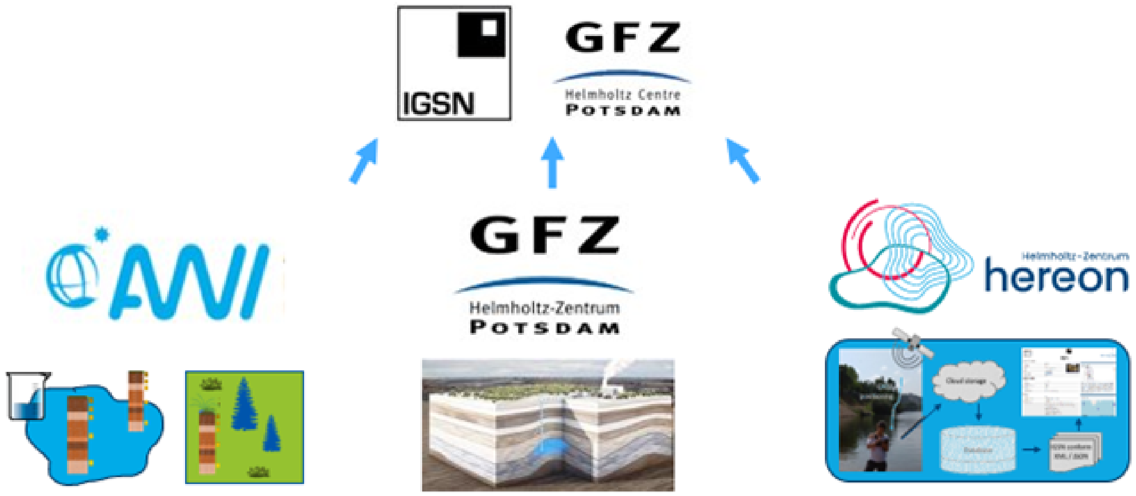
FAIR Workflows to establish IGSN for Samples in the Helmholtz Association

D1 - List of identified linked open data vocabularies to be included in IGSN metadata

(Hereon, GFZ)



Authors

Linda Baldewein (Hereon)

Alexander Brauser (GFZ)

Kirsten Elger (GFZ)

Birgit Heim (AWI)

Mareike Wieczorek (AWI)

Publication Date: June 30th, 2022

Contents

1. Introduction **3**
1.1. Purpose of this document 3
1.2. The role of samples and IGSN for Open Science 3
1.3. IGSN Metadata 4

2. Vocabulary search and integration strategy **6**
2.1. Aspects for the inclusion of controlled vocabularies 6
2.2. Selection of controlled vocabularies 7

3. Recommendations **7**
3.1. Sampling location and date 7
3.2. Sampling instrument / method description 8
3.3. Sample registration by sample type 9
3.4. Sample material 9

4. Controlled vocabulary resources **9**

5. Discussion **12**
5.1. Use case Hereon 12
5.2. Use case AWI expedition [RU-Land_2021_Yakutia expedition] 12
5.3. Use case Ketzin 13
5.4. Identified gaps in the use of controlled vocabularies 14

6. Conclusion and Outlook **14**

7. References **15**

8. Appendix **16**
8.1. List of recommended vocabularies 16
8.2. List of variables and vocabularies for the AWI use case 17
8.3. List of acronyms 24

1. Introduction

1.1. Purpose of this document

This document contains deliverable D1 *List of identified linked open data vocabularies to be included in IGSN metadata* of the project **FAIR W**orkflows to establish **I**GSN for **S**amples in the **H**elmholtz Association (FAIR WISH) funded by the Helmholtz Metadata Collaboration (HMC). Deliverable D1 is part of work package 2 *Identification and integration of linked open data vocabularies to be included in disciplinary metadata*.

Linked open data vocabularies allow for a concise and unambiguous language to describe samples in machine-actionable and interoperable form. They also ensure the possibility of interlinking and comparing samples for which similar or the same terms are used. Here we present the result of our search for open vocabularies, which seem suitable for the FAIR WISH project.

1.2. The role of samples and IGSN for Open Science

While the Berlin Declaration from 2003 was the starting point for Open Access to scholarly publications, Open Science reaches far beyond and represents collaborative, transparent and accessible research that includes all kinds of research results: scholarly literature, research data, software, samples, instruments, etc. In the FAIR WISH project, we focus on samples, as they play a crucial role in the data life cycle. Samples record unique events in history and are often not reproducible. At the same time, samples are essential for reproducing research results and deriving new results with new methodology. Consequently, the inclusion of sample metadata in the research results and digital data curation processes is an important step to provide the full provenance of research results.

