
[17] Norwegian Meteorological Institute

Recommended Citation: Bricher, P., Verhey, C., Duerr, R., Ingram, R., Pulsifer, P., Collins, J., Jones, M., de Bruin, T., Manley, W., Christoffersen, S., Budden, A., Van de Putte, A., Gaylord, A., Tacoma, M., Fremand, A., Aulicino, G., Gao, A., Minch, M., Tronstad, St., … Schneider, O. (2023). POLDER best practice guide to implementing schema.org for data discovery (1.0). Zenodo. https://doi.org/10.5281/zenodo.7787161

Executive Summary

This document aims to utilize schema.org for the Polar data management community. The community has agreed on and plans on implementing uniform best practices for documenting data, observing assets, and other entities. Extensive work has been conducted under both the Earth Science Information Partnership Science-on-schema.org group and the Research Data Alliance to develop approaches and guidelines for interoperable metadata practices. While the current Science-on-schema.org guidelines provide a strong technical basis for harmonised use of the schema.org vocabulary, they do not make specific requirements for minimal acceptable metadata nor for specific types of metadata requirements for key polar research use cases. This document assembles data discovery use cases and requirements for polar data discovery that represent a target set of features for the POLDER federated search system. Use cases specifically elucidate and prioritise the functional uses of the federated data discovery platform, which will in turn be used to articulate a specific list of metadata requirements needed to implement a discovery system that provides those features.

Table of Contents

1. Introduction

Previous experience with implementing metadata standards in the polar data management community has demonstrated that there are as many ways to write metadata as there are metadata authors. The end result of many parallel metadata writing efforts is difficulty in brokering the various flavours of each metadata standard. With schema.org (SDO) being a relatively young technology in the scientific data management field, we have the opportunity to collectively agree on some core principles of best practice before too many data centres have implemented it. Semantic markup is a core technology the research data management community has at our disposal. Adopting semantic markup and related technologies assists in automating our workflows. If you are new to the schema.org realm, there is a 'Schema.org for Research Data Managers: A Primer' (Payne, K., & Verhey, C., 2022) that was created to lay out the SDO landscape, where it came from, what is driving its uptake in research data management, and how it works in broad strokes. It was designed to introduce individuals with little technical knowledge to the benefits and importance of schema.org. It is aimed to be a 'one-step-back' to help set the landscape and equip you to better understand the various 'Best Practices' documents throughout the RDM community.

1.1 Related activities in the schema.org community

There are numerous initiatives currently working to harness the benefits of schema.org and produce valuable resources for others to follow suit. Many of these have informed POLDER's deliberations and therefore provide important context for this document. A few of these are described in the paragraphs below.

DataONE is a global community-driven network for harvesting schema.org and other structured metadata models to provide aggregated data services, including spatial, temporal, and semantic search, metadata FAIR reports, and data citation reports, all within customizable data portals (see https://search.dataone.org/portals/polderdemo/).

Science-on-schema.org (SOSO) is an Earth Science Information Partners (ESIP)-led cluster providing guidance for publishing schema.org in JSON-LD to the science community. Currently, the group has released the V1.3 set of their recommendations that help describe Datasets and Data Repositories. The science-on-schema guides include examples for 'variableMeasured', 'funding', 'identifier' and many more.

The United Nations has declared the 2020s to be the United Nations Decade of Ocean Science for Sustainable Development (UNDOS). The UNDOS Task Force has identified scientific priorities, including six cross-cutting challenges relevant to the data community, including "ensur[ing] capacity development and access to knowledge; Improv[ing] interdisciplinary capacity and knowledge integration; and facilitate[ing] transnational cooperation and complementarity". In response to these challenges, the Ocean Info Hub (OIH) was created. The OIH uses the gleaner.io tool that extracts JSON-LD from web pages. The OIH provides Gleaner with a list of ocean repositories to index and it will access and retrieve pages based on the

sitemap.xml of the domain(s). Gleaner can then compile the information into a form usable to drive a search interface.

NASA's International Directory Node (IDN), formerly known as the Global Change Master Directory (GCMD), has played a critical role in the development of the Antarctic data management community. The IDN underpins both the Antarctic Master Directory (AMD) and the Southern Ocean Observing System's (SOOS) metadata portal. The AMD presents all records in the IDN that were contributed by National Antarctic Data Centres, and contains a mix of unique records that were uploaded directly to the IDN and duplicates of records held by the NADCs. The AMD has been the key element around which the Standing Committee on Antarctic Data Management (a body of the Scientific Committee on Antarctic Research) has organised its data management activities over the past two decades. The SOOS metadata portal is a subset of all IDN holdings that overlap the geographic region below 40S, include keywords associated with oceans or coast, and with a few specific exclusions to remove terrestrial and sociological data.. The IDN has implemented schema.org as part of its ongoing development activities.

