Integrating Activities for Advanced Communities

D3.2 - How to get the data?
INTERACT Metadata Tag Pilot Technical Manual

Project No.871120– INTERACT

H2020-INFRAIA-2019-1

Start date of project: 2020/01/01 Duration: 23 months
Due date of deliverable: 2021/11/30 Actual Submission date: 2021/11/30

Lead partner for deliverable: 58 - INKODE
Author: Giorgio Resci, Raoul Nuccetelli

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the Consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

Table of Contents

Publishable Executive Summary 2

Introduction 4

Background 4

Purpose 4

Content 4

Metadata Tag Pilot Technical Manual 4

Metadata & Metadata standards 4

Metadata Harvesting 5

Data & metadata repository 6

Medatada discovery services 6

INTERACT Handbook 7

Publishable Executive Summary

INTERACT's 89 research stations collect and analyze a large amount of high-quality and valuable data important in various disciplines and especially for global climate change research. Both data value and costs are considerable due to the difficulties of access to the Arctic, thus making collected data open, free, easily searchable and reusable is a key point for the INTERACT III project. To achieve this, INTERACT project and Station Manager Forum have worked over the years producing deliverables that have involved the entire lifecycle of data management: *Data Management Plan*[1], *Report On Current Data Flows*[2], *INTERACT Field Guide To Data Repositories*[3], *INTERACT Data Policy*[4]. Continuing to improve data promotion and usability, this deliverable describes the approach used to collect and publish available metadata and provides standards and procedure on how to make metadata available, the annex document is intended as a manual for station's manager on how to organize and make metadata available.

[1] D4.1 Data Management Plan https://eu-interact.org/app/uploads/2018/11/D4.1.pdf
[2] D4.2 Report On Current Data Flows https://eu-interact.org/app/uploads/2018/11/D4.2.pdf
[3] D4.3 INTERACT Field Guide To Data Repositories https://eu-interact.org/app/uploads/2018/11/D4.3.pdf
[4] D4.4 – INTERACT Data Policy https://eu-interact.org/app/uploads/2019/02/D4.4.pdf

1. Introduction

1.1. Background

INTERACT Virtual Access Single Entrypoint, available at https://dataportal.eu-interact.org/, is a data portal that allows users to search and access data and information from the Arctic and beyond. The INTERACT Data Portal is based on metadata harvesting from several organizations and sites with end points to real data on various topics on e.g. earth sciences and ecosystems. The number of datasets presented in the portal is growing day by day, including both near real-time observation datasets and unique historical, retrospective data digitized and provided for open access.
INTERACT Virtual access single entrypoint works as a metadata repository and a search engine, it is powered by metadata discovery services provided by the research stations involved.

1.2. Purpose

INTERACT Virtual Access Single Entrypoint is basically a search engine for datasets produced by research stations.

Stations have the task of:

- collecting and cataloging data
- describing data through metadata standard
- publishing a repository where to host the data
- providing an interface for metadata discovery

The purpose of this manual is to help station managers on procedures and solutions found within the INTERACT Data Team and tested in the metadata tag pilot in order to be adopted by all INTERACT research stations.

1.3. Content

The manual touches on the inner workings of INTERACT Virtual Access Single Entrypoint and relevant key arguments that often appear in the context of metadata sharing. This allows the end user to gain a more informed perspective on the possible programmatic consumption of the metadata records provided by the portal and steps needed to be taken in setting up a data repository to integrate metadata records that a station may want to provide in the portal.

2. Metadata Tag Pilot Technical Manual

2.1. Metadata standards

Scientific data must be strictly accurate to be reliable and usable. Publishing data about e.g. CO_2 without a unit of measurement or a geographic location or a sampling period is almost useless. Metadata is "data that provides information about other data", and includes information such as author, title, description, contact information, UM, location and all other information that describes data.
Metadata standard is a requirement which establishes a common way of structuring and understanding data and includes agreeing on language, spelling, date format, etc in order to improve (re)-usability and

interoperability between different datasets, tools and services. There are many metadata standards[5] often discipline related.

In INTERACT we have adopted two schema metadata at the moment:

- Schema.org[6] / science-on-schema.org[7] over JSON-LD[8]
- DCAT-AP[9] over JSON-LD or RDF/XML[10]

Choice is guided by a survey among the early birds research stations, but in the future other schemes may be added.

2.2. Metadata Harvesting

Metadata Harvesting refers to gathering metadata from multiple places or archives and storing it in a central database.

INTERACT Virtual Access Single Entrypoint allows you to search for datasets through free search and attribute filters, filters are fed by metadata provided by the search stations following one of the two schemes previously indicated. In this table it is possible to compare the harvested metadata with the equivalent attribute in the metadata scheme.

