D12.4 PROTOTYPING A DISTRIBUTED SITE CATALOGUE

WORK PACKAGE 12 – A Framework for Environmental Literacy

LEADING BENEFICIARY: EAA

Author(s):	Beneficiary/Institution
Christoph Wohner	EAA (Environment Agency Austria)
Doron Goldfarb	EAA (Environment Agency Austria)
Johannes Peterseil	EAA (Environment Agency Austria)

To be accepted by: WP12 leader Florian Haslinger

Deliverable type: REPORT

Dissemination level: PUBLIC

Deliverable due date: 30.04.2019/M48

Actual Date of Submission: 08.07.2019/M51

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182

ABSTRACT

This deliverable describes the theoretical groundwork and development of a prototype for a distributed site catalogue.

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Ari Asmi	University of Helsinki
Magdalena Brus	ICOS ERIC

Document history:

Date	**Version**
18.06.2019	Version 1.0 sent to reviewer
28.06.2019	Received review
1.7.2019	Incorporated comments from review
8.7.2019	Document submitted to EC

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to Christoph Wohner (lead author; christoph.wohner@umweltbundesamt.at).

ENVRIplus PROJECT SUMMARY

ENVRIplus[1] is a Horizon 2020 project bringing together environmental and Earth system research infrastructures (RIs), projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonise policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

[1] http://www.envriplus.eu/

EXECUTIVE SUMMARY

The availability of quality controlled data is crucial for any data driven science. Long term monitoring and experimentation networks (MENs) and research infrastructures (RIs) are operating facilities at defined locations generating a wealth of data. Such site-based research infrastructures (RI) usually keep extensive documentation of their site network in dedicated catalogues in order to facilitate the management of the RI and its infrastructure.

However, even with extensive site information existing at the individual RI level, harmonisation and integration of site information across RIs is still scarce. Easy access and discovery for users to this information across different catalogues are still hampered due to a lack of integration. A federated site catalogue addresses these issues and focuses on harmonising and integrating such catalogues with the aim of providing a single access point for site information to foster environmental research and management.

This deliverable describes the work carried out in task 12.3. "Operational framework for RIs terrestrial ecosystem research related to biogeochemical cycles" about the development of a prototype that provides a centralised interface for searching across multiple site catalogues. It covers the fundamentals of interoperability, aggregating site information and the implementation of a prototype for a federated catalogue connected to three selected catalogue systems. These particular systems were chosen as they provide means of API-based data provision of site information as well as comprehensive records for each site. Based on the results of this aggregation, we also provide a set of recommendations for future work in this field.

TABLE OF CONTENTS

1 Introduction

The availability of quality controlled data is crucial for any data driven science. Long term monitoring and experimentation networks (MENs) and research infrastructures (RIs) are operating facilities at defined locations generating a wealth of data. In this respect, a site can be defined as "an in-situ observation or experimentation facility, delimited in space, but varying in size and complexity of the internal organisational and observational design, for the collection of data covering e.g. biogeophysical, biotic or socio-ecological characteristics" (Wohner et al. 2019).

The description of these facilities provides in most cases the organisational context of the observation including e.g. information on the design of the site selection as well as the main research and observation focus. Site-based research infrastructures (RI) usually keep extensive documentation of their site network in dedicated catalogues in order to facilitate the management of the RI and its infrastructure.

Even though providing a description of the site network is a core service of many site-based RIs, like eLTER, ICOS or ACTRIS, the integration and harmonisation of information between the infrastructures is still a challenge as dedicated standards and tools are missing. Simple access to this information and discovery across different catalogues for users is still hampered due to lacking integration. So far, the first steps towards a harmonisation of site descriptions have been made between networks like LTER Europe and ICP Forests (Kirchner et al. 2018). But even with such initiatives and the general tendency of site catalogues of becoming increasingly sophisticated in terms of their data models and functionality, a large-scale integration of site information that is bridging multiple systems and research infrastructures is still missing.

A federated site catalogue addresses these issues and focuses on harmonising and integrating such catalogues with the aim of providing a single access point for site information to foster environmental research and management.

Task 12.3 "Operational framework for RIs terrestrial ecosystem research related to biogeochemical cycles" focussed, among other topics, on enabling better integration and discovery of site information across environmental RIs in order to support user workflows. In the scope of this task, we selected and studied three established catalogue systems from different infrastructures (2.2 Selected catalogue systems). Based on the insights gained from these catalogues, we developed concepts for the integration of site information and built a working prototype and present our findings in this deliverable as well as a set of recommendations for existing catalogue systems and future work in this field.

1.1 Aspects of aggregating heterogeneous metadata catalogues

The realisation of a federated site catalogue raises important design questions at various levels. On the one hand, there are technical considerations to be made regarding issues such as interoperability, federation, aggregation and contextualisation. On the other hand, there are policy oriented questions regarding institutional responsibilities, if and how the workload is shared between data providers and the maintainers of the aggregation infrastructure.

1.1.1 Interoperability

In order to be able to aggregate site information, there is a need for interoperable systems that allow collecting their information to begin with. In the scientific literature different levels of interoperability between systems are defined, namely (a) Syntactic interoperability, and (b) Semantic interoperability (Veltmann 2001).

In a web-based system, **syntactic interoperability** describes the standardisation of the communication between a software client and a server (Schaeffer et al. 2012). In the context of this deliverable, syntactic interoperability describes the process of collecting information from multiple, pre-defined sources through the usage of standardised services and formats, such as "Catalogue Service for the Web" (CSW), "Representational State Transfer" (REST) or "SparQL Protocol and RDF Query Language" (SPARQL, recursive acronym). Such machine-readable services are essential in order to acquire site information.

The second level of interoperability - **semantic interoperability** - is the ability of information systems to exchange information on the basis of shared, pre-established and negotiated meanings of terms and expressions and is needed in order to make other types of interoperability work (Veltmann 2001) and is also a prerequisite for more high-level types of (cross-domain) interoperability. This translates to the usage of common metadata schemas which are described in chapter 2.1 "Relevant metadata schemas" and that are served using the aforementioned services.

However, even with a set of standardised metadata schemas, there are often gaps in interoperability due to language barriers of metadata records, e.g. English, French… the usage of different controlled vocabularies or thesauri across different environmental domains or even the different meanings of the same term, e.g. different forest definitions across Europe. Harmonising the values used to populate the different metadata fields is thus another important aspect of semantic interoperability.

1.1.2 Federation

One particularly fundamental question when aggregating information is the intended level of federation, i.e. the way information should be aggregated. This can range from "fully distributed" to "warehousing".

In a **Distributed Query Processing scenario**, shown in Figure 1 Distributed Query Processing, incoming queries are dynamically converted to the appropriate syntax and then redirected to the individual search APIs of the included source catalogues and the results subsequently collected, harmonised and displayed on the fly. The advantages of this approach are that the information queried from the individual catalogues represents their most up-to-date state and that there are no extensive data storage infrastructure requirements for running the query service. The disadvantages are that query processing times can differ based on current server load and network latency, leading to worst case scenarios where one sub-query slows down an otherwise finished distributed query. Moreover, in order to gather harmonised results, distributed queries are usually limited to relatively simple retrieval scenarios.

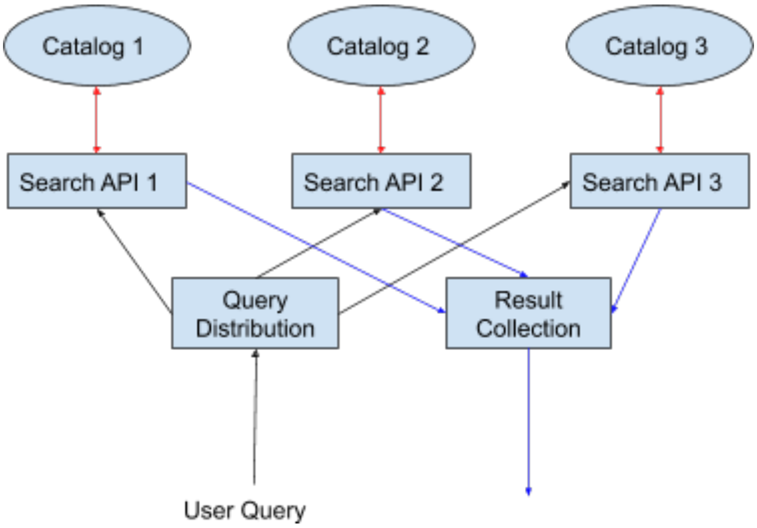


FIGURE 1 DISTRIBUTED QUERY PROCESSING

The other side of the spectrum is represented by warehousing approaches (see Figure 2 Warehousing) which are characterised by the local aggregation of remote content. Data from the different catalogues are downloaded, stored and harmonised locally, enabling to run "holistic" queries across the unified repository. Having complete (meta-)data collections in one storage enables complex, optimised queries which can be executed in a much more timely manner. The main disadvantages of warehousing approaches are issues regarding the up-to-dateness of the stored data and potentially high local data storage and processing requirements.

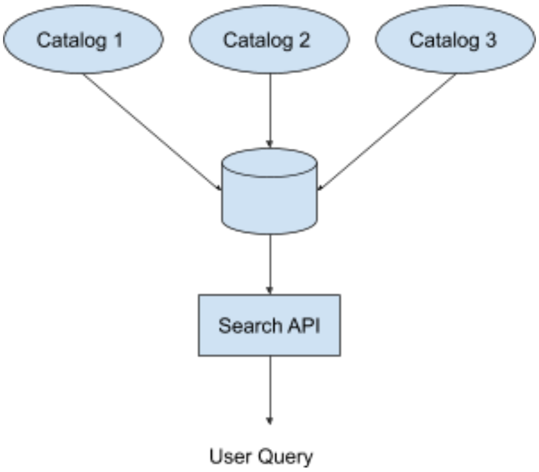


FIGURE 2 WAREHOUSING

1.1.3 Aggregation

Another important design decision regards the level of aggregation, i.e. how and to what extent the contributions of the different sources should be integrated with each other. Figure 3 Overlapping Metadata Information outlines the typical scenario where different catalogues use different metadata schemas for their holdings. Due to different thematic foci, diverging entity/attribute designations or even conceptual disagreement amongst the underlying data models, only a minimal set of metadata elements is expected to be present in all sources. This is usually limited to generic information such as ID, Title, Location, etc., while more specific metadata information is in turn often only present in a few or even only individual sources which themselves moreover often also contain records of varying richness in this regard.