The International Generic Sample Number (IGSN, www.igsn.org) is a globally unique and persistent identifier (PID) for physical samples with discovery functionality in the internet. IGSNs enable to directly link data and publications with samples they originate from. IGSN is governed by an international non-profit organisation (IGSN e.V.), which operates the central registration system based on the Handle.Net system. IGSNs resolve via a persistent link to IGSN landing pages with a digital sample description, managed by federated IGSN allocating agents (e.g. https://igsn.org/ICDP5054EHW1001).

GFZ is a founding member of IGSN e.V, and has been an active IGSN allocating agent for samples of scientific drilling projects in the framework of the International Continental

Scientific Drilling Program (ICDP) since 2015 (Conze et al., 2017). The IGSN metadata schema of GFZ was initially aligned with that already in use by the System for Earth Sample Registration (SESAR, www.geosamples.org) and was only extended for specific cases, like, e.g., drilling methods. For later projects, some new metadata elements were added when required. The aim for this strategy was to be as harmonised as possible across IGSN allocating agents and facilitate search options in the planned general IGSN catalogue. The IGSN metadata schema is described in more detail in section 1.3.

Within the FAIR WISH project, (1) standardised and discipline specific IGSN metadata schemes for different sample types within the research field Earth and Environment (EaE) and (2) workflows to generate machine-readable IGSN metadata from different states of digitisation will be developed. Deliverable D1 is the starting point of the first goal of the project.

In this deliverable, specific linked open data vocabularies are identified, which can be used to fill specific fields within the IGSN metadata schema that are either part of the description metadata schema or specific to the allocating agent GFZ. Using controlled vocabularies is an important step for harmonised metadata and crucial for avoiding typographic errors and differences due to different spelling of, e.g. country names in different languages. Some linked-data vocabularies are already in use for data management and publications at GFZ Data Services and Hereon. They provide the basis for our recommendations. Furthermore, these and further vocabularies are evaluated in this document for application within the IGSN metadata.

1.3. IGSN Metadata

The IGSN Metadata Schema is modular: The mandatory registration schema is complemented by the IGSN Description Schema and possibly additional extensions by allocating agents (Klump et al., 2021). This modular approach is delineated in Figure 1: The registration schema contains only four mandatory elements that are common to all IGSNs, the identifier, a registrant, related identifiers and a log / timestamp. The IGSN Description Schema is not mandatory across IGSN e.V. allocating agents, but contains many relevant metadata fields, such as the name of the sample, the sample collector and a location. GFZ uses the IGSN Description Schema (inner blue sphere) and builds upon it with further GFZ specific elements (outer blue sphere). The GFZ specific elements currently contain 82 metadata fields that can be optionally filled. Sample-type-specific metadata, which will be proposed within the FAIR WISH project, are added on top of the GFZ specific metadata (green, grey, orange and yellow spheres).

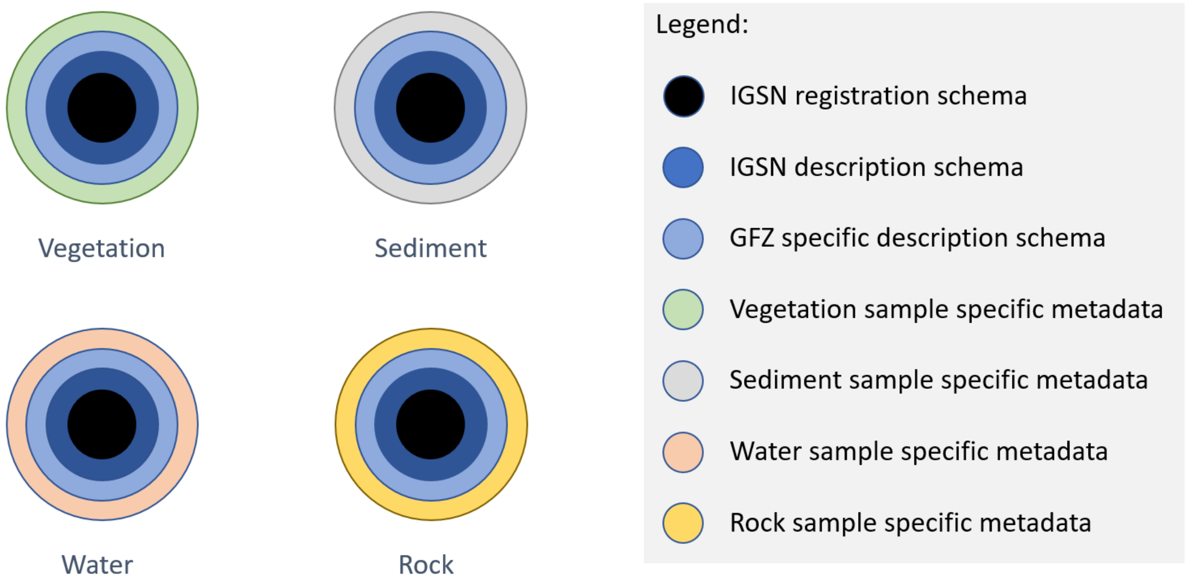

Vegetation
Sediment
Water
Rock
Legend:
IGSN registration schema
IGSN description schema
GFZ specific description schema
Vegetation sample specific metadata
Sediment sample specific metadata
Water sample specific metadata
Rock sample specific metadata