Type	Field Name	DCAT-AP	JSON
Text	name	dct:title	name
Text	description	dct:description	description
Text	keywords	dct:keyworks	keywords
URL	url	dct:landingpage	url
Text	citation	dct:identifier	citation
Text	creator	dct:creator	creator
Text	topicCategory	dct:theme	about
Text	temporalCoverage	dt:temporal	temporalCoverage
Text, Place	spacialCoverage	dct:spacial	spatialCoverage

[5] https://www.dcc.ac.uk/guidance/standards/metadata/list
[6] https://schema.org/
[7] https://github.com/ESIPFed/science-on-schema.org/
[8] https://json-ld.org/
[9] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/11
[10] https://www.w3.org/TR/rdf-syntax-grammar/

Text	datePublished	dct:issued	datePublished
Text	dateModified	dct:modified	dateModified
Text	publisher	dct:publisher	publisher
Text	provider	dcat:hadRole	provider
URL	license	dct:license	license
Text	variableMeasured	n/a	variableMeasured

2.3. Data and metadata repository

A repository is a service (usually a web application) that allows storing, cataloging, describing and publishing data and provides functionality to search and interact with said data.
There are advanced open source software solutions like CKAN, but research stations can also opt for custom solutions implemented to be hosted specifically on their data systems. Other than making data more orderly through categorization, one main goal usually is to expose one's data to the general public and making it more visible and accessible; All interactions between the data repository and the data consumers happen through the internet thanks to a variety of protocols, usually HTTP.

2.4. Metadata discovery services

The purpose of a metadata discovery service is to enable data consumers to search across a wide array of records and to make metadata harvestable through standard services. One of the most advanced protocols that enables these transactions is OAI-PMH[11], but also the use of a simple index can be an efficient solution, such as the XML sitemap[12] that lists all the available metadata

[11] http://www.openarchives.org/pmh/
[12] https://www.rd-alliance.org/system/files/documents/Guidelines%20for%20publishing%20structured%20data_V1.0_20210215.pdf

INTERACT GUIDEBOOK:

MAKING METADATA AVAILABLE FOR VIRTUAL ACCESS IN INTERACT DATA PORTAL

WRITTEN BY

Hannele Savela

Raoul Nuccetelli

Giorgio Resci

NOV 15, 2021

SUMMARY

Chapter	Topic	Page
1 - Basic Concepts		
1.1	What is data?	4
1.2	What is metadata?	4
1.3	What is data encoding?	5
1.4	What are metadata standards?	5
1.5	An overview of JSON	6
1.6	What is JSON-LD?	6
1.7	An overview on XML	7
1.8	What are DCAT and DCAT-AP?	7
1.9	What is an API?	8
1.10	The FAIR guiding principles	8
2 - Technical features of the INTERACT Data Portal		
2.1	Metadata guidelines	11
2.2	What fields are harvested to Interact Data Portal?	12
2.3	How is metadata harvested to Interact Data Portal?	13
2.4	Creating an entrypoint for metadata harvesting by Interact Data Portal	14

2.5	Solutions for hosting your metadata	16
3 - INTERACT Data Portal API endpoints		
3	API Documentation	17
	Acknowledgements	23
	References	23
	Appendix	23

TO THE READER

This guidebook is aimed to help the Data Managers, Station Managers and Station Contacts of INTERACT partner organizations providing Virtual Access to prepare their (meta)data and data repositories for harvesting by the INTERACT Data Portal. Altogether 29 partner organizations offer Virtual Access in INTERACT in 2020-2023, funded by the EU-H2020 (Grant Agreement No. 871120). The guidebook also provides practical information for data publishers and other organisations external to the INTERACT network who are interested in harvesting metadata from the INTERACT Data Portal.

Virtual Access means free and open on-line access to data and metadata from the research stations and partners operating the stations. In INTERACT, Virtual Access is provided via the INTERACT Virtual Access Single Entrypoint Data Portal, in short the INTERACT Data Portal.

The INTERACT Data Portal, where the INTERACT Virtual Access is available, is based on metadata harvesting from the partner organizations' repositories, with end points to real data on various topics e.g. earth sciences and ecosystems. The number of datasets presented in the portal is growing day by day, including both near real-time observation datasets and unique historical, retrospective data digitized and provided for open access.

You are warmly welcome to read this guidebook, and we hope you find it useful in initiating your institution's or organisation's Virtual Access provision via INTERACT Data Portal!