The Alaska Data Integration Working Group (ADIwg) pursued a related effort to share metadata through an ISO 19115/19110 compatible JSON standard. ADIwg published a GitHub archive to share schema documentation, an editor and translation tools. The US Fish and Wildlife Service's Science Applications Program, the US Geological Survey (USGS) and the US Bureau of Ocean Energy Management have continued to work in partnership to utilise this community adoption to describe projects and datasets. The USGS Alaska Science Center has also been working to include descriptions of collections of sampling sites (with expertise from ISO specialist Ted Habermann). The Arctic Research Mapping Application and Arctic Observing Viewer were early adopters of ADIwg's community specification for ISO XML and will continue to provide access to information for NSF projects through a new Arctic Operations Gateway API endpoint.

Several Interest and Working Groups of the Research Data Alliance (RDA) are focused on a range of metadata issues, including those pertinent to data discoverability, access, and interoperability. These groups include the Metadata Interest Group, Brokering Framework Working Group, Metadata Standards Catalog Working Group, and the Research Data Repository Interoperability Working Group. Of particular note is the set of guidelines published by the RDA Research Metadata Schemas Working Group. These guidelines do not advocate for any particular metadata schema when implementing schema.org, but instead are intended to ensure the consistent application of schema.org markup regardless of data source, and thus improve data discoverability over the long term.

The Canadian Consortium for Arctic Data Interoperability (CCADI) is an initiative to develop an integrated Canadian Arctic data management system (distributed) that will facilitate information discovery, establish sharing standards, enable interoperability among existing data infrastructures, and that will be co-designed with, and accessible to, a broad user base. The CCADI team is co-designing and implementing a multi-tiered architecture that includes establishment of metadata, data, vocabulary, media and other information services. Individual CCADI member centres are implementing schema.org metadata publication tools guided by

POLDER best practices. To support integration of metadata from different sources, crosswalk ontologies and a semantic mediator are under development. Standardised metadata will be served to end users through a metadata aggregator developed and hosted by the Polar Data Catalogue. Metadata and data services will be released throughout 2022.

WC3 provides a best practices document which can be found at https://www.w3.org/TR/dwbp/ . Section 8.2 recommends the following descriptive metadata fields:

The **title** and a description of the dataset.
The **keywords** describing the dataset.
The **date of publication** of the dataset.
The **entity responsible (publisher)** for making the dataset available.
The **contact point** for the dataset.
The **spatial coverage** of the dataset.
The **temporal period** that the dataset covers.
The **date of last modification** of the dataset.
The **themes/categories** covered by a dataset.
The **title** and a **description** of the distribution.
The **date of publication** of the distribution.
The **media type** of the distribution.

1.2 Polar data discovery as a use case

There are a wide range of use cases for including schema.org metadata on data landing pages. These range all the way from simply having your data show up on Google's Dataset Search to supporting the development of very advanced domain specific data discovery, access and analysis clients. Here the goal is intermediate between these two extremes - "To support the development of a federated polar data discovery portal", referred to hereafter as Polar Federated Search. Polar Federated Search is conceived as a single point of access that allows users to perform basic searches (text, time, space) on metadata records held in a large number of data catalogues that host polar-relevant data.

A note on scope: What do we mean by polar data?

Polar data is widely disparate in terms of formats, topics, and the people and institutions that collect them. Numerous forms of traditional knowledge are considered data. One example is the Inuit tactile driftwood maps with carved knobs and notches used to represent capes, islands, and inlets on the Greenlandic coast (*see figure 1 below*). Geographically, POLDER represents data managers working in the Arctic, Antarctic, and Southern Ocean. The latitudinal boundaries for these regions vary between institutions. This document refers to polar data as any data within the polar region. The polar region will be defined as above 50 degrees latitudinal north in North America, Scandinavia, Asia; 60 degrees latitude north in Europe and below 40 degrees latitude south in the southern hemisphere (acknowledging that this rough guide includes the southern tips of continents not usually thought of as polar) (Verhey, C., Minch, M., Payne, K., & PPFS Advisory Team, 2022).

