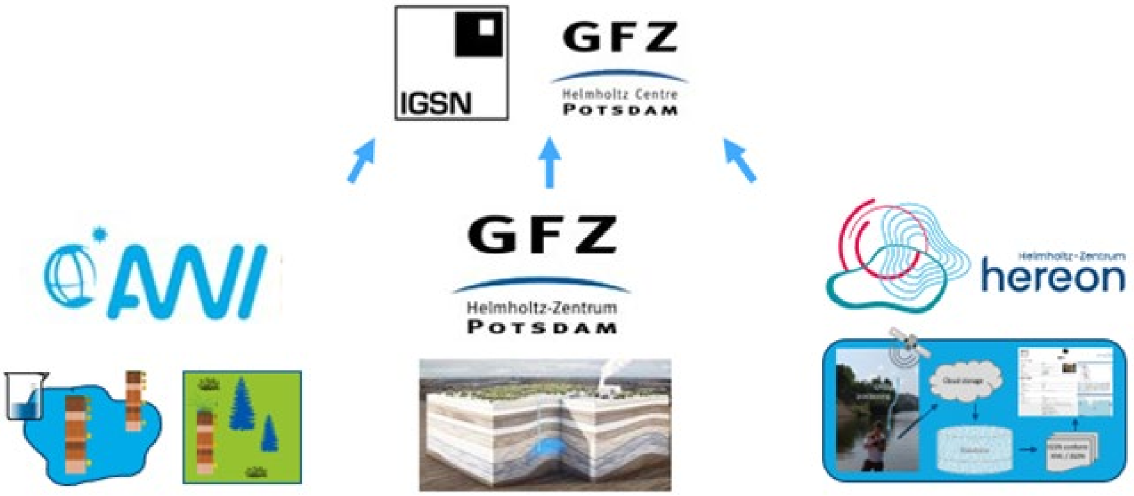
FAIR Workflows to establish IGSN for Samples in the Helmholtz Association

D5 – Mapping of database metadata to machine readable IGSN metadata (Use Case Hereon)

(Hereon, GFZ, AWI)



Authors

Linda Baldewein (Hereon)

Tim Leefmann (Hereon)

Ulrike Kleeberg (Hereon)

Alexander Brauser (GFZ)

Kirsten Elger (GFZ)

Simone Frenzel (GFZ)

Birgit Heim (AWI)

Mareike Wieczorek (AWI)

Publication Date: March 31st, 2023

Contents

1. Introduction

1.1. Purpose of this document

This document contains deliverable D5 *Mapping of database metadata to machine readable IGSN metadata (Use Case Hereon)* 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 D5 is part of work package 6 *IGSN registration workflows for biogeochemical sample databases*.

This deliverable builds upon D2 *Exemplary standardised metadata templates for Geo-Bio Samples (vegetation, water, sediment, rock samples) for all uses cases* (Wieczorek et al., 2022) and the FAIR WISH: Sample description template (Brauser et al., 2023) to develop the mapping of the existing Hereon campaign database EXPEDITIONSDATEN to machine actionable IGSN metadata.

1.2. The FAIR WISH Sample description template

The FAIR WISH Sample description template (Brauser et al., 2023) contains the suggested IGSN metadata elements for sample descriptions of different sample types relevant in the context of the FAIR WISH project and shall serve as the basis for batch uploading sample-metadata to the IGSN server. The modular template has been developed as an all-in-one solution for the varying use cases, with the possibility for users to select only the variables needed for their specific sample type (in addition to the core metadata required for all samples). This template can already be used to describe individual samples and will be further developed for hierarchical sample structures later during the project (i.e., vegetation plot > soil pit > soil horizon > soil sample).

The template will be the source for semi-automated generation of standardised XML files required for the IGSN registration and IGSN landing pages. This template strategy was explicitly followed to meet the practice of researchers who mainly organise their sample descriptions in tables. Furthermore, as often information for 100 and more samples are to be registered, submitting information via bulk uploads is favoured over submitting information for each sample individually via a webform. We developed the template for Excel, as this is widely used in the community. For the semi-automated XML generation, the information provided in the Excel Template will be exported as csv.

1.3. The Hereon expedition database

The Hereon expedition database has been developed since 2016 within the coastMap project (Baldewein et al., 2018). Its aim is to allow querying and downloading of campaign data collected during ship and land-based sampling campaigns at the Hereon Institute of Carbon Cycles and Hereon Institute of Coastal Environmental Chemistry.

The fully normalised relational database has been continuously expanded to fully describe the highly diverse metadata associated with biogeochemical campaign samples (long-tail data according to Heidorn, 2008). It currently consists of 34 tables, storing mostly metadata and the relationship between metadata elements. Eleven of these tables directly or indirectly store the sample metadata needed for IGSN registration. This includes tables storing the sample, information on the campaign, as well as those storing information on associated scientists, projects, and related publications.