15th November 2021,

Hannele Savela, INTERACT Virtual Access coordinator

Raoul Nuccetelli, INTERACT Data Portal system developer, Inkode

Giorgio Resci, INTERACT Data Portal system developer, Inkode

1 - BASIC CONCEPTS

1.1 WHAT IS DATA?

Data is the accessible expression of qualitative or quantitative variables about one or more persons, objects, facts or events, expressed in a consistent manner.

Data as a general concept refers to the fact that some existing information or knowledge is *represented* or *coded* in some form suitable for usage or processing.

In computing data is represented via a series of bits. A bit is the smallest unit of data that has a single binary value, either 0 or 1. Through the various combinations of bits we're able to have a wide series of ways people or machines can interact with more structured data, like for example videos, photos, documents and so forth.

1.2 WHAT IS METADATA?

Metadata is data that provides information about the data, specifically the what, where, how, when, by whom it was collected, its current location, and any access information.
Metadata facilitates the understanding, use, and management of data and is a means for networking and collaboration

There are three main types of metadata: descriptive, structural and administrative.

- Descriptive metadata adds information about who created a resource, and most importantly – what the resource is about, what it includes.
- Structural metadata includes additional data about the way data elements are organized – their relationships and the structure they exist in.
- Administrative metadata provides information about the origin of resources, their type and access rights.

The database in which metadata is stored is referred to as metadata registry or metadata repository. The degree to which the data or metadata is structured and detailed is referred to as its granularity. High granularity metadata implies higher quality and more structured information and enables greater level of technical manipulation.

1.3 WHAT IS DATA ENCODING?

All data in digitised systems is represented by a series of bits. A bit is the smallest unit of data in a computer and has a single binary value, either 0 or 1. The way our computer manages to convert these series into human readable characters is through character encoding, a mapping to associate a certain

series of bits to a certain character. INTERACT Data Portal encourages the use of the UTF-8 character encoding, as it makes certain processes less error-prone .

1.4 WHAT ARE METADATA STANDARDS?

A metadata standard is a requirement which is intended to establish a common understanding of the meaning or semantics of the data, to ensure correct and proper use and interpretation of the data by its owners and other users and applications. To achieve this common understanding, a number of characteristics, or attributes of the data are defined.

A metadata standard or schema outlines how data should be structured. Different schemas differ in the type and number of data elements, the designation of mandatory fields, encoding requirements, and the use of data content and value standards. Therefore, a decision about selecting a schema has implications for the quality and level of description.

These are the metadata standards that are currently supported by INTERACT Data Portal:

- schema.org / science-on-schema.org over JSON-LD
- DCAT-AP over JSON-LD or RDF/XML

1.5 AN OVERVIEW OF JSON

JSON is a standard way in which information is represented and encoded for storage in a computer file. It facilitates structured data interchange between different web applications or data consumers.

The way data is represented matches the syntax of objects, in a similar fashion to Javascript.

An "object" in this context is a collection of named properties (“key”), each of which have an assigned value that can be either a string, number, array (a collection of multiple values), booleans, or other objects.

1.6 WHAT IS JSON-LD

JSON is a lightweight data interchange format that makes it easy to parse and generate data. However, it is difficult to integrate JSON that has been harvested from different sources as the data may contain keys that conflict with other data sources. Furthermore, JSON has no built-in support for hyperlinks, which are a fundamental building block on the Web.

Linked Data is structured data which is interlinked with other data so it becomes more useful through queries of associative and contextual nature

Linked Data is a way to create a network of standards-based machine interpretable data across different documents and websites. It allows an application to start at one piece of Linked Data, and follow embedded links to other pieces of Linked Data that are hosted on different sites across the internet.

JSON-LD is a lightweight syntax to serialize Linked Data in JSON. Its design allows existing JSON to be interpreted as Linked Data with minimal changes. JSON-LD is primarily intended to be a way to use Linked Data in Web-based programming environments, to build interoperable Web services, and to store Linked Data in JSON-based storage engines.

When two people communicate with one another, the conversation takes place in a shared environment, typically called "the context of the conversation". This shared context allows the individuals to use shortcut terms, like the first name of a mutual friend, to communicate more quickly but without losing accuracy. A context in JSON-LD works in the same way. It

allows two applications to use shortcut terms to communicate with one another more efficiently, but without losing accuracy.

1.7 AN OVERVIEW OF XML

The Extensible Markup Language (XML) is a metalanguage, a language whose goal is to define other languages, called Object languages. It uses tags to label, categorize, and organize information in a structured way. When choosing the structure that defines the encoding rules, one can appeal to one of the many standards that have already been defined or create one more suitable for the use case.

This flexibility has made XML one of the most widely-used formats for sharing structured information today: between programs, between people, between computers and people, both locally and across networks.