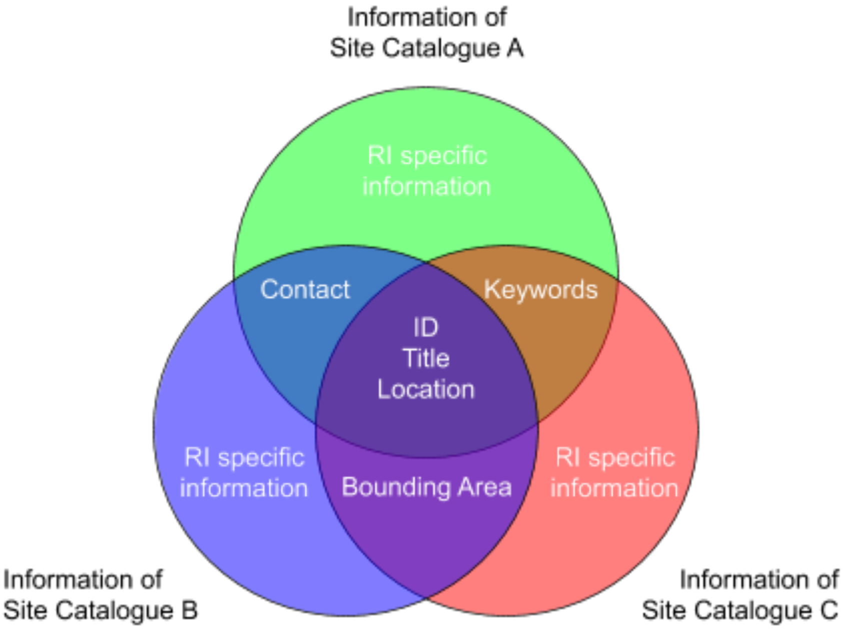


FIGURE 3 OVERLAPPING METADATA INFORMATION

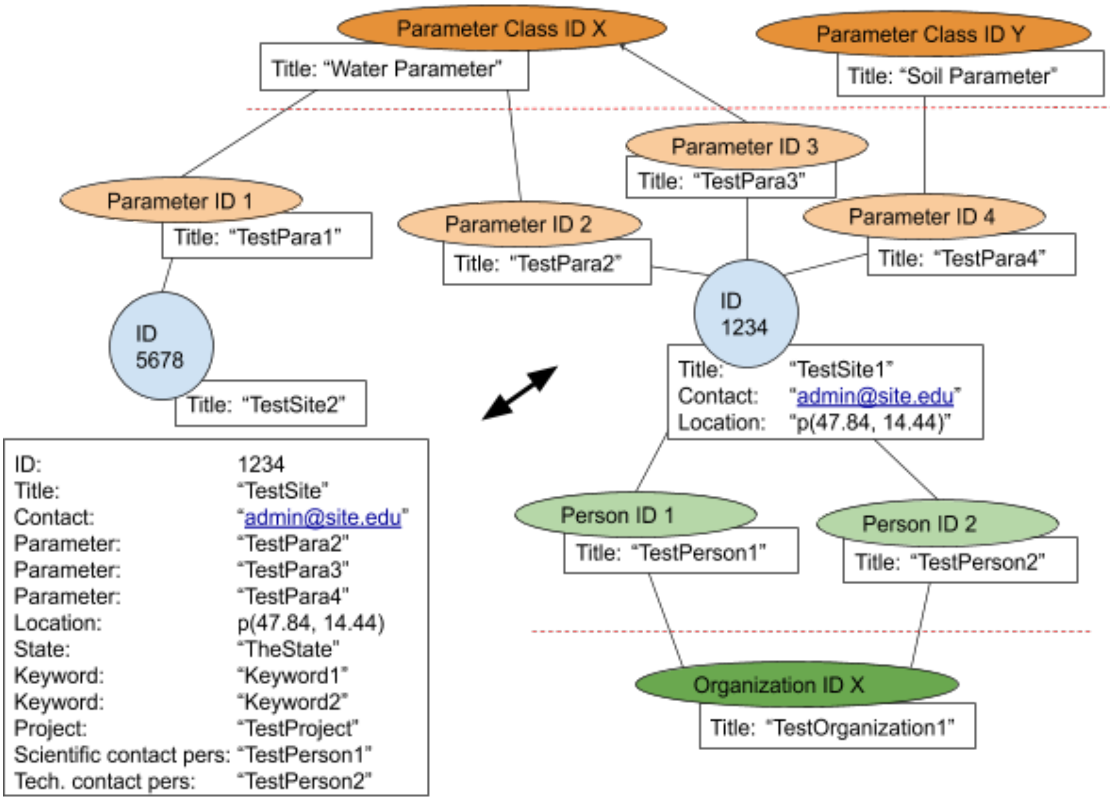
The situation shown in Figure 3 Overlapping Metadata Informationleaves room for three approaches, of which the first one a) is to aggregate the minimal overlapping subset only, the second one b) to identify and collect a selection of the available information, taking into account that it will only be partially available across the different sources, and the last one c) to aggregate "everything" that can be gathered. The main problems with the first approach are that there would be only few fields available for searching, making it difficult to use the aggregated catalogue for detailed enquiries, and that the overlapping subset would shrink with a growing number of included sources. The second approach has the advantage that a "curated" selection of metadata under a common intermediary schema would include the most relevant information for searching the aggregated catalogue, with the disadvantage of increased effort for curation and the process of deciding on a relevant selection of metadata elements to include in the aggregation schema. The "gather everything" approach would have the advantage that the full information from the individual catalogues would be available in the aggregated version as well. Since a common metadata schema would be unfeasible in this case, however, the disadvantage would be a potentially confusing mashup of different metadata collections. This could be relieved by following a more hybrid approach consisting of a harmonised core set of metadata elements surrounded by unconstrained, source-specific information. Considering the three approaches to aggregation, only variants b) or c) would enable meaningful search.

1.1.4 Contextualisation

Site records often contain references to external entities such as measured parameters or involved persons, which themselves are usually interrelated with similar entities, e.g. in form of taxonomic associations or organisational affiliations. In addition, site records sometimes also reference actual datasets. Such information often represents valuable contextual information that can be used for more complex queries for more individual results. Especially the combination of distinct but thematically interrelated catalogues potentially includes many common external context links which could serve as bridges across collections.

Although the source databases behind many catalogues often store contextual information to some extent, such as local taxonomies or organisational relationships amongst involved personnel, exported metadata records often represent “flattened” versions of this rich information, providing only the names of featured persons or parameters. The additional provision of externally dereferenceable identifiers would represent a first step towards a potential future use of such contextual information and should thus be actively proposed amongst site catalogue operators. Such externally dereferenceable identifiers could on the one hand lead to locally provided additional metadata about local taxonomy hierarchies or person relationships, on the other hand, and even more preferred, also to global IDs such as ORCID for persons or globally available domain taxonomies and increasingly also crowd-sourced repositories such as Wikidata.

Figure 4 outlines the difference between a standard “flat” metadata schema and its contextualized counterpart. The latter appears as a network of interrelated entities which are partly connected via external context information, represented in the figure as network nodes lying outside the area delineated with dashed red lines. Considering for example the parameters connected to site “1234”, they are subclasses of two main parameter types, one of which also the superclass for a parameter observed by another, otherwise unrelated site “5678”. Since the parameter taxonomy is not part of the immediate site metadata but part of its context, this relationship would become lost without it.



ID:	1234
Title:	"TestSite"
Contact:	"admin@site.edu"
Parameter:	"TestPara2"
Parameter:	"TestPara3"
Parameter:	"TestPara4"
Location:	p(47.84, 14.44)
State:	"TheState"
Keyword:	"Keyword1"
Keyword:	"Keyword2"
Project:	"TestProject"
Scientific contact pers:	"TestPerson1"
Tech. contact pers:	"TestPerson2"

FIGURE 4 “FLAT” METADATA VS GRAPH METADATA

Even if the original metadata does not provide dedicated identifiers linking to external context data sources, it is possible to try to match textual designations such as parameter or person names to existing registries. This does, however, usually also introduce wrong matches that must be taken into account by providing dedicated means to report and correct them.

1.2 Policies

Another important aspect of metadata aggregation is related to applied policies. This particularly refers to organisational issues such as the responsibility for the creation, the maintenance and the execution of required data transformations. Big and established metadata aggregators such as Europeana for the Cultural Heritage domain[2] or GBIF for Biodiversity provide clearly defined target data formats (“Europeana Data Model”, “Darwin Core” ...) for both immediate object metadata and contextual information. Providers are responsible for creating and executing appropriate mappings to that format and to deliver the result to the aggregator, where a small team checks the delivered data for quality issues and subsequently ingests it. Such a scenario divides the effort between the participants such that the **aggregator is responsible for running the aggregation infrastructure, providing consistent schemas and guidelines for data provision and performing quality control,** while the **providers are themselves responsible for handling the data**. Since such large-scale aggregation endeavours are based on extensive funding for both infrastructure and data providers, e.g. via dedicated project funds, a comparable setting is not feasible for smaller scale scenarios. In many cases, it is rather unlikely that data providers would be able to contribute significant effort to transform their data without receiving funding, transferring the responsibility to the aggregator in this regard. As far as station metadata are concerned, however, the number of existing schemas and also the record counts in the individual catalogues themselves are relatively low compared to the variety of sources contributing to large scale aggregators such as those mentioned above.

1.3 Data formats

Since large-scale aggregation approaches such as GBIF or Europeana utilise a common target data model and format which must be used by data providers for their submissions, the aggregation infrastructures can be designed to only accommodate conforming data. Considering situations where the responsibility of transforming source metadata into a suitable form remains with the aggregator, however, requires taking, at least initially, an increasingly diverse landscape of metadata provision approaches into account. This on the one hand includes a variety of possible data models, on the other hand also different, a priori incompatible data formats such as XML, JSON, CSV, and a variety of RDF based flavours or different API result formats.

The aggregation infrastructure must therefore be able to deal with this heterogeneous variety of data formats, which can be done in different ways. The most fundamental question is related to the level of federation, whether the aggregation should take place on the fly or rather be based on a warehousing approach. The former is limited to catalogues offering machine accessible search APIs whose different query syntaxes and the models/formats behind the respective results must be known a priori in order to perform the federation of the user query to the different API endpoints and the subsequent collection of the results. The latter usually only represents an extract of the underlying metadata and are thus potentially easier to aggregate. If a warehouse approach is followed, however, it is usually the original metadata which is collected first, opening up a number of different options regarding its subsequent processing. Assuming a situation where the aggregator must handle the source data by him- or herself, this can be:

[2] https://www.europeana.eu/portal/en

a) to define a suitable target metadata schema in a similar fashion as done in Europeana or GBIF, to subsequently map and transform the different heterogeneous source formats accordingly and aggregate only the target metadata or

b) to implement a multi-level approach trying to preserve as much of the original information as possible, initially following a “gather everything” approach and gradually harmonising the accumulated data in a stepwise manner, potentially allowing the co-existence of multiple target schemas for different application scenarios.

In both cases, unless the target metadata schema represents an existing standard already used by at least some of the data providers, the transformation from the original metadata format must be performed by the aggregator infrastructure.

2 State of the art

2.1 Relevant metadata schemas

There are a number of schemas containing site information that can be identified. In the following sections, there is a brief overview of these relevant schemas that include site descriptions and bear relevance for the work in this task:

- Observations and Measurements (O&M)
- Inspire Environmental Monitoring Facilities (Inspire EF)
- WMO Integrated Global Observing System (WIGOS) Metadata Standard

2.1.1 Observations and Measurements (O&M)

Observations and Measurements (O&M) is an international standard which defines a conceptual schema encoding for observations, and for features involved in sampling when making observations, i.e. the sensor or sensor station that was used to make the observations (Cox et al. 2011).

O&M defines a core set of properties for an observation including:

- feature of interest
    - Which is defined as “The thing whose property is being estimated or calculated in the course of an Observation to arrive at a Result, or whose property is being manipulated by an Actuator, or which is being sampled or transformed in an act of Sampling.”[3]
- observed property
- result
- phenomenon time – the real-world time associated with the result
- result time – the time when the result was generated
- valid time – the period during which the result may be used
- procedure – the instrument, algorithm or process used (which may be described using SensorML)

[3] http://www.w3.org/ns/sosa/FeatureOfInterest

SensorML is an approved Open Geospatial Consortium standard that provides standard models and an XML encoding for describing sensors and measurement processes[4]. A SensorML description of the used instrument or instrument station following the SensorML Sensor Web Enablement Lightweight SOS Profile[5] includes the following fields:

- SensorDescription
- TimePeriod
- keywords
- identifier
- classification
- contacts
- featuresOfInterest
    - which in the context of this Lightweight profile is "an identifier of the geometric feature (e.g. sensor station) to which the observation is associated; within the lightweight profile this is limited to sampling point[6]"
- outputs

O&M therefore has two aspects, one describing the actual observations and a second aspect providing information about the context of the observations including the actual site, station or sensor information. Even though none of the selected systems directly provide site information following O&M, it was included in this deliverable for the purpose of completeness.

2.1.2 WMO Integrated Global Observing System (WIGOS) Metadata Standard

WIGOS is a framework for all World Meteorological Organization (WMO) observing systems and for WMO contributions to co-sponsored observing systems in support of WMO Programmes and activities. An aspect of the WMO Integrated Global Observing System (WIGOS) implementation is ensuring maximum usefulness of WIGOS observations by providing context information for these observations.

For this purpose, two complementary types of metadata are required:

- Discovery metadata and
- Interpretation/description or observational metadata.

The discovery metadata and the interpretation/description or observational metadata enable data values to be interpreted in context and are the subject of the so called "WIGOS metadata standard". The WIGOS metadata should describe the "observed variable, the conditions under which it was observed, how it was measured, and how the data have been processed, in order to provide users with confidence that the data are appropriate for their application" (WIGOS Metadata Standard 2017).

Table 1 provides an overview of all site relevant fields of the WIGOS metadata standard. It should be noted that the WIGOS metadata standard is far more comprehensive and this selection solely focuses on the fields describing station information. Many of the metadata elements should be

[4] https://www.opengeospatial.org/standards/sensorml
[5] https://portal.opengeospatial.org/files/?artifact_id=52803
[6] https://portal.opengeospatial.org/files/?artifact_id=52803

populated using dedicated code lists which are available via a temporary registry[7] which is supposed to be made persistent[8] in near future[9].

Table 1. WIGOS Metadata Standard - Subset of site relevant information

Category	Name	Definition
3. Station/Platform	Region of origin of data	WMO Region
3. Station/Platform	Territory of origin of data	Country or territory name of the location of the observation
3. Station/Platform	Station/platform name	Official name of the station/platform
3. Station/Platform	Station/platform type	A categorization of the type of observing facility at which an observation is made
3. Station/Platform	Station/platform model	The model of the observing equipment used at the station/platform
3. Station/Platform	Station/platform unique identifier	A unique and consistent identifier for an observing facility (station/platform), which may be used as an external point of reference
3. Station/Platform	Geospatial location	Position in space defining the location of the observing station/platform at the time of observation
3. Station/Platform	Data communication method	Data communication method between the station/platform and some central facility
3. Station/Platform	Station operating status	Declared reporting status of the station
4. Environment	Surface cover	The observed (bio)physical cover on the Earth's surface in the vicinity of the observation
4. Environment	Surface cover classification scheme	Name and reference or link to document describing the classification scheme
4. Environment	Topography or bathymetry	The shape or configuration of a geographical feature, represented on a map by contour lines
4. Environment	Events at observing facility	Description of human action or natural event at the facility or in the vicinity that may influence the observation
4. Environment	Site information	Non-formalized information about the location and surroundings at which an observation is made and that may influence it
4. Environment	Surface roughness	Terrain classification in terms of aerodynamic roughness length

[7] http://test.wmocodes.info/wmdr

[8] http://codes.wmo.int/wmdr

[9] https://github.com/wmo-cop/wmo-oscar/issues/6

4. Environment	Climate zone	The Köppen climate classification of the region where the observing facility is located. The Köppen-Geiger climate classification scheme divides climates into five main groups (A, B, C, D, E), each having several types and subtypes
5. Instruments and methods of observation	Instrument specifications	Intrinsic capability of the measurement/observing method to measure the designated element, including range, stability, precision, etc.
5. Instruments and methods of observation	Instrument operating status	The status of an instrument with respect to its operation
5. Instruments and methods of observation	Vertical distance of sensor	Vertical distance of the sensor from a (specified) reference level, such as local ground, deck of a marine platform at the point where the sensor is located, or sea surface
5. Instruments and methods of observation	Maintenance party	Identifier of the organization or individual who performed the maintenance activity
5. Instruments and methods of observation	Geospatial location	Geospatial location of instrument/sensor
9. Ownership and data policy	Supervising organization	Name of organization who owns the observation
9. Ownership and data policy	Data policy/use constraints	Details relating to the use and limitations surrounding data imposed by the supervising organization
10. Contact	Contact (nominated focal point)	Principal contact (nominated focal point) for resource

Based on O&M, WIGOS also features the aspect of the actual observations accompanied by the context information including the site description.