For assigning IGSNs to samples, having meaningful sample metadata is a necessary requirement. At Hereon, we started using field apps to collect these metadata in June 2017 (Baldewein, Kleeberg, Möller, 2021). All samples taken before this date do not have any sample metadata associated with them and thus cannot receive IGSNs. Between June 2017 and December 2022, a total of 12041 samples and their associated sampling locations have been recorded using the field app. In December 2022, all these samples were assigned with IGSNs within the FAIR WISH project.

2. Mapping of metadata for Use Case Hereon

Using the guide of the FAIR WISH Sample Description Template (Brauser et al., 2023), the mandatory and optional metadata fields that have corresponding metadata entries in the expedition database were selected. In total, 33 different fields were identified as being suitable for some or all the samples within the Use Case Hereon. A differentiation is necessary to distinguish parent IGSNs from their children, as they are stored differently in the database. In case of the biogeochemical campaign samples, all samples have a parent IGSN of the related feature of interest (sample_type: site), describing the sampling site the sample was taken at.

For the sites described in the parent IGSN, the mapping is done as described in Section 6.1. “Database mapping for sample sites” of this document. The sample mapping is described in Section 6.2. “Database mapping for samples”.

Care was taken not only to assign IGSNs only to samples collected after June 2017, but also to exclude some other measurements described in the database from being assigned with IGSNs, because the sampling took place in-situ and thus no physical sample ever existed.

3. Recommendations

Based on the experiences gained in Section 2 of this document, the following recommendations for generalising mappings in other uses cases can be given:

1. Initially separate all samples based on their hierarchical relationship in parents and children and treat them independently, as they most likely will have varying metadata.
2. Use the ‘Guide’ of the template (second page in the spreadsheet) to find all potentially relevant metadata fields for your samples. It is easier to discard variables at the end, if not needed, rather than adding fields later in the mapping process.
3. Use rather more than less metadata fields to have a rich and comprehensive sample description.
4. Map your metadata, such as the sample type or the material, to the controlled vocabulary lists within the database.
5. Lists, such as several DOIs of articles or data publications related to a single sample, need to be comma separated. Generating lists within databases may be time consuming.
6. Know your database and the relationships within the database, as this will simplify the mapping process.

4. Conclusion

In this deliverable for the FAIR WISH project, we mapped database metadata to machine actionable IGSN metadata and provided recommendations on how to proceed in the mapping process for other use cases. The provided mapping is only exemplary but can be used as a guideline for other databases.

For the planned standard operating procedures, we will discuss the mapping of metadata further and expand on the generalisation.

5. References

Baldewein, L., Kleeberg, U., Lange, M., & Sauer, D. (2018). Coastal data portals to support marine science and management-the coastMap approach. Bollettino di Geofisica, 283.

Baldewein, L., Kleeberg, U., and Möller, L. (2021). Automation of (meta-)data workflows from field to data repository, EGU General Assembly 2021, online, 19–30 Apr 2021, EGU21-2521, https://doi.org/10.5194/egusphere-egu21-2521

Brauser, Alexander, Wieczorek, Mareike, Frenzel, Simone, Heim, Birgit, Baldewein, Linda, Kleeberg, Ulrike, & Elger, Kirsten. (2023). FAIR WISH: Sample description template. Zenodo. https://doi.org/10.5281/zenodo.7520016

Heidorn, B. P.. (2008). Shedding Light on the Dark Data in the Long Tail of Science. In Library Trends (Vol. 57, Issue 2, pp. 280–299). Project Muse. https://doi.org/10.1353/lib.0.0036

Wieczorek, Mareike, Heim, Birgit, Brauser, Alexander, Elger, Kirsten, & Baldewein, Linda. (2022). FAIR WISH D2 - Exemplary standardised metadata templates for Geo-Bio samples (vegetation, water, sediment, rock samples) for all use cases. Zenodo. https://doi.org/10.5281/zenodo.7147532

6. Appendix

6.1. Database mapping for sample sites

IGSN field name	Vari-ability	Database field	Example
igsn	variable	station.IGSN	GFHER63BAE
parent_igsn	constant		
name	variable	station.STATION_NAME	BOHAISEA2018Nov_Stat_Qingjinghuang_Drainage_Canal
sample_other_names	variable	station.STATION_LABEL	Qingjinghuang Drainage Canal
is_private	constant		0
sample_type	constant		Site
material	constant		NotApplicable
description	constant		
collection_start_date	variable	station.station_start_time	2018-11-01 10:24:00.0000000 +08:00
collection_end_date	variable	station.station_end_time	2018-11-01 10:24:00.0000000 +08:00
collection_date_time_zone	variable	Time zone info of station.station_start_time	UTC+08:00
collection_date_precision	constant		time
depth_min	constant		
depth_scale	constant		
collection_method	constant		
collection_method_descr	constant		
cruise_field_prgrm	variable	campaign.campaign_code	CE17013
patform_type	variable	If vessel.VESSEL_NAME starts with RV, then 'Ship', else leave blank	Ship
platform_name	variable	vessel.VESSEL_NAME	RV Celtic Explorer
latitude	variable	station.STATION_LATITUDE	57.8298
longitude	variable	station.STATION_LONGITUDE	7.9977
elevation	variable	station.STATION_BOTTOM_DEPTH	-520
elevation_unit	constant		m below sea level
collector	variable	pi.lastname, pi.firstname	Bento, Célia
givenName	variable	pi.lastname	Bento
familyName	variable	pi.firstname	Célia
affiliation	variable	institute.institute_long_name	Helmholtz-Zentrum Hereon