1.8 WHAT ARE DCAT AND DCAT-AP ?

DCAT is an RDF (a standard for data interchange that is used for representing highly interconnected data) vocabulary designed to facilitate interoperability between data catalogs (a collection of metadata, combined with data management and search tools, that helps analysts and other data users to find the data that they need and serves as an inventory of available data) published on the Web.

DCAT enables a publisher to describe datasets and data services in a catalog using a standard model and vocabulary that facilitates the consumption and aggregation of metadata from multiple catalogs. This can increase the discoverability of datasets and data services. It also makes it possible to have a decentralized approach to publishing data catalogs and makes federated search for datasets across catalogs in multiple sites possible using the same

query mechanism and structure. Aggregated DCAT metadata can serve as a manifest file as part of the digital preservation process

1.9. WHAT IS AN API ?

An API (Application Programming Interface) is a set of defined rules that explain how computers or applications communicate with one another. In contrast to a user interface, which connects a computer to a person, an application programming interface connects computers or pieces of software to each other. APIs sit between an application and the web server, acting as an intermediary layer that processes data transfer between systems.

Here's how an API works:

A client application initiates an API call to retrieve information—also known as a request. This request is sent from an application to the web server via the API's Uniform Resource Identifier (URI). The request includes a request verb, headers, and sometimes, a request body; all elements that indicate the type of action to be performed on or with the data. that differentiates the different types of possible responses

After receiving a valid request, the API makes a call to the external program or web server.

The server sends a response to the API with the requested information.

1.10 THE FAIR GUIDING PRINCIPLES

The FAIR guiding principles are best practice guidelines for management of scholarly data, comprised of 15 concise and measurable principles set to be adopted by the European Commission in Horizon 2020 projects. Data management principles help researchers to keep track of their data and to handle and store them in a sustainable manner throughout as well as after the project. The GoFAIR initiative provides a detailed description of every principle at their webpage.

FAIR data are:

Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier

F2. data are described with rich metadata

F3. (meta)data are registered or indexed in a searchable resource

F4. metadata specify the data identifier

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1. The protocol is open, free, and universally implementable

A1.2. the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

Reusable:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

These principles make data management measurable. They intend to enhance the findability of data for humans and machines and to increase data reuse. In order to implement the FAIR guiding principles in the workflow, it is recommended to create a data management plan prior to data collection at the beginning of the research project. The main purpose of a data management plan is to make research data created within a project FAIR. It should involve documentation of how data will be handled during and after the end of a research project, including information on the methodology and standards which will be applied and how data will be curated and preserved in the long-term. Its extent and content is varying depending on the funding agency, the size of the project and the amount and nature of the collected data. The implementation of the FAIR guiding principles is a process that starts before data collection and goes beyond their storage.

Making data FAIR: A step by step process

Planning: Already in the planning phase of the research project, it is helpful to think through the data life cycle, starting from how to collect them and ending on how to preserve them. As mentioned above, a data management plan can help to organize project data, keep track of them during the course of the project, as well as regulate data sharing and preservation during and after the project.

Collection: Prior to data collection, the creation of a data sampling plan may help to improve data collection by ensuring that data will be collected following standards set in the respective field or research. Usage of best practices for data sampling may reduce additional work to standardize data afterwards. During data collection, it is important to do thorough metadata documentation. The usage of controlled vocabulary, where applicable, is also recommended in order to raise the value of the data entry. The usage of open data formats, metadata standards and controlled vocabulary ensures the findability and interoperability of data.

Processing: Data processing implies its modification. Processing steps need to be included in the metadata. The documentation of data evolution is crucial for the quality of the data product, as well as for the reproducibility of the results. A data versioning and a regular data backup during processing counteract data loss. For backing up data, in addition to a personal device (computer, external hard drive etc.) an external storage system should be chosen, like for example a university server, or a cloud storage system. A good data backup plan also implies that data are stored in two places which are physically apart, like for example in the office and in the institutional server room.

Preservation: The preservation of research data is a decisive step in the data life cycle. The best practice is to store any relevant project data together with its metadata in a trusted data repository.

Publication: Data publication in a trusted data repository is a necessary step to ensure data accessibility. Either the data and its corresponding metadata, or just the metadata can be released for publication. INTERACT promotes free and open access to data in line with the European Open Research Data Pilot OpenAIRE. Embargos and licensing provide control over data access and usage. If the data set contains personal sensitive data, like for example names or birth dates, the open software Amnesia10 from OpenAIRE can be used in order to anonymize these data and thus make them suitable for publication. In some cases, access restrictions may apply to data, for example when data release would affect intellectual property rights. A listing of cases in which the INTERACT open data principles do not apply, can be found in the INTERACT data policy.