***Figure 1** IGSN description metadata schema, including proposed sample-type-specific metadata of the FAIR WISH project*

The upcoming merge of IGSN and DataCite will have an impact on the project. From summer 2022 on, IGSN handles will successively be registered as DataCite IGSN DOIs. This has several implications: (1) every DataCite member may register IGSNs using their own, or, ideally, a new namespace linked to their DataCite membership; (2) The IGSN DataCite Partnership Steering Group (PSG) is currently developing recommendations for namespace models and (3) for a mapping of the IGSN metadata schema to the DataCite schema. This mapping first focused on the mandatory fields in DataCite, but will further recommend fields that are “recommended for discovery” by DataCite. There is a general agreement to recommend using the DataCite Metadata Schema as complete as possible to support discoverability of sample descriptions beyond the institutional catalogues. Once the namespace and metadata discussions are completed, the transition will start. Due to this transition, the cost model of IGSN will also change and the registration of IGSN DOIs will follow the DataCite fee model (https://datacite.org/feemodel.html). As a member of the IGSN DataCite Partnership Steering Group, Kirsten Elger actively participates in these discussions, especially in the metadata mapping group and directly bridges the FAIR WISH project activities with the international development of IGSN.

2. Vocabulary search and integration strategy

To be compliant with the FAIR principles, we focus on linked data vocabularies that are presented in SKOS-compliant RDF format. We further prioritise those vocabularies that are already used by larger communities or our research centres and to use existing vocabularies before creating new ones. The rapid progress in research data management internationally, leads to a constantly increasing number of linked-data vocabulary registration services. Consequently, we gain an up-to-date overview on existing vocabulary services and identify whether the vocabularies they offer would be suitable for sample descriptions.

2.1. Aspects for the inclusion of controlled vocabularies

A large variety of vocabulary registration servers and vocabularies have been identified. In section 4 we provide a selection of suitable vocabularies for sample types relevant for FAIR WISH that fulfil the following criteria:

- RDF/SKOS format (interoperable, machine actionable)
- Active utilisation of (or need for) vocabulary by scientific community
- Revision cycle of the controlled vocabulary (annually / as needed)
- Long-term support / maintenance / governance
- Community acceptance / usability

The Resource Description Framework (RDF) is the semantic web standard for data interchange and the format of linked data vocabularies. If a linked data vocabulary follows the SKOS recommendations/guidelines, each term contains a link (URI) to its definition. This link is incorporated into the machine-actionable metadata (e.g. DataCite metadata).

Controlled vocabularies always represent a higher level of abstraction and can never yield the full richness of field observations. Consequently, we pursue the strategy to use controlled vocabularies and add fields for additional free text description whenever required. These are often organised in hierarchical form. An interesting result of the investigation of several existing vocabularies revealed that while we sometimes comply with the definition of a specific term in the vocabulary, we (partly) disagreed with the hierarchy leading to this term. This observation and the development of a strategy to cope with it will be further discussed in the project.

2.2. Selection of controlled vocabularies

The first selection of controlled vocabularies for EaE marine and terrestrial research fields that fulfil the criteria of section 2.1 are reviewed on keyword coverage and applicability for specific research fields and sample types. We discussed the tangible subselection of controlled vocabulary or part of it. The resulting list of recommended vocabularies can be found in section 8.1.

3. Recommendations

As described in section 1.3, the IGSN metadata contains a large number of optional metadata fields. In this deliverable, we focus on those fields, for which the usage of a controlled vocabulary leads to an unambiguous language and improves the standardisation of the IGSN metadata.

3.1. Sampling location and date

To be able to register samples which e.g. are subject to a confidentiality agreement, coordinates are not mandatory for sample description. Irrespective of this, it is recommended that the following information are provided:

- Geographic coordinates, elevation and locality
- Landscape description including the landform or physiographic feature (e.g. mountain, hill, plateau, slope, valley, plain…) and the surrounding biome (e.g. Forest, Tundra, Grassland…)
- And a finer definition of the water body (e.g. Glacial Lake, Tectonic Lake, Thermokarst Lake, River, Ocean, Stream…) if samples are of type water, sediment or aquatic flora and fauna.
- Date of sampling

We generally agreed to describe the location of the sample by geographical coordinates in decimal degrees in the World Geodetic System 1984 (WGS84) datum as well as the locality. The IGSN metadata schema further allows additional coordinates in different projections, like, e.g., UTM. The locality is described in the field “locality", which is defined as the name of the specific place where the sample was collected. We recommend using the controlled vocabulary provided by GeoNames (https://www.geonames.org/) whenever possible. Specific details not listed in GeoNames can be given in the a free text field, which still needs to be identified.