Figure 1: Inuit tactile driftwood map. Lightweight, made for kayak travel, specific to -feel- rather than sight (pictured above) Source: (Jakobsen, 2000)

1.3 Scope of this Document

A comprehensive guide to the implementation of schema.org for the earth science community is provided in Science-On-Schema.org (https://doi.org/10.5281/zenodo.6502539). This guide therefore is intended to complement that guide, not to override it. If there is a conflict between the SOSO guidance and the information given in this guide, SOSO should take precedence. This guide is written as a companion to the SOSO guides, providing specific guidance on metadata elements of key interest in the polar data community.

1.3.1 Aims

The aim of this document is to outline the required and recommended schema.org mark-up terms that repositories would need to implement in their metadata landing pages in order to be included in POLDERs Polar Federated Search. This document will cover any issues related specifically to the implementation of SDO in the context of polar data. The goal here is to describe the implementation of schema.org into the POLDER Federated Search Tool. Additionally, it is noted that a best practice guide for a collaboration such as Polar Federated Search must strike a balance between being implementable by the partner data centres and encouraging the use of tools that will improve the quality of metadata being shared.

2. POLDER Required fields

POLDER considers that for effective searching of most environmental, sociological, indigenous knowledge, and other polar datasets, more information is needed than Google demands. The following elements are considered mandatory for polar federated search. More detailed discussion of each element follows below.

Discovery
These fields are needed to support querying.

- Temporal coverage

- Spatial coverage
- Parameters/Variables

Bibliographic
These fields should be supplied to properly credit the products discovered as well as to facilitate access.

- Citation
- Creator
- Date Published
- Identifier
- Publisher
- Licence
- Distribution (i.e., how to get data)

Note on Google Dataset Search Requirements:

Google's requirements for having a schema.org record indexed by Google Dataset Search could be considered a bare minimum implementation; however, these requirements are not detailed enough to support the POLDER search we've agreed to. The fields required to support this minimal Google Dataset Search are:

- Title (schema:name)
- Description

Even with such a minimal schema, Google Dataset Search still rejects many potential records, as it requires the description to be > 50 and < 5000 characters, which trips up many providers as it is common for abstracts to be longer than 5000 characters.

2.1 Citation

General considerations

The data policies for the polar data committees (ADC, SCADM, and SOOS) strongly encourage all data producers to provide citation information and all data users to appropriately cite all data that they use. It is thus important that all catalogues providing metadata through schema.org provide citation information to support these imperatives.

Guidance

- Follow the SOSO guidance on citations.
- SOSO asks for citation components, rather than a citation string, but some people give both and they can be contradictory.
- If a PID is provided, a citation field or citation components can be optional.
- Citation should be optional but the components of it should not be optional.
    - If both are included and they differ then we recommend that repositories use the components. Citations to associated papers belong elsewhere.

- Additional guidance: ESIP Data Preservation and Stewardship Committee (2019): Data Citation Guidelines for Earth Science Data, Version 2. ESIP. Online resource. https://doi.org/10.6084/m9.figshare.8441816.v1

2.1.1 Creator

General Considerations

A citation which identifies a paper about or based on the dataset should not be supplied as a direct property of the dataset itself: this could be associated with other elements, such as *subjectOf*.

Debate within SOSO on these topics is described at https://github.com/ESIPFed/science-on-schema.org/issues/42 .

Notes on identifying authorship:
Some disagreement persists on the use of *creator* and/or *author* to identify the originator of a dataset. SOSO treats these terms as synonyms, so any combination could in principle be cited. However, a wider consensus may be forming on the preferred use of *creator*, based on usage by data-centres. Where multiple creators are cited, there is no intrinsic property through which to specify the order in which they are listed (and this is complicated further if both *creator* and *author* are supplied). Others may be listed using the *contributor* property, but this is considered a weaker degree of involvement. Use of *creditText* is not recommended, as the lack of internal structure in free-form text prevents machine-interpretability.

Note on use of citation-strings:
It is challenging for harvesters to parse conventional citation-strings programmatically, as the internal organisation of their components may vary between styles and implementations. Therefore, SOSO recommends that the various components (authors, title, date, DOI, etc) of the full citation are supplied in the corresponding separate schema.org elements: these should therefore be considered mandatory, and the full citation-string optional. If both the components of a citation and a full citation-string are included, it is important that these do not contradict one another. If they do, the components should be considered to provide the authoritative version.

Guidance

- For guidelines relating to citation components, refer to ESIP Data Preservation and Stewardship Committee (2019): Data Citation Guidelines for Earth Science Data, Version 2. ESIP. Online resource. https://doi.org/10.6084/m9.figshare.8441816.v1

2.1.2 Date published

General Considerations

The date the metadata was published.