Reuse: For the reuse of data, they need to be made accessible in an interoperable format and documented in a self-explanatory way. Rich and standardised metadata and open data formats increases the value of data and the likelihood for its reuse

Redacted from D4.3_INTERACT_Field_guide_to_data_repositories

2 - TECHNICAL FEATURES OF THE INTERACT DATA PORTAL

2.1 METADATA GUIDELINES

Metadata guidelines are documented agreements regarding the representation, format, definition, structuring of metadata in a context of a data sharing network. Metadata guidelines regulate how data can be shared across systems providing a context in which to work predictably, allowing a streamlined and potentially fully automated operation to be put in place enabling wider access and reusability, as common clear defined meanings for data encourage data of higher quality that can be consumed by both humans and machines for multiple purposes. They enable consistent results during data retrieval and processing, regardless of the source that is providing such data and encourage good practices that make data more FAIR -findable, accessible, interoperable and reusable.

INTERACT has created its own set of guidelines to be used in relation to Virtual Access provision via INTERACT Data Portal that can be found in this document's appendix. In addition to defining the modalities available in sharing metadata with the portal, these guidelines also define which fields must be

present in the metadata record and how these fields should be formatted. Compliance with the guidelines is a mandatory requirement if metadata is to be shared with the INTERACT Data Portal.

2.2 WHAT FIELDS ARE HARVESTED TO INTERACT DATA PORTAL?

The fields that are currently being harvested are shown in the following table.

Table 2.2 Metadata fields harvested to INTERACT Data Portal.

Type	Field Name	DCAT-AP	JSON
Text	name	dct:title	name
Text	description	dct:description	description
Text	keywords	dct:keyworks	keywords
URL	url	dct:landingpage	url
Text	citation	dct:identifier	citation
Text	creator	dct:creator	creator
Text	topicCategory	dct:theme	about
Text	temporalCoverage	dt:temporal	temporalCoverage
Text, Place	spacialCoverage	dct:spacial	spatialCoverage
Text	datePublished	dct:issued	datePublished
Text	dateModified	dct:modified	dateModified
Text	publisher	dct:publisher	publisher
Text	provider	dcat:hadRole	provider
URL	license	dct:license	license
Text	variableMeasured	n/a	variableMeasured

2.3 HOW IS METADATA HARVESTED TO INTERACT DATA PORTAL?

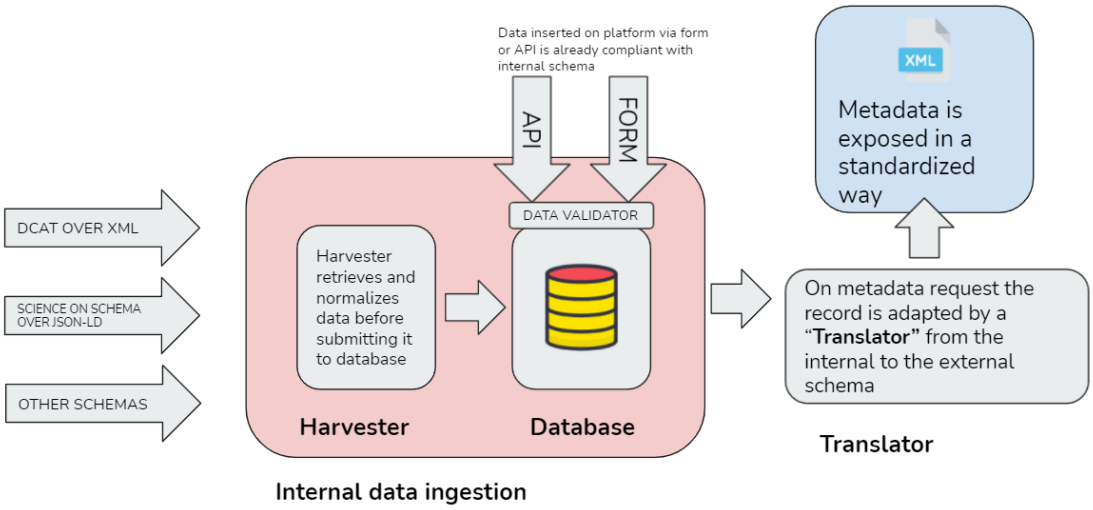
Once an entrypoint exposing metadata compliant with INTERACT metadata guidelines has been provided to the system, it will start periodically analyzing the available resources to process and import them into the system.

We can divide this process into three stages:

In the **Gather stage** the map of resources is parsed and a harvesting process is created for each single resource. This new process will contain the information necessary for the next stage to execute, like the location of the metadata and the type of operation that needs to be executed. Once each resource has a corresponding assigned process, these processes are collected and passed in bulk to the next stage.