A physiographic feature describes the physical feature that the sample was collected from, i.e. the landform. It is described in the fields “primary_location_type” and "primary_location_name". For many samples, this field is also needed to describe the biome setting. We recommend the use of EnvO for describing the primary_location_type in this case. Some samples, e.g. marine water samples, do not necessarily have a physiographic feature. For more generic physiographic features, SESAR provides a broad yet expressive vocabulary for the landform setting of a sample location (https://www.geosamples.org/vocabularies), which we alternatively recommend to use, even though the lists are not provided in RDF format at the moment.

The field with a date for the sample generation should also be recommended whenever this is existing: it can be filled out fine-grained with high temporal resolution (seconds, minutes, hours, day), e.g. according to ISO 8601 (YYYY-MM-DD) or indicating the year of sampling only.

Table 3.1.1 *Overview of variables to describe sampling location, date and format in which they should be provided*

Variable	**Format**
Coordinates of sample location	Geographic coordinates in decimal degrees, WGS 84
Country (if applicable) 'locality_description'	https://www.geonames.org/
Administrative Division (if applicable) 'locality_description'	E.g. https://www.geonames.org/DE/administrative-division-germany.html
landscape/ physiographic feature (if applicable) 'primary_location_type'	E.g. https://www.geosamples.org/vocabularies/physiographic-feature or https://obofoundry.org/ontology/envo.html
Sampling Date	ISO 8601 (YYYY-MM-DD)
Elevation (if applicable)	Altitude above sea level (a.s.l.)

3.2. Sampling instrument / method description

The instrument or method used to collect the sample is described in the field "collection_method". We recommend the usage of the following controlled vocabularies:

- Marine Samples, but also applicable for other sample types:
    - SeaDataNet device categories (Terms used to classify groups of sensors, instruments, sources of algorithmically computed data (numerical models) or samplers (collectors of water, suspended particulate matter (SPM), sediment, rock, air or biota samples).) -> http://vocab.nerc.ac.uk/collection/L05/current/

    - SeaVoX Device Catalogue (Terms for distinct sampling or measuring devices that may be identified in the real world in terms of manufacturer and model number.) -> http://vocab.nerc.ac.uk/collection/L22/current/
- Sampling devices used in the field (piston corer, gravity corer, sediment trap, manual, etc.) can also be described by the SESAR Collection Method -> https://www.geosamples.org/vocabularies/collection-method. Missing terms like 'tree corer' or 'bottle' may be extended to the list.

3.3. Sample registration by sample type

The sample type is described in the field "sample_type". SESAR has a controlled list that is already in use. The list contains generic terms, such as "Individual Sample", "Site" and "Other", but also highly specific terms like "Squeeze Cake" and "Toothpick". For the use cases within FAIR WISH, many samples can currently be only described as "Individual Sample" or "Other". We thus recommend to SESAR to expand the list by the following terms (the list might be extended depending on user demand):

- Rock and mineral samples
- Sediment, particle and suspended particulate matter samples
- Soil samples
- Water samples (including fresh and marine waters, porewaters, precipitation water), snow samples, ice samples
- Vegetation samples (including plant samples of whole plants and plant organs, and information on plant species)
- Air samples (pollen, spores, aerosols, dust, gases)
- Just as "Hole" is registered as parent for cores, we furthermore recommend to register "vegetation plot", "lake" and further types of sampling features with multiple samples as parent sample, to have all samples from one study site combined as siblings.

3.4. Sample material

The material of the sample is described in the field "material". We recommend using the controlled list of SESAR.

4. Controlled vocabulary resources

In section 2.1 we described the criteria for controlled vocabularies that we evaluated for usage within the IGSN metadata. The vocabulary websites can be divided into two groups:

- Comprehensive vocabulary list servers (Table 4.1)

- Single vocabularies (Table 4.2)