Even though none of the selected catalogue systems provide site information following the WIGOS standard, it was included in this deliverable as it is implemented in the so called OSCAR system mentioned in chapter 5.2 Complementing site information with WMO station data.

2.1.3 Inspire Environmental Monitoring Facilities (EF)

INSPIRE (Infrastructure for Spatial Information in the European Community) is an EU initiative to establish a spatial data infrastructure in Europe that is intended to make spatial or geographical information more accessible and interoperable for a wide range of purposes supporting sustainable development and environmental policies and activities which may affect the environment. The INSPIRE Directive entered into force on 15 May 2007 has to be implemented by the EU member states until October 2020 (INSPIRE Directive 2007).

The directive addresses 34 spatial data themes needed for environmental applications. One of these themes is “Environmental monitoring facilities (EF)” which is defined as “Location and operation of environmental monitoring facilities includes observation and measurement of emissions, of the state of environmental media and of other ecosystem parameters (biodiversity, ecological conditions of vegetation, etc.) by or on behalf of public authorities.” (INSPIRE EF 2019).

The theme scope includes two main aspects:

- The environmental monitoring facility as a spatial object and
- Data obtained through observations and measurements taken at this facility, encoded using the ISO 19156 standard. This information is complemented by further information such as networks the facility is part of or programmes the facility provides data to (INSPIRE EF 2019).

Analogous to O&M and WIGOS, Inspire EF features two aspects (the procedure/infrastructure used to gather measurements/observations and the observation itself). In this case, however, the encoded information is further complemented by network information.

In table 2 all site relevant fields and their definitions of Inspire EF are listed.

TABLE 2 INSPIRE EF - SUBSET OF SITE RELEVANT INFORMATION

Name	**Definition**
inspireIdIdentifier	External object identifier
name	Plain text denotation of the EnvironmentalMonitoringFacility
additionalDescription	Plain text description of additional information not fitting in other attributes
mediaMonitored	Monitored environmental medium
responsibleParty	Responsible party for the EnvironmentalMonitoringFacility
geometry	Geometry associated to the EnvironmentalMonitoringFacility. For mobile facilities the geometry represents the area the facility is expected to measure in
onlineResource	A link to an external document providing further information
ObservingCapability	Observations and Measurements acquired.
representativePoint	Representative location for the EnvironmentalMonitoringFacility
mobile	Indicate whether the EnvironmentalMonitoringFacility is mobile (repositionable) during the acquisition of the observation.
operationalActivityPeriod	Lifespan of the physical object (facility).
belongsTo	A link pointing to the EnvironmentalMonitoringNetwork(s) this EnvironmentalMonitoringFacility pertains to. The association has additional properties as defined in the association class NetworkFacility.

Out of the selected systems Inspire EF is used by DEIMS-SDR to expose site information. Additional information about the mapping of the DEIMS-SDR site metadata model to the Inspire model can be found in the respective chapter (2.2.1 DEIMS-SDR).

2.2 Selected catalogue systems

This section provides an overview of the selected systems, their supported services and formats and underlying data models. All of the selected systems provide means of API-based data provision of site information as well as comprehensive records for each site, which were the reasons for their selection for this deliverable. In order to compile a selection of systems, a first screening for site catalogues was made in preparation of the task. The results are provided in 10.1 List of relevant site catalogues. The site catalogues that were finally selected for the development of the concept of the federated site catalogue are described in the following chapters.

2.2.1 DEIMS-SDR

DEIMS-SDR (Dynamic Ecological Information Management System - Site and dataset registry, https://deims.org) is an information management system that allows discovering long-term ecosystem research sites around the globe, along with the data gathered at those sites and the people and networks associated with them. DEIMS-SDR describes a range of sites, providing information about each site's location, ecosystems, facilities, parameters measured and research themes. Within the International Long Term Ecological Research Network (ILTER[10]) as well as its regional group LTER-Europe[11], DEIMS-SDR is used as the central site catalogue and is also developed and maintained by these groups. Regardless of being a system developed and used by the LTER community, it is also open for users outside LTER and stores an increasing number of non-LTER sites and datasets. DEIMS-SDR has been used as both a site registry and metadata editor for measurements in H2020 projects, such as ENVRI plus, eLTER[12] and EcoPotential[13].

DEIMS-SDR supports a number of standardised services (Web Mapping Service (WMS) for serving georeferenced images and tiles allowing to implement site information in maps; Web Feature Service (WFS) for providing information for each site object; Catalogue Service for the Web (CSW) and Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to serve metadata records) and formats (Inspire EF and ISO 19139) for serving site information. Metadata published by DEIMS-SDR is provided by CC-BY-NC 4.0 International license to enable a free use of the information (Wohner 2019). This openly available data has also already been used for scientific analyses and subsequent publications (Mollenhauer et al. 2018, Zilioli et al. 2019) illustrating a certain maturity of the system and its data.

The underlying site metadata model allows capturing information about the organisation (e.g. contact, information and networks), the location, the observation characteristics (e.g. climate, habitats) or available equipment. Additionally, there are fields about the focus and design of a site, network affiliation and information about data policies and data management[14].

Apart from geographic and environmental characteristics, the model also supports indicating which networks and projects a site belongs to, contact points for a site (e.g. institution or a person) as well as site hierarchies, e.g. umbrella sites consisting of subsites that may in turn consist of plots. It is also possible to reference external data collections and data portals of sites.

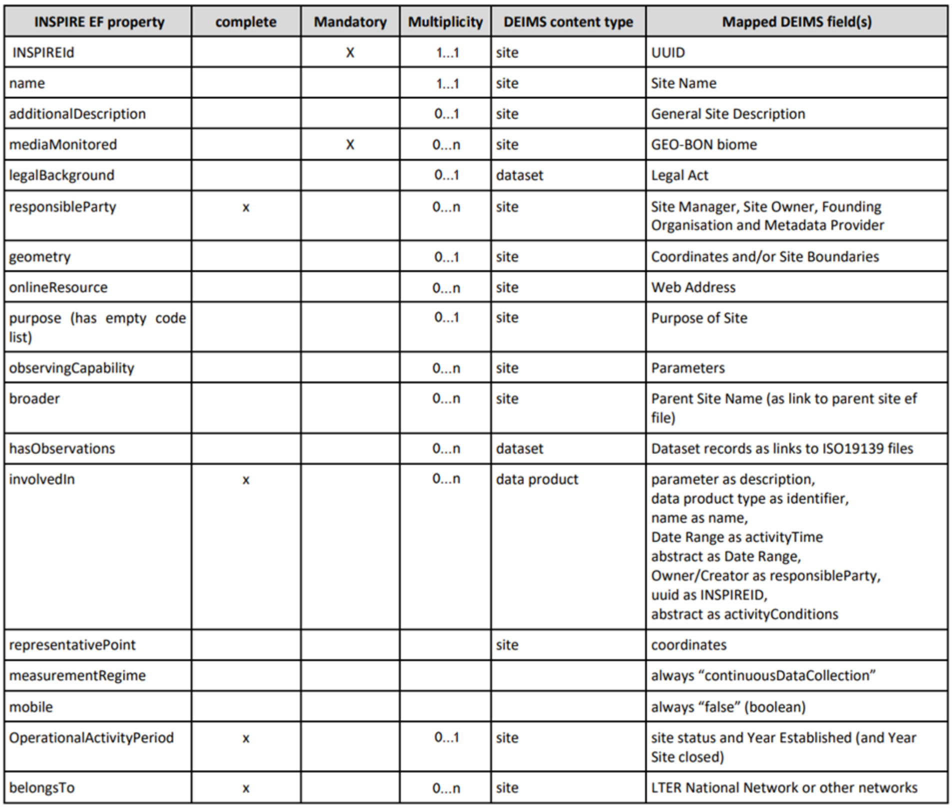
In addition to the mandatory fields (site name, site manager, country and network/project/RI affiliation), a set of recommended fields was defined that forms the basis for calculating the completeness of a site record as a percentage, based on the amount of information provided for the recommended fields, which allows easier assessment of the quality of any given site record. Currently, the following types of information are recommended to be provided (Wohner 2019):

[10] https://www.ilter.network/
[11] https://www.lter-europe.net/
[12] https://cordis.europa.eu/project/rcn/194957/factsheet/en
[13] https://cordis.europa.eu/project/rcn/196809/factsheet/en
[14] https://deims.org/models

- General characteristics (‘General Site Description’, ‘Coordinates’, ‘Site Type’, and ‘Size’)
- Climatic characteristics (‘Mean Annual Air Temperature’ and ‘Sum Annual Precipitation’)
- Topographic characteristics (‘Elevation Range (minimum – maximum)’)
- Ecosystem characteristics (‘Biome’ and ‘Ecosystem and Land Use’)
- Scientific characteristics (‘Purpose of Site’, ‘Research Topics’, ‘Design of Observation’, ‘Scale of Observation’, ‘Design of Experiments’, ‘Scale of Experiments’ and ‘Observed parameters’)
- Operation characteristics (‘Year Site was established’, ‘Site Status (active, inactive, closed)’, ‘Permanent Operation’, ‘Accessible All Year’, and ‘Permanent Power Supply’)
- Data management (‘Data Request Format’ and ‘Data Storage Location’)
- Metadata information (‘Metadata provider’)

In order to better provide site and other associated information, DEIMS-SDR offers a dedicated Inspire EF export, created by pulling together a site record and associated dataset, data product and person records. A description of the mapping can be seen in Tab. 3 Mapping of DEIMS fields to Inspire EF (Poursanidis et al. 2016). The Inspire EF exports for each site can be accessed programmatically as described in table 7 “Selected systems and access point descriptions”.

TABLE 3 MAPPING OF DEIMS FIELDS TO INSPIRE EF (POURSANIDIS ET AL. 2016)



INSPIRE EF property	complete	Mandatory	Multiplicity	DEIMS content type	Mapped DEIMS field(s)
INSPIREId		X	1...1	site	UUID
name			1...1	site	Site Name
additionalDescription			0...1	site	General Site Description
mediaMonitored		X	0...n	site	GEO-BON biome
legalBackground			0...1	dataset	Legal Act
responsibleParty	x		0...n	site	Site Manager, Site Owner, Founding Organisation and Metadata Provider
geometry			0...1	site	Coordinates and/or Site Boundaries
onlineResource			0...n	site	Web Address
purpose (has empty code list)			0...1	site	Purpose of Site
observingCapability			0...n	site	Parameters
broader			0...n	site	Parent Site Name (as link to parent site ef file)
hasObservations			0...n	dataset	Dataset records as links to ISO19139 files
involvedIn	x		0...n	data product	parameter as description, data product type as identifier, name as name, Date Range as activityTime abstract as Date Range, Owner/Creator as responsibleParty, uuid as INSPIREID, abstract as activityConditions
representativePoint				site	coordinates
measurementRegime					always “continuousDataCollection”
mobile					always “false” (boolean)
OperationalActivityPeriod	x		0...1	site	site status and Year Established (and Year Site closed)
belongsTo	x		0...n	site	LTER National Network or other networks

2.2.2 ICOS Station Catalogue

The station catalogue of the Integrated Carbon Observation System (ICOS), in this deliverable commonly referred to as “ICOS Station Catalogue”, features all sites and stations that are part of the ICOS network. ICOS is a distributed research infrastructure comprising three coordinated, complementary operational observation networks focusing on atmospheric observatories of concentrations of CO2, CH4, N2O and other greenhouse gases, terrestrial flux tower sites to measure the ecosystem exchange of CO2, water vapour and energy, and oceanographic observation platforms monitoring air-sea fluxes[15].