current_archive	constant		Helmholtz Coastal Data Center (HCDC), Helmholtz-Zentrum Hereon, Max-Planck-Straße 1, 21502 Geesthacht, Germany, please note: The samples described here are usually not stored in long-term archives. For potential exceptions, please contact us.
current_archive_contact	constant		hcdc_support@hereon.de
sampleAccess	constant		
relatedIdentifier	variable	DOI_list.DOI	https://doi.org/10.1594/PANGAEA.940860
relatedIdentifierType	constant		DOI
relationType	constant		References

6.2. Database mapping for samples

IGSN field name	Vari-ability	Database field	Example
igsn	variable	sample.IGSN	GFHERE5598
parent_igsn	variable	station.IGSN	GFHER75864
name	variable	sample.SAMPLE_NAME	p_CE17013_Stat_040_a_0.01
sample_other_names	variable	sample.SAMPLE_LABEL	
is_private	constant		0
sample_type	variable	Either depending on SAMPLING.SAMPLING_NAME, SAMPLING.SAMPLING_STANDARDNAME or PARAMETERSITE.PARAMETERSITE_NAME	CTD
material	variable	PARAMETERSITE.PARAMETERSITE_NAME	Sediment
description	variable	sample.SAMPLE_INFO	
collection_start_date	variable	sample.DATA_DATETIME	2020-03-23 09:26:00.0000000 +00:00
collection_end_date	variable	sample.DATA_DATETIME	2020-03-23 09:26:00.0000000 +00:00
collection_date_time_zone	variable	Time zone info of sample.DATA_DATETIME	UTC+00:00
collection_date_precision	constant		time
depth_min	variable	Depending on PARAMETERSITE.PARAMETERSITE_NAME, the columns sample.WATER_DEPTH, sample.SEDIMENT_DEPTH or	3

		sample.ATMOSPHERE_HEIGHT are used	
depth_scale	variable	'm below water level' used for water and suspended particulate matter samples, 'cm below ground' used for porewater and sediment samples, 'm above water level' used for atmosphere and snow samples	m below water level
collection_method	variable	sampling_standardname. sampling_standardname	unconsolidated sediment corers
collection_method_descr	variable	sampling.sampling_name	Multicorer
cruise_field_prgrm	variable	campaign.campaign_code	CE17013
patform_type	variable	If vessel.VESSEL_NAME starts with RV, then 'Ship', else leave blank	Ship
platform_name	variable	vessel.VESSEL_NAME	RV Celtic Explorer
latitude	variable	sample.LATITUDE	57.8298
longitude	variable	sample.LONGITUDE	7.9977
elevation	variable	station.STATION_BOTTOM_DEPTH	-520
elevation_unit	constant		m below sea level
collector	variable	pi.lastname, pi.firstname	Bento, Célia
givenName	variable	pi.lastname	Bento
familyName	variable	pi.firstname	Célia
affiliation	variable	institute.institute_long_name	Helmholtz-Zentrum Hereon
current_archive	constant		Helmholtz Coastal Data Center (HCDC), Helmholtz-Zentrum Hereon, Max-Planck-Straße 1, 21502 Geesthacht, Germany, please note: The samples described here are usually not stored in long-term archives. For potential exceptions, please contact us.
current_archive_contact	constant		hcdc_support@hereon.de
sampleAccess	constant		destroyed
relatedIdentifier	variable	DOI_list.DOI	https://doi.org/10.1594/PANGAEA.940860
relatedIdentifierType	constant		DOI
relationType	constant		References

6.3. List of acronyms

AWI	Alfred Wegener Institute for Polar and Marine Research
FAIR	Guiding Principles for Findable, Accessible, Interoperable and Reusable research data (Wilkinson et al., 2015, https://doi.org/10.1038/sdata.2016.18)
FAIR WISH	FAIR Workflows to establish IGSN for Samples in the Helmholtz Association (https://zenodo.org/communities/fair_wish/)
GFZ	German Research Centre for Geosciences (https://gfz-potsdam.de)
IGSN	International Generic Sample Number (https://www.igsn.org)