Each of these groups can be either thematically specific or broad

Table 4.1 *Overview of the most relevant list servers.*

Name	**URL**	**Specific / broad**	**Topic**	**Comment**
ARDC Australian Research Data Commons	https://vocabs.ardc.edu.au/	broad	Large collection of different vocabularies. New vocabularies are developed, discussed and expanded	
BGS Vocabularies	https://www.bgs.ac.uk/information-hub/dictionaries/vocabularies/	specific	Lithology, lithostratigraphy, names of mapped rock units, names of maps	
GBIF - Global Biodiversity Information Facility	https://www.gbif.org/	specific	Biodiversity	Used by USGS
NERC Vocabulary Server (NVS) by BODC	http://vocab.nerc.ac.uk/collection/	specific	Oceanography	Used by Hereon, PANGAEA
OLS Ontology Lookup Service	https://www.ebi.ac.uk/ols/index	specific	Biomedical ontologies	
Open Biological and Biomedical Ontology (OBO) Foundry	https://obofoundry.org/	specific	Biological and Biomedical ontologies	
ODM2 Controlled Vocabularies	http://vocabulary.odm2.org/	specific	Observations and measurements in EaE	

From these list servers, we further evaluated the following vocabulary resources.

Table 4.2 *provides an overview of the most relevant vocabularies*

Name	URL	Found on list server	Specific/ broad	Topic	Comment
BCO Biological Collections Ontology	https://obofoundry.org/ontology/bco.html	Open Biological and Biomedical Ontology (OBO) Foundry, OLS Ontology Lookup Service	specific	Biodiversity	Used by GFBIO
EnvO - The Environment Ontology	https://obofoundry.org/ontology/envo.html	Open Biological and Biomedical Ontology (OBO) Foundry, OLS Ontology Lookup Service	specific	Environments	Used by GFBIO, PANGAEA
CGI Simple Lithology	http://resource.geosciml.org/classifier/cgi/lithology	geoSciML	specific	lithology	Used by GFZ
GeoEra	https://github.com/schmar00/project-vocabularies	Project website	specific	Geo-energy, groundwater, raw materials, information platform	
ITIS - Integrated Taxonomic Information System	https://www.itis.gov/	GBIF - Global Biodiversity Information Facility	specific	Taxonomy	OLS, Used by US institutions, GFBIO, PANGAEA
ODM2 - Observations Data Model 2	http://vocabulary.odm2.org/	ODM2 website	specific	Earth observations	
The Plant Ontology (PO)	https://www.ebi.ac.uk/ols/ontologies/po	Open Biological and Biomedical Ontology (OBO) Foundry,OLS Ontology Lookup Service	specific	Plant anatomy, morphology	
SeaDataNet device categories	http://vocab.nerc.ac.uk/collection/L05/current/	NERC Vocabulary Server (NVS) by BODC	specific	Oceanographic instruments	Used by Hereon

Name	URL	Found on list server	Specific/ broad	Topic	Comment
SeaVoX Device Catalogue	http://vocab.nerc.ac.uk/collection/L22/current/	NERC Vocabulary Server (NVS) by BODC	specific	Very specific oceanographic instruments	Used by Hereon
SESAR	https://www.geosamples.org/vocabularies	SESAR website	specific	IGSN specific	No RDF. plain lists

The vocabulary used by SESAR is partly aligned with ODM2, but both vocabulary lists contain terms that are not included in the other one. To date, SESAR is not providing their vocabularies in machine-actionable form. However, the relevance of these vocabularies is high, because they are frequently used by IGSN allocating agents across the globe.

5. Discussion

5.1. Use case Hereon

The biogeochemical campaign database at Hereon contains thousands of samples from dozens of campaigns, most of which were ship-based. The data is stored in a relational database and publicly accessible. For standardisation purposes and for linking the database with other repositories, controlled vocabularies have already been implemented to describe the samples within the database. The sample types are described by the BODC parameter semantic model sphere names (http://vocab.nerc.ac.uk/collection/S21/current/), which can be easily mapped to the SESAR controlled list. The collection method is mapped to the SeaDataNet device categories (http://vocab.nerc.ac.uk/collection/L05/current/) and the SeaVoX Device Catalogue (http://vocab.nerc.ac.uk/collection/L22/current/), which is in accordance to the recommendations of the World Data Center PANGAEA and the German Marine Research Alliance.

5.2. Use case AWI expedition [RU-Land_2021_Yakutia expedition]

AWI arctic land expeditions collect a variety of samples and subsamples, beginning with a ‘sampling container type’ that are e.g., vegetation plots, cliff sections or lakes at the highest sampling level down to e.g. needles and sediment core sections. At the moment, controlled vocabularies are not yet used for the samples of the most recent expedition in summer 2021 [RU-Land_2021_Yakutia], but there is a high potential to standardise parts of the sample descriptions. We identified a first set variables (see Appendix 8.2) to which controlled vocabularies can be applied and existing ontologies which can be implemented, mainly the environment ontology (EnvO), the plant ontology (PO) and the integrated taxonomic information system (ITIS), which are also used by GFBIO and/or PANGAEA. However, the vocabulary needed to describe the vegetation samples has a large overlap in terms between EnvO and PO, with most of the terms being defined in both ontologies and only a few, e.g. “Strobilus bud” (i.e. cones) being defined only in PO. We thus plan to implement only EnvO and see if it is sufficient for our purpose. Being a community project, it is furthermore possible to make suggestions for additional terms. We did that already for the term ‘subpolar deciduous needleleaf forests’, which are not defined in EnvO and might do so for further terms. It is, however, up to the developers if suggestions are implemented.