Sites and their records are grouped according to these themes ("Atmospheric[16]", "Ecosystem[17]" and "Oceanic [18] "). Each station is identified by a URL (e.g. https://meta.icos-cp.eu/ontologies/stationentry/AS/ATM-HYY), featuring a landing page with information about the station, including fields about:

- accessibility
- address
- already operational
- anthropogenic density
- application status
- associated files (e.g. images featuring site maps or photos)
- country code
- decimal latitude
- decimal longitude
- elevation above ground, meters
- elevation of ground above sea, meters
- estimated date when operational
- existing infrastructure
- funding for construction
- funding for operation
- has pre-ICOS measurements (true/false)
- ICOS station class
- kind of station (TT, aircraft, atm, ground, eco, ocean)
- long name
- name list of the networks station belongs to
- principal investigator
- responsible institution name
- short name
- site type
- thematic center's ID
- vegetation
- website URL

There is a dedicated ontology for site and station information[19]. Site records are created using two interfaces: the "station entry" service [20] and labelling service [21] . These provisional stations metadata eventually become part of the Carbon Portal metadata. The results of this metadata

[15] https://www.icos-ri.eu/
[16] https://meta.icos-cp.eu/ontologies/cpmeta/AS
[17] https://meta.icos-cp.eu/ontologies/cpmeta/ES
[18] https://meta.icos-cp.eu/ontologies/cpmeta/OS
[19] https://github.com/ICOS-Carbon-Portal/meta/blob/master/src/main/resources/owl/stationEntry.owl
[20] https://meta.icos-cp.eu/edit/stationentry/
[21] https://meta.icos-cp.eu/labeling/

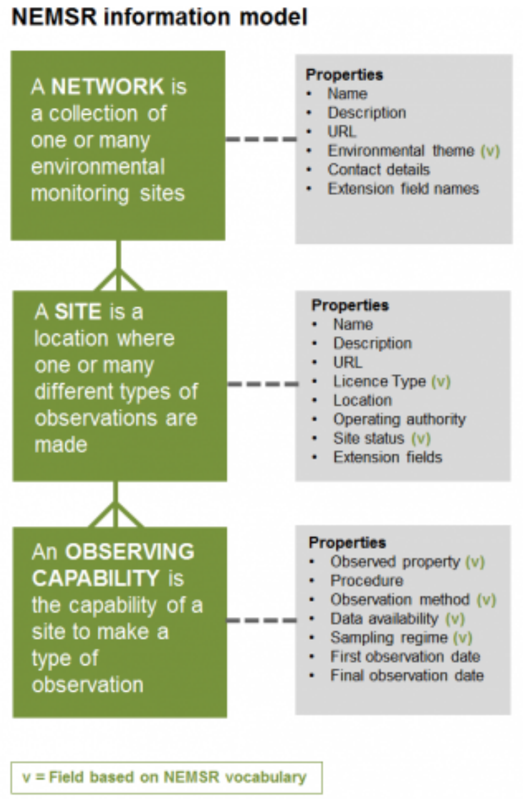
input can be viewed on the stations map[22] and the stations table[23] and also be queried using a SparQL endpoint which is further described in table 7.

2.2.3 NEMSR

The National Environmental Monitoring Sites Register (NEMSR) of the Australian National Environmental Information Infrastructure (NEII) provides a consolidated overview of Australia's environmental monitoring sites. It brings together a range of networks across environmental domains, including seismic monitoring stations, ocean radars, long-term weather observation sites, flux stations, and ground cover reference sites. The NEMSR information model allows covering sites with a set of properties as well as networks as separate data entities (Figure 5 NEMSR information model (NEMSR 2019)).

For the inclusion of environmental monitoring sites in NEMSR, they should meet the following general criteria:

- The sites are a product of operational monitoring programmes thus sites, data and user support are readily available.
- The observational data that the monitoring sites point to are publicly accessible in an electronic format.
- Monitoring methods and protocols are well described to enable re-use of data (NEMSR 2019).



[22] https://static.icos-cp.eu/share/stations/
[23] https://www.icos-cp.eu/node/83

FIGURE 5 NEMSR INFORMATION MODEL (NEMSR 2019)

As NEMSR itself already integrates site information from a number of Australian networks, it features a more generalised set of fields for the description of sites compared to DEIMS-SDR and the ICOS station catalogue (Table 4).

TABLE 4 NEMSR SITE DATA FIELDS[24]

FIELD	DESCRIPTION
id	The unique identifier of the site. Data type - varchar(255)
name	The name of the site. Data type - varchar(255)
siteDescription	The description of the site. Data type - varchar(500)
siteLicence	The type of licence that applies to the site metadata. Datatype - controlled list.
siteURL	A URL of a web page or resource that contains information about the site.
srsName	The EPSG code of the spatial referencing system used to locate the geographic entity. Datatype - controlled list.
latitude/longitude	The latitude, latitude of the site. encoded using the GeoJSON
elevation	The elevation of the site. Datatype - number
operatingAuthority/name	The organisation that is the operating authority for the site.
operatingAuthority/url	A URL of a web page or resource that contains information about the operation authority.
siteStatus	The operating status of the site. Datatype - controlled list.
extensionFieldValue(s)	The value of the extension field. For example, the network may store a WMO ID for sites in an extensionField. Data type - varchar(500)

For some of the fields in the information model, NEMSR utilises controlled vocabularies to ensure standardised information, e.g. “network id” and “environmental theme” for networks and “site status”, “observed property” and “site licensing” for sites.

In addition to access through a web based viewer, NEMSR data can also be accessed programmatically through web service endpoints using a REST-API serving site records as JSON objects provided separately for each registered network, e.g. Seismic, National Geomagnetic Observatories, Air Quality Monitoring Network, etc. All of these URLs or just a subset can be embedded and queried in applications, such as a federated site catalogue.

[24] http://www.neii.gov.au/nemsr/documentation/1.0/data-fields/site

3 Designing a federated site catalogue

3.1 Use Cases

Site catalogue information is of use for a number of user groups. Such user groups can be defined using site information to either conduct science looking for data or using the sites to implement e.g. in-situ experiments. The **Research community** would be the most important one focusing either on data provision or using sites along gradients for dedicated research. A second group are **RI Managers** focusing on a comparison of facilities across different RIs. **National funders and Joint Programme Initiatives (JPIs)** would also be an additional user group. The **Earth Observation Community** looking for facilities to retrieve or generate validation data for EO campaigns or products are also an important community. Common to all communities is the (a) focus on the site documentation and (b) the need for an unambiguous identification of sites. Proper integration of information provided by the different catalogues, including both syntactic and semantic harmonisation, is needed in order to utilise the full benefit of a distributed site catalogue.

Multiple basic use cases can be defined for the integration of site information from a range of different catalogues. Since the available means of discovery are limited by the level of aggregation applied, the current use case definitions assume a rather basic scenario in this regard which can be easily extended at a later stage.

The selection of use cases is based on the experience from the different RIs participating in ENVRI PLUS as well as on discussions raised in the context of GEO (GEO, 2017) focusing on a consistent documentation of in-situ sites in the Biodiversity and Ecosystem domain. In order to derive additional use cases, a case study was taken from the site selection for the proposal of the H2020 LTER Plus project. The selection for this particular project capitalised on the existing documentation of sites in DEIMS-SDR. The selection process only considered formally acknowledged LTER sites from the 26 European LTER networks with complete site metadata sets and a wide range of accessory information, including co-location with other RIs (e.g. ICOS) and networks such as the UNECE ICPs, as well as responsiveness and reliability of sites as data providers during previous projects.

The list of use cases defined in the scope of this deliverable is not exhaustive but can be used as a first start for the development of the concept of a federated site catalogue.

TABLE 5 DEFINED USE CASES

UC 1	A user wants to find sites across a range of site catalogues with defined measurements (e.g. air temperature) in a selected biogeographical region (e.g. the Boreal region)
UC 2	A user wants to find sites across a range of site catalogues located in a defined country (e.g. Belgium)
UC 3	A user wants to select sites in a defined are distance to another site in order to address the subject of co-location
UC 4	A user wants to get an overview (on a given spatial scale) what sites are operated by which network or RI
UC 5	A user wants to identify if a site with a different research focus (e.g. meteorological) is near to my site or a set of sites at a given scale.

In a second process, the defined use cases can be mapped to the user communities defined for the scope of the federated catalogue. Table 6 provides and overview of the current mapping. The defined use cases are shared by different user communities.

TABLE 6 MAPPING OF USE CASES TO THE DIFFERENT USER COMMUNITIES

	UC 1	UC 2	UC 3	UC 4	UC 5
RI Manager			X	X	X
Researcher	X	X	X	X	X
EO analyst	X	X			
National funders and JPIs		X		X	

In order to limit the scope of the prototype development of the federated site catalogue, **UC 1** and **UC 2** were selected for the prototype to be addressed. Based on the requirements defined by the use cases and research questions we developed outlines of concepts for a federated site catalogue.

3.2 Technical options

As outlined in the discussion of the relevant aspects in the introduction (1.1 Aspects of aggregating heterogeneous metadata catalogues), there is a variety of options to realise a distributed site catalogue. The most straightforward would be to establish a data warehouse approach based on a core set of metadata fields defined for the different catalogues to be integrated.

Extended functionality would be made possible by taking contextual information into account, allowing finding commonalities between otherwise separated sites via pathways located in their wider context, i.e. via person affiliations or taxonomic groupings of parameters. From that point on, however, the structure of the aggregated information would be more like a knowledge graph, which would call for appropriate technologies such as graph oriented databases, allowing for more path oriented queries.

Establishing means to distinctly identify the same site across different catalogues would be an enhanced functionality. The current chapter describes the different technical options for the implementation of a federated site catalogue in short. This is intended to be the basis for further developments and discussions.

3.2.1 Basic integration

Based on an analysis of the data structure of each of the sources to be included an individual ETL procedure can be written to conduct their integration based on a “flat” metadata approach. This would concentrate on metadata elements expected to be present in most of the integrated records, such as title, location, description, keywords, parameters, etc. This would result in a compact target metadata structure which could be made searchable via faceted search platforms such as e.g. Solr.

An additional layer of functionality could be provided by including means to harmonise relevant metadata element values, such as parameter or contact point names, which would enable more efficient and meaningful information retrieval. Existing taxonomies or ontologies for representing the former in a structured form, such as the EnvThes[25] controlled vocabulary, could serve as a useful reference in this regard. However, since automated mappings tend to only partially lead to correct results, efficient means for user feedback or even crowd-sourced correction would have to be provided too.

3.2.2 Integration based on Linked Data

Taking the discussion from chapter 1.1 "Aspects of aggregating heterogeneous metadata catalogues" into account; this section presents considerations for a flexible approach to aggregate information from site catalogues, focusing on the inclusion of contextual information and the resulting data representation as a knowledge graph. It is understood as a potential future roadmap beyond ENVRIplus activities and proposes a multi-level approach seeking to preserve as much of the original information as possible, holding it at disposal for different aggregation scenarios. An outline of the concept is shown in Figure 6 RDF-based multi-level catalogue metadata aggregation and the individual aspects are discussed in the following paragraphs.

Multi-level aggregation

In this scenario, multi-level aggregation refers to the approach to perform the aggregation of different metadata sources in different steps. Each level represents a specific transformation "away" from the original metadata, which potentially introduces a loss of information but in turn makes the different contributions more interoperable with each other.

Storing original metadata records

The basic layer of integration stores the to-be-aggregated metadata in its native format as provided by the source catalogue. In the most straightforward way, the individual files can be stored in a local filesystem, where a versioning system such as Git could be used to track changes and trigger updates in the transformation pipeline. Alternatively, the source metadata could as well be stored in dedicated database systems, in which case multi-model databases able to store the full variety of different source formats would be the only feasible solution.

RDF as "lingua franca"

The flexibility of RDF to represent heterogeneous information assets suggests it as a useful framework for integrating sources based on different formats, some of them even natively provided in RDF. As the first - syntactically interoperable - aggregation level, harvested catalogue data presented in other formats should therefore initially be transformed to RDF in order to reach a common data representation. This transformation should be preserving as much information as possible and therefore not focus on the modification of the underlying data models but only on format conversions and the required modifications in this regard. Existing approaches to map different source data to RDF should be considered, including XSLT for XML and JQ for JSON but also more specific declarative languages such as RML[26] or LIXR[27]. The transformed RDF representations should be stored in a dedicated Triple store enabling one-stop access to the collected data.

Identifiers

[25] http://vocabs.lter-europe.net/EnvThes

[26] http://rml.io/

[27] https://github.com/liderproject/lixr

One crucial aspect of RDF representations is the use of http://-URIs as identifiers. While sources providing native RDF data already include identifiers in the required form, data from other sources must be augmented accordingly. Existing URLs pointing to the Web representations of site metadata in the different catalogues could be re-used for this purpose. In both cases, however, these existing URLs would be in the scope of the data providers and controlled by their infrastructure, their dereferencing would thus lead to the original representations/web presence and not to their aggregated catalogue representation. Especially for non-native RDF resources this would be problematic, since no RDF representation could be retrieved via the original URls. A better approach would thus be to define a dedicated URL namespace for the aggregated RDF representation and create new identifiers for all the aggregated entities, linking them, where available, to their original counterparts via dedicated metadata entries.