Guidance

- Follow SOSO guidance on Dates
    - use ISO 8601 time format

2.1.3 Identifier

General Considerations

For many newer initiatives, digital object identifiers (DOIs) have become the standard way to persistently identify (PID) a dataset and POLDER supports the use of DOIs. It also supports other PIDs that have resolution services.

However, many polar repositories developed metadata systems before the advent of DOIs and other PIDs and these data repositories need to be represented in a Polar Federated Search, so we do not make DOIs or other PIDs mandatory. .

Guidance

- Follow the SOSO guidance (which recommends using DOIs)
    - the identifiers.org registry has examples of how to use property values to identify an identifier
- Every record should have a globally unique PID that resolves to a landing page. If it has a PID, then a citation field is optional.
- If a record does not have a globally unique PID, then any identifiers attached to that record should be included. In schema, that PID could be a URL, a DOI, or text.
- The identifier may be represented by the schema.org PropertyValue type (https://schema.org/PropertyValue): this provides many elements through which to provide further descriptive details of the supplied identifier.
- Original identifiers should NEVER be deleted or replaced. We are surprised that we need to state this explicitly, but apparently we do
- If a record has multiple identifiers, make this a repeating element and keep all identifiers.
    - To see specific examples of multiple identifiers, see section 6.1.1 'Examples' below.

2.2 Temporal coverage

General Considerations

Temporal coverage is defined as "the time period during which data was collected or observations were made; or a time period that an activity or collection is linked to intellectually or thematically (for example, 1997 to 1998; the 18th century)" (ARDC RIF-CS).

Temporal coverage is distinct from the publication or modification date of the dataset.

Guidance

- See SOSO temporal coverage for details of the Science on schema.org guidance
- For temporal coverage that is well described by dates, use ISO 8601 time interval format.
    - Uncontrolled plain text date strings (e.g. “January 20, 2017”) are uninterpretable by machines and so do not fulfil POLDER’s requirement for describing temporal coverage.
- Similarly, uncharacterized URLs also do not meet POLDER’s requirements for temporal coverage. A URL used to describe temporal coverage should be dereferenceable, containing both machine- and human-understandable definitions.

2.2.1 Dynamically updated datasets

General Considerations

In many catalogues, datasets are labelled as “ongoing” when they are created, as the intention is to continue to add to them. However, when the data collection ends, the end date is often not updated to provide the true end date. Thus, the dataset will be found by any temporal search that includes any time after the start date, even many years after the dataset was last updated. Metadata records are left without a true end date sufficiently often that POLDER believes that data centres should never label a dataset as “ongoing”.

Guidance

- Repositories should update the end date routinely, though not more frequently than once per day. The frequency of the update should be based on the frequency of data acquisition by the repository.
- Thus, if a dataset is being updated automatically, associated metadata updates should be automated as well.
- For both manually and automatically updated datasets, the end date specified in the metadata should always match the date of the latest observation in the dataset as currently stored in the repository, even if more data is likely to be added later.
- The end date field must never have a text value like “ongoing”.
- The end date field should not be left blank.
- More general information on this topic is provided by *Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data* https://hdsr.mitpress.mit.edu/pub/si7wzxxa/release/2?readingCollection=1ccd159a

Note: We expect that Polar Federated Search and other search tools will ignore the temporal field for a dataset that has an empty end date when a user uses a temporal filter. In other words, if your data doesn’t have an end date, it will not show up whenever an end date is queried!

2.2.2 Discontinuous, cyclical or seasonal data

General Considerations

Seasonal data are important for many environmental questions, so discontinuous time ranges are common in many kinds of data (e.g., intermittent data gathering efforts). As a result, representing these discontinuous time ranges is important. While schema.org allows for associating multiple time ranges to a dataset, named seasons (e.g., moose season, winter, ice breakup) cannot be easily turned into machine-readable date-time ranges. Data should not be described by a named season, unless it is linked to an ontology that defines the season as an ISO 8601 time interval. A data creator may need to create the relevant ontology if one does not exist. In that case, the SOSO recommendation is simply to use the IRI of the ontology term, as has been done to describe geologic time in this example:

"temporalCoverage": "http://sweetontology.net/stateTimeGeologic/Paleocene"

There is considerable discussion on the best ways to represent data from non-Gregorian calendars or with recurrent collection intervals, which may inform ongoing discussions about how best to describe and manage these datasets. Key texts in this field include De Souza et al. (2014) discussing time series provenance; Carriera et al. (2021) on recurrent situation series; and Cox 2016) on time ontologies for non-Gregorian calendars.