The **Fetch stage** will then receive the information from the previous stage and will be responsible for fetching the contents of the metadata record from the remote server and storing them in the database for later use.

In the **Import stage** the raw metadata is retrieved from where it was stored by the previous stage and a series of checks on the nature of the received metadata are performed. Initially the validity of the structure of the record is analyzed and normalized to the internal data structure. Once the record is finally ready to be imported a check is made to understand if the record is to be saved as a new entity or if it is a more updated version of a record that has already been imported in the system in the past and the record is saved in the database.



2.4 CREATING AN ENTRYPOINT FOR METADATA HARVESTING BY INTERACT DATA PORTAL

Once all the metadata records that need to be exposed (see Table x) have been properly formatted in compliance with INTERACT Metadata Guidelines you can disclose the location of each resource in a single XML sitemap so that the harvester knows where it has to fetch the record.

An example of a XML sitemap is shown below:

```
▼<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  ▼<url>
     <loc>https://meta.fieldsites.se/objects/cpCCneLWaNjRZLq6i_Pih0vx</loc>
   </url>
  ▼<url>
     <loc>https://meta.fieldsites.se/objects/dJnKcyGN_CHr2uyasm8ZrOfv</loc>
   </url>
  ▼<url>
     <loc>https://meta.fieldsites.se/objects/3FXRUoYvVHXPYVqU-h6XGruk</loc>
   </url>
  ▼<url>
     <loc>https://meta.fieldsites.se/objects/z-qsKugEvwswf1duxmlKvHZY</loc>
   </url>
</urlset>
```

The resource to which each URL is pointing must be either a DCAT-AP XML file, JSON file or text file.

An example of a compliant JSON file (extract, the whole file not presented here):


```
  "name": "Water balance - stream water level from Röbäcksdalen Catchment, Sampling point 4, 2020-05-08–2021-05-26",
  "provider": {
    "name": "Röbäcksdalen Field Research Station",
    "email": "data@fieldsites.se",
    "sameAs": "https://meta.fieldsites.se/resources/stations/Robacksdalen",
    "@type": "Organization",
    "@id": "https://meta.fieldsites.se/resources/stations/Robacksdalen"
  },
  "publisher": {
    "@id": "data.fieldsites.se",
    "@type": "Organization",
    "logo": "https://static.icos-cp.eu/images/sites-logo.png",
    "name": "SITES data portal",
    "url": "https://data.fieldsites.se"
  },
  "temporalCoverage": "2020-05-08T11:00:00Z/2021-05-26T07:00:00Z",
  "url": "https://meta.fieldsites.se/objects/cpCCneLWaNjRZLq6i_Pih0vx",
  "variableMeasured": [{
    "@type": "PropertyValue",
    "description": "Time instant",
    "name": "TIMESTAMP",
    "unitText": null
  }, {
    "@type": "PropertyValue",
    "description": "Stream level",
    "name": "SL",
    "unitText": "m"
  }, {
    "@type": "PropertyValue",
    "description": "Water temperature",
    "name": "TW",
    "unitText": "°C"
  }]
}
```


An example of a DCAT XML file (extract, the whole file not presented here):

```xml
<?xml version="1.0"?>

<dcat:Dataset>
    <dct:title>Zimbabwe Regional Geochemical Survey.</dct:title>
    <dct:description>During the period 1982-86 a team of geologists from the British Geological Survey ...</dct:description>
    <dcat:landingPage rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">http://dataset.info.org</dcat:landingPage>
    <dcat:keyword>exploration</dcat:keyword>
    <dcat:keyword>geochemistry</dcat:keyword>
    <dcat:keyword>geology</dcat:keyword>
    <dct:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2012-05-10</dct:issued>
    <dct:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2012-05-10T21:04</dct:modified>
    <dct:identifier>9df8df51-63db-37a8-e044-0003ba9b0d98</dct:identifier>
    <owl:versionInfo>2.3</owl:versionInfo>
    <adms:versionNotes>New schema added</adms:versionNotes>
    <dct:language>en</dct:language>
    <dct:language>es</dct:language>
    <dct:language>ca</dct:language>
    <dcat:theme rdf:resource="http://eurovoc.europa.eu/100142"/>
    <dcat:theme rdf:resource="http://eurovoc.europa.eu/209065"/>
    <dcat:theme>Earth Sciences</dcat:theme>
    <dct:temporal>
        <dct:PeriodOfTime>
            <schema:startDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1905-03-01</schema:startDate>
            <schema:endDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2013-01-05</schema:endDate>
        </dct:PeriodOfTime>
    </dct:temporal>
    <dct:publisher>
        <foaf:Organization rdf:about="http://orgs.vocab.org/some-org">
            <foaf:name>Publishing Organization for dataset 1</foaf:name>
            <foaf:mbox>contact@some.org</foaf:mbox>
            <foaf:homepage>http://some.org</foaf:homepage>
            <dct:type rdf:resource="http://purl.org/adms/publishertype/NonProfitOrganisation"/>
        </foaf:Organization>
    </dct:publisher>
</dcat:Dataset>
```