Accessing, Viewing and Querying the raw RDF graph

The raw RDF graph should be made available in machine accessible form via one or more established Linked Data provisioning mechanisms such as data dump, per-entity dereferencing, SPARQL and/or a Linked Data API. On top of that, a raw RDF viewer such as Pubby[28] or BRWSR[29] should provide direct human readable access to the accumulated triples. More targeted browsing/querying of the accumulated data should be enabled via a dedicated faceted browser, allowing drilling down on the available content.

Semantic mapping / Unified Schema

The raw collection of RDFised metadata should serve as a foundation for the second, semantically interoperable level providing a harmonised representation and access to the aggregated data. Harmonisation should take place along two main directions, on the one hand as a schema alignment by mapping similar properties across the different metadata schemas, on the other hand as an instance alignment by mapping similar property values accordingly. Assuming situations where differently named properties from different catalogues share similar data values (or types thereof, such as “only values of type parameter”), the schema alignment can thus potentially be supported by the instance alignment.

The property as well as instance alignment could be performed using established platforms such as the SILK[30] link discovery framework. Moreover, the faceted browser for the raw RDF data could be adapted to serve as additional means for reconciliation, allowing users to establish similarity links between featured property facets and/or values. Since both automated as well as crowd-sourced alignments are not guaranteed to be error-free, the faceted browser could also serve as quality control platform allowing annotating wrong matches. A specific focus should be put on the alignment of contextual information such as persons, locations and parameters with globally available authorities.

Basic inferences

The alignment of specific types of entities with structured contextual information could enable basic inference functionality taking context links into account. Connecting, for example, observed parameters with external taxonomies such as outlined in (1.1.4 Contextualisation) would allow the retrieval of all sites measuring parameters having a common superclass.

[28] http://wifo5-03.informatik.uni-mannheim.de/pubby/
[29] https://github.com/Data2Semantics/brwsr
[30] http://silkframework.org/

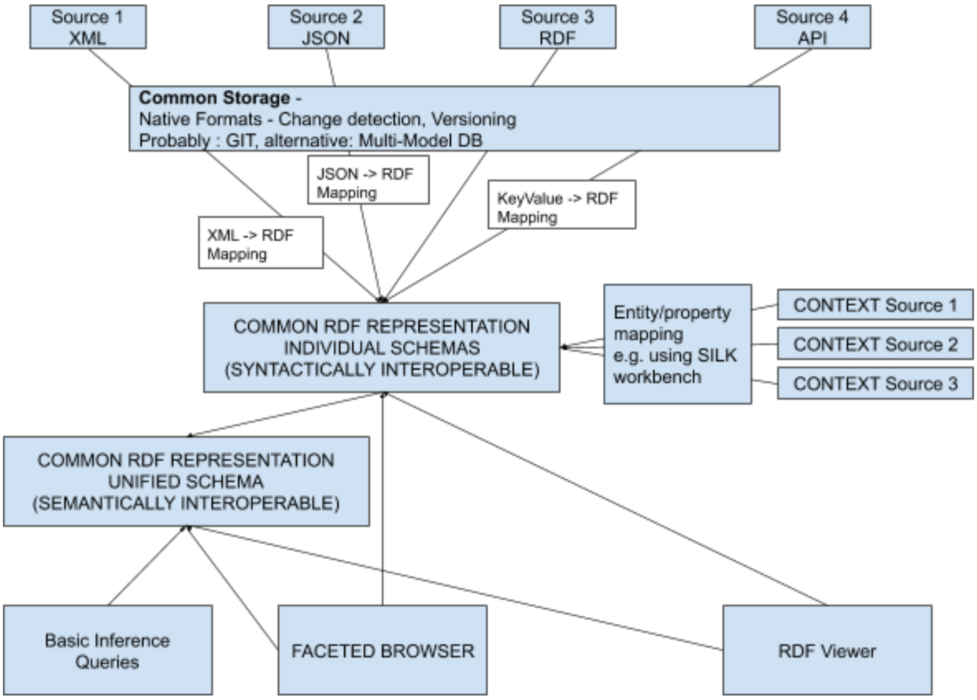
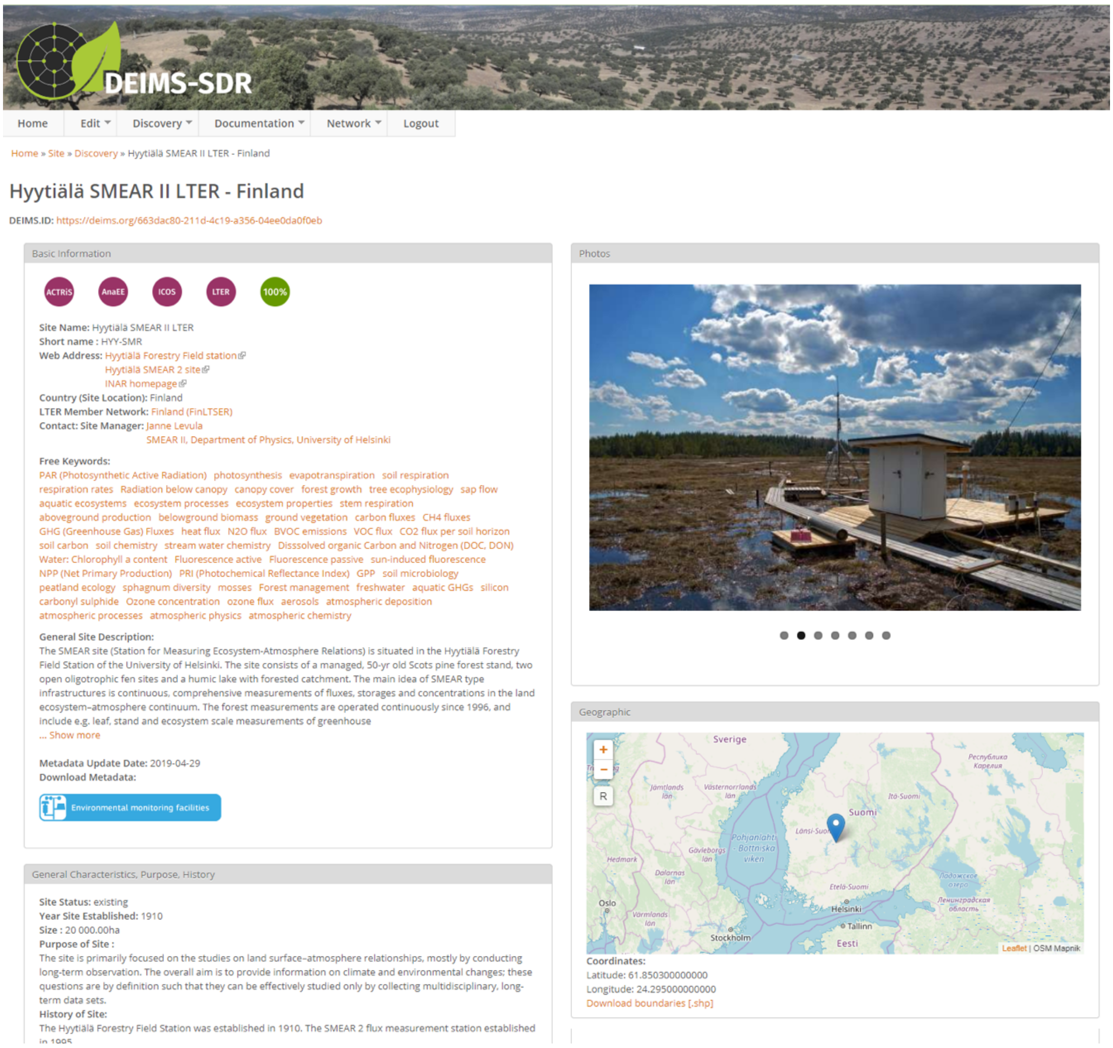


FIGURE 6 RDF-BASED MULTI-LEVEL CATALOGUE METADATA AGGREGATION

3.2.3 Integration based on a centralised registry

An important aspect, which emerged from the discussion across the different environmental RIs, is the aspect of co-location and related to this also the unambiguous identification of sites across catalogues. A common site registry providing the means of distinct identification would aid the process of integrating basic information on sites further. This would be an important extension to the federated catalogue based on the concepts presented in the previous chapter.

DEIMS-SDR for instance already issues unique, resolvable (RDF compliant) identifiers for site records from multiple research infrastructures, allowing their cross-RI identification (see Figure 7 DEIMS-SDR record of co-located site Hyytiälä).



Home » Site » Discovery » Hyytiälä SMEAR II LTER - Finland

Hyytiälä SMEAR II LTER - Finland

DEIMS.ID: https://deims.org/663dac80-211d-4c19-a356-04ee0da0f0eb

Basic Information

Site Name: Hyytiälä SMEAR II LTER
Short name : HYY-SMR
Web Address: Hyytiälä Forestry Field station
Hyytiälä SMEAR 2 site
INAR homepage
Country (Site Location): Finland
LTER Member Network: Finland (FinLTSER)
Contact: Site Manager: Janne Levula
SMEAR II, Department of Physics, University of Helsinki

Free Keywords:
PAR (Photosynthetic Active Radiation) photosynthesis evapotranspiration soil respiration respiration rates Radiation below canopy canopy cover forest growth tree ecophysiology sap flow aquatic ecosystems ecosystem processes ecosystem properties stem respiration aboveground production belowground biomass ground vegetation carbon fluxes CH4 fluxes GHG (Greenhouse Gas) Fluxes heat flux N2O flux BVOC emissions VOC flux CO2 flux per soil horizon soil carbon soil chemistry stream water chemistry Disssolved organic Carbon and Nitrogen (DOC, DON) Water: Chlorophyll a content Fluorescence active Fluorescence passive sun-induced fluorescence NPP (Net Primary Production) PRI (Photochemical Reflectance Index) GPP soil microbiology peatland ecology sphagnum diversity mosses Forest management freshwater aquatic GHGs silicon carbonyl sulphide Ozone concentration ozone flux aerosols atmospheric deposition atmospheric processes atmospheric physics atmospheric chemistry

General Site Description:
The SMEAR site (Station for Measuring Ecosystem-Atmosphere Relations) is situated in the Hyytiälä Forestry Field Station of the University of Helsinki. The site consists of a managed, 50-yr old Scots pine forest stand, two open oligotrophic fen sites and a humic lake with forested catchment. The main idea of SMEAR type infrastructures is continuous, comprehensive measurements of fluxes, storages and concentrations in the land ecosystem–atmosphere continuum. The forest measurements are operated continuously since 1996, and include e.g. leaf, stand and ecosystem scale measurements of greenhouse
... Show more

Metadata Update Date: 2019-04-29
Download Metadata:

General Characteristics, Purpose, History

Site Status: existing
Year Site Established: 1910
Size : 20 000.00ha
Purpose of Site :
The site is primarily focused on the studies on land surface–atmosphere relationships, mostly by conducting long-term observation. The overall aim is to provide information on climate and environmental changes; these questions are by definition such that they can be effectively studied only by collecting multidisciplinary, long-term data sets.
History of Site:
The Hyytiälä Forestry Field Station was established in 1910. The SMEAR 2 flux measurement station established in 1995

Photos

Geographic

Coordinates:
Latitude: 61.850300000000
Longitude: 24.295000000000
Download boundaries [.shp]

FIGURE 7 DEIMS-SDR RECORD OF CO-LOCATED SITE HYYTIÄLÄ

The management of the respective site records would still be done in the original catalogue systems with DEIMS-SDR only providing the DEIMS.ID for unique identification. This would allow maximum freedom for the original catalogues, while also ensuring the distinct and unambiguous identification of sites needed for a federated catalogue.

Example: Hyytiälä SMEAR II - Finland

- DEIMS.ID: https://deims.org/663dac80-211d-4c19-a356-04ee0da0f0eb
- ICOS Site Code: http://meta.icos-cp.eu/ontologies/stationentry/AS/ATM-HYY

4 Implementation

Based on the presented concepts, formats and systems and also looking at the limited amount of resources and limited timeframe, we implemented a warehousing concept focusing on only a minimum set of information and developed viable workflow to collect, aggregate and provide site information through a single access point. In addition, the aggregated site records are enriched with rudimentary information to facilitate the search interface.

The harvesting and processing scripts were implemented in Python with the code being available online (see 10.2 Code repositories). The workflow implemented for the harvesting, aggregation and enrichment implemented by those scripts are visualised Figure 8 Workflow for harvesting, aggregation, enrichment and provision of site information.