Note also that SOSO recommendations for temporal coverage will change in the v1.3 release: they will expand guidance on representing geological time-intervals, with additional structures for pre-calendar coverages. Handling of Gregorian coverages will not change. See https://github.com/ESIPFed/science-on-schema.org/pull/181 for details.

Guidance

- Follow the SOSO recommendations for associating the season name to its ontology.
- A dataset with discontinuous time ranges should have multiple, repeating temporal coverage records - one for each time range;
- Avoid using seasonal names unless there is an existing ontology defining them available or you create one using internationally recognised guidelines for ontology creation (E.g. https://obofoundry.org/)

2.2.3 Uncertain dates

General Considerations

In many collections, the temporal coverage of a dataset may be uncertain. For example, a scrapbook of photos from a historic field program may be identified only within a decade or two.

Guidance

- Follow SOSO Guidance (OWL time guidance) around uncertainty (see example 3 in the linked section)

- If having those data show up in a temporal search is important to your repository, include the presumed date range in your markup to an appropriate level of precision (e.g. year(s) or decade(s).

For example, if a scrapbook of Arctic photos was from the 1950s, the temporal coverage could be described as:

"temporalCoverage": "1950/1959"

2.2.4 Deep time and chronometric dates

General Considerations

While supporting deep time and chronometric dates is generally agreed to be important for polar federated search, this is a challenge in many parts of the globe beyond polar regions and is being actively discussed within SOSO. SOSO guidance suggests the use of ontologies to refer to named time periods within the `owl:Time` approach for representing structured times. For example,`icsc:Triassic` is an example time period from the recommended International Chronostratigraphic Chart vocabulary that could be used in a temporal coverage field for deep time:

```
"time:hasBeginning": {
     "@type": "time:Instant",
     "time:inTimePosition": {
         "@type": "time:TimePosition",
         "time:hasTRS": {"@id": "ts:gts2020"},
         "time:NominalPosition": {
             "@value": "icsc:Triassic",
             "@type": "xsd:anyURI"
         }
     }
 }
```

Guidance

- For full guidance on representing deep time, POLDER advises that catalogues follow developments in the SOSO guidance on managing deep time. See Geologic Time examples in SOSO V1.3.

2.3 Spatial coverage

General considerations

There is extensive treatment of spatial coverage in the SOSO guidance on this topic, and this document does not attempt to duplicate that, but instead to expand on particular issues for polar regions.

Bounding boxes based on global map projections (e.g. Mercator) are problematic for describing data in polar regions, especially for data that cover both poles (but not the equator); for scattered sites; ship or buoy tracks; or other large geographic ranges. Bounding boxes can also be problematic when crossing the dateline, as they can be misinterpreted to wrap around the entire globe. Despite these limitations, bounding boxes are widely used in polar-relevant data centres due to the simplicity of implementation and long history of their use. Where possible, accommodating lists of unprojected latitude/longitude pairs for scattered locations would be preferred to auto-generating centroid coordinates for large bounding boxes.

Therefore, Polar Federated Search will likely need to support single points and bounding boxes as forms of spatial coverage but we recommend data centres store spatial coverage as a list of GeoShapes or GeoCoordinates, especially to enable the I in FAIR (Findable, Accessible, Interoperable, Re-usable).

Guidance

- Follow SOSO Guidance for Spatial Coverage
- Use a list of GeoShapes or GeoCoordinates for your spatial coverage where possible

2.3.1 Place names

General Considerations

Where place names are associated with detailed spatial boundaries (e.g., administrative boundaries, watersheds, etc.), these may often provide much better query results than a roughly drawn polygon, which may hit multiple defined places inadvertently. However, place names must be attached to an ontology that includes detailed spatial boundaries to be more useful than as simply a text string.

Guidance

- Follow SOSO guidelines for associating place names with their detailed spatial boundaries in existing controlled vocabularies/gazetteers
- The schema:name property should be used for the human-readable name of a place

2.4 Parameters/Variables Measured:

General Considerations

In schema.org, the repeatable property variableMeasured can indicate the variables that are measured in some dataset, either described as text or as pairs of identifiers and description using PropertyValue."