5.3. Use case Ketzin

At the CO2 pilot site Ketzin, more than 200 m core material were collected from 5 boreholes in the form of subsampled elements in order to study the suitability of the geological formations for a potential CO2 storage. The 3-m long core barrels were cut into 1-m sections on the drill site and stored in wooden boxes and intensively analysed later on. The core samples are linked to a wide and unique range of measurements such as permeability measurements, baseline seismicity, densitometry, and x-ray diffraction (XRD) for geochemical analyses.

Cores from the reservoir formations and the cap rock formation were taken by using wireline coring, a method that is not yet mentioned in SESAR “Collection Method” or CGI “Borehole Drilling Method” vocabularies. We therefore recommend to include the term ‘wireline coring’ to the respective vocabularies.

Recommended definition of ‘wireline coring’ (by Ben Norden):
“In wireline coring, the drilled core enters a core barrel which can be removed from the drill string without dismounting of the drill string by raising a wire. As the drill string needs not to be dismounted, coring using the wireline technique can provide rig time savings of about 25% or more compared to conventional coring allowing faster penetrating rates.”

For rock sample descriptions, we recommend the use of CGI Simple Lithology as this vocabulary is already in use by GFZ Data Services.

5.4. Identified gaps in the use of controlled vocabularies

Controlled vocabularies are inherently more generic than what is theoretically possible in free text descriptions. Lists of hundreds of thousands of terms within controlled vocabularies are possible, but hinder their application, as finding the correct term becomes increasingly more difficult the longer the vocabulary. We acknowledge that not all specific scenarios can be described using the recommended vocabularies. We suggest mapping each detailed term to the closest fitting generic term in the controlled vocabularies. In cases where this is not possible, we recommend contacting the owners of the controlled vocabularies and requesting an addition to the vocabulary. For our IGSN metadata, we consider adding an additional field for more specific information.

6. Conclusion and Outlook

In this deliverable for the FAIR WISH project, we identified a series of vocabularies that will be applied for the usage within the IGSN metadata. The list is dynamic, and will be growing as more samples are being registered. The vocabularies themselves may lack in some cases specific terms, which then can be registered with the regulating body of the vocabulary.

As expected, not all cases can be described in detail or explicitly in the selected vocabularies. We need to find a balance between a variety of different specific and fewer general vocabularies, as the more vocabularies used, the more complex and less user-friendly the system becomes. For specific descriptions, the free text fields within the IGSN metadata schema are used to overcome the limitations of the vocabularies.

To connect samples and measurements, it would be useful to not only reference the sample IGSN in data publications or scholarly literature, but also to link the measurements and analyses made within the sample’s metadata. At the moment, this feature is not implemented in the IGSN schema and it has to be discussed within the team, how a scalable approach could look like, especially if IGSNs are registered long before any analyses are conducted. From the provenance perspective, for samples it would be helpful to have references to the data and scholarly literature already in the sample metadata..

During the planned user workshops, we will discuss the linking of samples and measurements with the participants and additionally address the need for more terms to describe a sample, when including further sample types and/or locations.

7. References

Conze, R., Lorenz, H., Ulbricht, D., Elger, K., Gorgas, T. (2017): Utilizing the International Geo Sample Number Concept in Continental Scientific Drilling During ICDP Expedition COSC-1. - Data Science Journal, 16, 1, 1-8. https://doi.org/10.5334/dsj-2017-002

Klump, J., Lehnert, K., Ulbricht, D., Devaraju, A., Elger, K., Fleischer, D., Ramdeen, S., Wyborn, L. (2021): Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. - Data Science Journal, 20, 1, 1-16. https://doi.org/10.5334/dsj-2021-033

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In Scientific Data (Vol. 3, Issue 1), https://doi.org/10.1038/sdata.2016.18

8. Appendix

8.1. List of recommended vocabularies

Table 8.1.1 *List of recommended vocabularies for different sample sources and IGSN fields.*