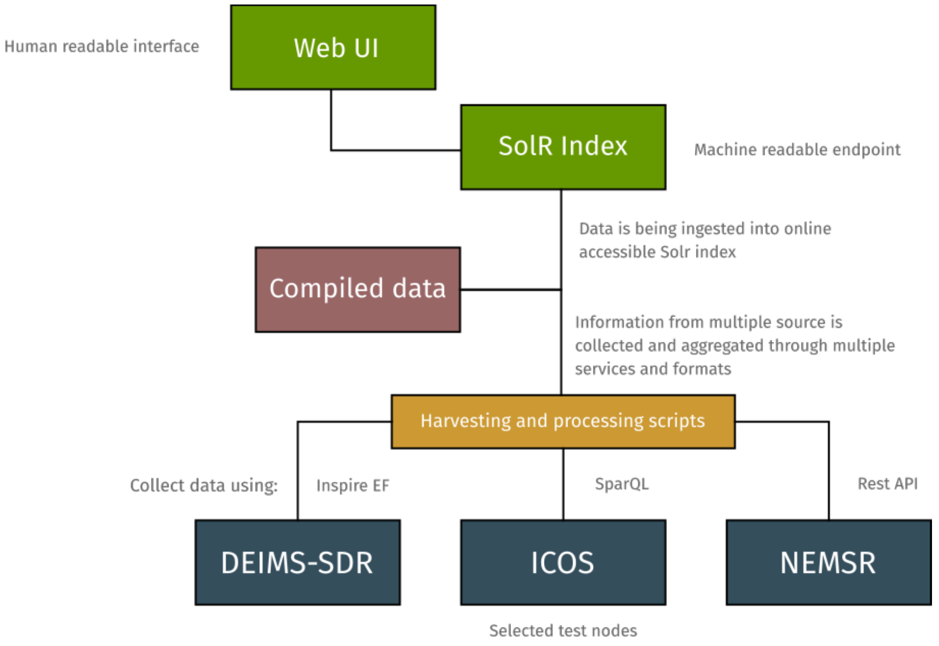

Human readable interface
Web UI
SolR Index
Machine readable endpoint
Data is being ingested into online accessible Solr index
Compiled data
Information from multiple source is collected and aggregated through multiple services and formats
Harvesting and processing scripts
Collect data using:
Inspire EF
SparQL
Rest API
DEIMS-SDR
ICOS
NEMSR
Selected test nodes

FIGURE 8 WORKFLOW FOR HARVESTING, AGGREGATION, ENRICHMENT AND PROVISION OF SITE INFORMATION

Based on the initial evaluation of available site catalogues, a small number of catalogues was selected to be used of the implementation of the prototype. Availability of machine-readable data access as well as basic support by the different user communities and catalogue providers were the main criteria for the selection. Table 7 provides an overview on the selected catalogues.

TABLE 7 SELECTED SYSTEMS AND ACCESS POINT DESCRIPTIONS

Catalogue	URL	Type	Data format	Remarks
DEIMS-SDR	https://deims.org/emf/harvest_list	Harvest List	Inspire EF as XML	Additional endpoints with other formats available as well
ICOS	https://meta.icos-cp.eu/sparqlclient/?type=CSV	SparQL	JSON (and others)	
NEMSR	http://www.neii.gov.au/nemsr/using	REST API	geojson with proprietary format	

For each of the selected catalogue systems a dedicated harvesting routine was implemented:

- For **DEIMS-SDR** all available Inspire EF records provided by a harvest list were harvested. The respective script then parsed the XML records and extracted the relevant information (identifier, name, location, textual description and DEIMS.ID/URL) and stored it in a buffer store.

- For the **ICOS Station Catalogue** a SparQL query was written that extracted all relevant site information as JSON, which was also parsed and stored in the same buffer store.
- **NEMSR** was harvested by querying their Rest API, which returned all relevant site information as JSON, which was also parsed and stored in the same buffer store.

The common harvested site information is limited to "site name", "site description", "url" and centre coordinates, with the URL linking to the original site record in the catalogue of origin. Table 8 provides an overview of these common fields and their suitability for the implementation.

TABLE 8 COMMON SITE FIELDS ACROSS CATALOGUES

DEIMS-SDR	ICOS Station Catalogue	NEMSR
Site name	long name	name
Site description	site type (ICOS specific)	siteDescription
DEIMS.ID	website URL	siteURL
Center coordinates/Site Boundaries	decimal latitude, decimal longitude	latitude/longitude
country (part of metadata model, but not part of Inspire EF export)	country code	(country not part of export, but always Australia)
networks/projects	name list of the networks station belongs to (hardly filled in)	(main network can be derived from the respective API URL, other potential networks, i.e. co-location, of site not part of export)

The aggregated site information in the buffer store is then enriched by intersecting each site record with the "Biogeographical regions[31]" dataset based on the sites centroid coordinates. The "Biogeographical regions" dataset is provided by the European Environment Agency and contains the official delineations used in the Habitats Directive (92/43/EEC).

It is used in order to have a common biome classification across the RIs. This step can't be applied to all sites that are not within the extent of the dataset, i.e. sites outside continental Europe. This was the case for all NEMSR records, with these sites being located in Australia, as well as LTER sites in Israel, ICOS sites in French Guiana and others.

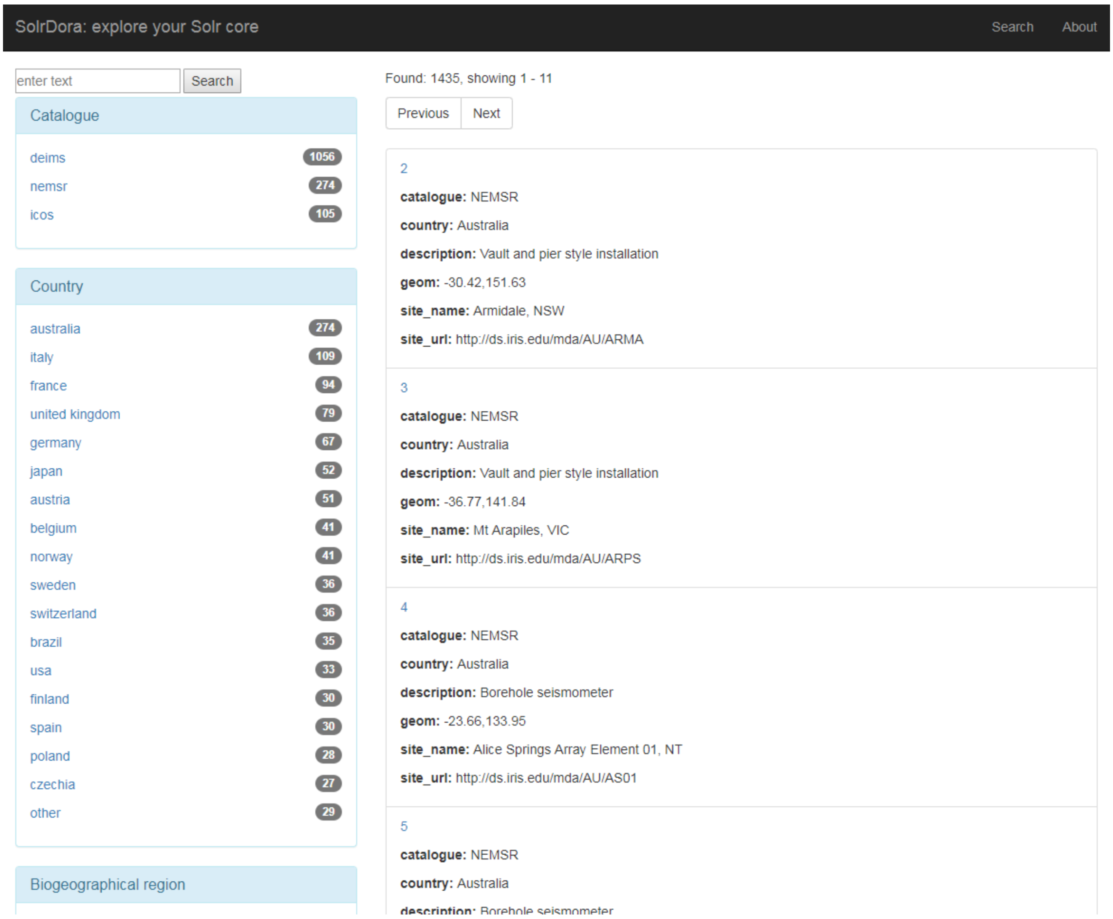
In addition, site records are reverse-geocoded based on the provided site coordinates using the Nominatim API of OpenStreetMap to add the country a site is situated in to each record in order to improve the search interface for users and allow easier querying. All site records that don't contain any coordinates are omitted from this and the following steps. Measures for calculating moving stations, such as research vessels, have not been implemented, but could be solved by allocating multi-values based on the trajectory of a moving station.

In the next step, these aggregated and enriched site records are imported into a Solr 8 index[32], an open-source enterprise-search platform that also allows API based access. In order to make the

[31] https://www.eea.europa.eu/data-and-maps/data/biogeographical-regions-europe-3

[32] https://lucene.apache.org/solr/

final site information accessible and searchable in a user friendly environment we added an instance of SolrDora[33] to the stack (Figure 9 SolrDora search interface).



Found: 1435, showing 1 - 11

Catalogue	
deims	1056
nemsr	274
icos	105

Country	
australia	274
italy	109
france	94
united kingdom	79
germany	67
japan	52
austria	51
belgium	41
norway	41
sweden	36
switzerland	36
brazil	35
usa	33
finland	30
spain	30
poland	28
czechia	27
other	29

Biogeographical region

2

catalogue: NEMSR
country: Australia
description: Vault and pier style installation
geom: -30.42,151.63
site_name: Armidale, NSW
site_url: http://ds.iris.edu/mda/AU/ARMA

3

catalogue: NEMSR
country: Australia
description: Vault and pier style installation
geom: -36.77,141.84
site_name: Mt Arapiles, VIC
site_url: http://ds.iris.edu/mda/AU/ARPS

4

catalogue: NEMSR
country: Australia
description: Borehole seismometer
geom: -23.66,133.95
site_name: Alice Springs Array Element 01, NT
site_url: http://ds.iris.edu/mda/AU/AS01

5

catalogue: NEMSR
country: Australia
description: Borehole seismometer

FIGURE 9 SOLRDORA SEARCH INTERFACE

SolrDora is a small programme written in GO to explore data in a Solr core. It features only limited functionality, but is fast and efficient way to make the contents of a Solr core easily accessible. Other programmes or tools to visualise the Solr content would have been a viable solution for this case as well.

These scripts are designed to be run periodically and update the site records in the index on each run with the aim to keep the site information as recent as possible depended on the defined update interval.

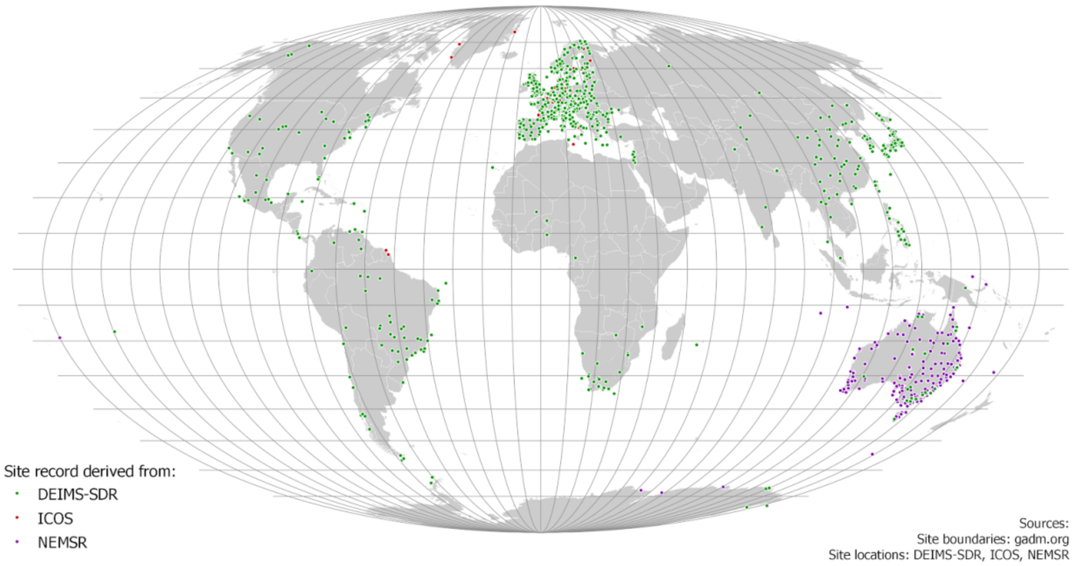
5 Results

5.1 Integration of site description

As a result of the harvesting process, a total of 1489 site records were harvested from the three catalogue systems. The sites are visualised in Figure 10 Map of harvested site records. 1085 site

[33] https://github.com/hectorcorrea/solrdora

records were harvested from DEIMS-SDR, 130 from the ICOS station catalogue and 274 from NEMSR.



Sources:
Site boundaries: gadm.org
Site locations: DEIMS-SDR, ICOS, NEMSR

FIGURE 10 MAP OF HARVESTED SITE RECORDS

In order to evaluate the usefulness of the aggregated site information, a distance matrix was calculated for the harvested sites located in Europe in order to find sites located in close proximity to each other. The three most significant examples of research sites closely located to each other, but situated in different countries are presented next:

- The distance between “Zofin natural forests” in the Czech Republic (https://deims.org/8808a392-5f31-4760-8faf-a6a7bac80f73) and “Weitra” in Austria (https://deims.org/9d008e2a-5b49-4ffe-b73a-87da4ab8ee31) is 2km.
- The distance between “Arctic-alpine tundra” in the Czech Republic (https://deims.org/48660128-e478-42a4-91f8-b83a8735ba50) and “Karkonoski National Park” in Poland (https://deims.org/24901777-6ad9-4c07-b570-595cb9446482) is 5.7 km.
- The distance between “TERENO - Rollesbroich” in Germany (https://deims.org/356417de-5a3c-429d-82c1-08a4e924ab3b) and “Baelen” in Belgium (https://deims.org/5322912e-bd69-4cda-91b7-e7a9e45c782a) is 15km.