Variables in a dataset present various kinds of information. Examples given include dataset identifiers such as keys, date of creation/update, source, temperature value, colour, species, and metadata information. VariableMeasured can be a text string, but a PropertyValue type is preferred. These are not particularly adaptable to humanities and social science data. For example, oral place names, traditional songs and stories, petroglyphs and Inuit driftwood tactile maps can all be considered types of data (Fig. 1), but do not readily fit within the scope of schema's PropertyValue. Although schema also recommends using QUDT ontology with the DataType class for such instances, the user then essentially ends up creating their own class within the text field, i.e. "oral place name" (as opposed to, for example, "written place name") or "driftwood tactile map" (as opposed to "Marshall Island stick map"). Describing these data using a simple text string makes it possible to utilise schema for this type of information, but essentially creating a class "by hand" for each entry may not be adequate enough to warrant utilisation of schema for such a purpose. Given the wealth of information about the Arctic from the Indigenous peoples of the Arctic which traditionally comes in forms other than written Roman alphabet, or "objectively measurable instrumental data", it seems clear that this is an issue that deserves further discussion given the drive by Inuit- and Indigenous-led and organisations to include such data.

Guidance

- Follow SOSO Guidance on Variables Measured
- Where a dataset includes variables that can be measured or directly calculated from measurements, use variableMeasured
- POLDER is working with schema.org and ESIP to identify best practice ways to describe datasets for which the property variableMeasured is inappropriate

2.5 Licensing

General Considerations

Data must be labelled as reusable. Open data access and FAIR legal interoperability require that the rights to reuse the data are made clear to the user. As copyright legislation and specific requirements and obligations tied to licensing vary across jurisdictions, the rights and obligations of the data originator and the data user may be declared by attaching a rights waiver, a public domain statement, or a non-restrictive, internationally recognised data licence to the dataset. A URL that identifies the pertinent legal document can be provided in the schema:licence property.

Guidance

- Follow the SOSO guidelines on licensing.
- Where a data centre has been assigned a licence or public domain mark it must be represented in schema.org using the licence variable
- Using CC 0 or the public domain mark for works in the public domain
    - CC 0 is used by a rightsholder to assign a work that is not already in the public domain to the public domain

    - The Public Domain Mark is used to indicate that a work is already in the public domain

2.6 Pointing to full metadata records

General Considerations

All schema.org records should resolve to a full metadata record (or more than one). Follow SOSO guidance on how to do this.

Guidance

- Dealing with conflicts between a SOSO record and a full metadata record is hard, so avoid doing that if at all possible (e.g., Title can be used in entirely different ways).
    - It should be noted that DataONE has asked each centre to define whether their SO or ISO record is primary.

2.7 contentUrl link

General Considerations

Automating data access through structured information that explicitly contains information about where to download the actual data can improve data accessibility (the A in FAIR), and make it more amenable to use in multiple applications.

Guidance

- Follow the SOSO guidance on providing download information through `schema:distribution` using the contentUrl field. Depending on how it is applied, the contentUrl can apply to a distribution of the whole dataset, or to individual DataDownload objects that represent either parts of a dataset or different serialisations of the dataset.

3. POLDER Optional fields:

In addition to the fields that POLDER considers essential for making a record searchable through Polar Federated Search, the following fields are recommended wherever they are relevant.

3.1 Same As

General Considerations

As the FAIR principles are increasingly applied across the data management community, and individual observations and datasets are harvested and republished across multiple data sharing systems, the sameAs property is a valuable way to help identify duplicate datasets and metadata records.

Guidance

- Where a metadata catalogue is harvesting records from another catalogue, the sameAs property should be used to identify the authoritative version.

3.2 Authoritative Sources

General Considerations

Being addressed somewhat in SOSO. What is an authoritative source? Is it the original host repository? SOSO issue 37 covers this. SOSO has delayed resolution because we need to work out how Google decides what's authoritative. This refers to pointing to the original archive that is really responsible for these data.

Guidance

- Follow the SOSO guidance

4. Publishing dataset landing pages

4.1 Sitemap

General Considerations

Sitemaps provide guidance to a harvester on where to find metadata records, considerably reducing the load on the harvester and the intrusiveness of the search. Therefore, a metadata catalogue would not be able to be directly included in a Polar Federated Search if it does not provide a sitemap.

Guidance

- Follow the SOSO guidance on how to provide a sitemap.xml file.

4.2 The challenges of catalogues with dynamically generated landing pages

General Considerations

JSON-LD is generally inserted into (landing) pages in two ways. One is to include JSON-LD along with the page contents as the page is being served to an end user who loads it in. In this circumstance, the JSON-LD is readily available within the page from the start.

The other method is for the server to send out the basic information of a page needed, then client-side JavaScript will continue to run as the page is loading in for the end user. This pre-processing step could include a variety of actions which include fetching the respective JSON-LD metadata records and inserting it within the page.