2.5 SOLUTIONS FOR HOSTING YOUR METADATA

As a station with fewer technical resources and a smaller budget it is usually recommended that you rely on bigger data publishers or consortiums to host your data and/or metadata, as they already have the infrastructure to serve your metadata effortlessly in place. If hosting your metadata at your own repository is the solution that works best for your needs, a simple server configuration will most likely offer enough functionality to serve your metadata over to INTERACT Data Portal, though the setup and maintenance of said system can be a strenuous process. Any additional features would have to be developed and deployed and that is an effort that must be counted in when taking such decisions. Alternatively once could rely on existing open source or proprietary services / frameworks. INTERACT Data Portal currently relies on the CKAN framework to power many functionalities

CKAN is a powerful open-source data hub / data management system for publishing, sharing and finding data. It's trusted by many government agencies as its main public data repository such as the USA, the UK, Australia, Canada and the European Data Portal. It aims to provide a stable platform that's both simple, easy to build on and extend, as to use and interact with, that provides a way for people to find, share and reuse data. The core of CKAN is a robust registry / catalog system designed for machine interaction so that tasks like registering and acquiring datasets can be automated. It's internal model is used to store metadata about different records, and present it on a web interface with user usability in mind, that allows users to browse, search and filter this metadata. It's flexibility allows it to be highly customizable to the needs that we outlined in this presentation.

3- INTERACT DATA PORTAL API ENDPOINTS

This section documents INTERACT Data Portal's API for any organization which may want to programmatically access INTERACT Data Portal and the metadata hosted on it.

INTERACT Data Portal's API is based on CKAN's Action API, a powerful, RPC-style API. All of a CKAN website's core functionality (everything you can do with the web interface and more) can be used by external code that calls the API

To call the API, post a JSON dictionary in an HTTP POST request to one of the API endpoints through the software you are developing or alternatively with tools like curl or postman. The required and optional parameters for the API function should be given in the JSON dictionary as described in the endpoint's documentation. The returned response will also be in a JSON dictionary comprised of three keys:

SUCCESS

Key	Description	Value
“success”	Whether the function you called executed successfully	Boolean: `true` or `false`

The API aims to always return 200 OK as the status code of its HTTP response, whether there were errors with the request or not, so it’s important to always check the value of the "`success`" key in the response dictionary and if it’s `false` check the value of the "`error`" key.

RESULT

Key	Description	Value
“`result`”	The returned result from the function you called.	The type and value of the result de

If there was an error responding to your request, the dictionary will contain an "error" key with details of the error instead of the "result" key.

Key	Description	Value
“`error`”	A report of what went wrong in processing your request.	Object.

A response dictionary containing an error will look like this:

HELP

Key	Description	Value
“help”	The documentation string for the function you called.	String

An example of a typical successful response:


```
{
    "help": "https://dataportal.eu-interact.org/api/3/action/help_show?name=organization_list",
    "success": true,
    "result": [
        "aarhus-university-au-partner-5",
        "alfred-wegener-institute-for-polar-and-marine-research-awi-partner-7",
        "arctic-institute-of-north-america-aina-partner-38",
        "aurora-research-institute-ari-partner-37",
        "canadian-high-arctic-research-station-polar-partner-46",
        "centre-detudes-nordiques-cen-centre-for-northern-studies",
        "churchill-northern-studies-centre-cnsc-partner-55",
        "consiglio-nazionale-delle-ricerche-cnr-partner-31",
        "finnish-meteorological-institute-fmi-partner-34",
        "greenland-institute-of-natural-resources-ginr-partner-16",
        "institute-of-geography-and-spatial-organisation-of-polish-academy-of-sciences-igso-pas-partner-30",
        "institute-of-geography-and-spatial-organization-polish-academy-of-sciences",
        "interact",
        "jardfeingi-jf-partner-27",
        "moscow-state-university-msu-partner-22",
        "natural-resources-institute-finland-luke-partner-20",
        "norwegian-institute-of-bioeconomy-research-nibio-partner-13",
        "stockholm-university-su-partner-14",
        "swedish-polar-research-secretariat-sprs-partner-12",
        "swedish-university-of-agricultural-sciences-slu-partner-23",
        "tomsk-state-university-tsu-partner-10",
        "university-of-alaska-fairbanks-uaf-partner-32",
        "university-of-copenhagen-ucph-partner-3",
        "university-of-helsinki-uh-partner-15",
        "university-of-innsbruck-acinn-partner-25",
        "university-of-oulu-uoulu-partner-4",
        "university-of-turku-utu-partner-18",
        "unesco-chair-on-environmental-dynamics-and-climate-change-at-the",
        "zentralanstalt-fur-meteorologie-und-geodynamik-zamg-partner-24"
    ]
}
```


FREQUENTLY USED GET FUNCTIONS

The URL to query is the result of chaining the base URL with the requested action.