Even though the three listed examples originate from the same catalogues, this nevertheless illustrates one of the potential upsides of a federated site catalogue as this can be used to further derive redundancies and overlaps in the coverage of European ecosystems and biomes, which might be useful for the efficient allocation of resources. The datasets used for the biome classification could vary depending on the geographic and thematic context of each query.

5.2 Complementing site information with WMO station data

As an additional measure for quality assurance, WMO station data from the "Observing Systems Capability Analysis and Review Tool (OSCAR[34])" can be used to further enrich existing site information. WMO station records can be retrieved using the OSCAR API[35]. These records can then be matched and compared to the previously harvested site records based on their geographic locations.

Since OSCAR records feature detailed parameter information, this can be used to complement site records with incomplete or non-existing parameter information. To illustrate this, we selected a small number site records from DEIMS-SDR without any parameter description that were located near a WMO station (Table 9).

TABLE 9 DEIMS-SDR AND WIGOS COMPLEMENTARY SITE RECORDS

DEIMS-SDR site name	OSCAR Site name	Country	DEIMS.ID	OSCAR URL	Calculated distance based on site records
Mt. Namsan	SINYONGSAN	South Korea	https://deims.org/f6cb2296-2f80-46ee-adad-854fa564c0ac	https://oscar.wmo.int/surface/#/search/station/stationReportDetails/7044	~30 m
Volbu	LOKEN I VOLBU	Norway	https://deims.org/a7e1e6e2-6275-4cb6-855e-f2f7aa79cfa1	https://oscar.wmo.int/surface/#/search/station/stationReportDetails/1234	~272 m
Kaarvatn	Kårvatn / Kaarvatn	Norway	https://deims.org/cac466c8-ee2a-4133-afd2-4497539e25a1	https://oscar.wmo.int/surface/#/search/station/stationReportDetails/189	~405 m
Predeal-spruce	PREDEAL	Romania	https://deims.org/ba4963e3-0164-4448-a53c-6951c10e9cd0	https://oscar.wmo.int/surface/#/search/station/stationReportDetails/3596	~449 m

Due to the calculated proximities of these sites, it can be assumed that the two corresponding site records actually describe the same site. Especially in the case of "Volbu" it can be assumed that the geographic distance stems from the fact that the DEIMS-SDR record contains coordinates with only two decimal digits (61.12° N, 9.06°E) compared to a more exact location description in the OSCAR record (61.1219444444°N, 9.0630555556°E). In such cases, the observing parameter information from the OSCAR records can therefore be used to complement the DEIMS-SDR records.

As a side note, it should be noted that even though the WMO stations catalogue (OSCAR) also fulfils the catalogue requirements defined earlier, we decided to omit this system from the actual harvesting process due to the different scope of WMO stations and the sites documented in the other selected site catalogues.

[34] https://oscar.wmo.int/surface/#/

[35] https://oscar.wmo.int/surface/rest/api/stations/download/stationsAsKML

5.3 Co-location of sites and resulting duplicate records

Many of the evaluated research sites are used by different networks or research infrastructures since basic infrastructure components (e.g. permanent energy supply or measurement towers) are needed in order to implement observation campaigns. Colocation takes place e.g. at the 'Hyytiälä' station[36] in central Finland, which is listed as both as part of LTER Europe and ICOS[37], as well as other networks (WMO[38], ACTRIS[39] …). As a consequence, this site and others are listed in each of the networks' site catalogues resulting duplicate records in the federated site catalogue.

Another case of site record duplication involves the Australian "Terrestrial Ecosystem Research Network" (TERN). All sites that are part of TERN are also part of the "International Long Term Ecological Research Network" (ILTER) and are registered on DEIMS-SDR. NEMSR and DEIMS-SDR therefore have overlapping site records. However, the way NEMSR records are structured and grouped by networks, with TERN being one of those[40] allows easy identification of TERN sites. As a result, the overlapping sites can easily be identified and duplicate records excluded from being imported into the system by omitting the respective records from the harvesting process[41,42].

In some cases, sites are co-located but feature diverging documentation, e.g. exact coordinates or geographic extent of the site, e.g. Svartberget in Sweden, with its records in the ICOS station catalogue (https://meta.icos-cp.eu/ontologies/stationentry/AS/SE-Sva) and DEIMS-SDR (https://deims.org/c0705d0f-92c1-4964-a345-38c0be3113e1).

It is not possible to easily identify the reason for diverging records, whether it being due to a mistake during the creation of the respective site record or a purposeful difference due to the network setup. Automatically selecting one record over another or even merging those two records would therefore inevitably bear inaccuracies.

In other cases, it is not clear whether two site records actually are describing the same, co-located site at all, such as in the case of "HOBE" (https://deims.org/ce71c6e9-6fcf-401a-9128-db4ac5a355b9) and "Gludsted" (https://meta.icos-cp.eu/resources/stationentry/DK-Gds) in Denmark as the records are similar in some regards, but diverge in others.

Nevertheless, a total of 19 sites that are clearly co-located LTER and ICOS sites could be identified (Table 10). This table of co-located sites allows excluding records from the harvesting process and further minimising the amount of duplicate records.

TABLE 10 LIST OF CO-LOCATED SITES MANAGED BY ELTER AND ICOS

Site Name	DEIMS.ID	LTER Site Code	ICOS Station Code

[36] https://deims.org/663dac80-211d-4c19-a356-04ee0da0f0eb
[37] http://meta.icos-cp.eu/ontologies/stationentry/AS/ATM-HYY
[38] https://oscar.wmo.int/surface/#/search/station/stationReportDetails/926
[39] https://www.actris.se/node/12
[40] http://www.neii.gov.au/nemsr/using
[41] http://neii.bom.gov.au/cgi-bin/nemsr/get_geojson.py?network_id=8
[42] http://neii.bom.gov.au/cgi-bin/nemsr/get_geojson.py?network_id=9

Lonzée	https://deims.org/c3c8a84f-ff66-4d19-8c28-42c7ed63b43d	LTER_EU_BE_33	https://meta.icos-cp.eu/ontologies/stationentry/ES/IT-Tor
TERENO - Rollesbroich	https://deims.org/356417de-5a3c-429d-82c1-08a4e924ab3b	LTER_EU_DE_023_002	https://meta.icos-cp.eu/ontologies/stationentry/ES/DE-RuS
Hohes Holz, Germany	https://deims.org/ddd2e8d2-44db-420e-8fa4-6b4fe1b00c78	DE-HoH-1	https://meta.icos-cp.eu/ontologies/stationentry/ES/DE-HoH
Brasschaat - De Inslag	https://deims.org/68e6a8e5-d6d2-4c8c-91c4-10e7f87ac556	LTER_EU_BE_001	https://meta.icos-cp.eu/ontologies/stationentry/ES/BE-Bra
Birkenes	https://deims.org/68af6e55-e241-4afe-a3a6-32e79eef12fb	LTER_EU_NO_003	https://meta.icos-cp.eu/ontologies/stationentry/AS/BIR
OZCAR-RI PEATLAND La Guette	https://deims.org/331c1b2b-2283-4396-8e8b-d9d3d040e3cd	CZO_EU_FR_243	https://meta.icos-cp.eu/resources/stationentry/FR-LGt
Simon Stevin Research Vessel	https://deims.org/64428d5d-9c8c-4f9a-9d55-e866c80ca342	LTER_EU_BE_10_002	https://meta.icos-cp.eu/ontologies/stationentry/OS/Simon+Stevin
VLIZ Thornton Bouy	https://deims.org/177ff4a8-9481-495e-a55a-ec7d32bf6e30	LTER_EU_BE_10_001	https://meta.icos-cp.eu/ontologies/stationentry/OS/Lotto+buoy
Observatory HAUSGARTEN	https://deims.org/f6d9ed12-6bc1-47fb-8e81-ef24e9579596	LTER_EU_DE_014	https://meta.icos-cp.eu/ontologies/stationentry/OS/HG
Svartberget Field-research Infrastructure	https://deims.org/c0705d0f-92c1-4964-a345-38c0be3113e1	LTER_EU_SE_006	https://meta.icos-cp.eu/ontologies/stationentry/ES/SE-Sva
Torgnon grassland Tellinod (IT19 Aosta Valley)	https://deims.org/a03ef869-aa6f-49cf-8e86-f791ee482ca9	LTER_EU_IT_077	https://meta.icos-cp.eu/ontologies/stationentry/ES/IT-Tor
Renon BOL1	https://deims.org/5d32cbf8-ab7c-4acb-b29f-600fec830a1d	LTER_EU_IT_029	https://meta.icos-cp.eu/ontologies/stationentry/ES/IT-Ren
OZCAR-RI Aurade Experimental Catchment	https://deims.org/6f4ee641-2339-4006-b815-6e2ca6c6b0bf	CZO_EU_FR_040	https://meta.icos-cp.eu/ontologies/stationentry/ES/FR-AUR

OZCAR-RI Regional Spatial Observatory in the South West France	https://deims.org/bf457b49-2074-4d2e-bd57-adc3a9cffa9a	CZO_EU_FR_250	https://meta.icos-cp.eu/ontologies/stationentry/ES/FR-AUR
Värriö Research Station	https://deims.org/b471311f-e819-4f6f-bbae-1ac86cd9777f	LTER_EU_FI_020	https://meta.icos-cp.eu/ontologies/stationentry/ES/ECO-SMEAR+I
TERENO - Selhausen	https://deims.org/0a006b69-5134-4c0a-864c-f86c0c61288f	LTER_EU_DE_023_001	https://meta.icos-cp.eu/ontologies/stationentry/ES/DE-RuS
Davos Seehornwald	https://deims.org/a547dab2-859a-414c-b148-0e7df8de5773	LTER_EU_CH_006	https://meta.icos-cp.eu/ontologies/stationentry/ES/DAV
Vielsalm Terrestrial Observatory	https://deims.org/c4c1c0ca-5a43-4d19-ab50-34dad0af44e8	LTER_EU_BE_16	https://meta.icos-cp.eu/ontologies/stationentry/ES/BE-Vie
Hyytiälä SMEAR II	https://deims.org/663dac80-211d-4c19-a356-04ee0da0f0eb	LTER_EU_FI_007	https://meta.icos-cp.eu/ontologies/stationentry/AS/ATM-HYY

6 Conclusion

In successfully building the prototype, we were also able to come up with a viable workflow for the aggregation of site information. We identified a set of common fields across the different catalogues, addressed the key issues and formulated a set of recommendations for site catalogues in general as well as future work in this field.

Looking at the metadata schemas presented in chapter 2.1 Relevant metadata schemas, as well as the data models implemented by the three site catalogues, a set of common attributes could be identified, namely:

- site name
- location
    - either point based providing the centroid coordinates
    - or the site boundaries or bounding box
- network specific identifier
- website URL
- textual description
- And to a lesser degree also “network affiliation”

Even though information about network affiliation is part of the site data models in all three selected catalogue system the actual network information is limited. While for instance in DEIMS-SDR a total of 19 sites indicated ICOS affiliation, none of the corresponding site records in the ICOS station catalogue mentioned “LTER Europe” or any national LTER network as part of their network affiliation. Albeit network information is one of the core features of the NEMSR system, co-location with other networks is actually difficult to identify as site information is organised hierarchically based on the affiliation with a particular network.

In this particular prototype implementation, it was easy to identify the duplicate site records in DEIMS-SDR and NEMSR due to the existing in-depth knowledge of the ILTER network, but in the case of duplicate records in DEIMS-SDR and ICOS this already posed a much greater challenge. Programmatically identifying co-location therefore remains a challenge for further harmonisation, e.g. in case of an implementation on a bigger scale with additional catalogues (see 10 Annex).

6.1 Possibilities for extensions of the current prototype

The considerations regarding the notion of a federated site catalogue and the preliminary results acquired from the development of the initial prototype suggest possible extensions along various dimensions.

Adding more site catalogues, preferably those featuring non-European sites (see 10 Annex), would be an obvious extension. As mentioned in (1.1.3 Aggregation), however, a growing number of aggregated site catalogues would likely result in a smaller overlap of similar metadata fields present in all of them, requiring additional effort to identify and map them to each other and/or to design an appropriate common data schema for aggregation.

As outlined in 1.1.4 Contextualisation, another valuable extension would be to contextualise the aggregated metadata with both provider specific as well as global context sources, such as domain taxonomies or person registries, the latter requiring special attention due to GDPR related issues. Provider specific resources would have to be made available by them for retrieval and their metadata ideally already include unambiguous links to the entities featured there, since automated mappings to both local and global resources would require a separate layer of quality control.

Especially the enrichment with additional contextual data sources would embed the accumulated metadata into a larger knowledge graph, which suggests shifting the data representation of the aggregation infrastructure towards a graph structure and related storage technologies, such as for example existing in the context of RDF.

Advanced matching functionalities to identify related sites across catalogues would be another important extension. This could include automated co-location checks based on geographic information and other relevant attributes, as well as groupings based on contextual information.