Gleaner can handle both cases, if the landing page is static and thus already includes the JSON-LD, then it is retrieved normally. Normally refers to if it's done as a programmatic web request that does not require a webpage to be rendered. The HTML for the page is retrieved and the JSON-LD is found in it. If the landing page is dynamic, a browser without a user interface ("headless") is used to render the page (which also asks to run all of the client-side pre-processing) and then retrieves the inserted JSON-LD. Headless metadata retrieval is less desirable because it is more time-consuming, and occasionally less reliable. In addition to the usual time that Gleaner waits for an indexing request, which retrieves plain HTML, Gleaner then has to wait for the headless browser to render that HTML, and complete any other attendant actions (like fetching the metadata records) as part of fully loading the page.

5. Recommendations on how to document repository information

General recommendations

Follow SOSO Guidelines on providing a description of your repository or catalog.

6. Implementation Stories

The POLDER WG has held two workshops to help interested repositories complete the POLDER schema.org best practices as outlined in the above sections. Workshop recordings can be found here:

- Jan 26th Workshop 1 (Integrated in the Polar to Global Hackathon): https://drive.google.com/file/d/1QXXhtLEhGzRcPGs9sakrteTWYrgOLg0y/view
- Feb 7th Workshop 2: Recording - https://www.youtube.com/watch?v=ufgx3YViOFM ; agenda: https://drive.google.com/file/d/1q0DhBZEoowpsVnMfZg7EW4RxPJxBK5NF/view?usp=sharing

The following are the two organisations that allowed the POLDER WG to use their dataset landing pages as examples at the above workshops. This entailed using their examples and walking through the implementation of the SDO in their example landing pages. These two repositories are good examples of the implementation of SDO at various repository maturity levels. Antarctica New Zealand was a new repository, at the time of this writing, and completing both the actual creation of their metadata catalogue and completing the SDO implementation

alongside their development. Where the Norwegian Polar institute was a well established existing repository that is implementing the SDO recommendations at a later date.

Norwegian Polar Institute

In this case the implementation was done in-house over the course of about one week, including unit tests and some end-to-end testing. As reported by the developer the process was straightforward, following the POLDER SDO Best Practice Guide, the SOSO dataset description guide, and the Sitemaps XML format. Parameters/variables constitute an exception, as these are not included in the local data model. Overall, the implementation of schema.org using a sitemap and embedded JSON-LD was experienced as “surprisingly easy”, especially as compared to OAI-PMH.

Antarctica New Zealand

This repository was able to complete the implementation of the POLDER SDO best practices alongside the actual establishment of their database and done by a third party. They utilised this document in the SOSO guidance documentation to implement it, coupled with the 1-to-1 help at the workshop above, they were able to have their implementation reviewed and validated by the WDS-ITO dedicated web developer and overcame some minor inconsistencies.

7. Document Maintenance plan

We intend to update this document as needed. This includes anytime the POLDER community feels there has been updated information. It would also be good practice for a general review after any SOSO major release to ensure any links and/or content remains aligned. We also plan to ensure that we adhere to any of the ocean best practices guidelines for maintenance.

This document will be updated and properly versioned in both Zenodo and the Ocean Best Practices (OBPS).

8. Other sources of recommendations (for things other than datasets)

For Polar observing assets – infrastructure and activities, e.g. research and monitoring sites, individually funded research projects, and more – see evolving guidance from the SAON Polar Observing Assets Working Group (POAwg).

Alignment of polar data policies (DOI: `10.5281/zenodo.5734900`) is a report published in 2021 by a working group under the Arctic Data Committee (ADC), Arctic Spatial Data Infrastructure (ASDI), Southern Ocean Observing System (SOOS) and Standing Committee on Antarctic Data Management (SCADM). The report develops common principles for the management of scientific data from the polar regions, some of which pertain to dataset descriptions. The

principles establish, inter alia, that the FAIR Principles should be applied to the greatest extent practicable; that all data must be accompanied by a complete set of metadata, have persistent and globally unique identifiers, be labelled as reusable, and that data sources should be attributable and attributed.

The Earth Science Information Partners hosts several working groups and clusters related to the semantic web and data discovery. A prevalent group to watch is the ESIP Semantics Harmonization Cluster which stems from the Semantic Technologies Committee. The Semantics Harmonization cluster is currently working on writing up a suite of leading semantics harmonisation practices documentation based on findings from their GCW-ENVO-SWEET cryospheric vocabulary work. The alignment of approximately 43 ice terms have been harmonised through the work of this group and monumentous efforts of Ruth Duerr. The ESIP WG also hosts groups related to citation, such as the Research Object citation cluster. The ESIP website hosts a telecon calendar outlining the various WG and their meeting times for any interested parties, the calendar can be found at: https://www.esipfed.org/get-involved/community-calendar.