Example:

```
BASE URL = https://dataportal.eu-interact.org/api/action/

ACTION = tag_list

URL TO QUERY = https://dataportal.eu-interact.org/api/action/tag_list/tag_list
```

The request will fail if the required parameters are not passed.

Action	Description	Parameters
organization_list	Return a list of the names of the site's organizations. **Return type**: list of strings	**sort** (string)(optional) – sorting of the search results.. Default: "`title asc`" string of field name and sort-order. The allowed fields names are '`name`', '`package_count`' and '`title`'. The allowed sort orders are `asc` and `desc`. **limit** (int)(optional) – the maximum number of organizations returned Default value: `1000` when all_fields=`false`,: `25` when all_fields=`true` **offset** (int) – when limit is given, the offset to start returning organizations from **organizations** (list of strings)(optional) – a list of names of the groups to return, if given only groups whose names are in this list will be returned **all_fields** (bool)(optional) – return group dictionaries instead of just names. Default value: `false` **include_dataset_count** (bool)(optional) – if all_fields, include the full package_count Default: `true` **include_extras** (bool)(optional) – if all_fields, include the organization extra fields Default value: `false` **include_tags** (bool)(optional) – if all_fields, include the organization tags. Default value: `false` **include_groups** (bool)(optional) – if all_fields, include the organizations the organizations are in Default value: `false`

		include_users (bool)(optional) – if all_fields, include the organization users Default value: `false`
organization_show	Return the details of an organization. **Return type**: dictionary	**id** (string) – the id or name of the organization **include_datasets** (bool)(optional) – include a truncated list of the organization's datasets. Default value: `false` **include_dataset_count** (bool)(optional) – include the full package_count Default value: `true` **include_extras** (bool)(optional) – include the organization's extra fields. Default value: `true` **include_tags** (bool)(optional) – include the organization's tags. Default value: `true` **include_followers** (bool) – include the organization's number of followers Default value: `true`
tag_list	Return a list of the site's tags **Return type**: list of dictionaries .	**query** (string)(optional) – a tag name query to search for, if given only tags whose names contain this string will be returned **vocabulary_id** (string) (optional) – the id or name of a vocabulary, if give only tags that belong to this vocabulary will be returned **all_fields** (bool)(optional) – return full tag dictionaries instead of just names Default value: `false`
package_show	Return the metadata of a dataset and its resources. **Return type**: dictionary	**id** (string) – the id or name of the dataset **use_default_schema** (bool) – use default package schema instead of a custom schema Default value: `false` **include_tracking** (bool) – add tracking information to dataset and resources Default value: `false`
package_list	Return a list of the names of the site's datasets. **Return type**: list of dictionaries	**limit** (int)(optional) – if given, the list of datasets will be broken into pages of at most limit datasets per page and only one page will be returned at a time **offset** (int) – when limit is given, the offset to start returning packages from

package_search	Return a list of the names of the site's organizations. **Return type**: dictionary	View all parameters in attached documentation.

Any additional functions and features are detailed in the documentation of CKAN's API

Acknowledgements

We warmly acknowledge the INTERACT partner organizations, and their Data Managers, Station Managers and Station Contact Persons for their support and collaboration in the development of the INTERACT Metadata Guidelines and the INTERACT Data Portal. We also wish to thank and acknowledge the several data publishers, such as PANGAEA, SITES, GEM Database, DEIMS, and Nordicana D- for their collaboration, enabling the harvesting of metadata from several INTERACT VA partner organisations and sites to INTERACT Data Portal from their repositories. The INTERACT Data Team and the INTERACT Data Safeguard, Øystein Godøy from the Norwegian Meteorological Institute, are acknowledged for their advice and fruitful discussions during the development of metadata guidelines and data portal. This work, as well as the whole Virtual Access provision in INTERACT III, was made possible with funding received from EU-H2020 (GA No. 871120), which is highly appreciated and acknowledged.

References

2.3 Wilkinson et al. in 2016

Appendix 1. INTERACT Metadata Guidelines

Metadata guidelines