The overall usefulness of a site records is not only defined by the extent of the site data model, but also by the level of completeness for each record. The complementation of site information with WMO station data therefore promises to be a useful measure in order to refine existing site information. It should be pointed out though, that this is only one minor measure in order to achieve harmonisation as measures for semantic harmonisation are needed in order to ensure the usability of the complemented site records. This would also address the harmonisation of used thesauri and vocabularies and therefore require a substantial amount of work.

7 Recommendations for harmonisation across RIs

One of the biggest issues if not the single biggest issue of a federated catalogue is the distinct identification of co-located research sites across different research infrastructures. Without such identification and subsequent measures to merge and filter records describing the same site a federated catalogue will feature duplicate records and potentially also mismatching information in those records and therefore fail to achieve one of its core goals to provide a viable single access point for site information.

Therefore, the usage of a resolvable, unique and persistent, cross-RI site identifier, analogous to a DOI or an ORCID, is highly recommended for site records. Otherwise, a federated catalogue will have to identify co-location based on geographic proximity, similar site names and/or network affiliation, which are prone to errors as laid out in chapter 5.3 "Co-location of sites and resulting duplicate records". The usage of a cross-RI identifier for catalogues in the environmental domain would facilitate the automated deduction of network affiliations for sites as well as easily minimise the number of site record duplicates. However, this does not apply to the WMO stations as those stations focus on meteorological measurements and therefore have a different scope than ICOS or LTER Europe sites. Other environmental RIs, that weren't considered for this deliverable, such as ACTRIS or INTERACT would also greatly benefit from using such identifiers.

As already suggested in chapter 3.2.3 "Integration based on a centralised registry", DEIMS-SDR could act as an issuing service for such cross-RI identifiers as it already issues the "DEIMS.ID" for all of its site records. Furthermore, DEIMS-SDR is already recommended as a service to generate site identifiers in the technical specifications document for the (NEC) Directive, Article 10 (4a) (NEC Directive Technical Specifications, 2018). A successful implementation of the DEIMS.ID is also illustrated by its use for Non-LTER protected areas in the H2020 project EcoPotential.

In addition to the usage of such an identifier, a common standard for site information should also be used. While Inspire EF or WIGOS could serve as a common format, those formats unfortunately have the downside of featuring only a limited set of site information. The willingness of research infrastructures and site catalogue developers to implement such standard is also uncertain. So even though the usage of a common metadata format is recommended, it is unlikely to cover all necessary aspects that are needed for the successful implementation of a federated site catalogue.

As a consequence, we would instead like to propose a set of minimal information that a site catalogue should always include regardless of the implemented data model or supported data formats. This set of minimal information includes:

- **Site Name**
- **Location**
    - Using centroid coordinates and/or boundary information
- **Textual description**
- **A unique, resolvable, persistent identifier that is independent from any affiliation**
    - analogous to a DOI or an ORCID
- **Network/project/RI affiliation**
    - In a first step, the name of networks or projects as a free text would be sufficient, but eventually persistent identifiers for networks or projects should also be adopted. For Research Infrastructures a dedicated ORCID could already be used.

    - In addition, any network or RI-related site identifier could also be listed here (LTER Site Code, ICOS station code, INTERACT station code …).
- **Standardised information about deployed infrastructure** (e.g. measurement **equipment) and observed properties**
    - Documentation on infrastructure and observed properties should be using controlled vocabularies and thesauri.
    - Such vocabularies and thesauri should ultimately be mapped to each other or even harmonised further.

Furthermore, site catalogues should always have an API that allows querying the entirety of available site information or at least a subset of site information that includes the proposed set of minimal information to further facilitate the exchange of site information.

Implementing these suggestions would already foster interoperability between the research infrastructures and their catalogues massively without posing too much of a burden for the RIs and the catalogue developers.

These steps towards harmonisation between research infrastructures would quickly reduce the administrative efforts necessary for providing documentation about research sites that is often necessary for project proposal and subsequent project administration and reporting as well as allowing easier inventories of research infrastructures.

8 Acknowledgements

- Australian Bureau of Meteorology
- Australian National Environmental Information Infrastructure
- Australian Terrestrial Ecosystem Research Network
- eLTER
- ICOS
- WMO

We would especially like to express our sincere thanks to the staff of the Australian National Environmental Information Infrastructure for their voluntary aid in adding NEMSR to the selected test systems for this deliverable and the overall fruitful and pleasant collaboration.

9 REFERENCES

Cox, Simon Jonathan David (2011). "ISO 19156:2011 Geographic information – Observations and measurements". doi:10.13140/2.1.1142.3042. Retrieved 2019-06-03.

GEO (2017) 2017-2019 GEO Work Programme. Document GEO-XIV-5.4_rev. (Online https://www.earthobservations.org/documents/work_programme/geo_2017_19_Work_Programme.pdf) [Last accessed 18.6.2019]

INSPIRE Directive (2007) (https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=OJ:L:2007:108:TOC Retrieved 2019-06-03

INSPIRE EF (2019) http://inspire.ec.europa.eu/theme/ef Retrieved 2019-06-03

Kirchner, T., Ryl, R., Wohner, C., & Peterseil, J. (2018). ICP Forests and LTER – a first handshake between data infrastructures. Unpublished. https://doi.org/10.13140/rg.2.2.13963.13604

NEC Directive Technical Specifications, 2018. http://ec.europa.eu/environment/air/pdf/Technical%20Specifications%20NEC%20Article%209%20location%20and%20indicators%20final.docx, Accessed date: 21 November 2018.

NEMSR (2019) http://www.neii.gov.au/nemsr Retrieved 2019-06-03

Poursanides et al. (2016) Metadata for pre-existing datasets. ECOPOTENTIAL Project Deliverable D5.2. 116pp.

Schaeffer, B., Baranski, B., Foerster, T., & Brauner, J. (2011). A Service-Oriented Framework for Real-Time and Distributed Geoprocessing. In Lecture Notes in Geoinformation and Cartography (pp. 3–20). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10595-1_1

Veltman, K. H. (2001). Syntactic and semantic interoperability: New approaches to knowledge and the semantic web. New Review of Information Networking, 7(1), 159–183. https://doi.org/10.1080/13614570109516975

WIGOS Metadata Standard 2017, https://library.wmo.int/doc_num.php?explnum_id=3653 Retrieved 2019-06-03.

Wohner, C., Peterseil, J., Poursanidis, D., Kliment, T., Wilson, M., Mirtl, M., & Chrysoulakis, N. (2019). DEIMS-SDR – A web portal to document research sites and their associated data. Ecological Informatics, 51, 15–24. https://doi.org/10.1016/j.ecoinf.2019.01.005

Zilioli, M., Oggioni, A., Tagliolato, P., Pugnetti, A., & Carrara, P. (2019). Feeding Essential Biodiversity Variables (EBVs): actual and potential contributions from LTER-Italy. Nature Conservation, 34, 477–503. https://doi.org/10.3897/natureconservation.34.30735

10 Annex

10.1 List of relevant site catalogues

In the preparation phase for the development of the concept for a distributed site catalogue a screening of available online site catalogues was carried out. This overview does not claim to be an all-encompassing list of catalogues but instead provide an overview of the systems and site catalogues that were considered and examined for this work.

TAB 11 LIST OF SITE CATALOGUES

Research Infrastructure (RI)	Short description RI	Link site catalogue	Unified data model available online	Format
DEIMS-SDR (=Dynamic Ecological Information Management System - Site and dataset registry)	information about sites and datasets of networks dealing with ecological long term observation and experimentation in Europe and globally	https://deims.org	yes, https://deims.org/models	CSV; ISO 19115-2 North American Profile and ISO 19139 Inspire Profile, BDP, EML 2.1.1 and Inspire EMF
ECOMET (= economic interest grouping of the National Meteorological Services of the European Economic Area)	26 Members in all Europe; envisages the widest availability of basic meteorological data for re-use applications;2 different data sets (RBSN [=Regional Basic Synoptic Network] & European GTS data set)	http://www.ecomet.eu/ecomet-catalogue/download-facilities	no	various formats
German Leibniz-Institute of Freshwater Ecology and Inland Fisheries	located in 120 countries, global network for long-term environmental monitoring and research, education, and public information	http://bfs.igb-berlin.de/index.php/station-catalogue-biological-field-stations.html	no	html
GLOBE (= Global Collaboration Engine)	interdisciplinary research group based at UMBC (= University of Maryland, Baltimore County); explores the ecology of populated landscapes at local, regional and global scales towards the goal of making ecosystem management more sustainable. The project is over	http://ecotope.org/projects/globe/	no	PDF, mxd, zip

GMP DWH (= Global Monitoring Plan Data Warehouse)	online tool developed for handling persistent organic pollutants (POPs) monitoring data generated in the frame of the Global Monitoring Plan under the Stockholm Convention on POPs; GMP DWH is a tool for storage, analysis and visualization of global POPs data online; contains information on POPs on centrations in ambient air, human tissues (breast milk and maternal blood) and surface water for water-soluble POPs (perfluorooctane sulfonic acid, its salts and perfluorooctane sulfonyl fluoride) collected in the frame of the GMP and validated by the regional organization groups of the five UN regions	http://visualization.pops-gmp.org/2014/spatial-distribution/8e5c75dac947bbafa17ee8aefa887138/@@@DSI0/default/none/	no	CSV
Government of Canada	lists all Canadian surface and upper air stations providing synoptic meteorological reports, their operational activities and essential detailed information	https://open.canada.ca/data/en/dataset/9764d6c6-3044-450c-ac5a-383cedbfef17	no	html and csv
GRDC (= Global Runoff Data Centre)	International data centre operating under the auspices of the World Meteorological Organization (WMO), support the research on global and climate change and integrated water resources management	https://www.bafg.de/GRDC/EN/02_srvcs/21_tmsrs/212_prjctlgs/project_catalogue_node.html	no	XLSX, PDF
ICOS	pan-European research infrastructure for quantifying and understanding the greenhouse gas balance of Europe and its neighbouring regions; 12 member countries and involves more than 120 measurement locations where greenhouse gas	https://www.icos-cp.eu/node/83	yes, https://github.com/ICOS-Carbon-Portal/meta/blob/master/src/main/resources/o	

	concentrations and fluxes are measured		wl/station Entry.owl	
Interact	founded by the EU; has 83 terrestrial field bases throughout the Arctic; main objective to build capacity for identifying, understanding, predicting and responding to diverse environmental changes throughout the wide environmental and land-use envelopes of the Arctic	https://eu-interact.org/field-sites/	yes	
IR (=International Seismograph Station Registry)	result of a special effort to adapt and substantially extend and improve currently existing bulletin data for large earthquakes (magnitude 5.5 and above) to serve the requirements of the specific user group who assess and model seismic hazard and risk; also multidisciplinary use in a wide range of other areas such as studies of global seismicity, tectonics, inner structure of the Earth, nuclear test monitoring research, rapid determination of hazard etc.;	http://www.isc.ac.uk/iscgem/download.php	yes	CSV
MWRnet (= International Network of Ground-based Microwave Radiometers)	set up of an operational network sharing knowledge, software, procedures, formats, quality control, etcetera, similarly to other successful networks such as CWINDE, EARLINET, AERONET	http://cetemps.aquila.infn.it/mwrnet/MWRnetmap.html	no	
NOAA (= National Centers For Environmental Information; National Oceanic and Atmospheric Administration)	Global Observing Systems Information Center (GOSIC)	https://www.ncdc.noaa.gov/gosic	no	

NEMSR (National Environmental Monitoring Sites Register)	The National Environmental Monitoring Sites Register (NEMSR) provides a consolidated view of Australia's environmental monitoring sites.	http://www.neii.gov.au/nemsr	yes http://www.neii.gov.au/nemsr/information-model	json
Network for the Detection of Atmospheric Composition Change	The international Network for the Detection of Atmospheric Composition Change (NDACC) is composed of more than 70 globally distributed, ground-based, remote-sensing research stations with more than 160 currently active instruments providing high quality, consistent, standardized, long-term measurements of atmospheric temperatures and trace gases, particles, spectral UV radiation.	http://www.ndaccdemo.org/	no	txt
UK Environmental Change Network (ECN) monitoring sites	This dataset provides the location details of Environmental Change Network (ECN) sites from which data are collected. There are 12 terrestrial sites and 45 freshwater sites. Sites range from upland to lowland, moor land to chalk grassland, small ponds and streams to large rivers and lakes. ECN is the UK's long-term environmental monitoring programme.	https://catalogue.ceh.ac.uk/documents/813712d4-d162-4ede-aff8-cf1c337bdc27	no	Shapefile

WMO (= World Meteorological Organization)	OSCAR/Surface is the official repository of metadata on surface-based meteorological and climatological observations that are required for international exchange. These observing systems are integrated under the WMO Integrated Global Observing System (WIGOS) framework. OSCAR/Surface is one of the components of the WIGOS Information Resources	https://oscar.wmo.int/surface/#/	yes	json, kml, ...

10.2 Code repositories

Harvesting scripts: https://github.com/stopopol/dist_site_catalogue

SolrDora: https://github.com/hectorcorrea/solrdora