9. Annexes[1]

9.1 Acronyms

ADC	Arctic Data Committee
ADIwg	Alaska Data Integration Working Group
AMD	Antarctic Master Directory
ARDC	Australian Research Data Commons
CCADI	Canadian Consortium for Arctic Data Interoperability
DOI	Digital Object Identifiers
EBV	Essential Biodiversity Variables
ECV	Essential Climate Variables
ESIP	Earth Science Information Partners
EOV	Essential Ocean Variables

[1] For more information on Glossary and terminology used, please reference the “Research Data Management Terminology: Scope of the Terminology” page found here, https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/rdm-terminology/

FAIR	Findable, Accessible, Interoperable, Re-usable
GCMD	Global Change Master Directory
HTML	HyperText Markup Language
IDN	International Directory Node
IRI	Internationalized Resource Identifier
ISO	International Organization for Standardization
JSON-LD	JavaScript Object Notation for Linked Data
NASA	National Aeronautics and Space Administration
NADC	National Antarctic Data Centres
NSF	National Science Foundation
POLDER	Polar Data Discovery Enhancement Research
RDA	Research Data Alliance
RDM	Research Data Management
RIF-CS	Registry Interchange Format - Collections and Services
OBS	Ocean BEst Practices
OIH	Ocean Info Hub
PFS	Polder Federated Search
PID	Persistent Identifier
POAwg	Polar Observing Assets Working Group
SCADM	Standing Committee for Antarctic Data Management
SDO	Schema.org
SOOS	Southern Ocean Observing System
SOSO	Science-On-Schema.org
UNDOS	United Nations Decade of Ocean Science for Sustainable Development
URI	Uniform Resource Identifier

URL	Uniform Resource Locator
USGS	United States Geological Survey
W3C	World Wide Web Consortium
XML	Extensible Markup Language

9.2 References

Bacon, S. (2017). Temporal coverage. Australian Research Data Commons, 3 Aug 2017. Retrieved from https://documentation.ardc.edu.au/display/DOC/Temporal+coverage. (Last accessed 12 Sept 2022).

Carriero, V., Gangemi, A., Nuzzolese, A., & Presutti V. (2021). Chapter 10. An Ontology Design Pattern for Representing Recurrent Situations. Volume 51: Advances in Pattern-Based Ontology Engineering, p 166 - 182. https://www.doi.org/10.3233/SSW210013

Cox. S. (2016). Time Ontology Extended for Non-Gregorian Calendar Applications. Semantic Web 7(2), pp 201–209. https://doi.org/10.3233/SW-150187

Edwards, W. (2022). Australia And Oceania: Human Geography. National Geographic Resource Library. Accessed from https://education.nationalgeographic.org/resource/oceania-human-geography). [Last accessed 24 Jun 2022].

ESIP Data Preservation and Stewardship Committee (2019): Data Citation Guidelines for Earth Science Data, Version 2. ESIP. Online resource. https://doi.org/10.6084/m9.figshare.8441816.v1

Jakobsen, B. H. (2000). Topografisk Atlas Grønland. C.A. Reitzels Forlag.

Payne, K., and Verhey, C. (2022) 'Schema.org for research data managers: a primer', Int. J. Big Data Management, https://doi.org/10.1504/IJBDM.2022.10048569.

Rauber, A., Gößwein, B., Zwölf, C. M., Schubert, C., Wörister, F., Duncan, J., Flicker, K., Zettsu, K., Meixner, K., McIntosh, L. D., Jenkyns, R., Pröll, S., Miksa, T., & Parsons, M. A. (2021). Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data. Harvard Data Science Review, 3(4). https://doi.org/10.1162/99608f92.be565013.

Research Data Management Terminology Working Group. (2021). Scope of the terminology, CODATA. Retrieved from https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/rdm-terminology/. [Last accessed 2 Dec 2022].

Shepherd, A., Jones, M., Richard, S., Jarboe, N., Vieglais, D., Fils, D., Duerr, R., Verhey, C., Minch, M., Mecum, B., & Bentley, N. (2022). Science-on-Schema.org v1.3.0 (1.3.0). Zenodo. https://doi.org/10.5281/zenodo.6502539

Souza, L. & Gomes Vaz, Maria Salete & Sunyé, Marcos. (2014). Domain ontology for time series provenance. ICEIS 2014 - Proceedings of the 16th International Conference on Enterprise Information Systems. 2. 217-224.