Sample source	IGSN field	Vocabulary	Vocabulary URL	Comment
Marine	collection_method	SeaDataNet device categories	http://vocab.nerc.ac.uk/collection/L05/current/	Generic terms
Marine	collection_method	SeaVoX Device Catalogue	http://vocab.nerc.ac.uk/collection/L22/current/	Limited, specific devices
All	collection_method	SESAR	https://www.geosamples.org/vocabularies/collection-method	Generic terms
All	sample_type	SESAR	https://www.geosamples.org/vocabularies/sample-type-object	Missing sample types, will be expanded by FAIR WISH
All	locality	GeoNames	https://www.geonames.org/	As specific as possible
All	material	SESAR	https://www.geosamples.org/vocabularies/material	The list is already extensive.
All	primary_location_type	SESAR/EnvO	https://www.geosamples.org/vocabularies/physiographic-feature	Physiographic feature
All	primary_location_type	EnvO	https://obofoundry.org/ontology/EnvO.html	Biomes; Water Body

8.2. List of variables and vocabularies for the AWI use case

These tables provide an overview of vocabulary that we see is beneficial for samples (water, vegetation, soil) similar to the AWI use case of arctic land expeditions and repositories where we found the vocabulary (sometimes with a different name). This list is not meant to be comprehensive, it might as well be that a vocabulary word is found in sources not mentioned here and the list of variables will be extended when including further sample types and use cases.

Table 8.2.1 *Variable names needed to describe sampling locations (here: landform) of AWI arctic land expeditions, their differing name in the EnvO-vocabulary (if applicable), the link to the vocabulary definition in EnvO, information on further ontologies listing the variable and comments.*

Location Variable	Different name in EnvO	EnvO Link	Also available in	Comment
Landform				Possible IGSN field: *primary_location_type*
Mountain		http://purl.obolibrary.org/obo/ENVO_00000081	SESAR	
Hill		http://purl.obolibrary.org/obo/ENVO_00000083		
Plateau		http://purl.obolibrary.org/obo/ENVO_00000182	SESAR	
Plain		http://purl.obolibrary.org/obo/ENVO_00000086	SESAR	
Slope		http://purl.obolibrary.org/obo/ENVO_00002000		
Valley		http://purl.obolibrary.org/obo/ENVO_00000100	SESAR	
etc.				

Table 8.2.2 *Variable names needed to describe sampling locations (here: Biome/Vegetation type) of AWI arctic land expeditions, their differing name in the EnvO-vocabulary (if applicable), the link to the vocabulary definition in EnvO, information on further ontologies listing the variable and comments.*

Location Variable	**Different name in EnvO**	**EnvO Link**	**Also available in**	**Comment**
Biome/Vegetation type				Possible IGSN field: *primary_location_name*
Boreal Forest	Subpolar coniferous forest biome	http://purl.obolibrary.org/obo/ENVO_01000250		Term „Subpolar deciduous needleleafed forest" suggested to ENVO
Forest tundra	Tree-line ecotone	http://purl.obolibrary.org/obo/ENVO_01000953		
Shrub Tundra	Area of dwarf scrub	http://purl.obolibrary.org/obo/ENVO_01000861		Shrub tundra would be most appropriate to define the climatic conditions of the "area of dwarf scrub" in our use case, but is not available in EnvO. We will suggest the term and until then recommend to use "shrub tundra" as free text keyword if applicable.
Tundra	Area of polar tundra	http://purl.obolibrary.org/obo/ENVO_3400002		"Polar tundra ecosystem" http://purl.obolibrary.org/obo/ENVO_01001625 as alternative name in EnvO
Polar desert biome		http://purl.obolibrary.org/obo/ENVO_01000186		
Grassland	Grassland area;	http://purl.obolibrary.org/obo/ENVO_00000106		„Grassland ecosystem" http://purl.obolibrary.org/obo/ENVO_01001206 or „Grassland biome" http://purl.obolibrary.org/obo/ENVO_01000177 as alternative names in EnvO
etc.				

Table 8.2.3 *Variable names needed to describe sampling locations (here: water body) of AWI arctic land expeditions, their differing name in the EnvO-vocabulary (if applicable), the link to the vocabulary definition in EnvO, information on further ontologies listing the variable and comments.*

Location Variable	**Different name in EnvO**	**EnvO Link**	**Also available in**	**Comment**
Water body				Possible IGSN field: *primary_location_type*
Glacial lake		http://purl.obolibrary.org/obo/ENVO_00000488		
Fluvial plain lake		NA		
Thermokarst lake		http://purl.obolibrary.org/obo/ENVO_03000082		
Meterorite lake		http://purl.obolibrary.org/obo/ENVO_01001065		
Tectonic lake		http://purl.obolibrary.org/obo/ENVO_01001092		
Lagoon lake	Lagoon	http://purl.obolibrary.org/obo/ENVO_00000038		
Pond		http://purl.obolibrary.org/obo/ENVO_00000033		
Swamp lake	Swamp ecosystem	http://purl.obolibrary.org/obo/ENVO_00000233		
River		http://purl.obolibrary.org/obo/ENVO_00000022		
Stream		http://purl.obolibrary.org/obo/ENVO_00000023	SESAR/ODM2	
Lake		http://purl.obolibrary.org/obo/ENVO_00000020		
etc.				

Table 8.2.4 *Variable names needed to describe sampling locations (here: soil horizon and water sampling depth) of AWI arctic land expeditions, their differing name in the EnvO-vocabulary (if applicable), the link to the vocabulary definition in EnvO, information on further ontologies listing the variable and comments.*

Sampling Variable	Different name in EnvO	EnvO Link	Also available in	Comment
*Horizon**				
Litter	Litter layer	http://purl.obolibrary.org/obo/ENVO_01000338		
Organic	Organic horizon	http://purl.obolibrary.org/obo/ENVO_01000338		
Mineral	Mineral horizon	http://purl.obolibrary.org/obo/ENVO_03600011		
Water sampling depth				
Surface water		http://purl.obolibrary.org/obo/ENVO_00002042		
Bottom water		NA		

*The vocabulary for the soil horizon is the ENVO vocabulary for soil samples, used by the Bioscience communities. In contrast, the Geoscience communities apply different terms for soils which will be also tested in future in close cooperation with the University of Göttingen (Prof. Dr. Elisabeth Dietze).

Table 8.2.5 *Variable names needed to describe sampling type (here: land related samples) of AWI arctic land expeditions, their differing name in the EnvO-vocabulary (if applicable), the link to the vocabulary definition in EnvO, information on further ontologies listing the variable and comments.*

Sampling Variable	**Different name in EnvO**	**EnvO Link**	**Also available in**	**Comment**
Sample type				
Vegetation plot		NA	ODM2 as „Sitetype>Land“	
Soil		http://purl.obolibrary.org/obo/ENVO_00001998	SESAR/ODM2	
Soil Pit		NA		
Soil pit section		NA	ODM2	
Whole plant		http://purl.obolibrary.org/obo/PO_0000003	PO	
Plant organ		http://purl.obolibrary.org/obo/PO_0009008	PO SESAR as **„**Plant Structure“	
• Leaf/Needle	Leaf	http://purl.obolibrary.org/obo/PO_0025034	PO as „Vascular leaf“	
• Cones		NA	PO as „Strobilus bud“	Link to PO: http://purl.obolibrary.org/obo/PO_0025085
• Stem	Shoot axis	http://purl.obolibrary.org/obo/PO_0025029	PO	
• Branch		NA	PO	Link to PO: http://purl.obolibrary.org/obo/PO_0025073
• Flower	collective plant organ structure	http://purl.obolibrary.org/obo/PO_0025007	PO	Link to PO: http://purl.obolibrary.org/obo/PO_0009046
• etc				
Species			ITIS	This field will contain the species’ scientific name
etc.				

Table 8.2.6 *Variable names needed to describe sampling type (here: water related samples) of AWI arctic land expeditions, their differing name in the EnvO-vocabulary (if applicable), the link to the vocabulary definition in EnvO, information on further ontologies listing the variable and comments.*

Sampling Variable	**Different name in EnvO**	**EnvO Link**	**Also available in**	**Comment**
Sample type				
Lake		http://purl.obolibrary.org/obo/ENVO_00000020		
River		http://purl.obolibrary.org/obo/ENVO_00000022		
Water		http://purl.obolibrary.org/obo/CHEBI_15377	SESAR/ODM2 as „Liquid Aqueous“	
Sediment		http://purl.obolibrary.org/obo/ENVO_00002007	ODM2	
Sediment Core		NA	SESAR/ODM2 as „Core“	
• Core Section		NA	SESAR/ODM2	
Ice		http://purl.obolibrary.org/obo/ENVO_01001125	SESAR/ODM2	
Snow		http://purl.obolibrary.org/obo/ENVO_01000406	ODM2	
etc.				

8.3. List of acronyms

ARDC	Australian Research Data Commons
AWI	Alfred Wegener Institute for Polar and Marine Research
CGI	IUGS (International Union for Geological Sciences) Commission for the Management and Application of Geoscience Information
EaE	Earth and Environment
EnvO	Environmental Ontology
FAIR	Guiding Principles for Findable, Accessible, Interoperable and Reusable research data (Wilkinson et al., 2015)
FAIR WISH	FAIR Workflows to establish IGSN for Samples in the Helmholtz Association
GFBIO	German Federation for Biological Data
GFZ	German Research Centre for Geosciences
IGSN	International Generic Sample Number
ITIS	Integrated Taxonomic Information System
ODM2	Observations Data Model 2
PID	Persistent Identifier
PO	Plant Ontology
RDF	Resource Description Format
SESAR	System for Earth Sample Registration
SKOS	Simple Knowledge Organization System
URI	Uniform Resource Identifier