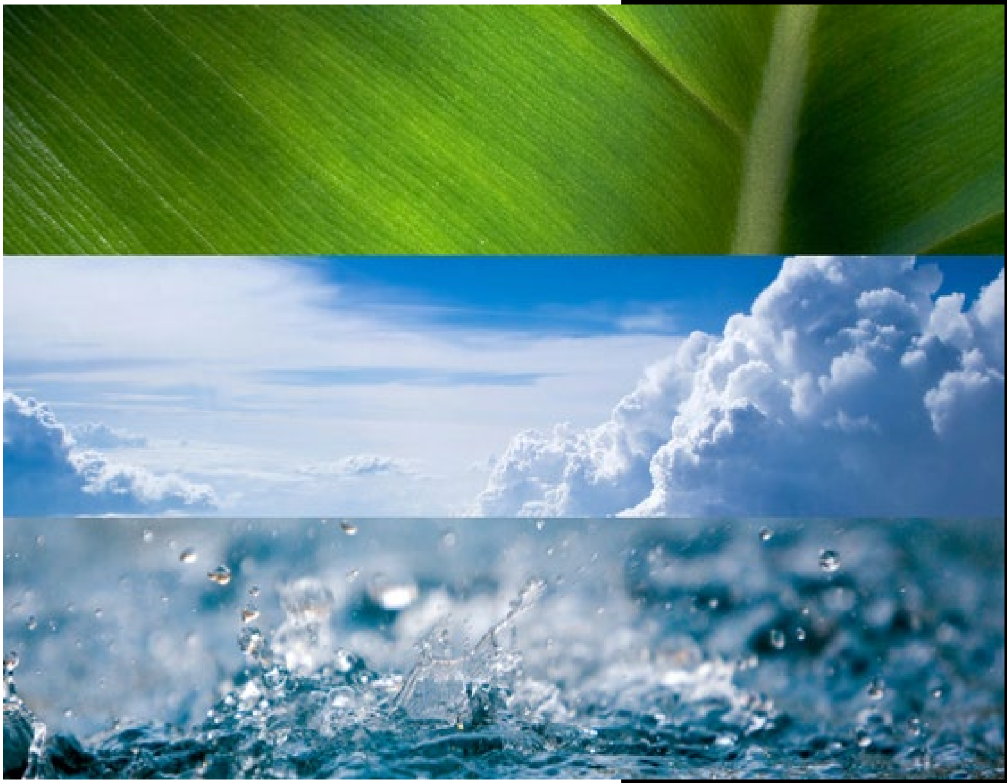
RINGO | Readiness of ICOS

Readiness of ICOS for Necessities of integrated Global Observations

D4.1

ICOS data type-registry and unified meta-database



RINGO (GA no 730944)

Public Document

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730944

Deliverable: D4.1 ICOS data type-registry and unified meta-database

Author(s): Alex Vermeulen, Oleg Mirzov, Harry Lankreijer, Maggie Hellström, Claudio D'Onofrio, Dario Papale, Leo Rivier, Lynn Hazan, Jerôme Tarniewicz, Benjamin Pfeil, Steve Jones

Date:	30 November 2020
Activity:	WP4 Task 1
Lead Partner:	ICOS ERIC
Document Issue:	
Dissemination Level:	Public
Contact:	alex.vermeulen@icos-ri.eu

	Name	Partner	Date
From	Alex Vermeulen	ICOS ERIC	30/11/2020
Approved by	Elena Saltikoff	ICOS ERIC	22/12/2020

Version	Date	Comments/Changes	Author/Partner

Deliverable Review Checklist

A list of checkpoints has been created to be ticked off by the Task Leader before finalizing the deliverable. These checkpoints are incorporated into the deliverable template where the Task Leader must tick off the list.

Appearance is generally appealing and according to the RINGO template. Cover page has been updated according to the Deliverable details.	x
The executive summary is provided giving a short and to the point description of the deliverable.	x
All abbreviations are explained in a separate list.	x
All references are listed in a concise list.	x
The deliverable clearly identifies all contributions from partners and justifies the resources used.	x
A full spell check has been executed and is completed.	x

DISCLAIMER

This document has been produced in the context of the *project* Readiness of ICOS for Necessities of integrated Global Observations (RINGO) . The Research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730944. All Information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view. Amendments, comments and suggestions should be sent to the authors.

ABSTRACT

As part of the RINGO project (Deliverable 5.5) we defined and started to implement a comprehensive unified metadata flow from Thematic Centres to the Carbon Portal. The design criteria of this system were to integrate as much as possible the operational (legacy) database systems at the TCs with the data portal, thereby preserving the investments in the robust and proven QA/QC and database systems at the TCs and combining these with the benefits of a linked open data system with connected data licence check, usage tracking and dynamic machine operable data and metadata based on a versioned RDF triple store.

The Data Type Registry (DTR) is a concept that at the moment is under development in the framework of the Fair Digital Objects (FDO[1]) architecture. As the ICOS Carbon Portal already implements many of the FAIR principles and is based on the same ideas as FDO it is ideally suited to consider to support this. However changing an operational data system while in use is not without dangers, so ICOS is following the developments very closely and meanwhile makes sure that the Carbon Portal in the future can be one of the first repositories to adopt the FDO structures. Already CP only accepts data for which it has a definition of the data type, that also specifies not only the required and optional metadata and thus the relations with the ICOS ontology, but also the contents of the data and its structure. ICOS also already mints Persistent Identifiers (PIDs) at the highest possible granularity, i.e for every digital object through the Handle system, which makes registration of the Object Type in the Handle system the next logical step.

The work performed within RINGO concerned the use of stable and unique identifiers throughout the whole data lifecycle in the ICOS Thematic Centers and the harmonisation of this metadata at the central level of the Carbon Portal. This has now been implemented for metadata concering stations, persons and organisations and will be extended to instruments, documentation like protocols, and software. ICOS also actively contributed to the development of an FDO conceptual standard for instrument PIDs in the framework of RDA.

[1] https://fairdo.org (under development)

Contents

1 Fair Digital Objects and Data Type Registries

(adapted from: Schwardmann, 2020[2])

The major obstacle for automation of data processing and use of data in the scientific process is the heterogeneity and complexity of data. Abstraction is a generic way to hide this heterogeneity and complexity by encapsulation and virtualization.

Data, metadata, software, semantic assertions, and many other elements can be seen as some kind of data, for example as files in a filesystem, that is copied, changed or deleted. At that layer all operations do not distinguish between metadata and data. On the data management and reuse layer a distinction is however necessary, and metadata must be used to govern the management operations on data.

By virtualization one substitutes objects by their logical representation. The most abstract way of such a logical representation is the pointer that leads to the object, a classical and often used approach in Computer Science, hiding all complexity behind a pure reference to the object.

The first step of abstraction of data is the identification of minimal elements that are to some degree atomic from the perspective of data management and reuse. Already more than 25 years ago these elements have been called digital objects, as described in a reprint of an article from 1995 (Kahn, Wilensky 2006), which was reused and adapted by the RDA (RDA 2019) working group on "Data Foundations and Terminology" (Berg-Cross 2015). They can be thought as some generalization of files in local file systems or streams of streaming providers for instance, and they are embedded in a structure of other important data concepts as one can see in Figure 1 below.

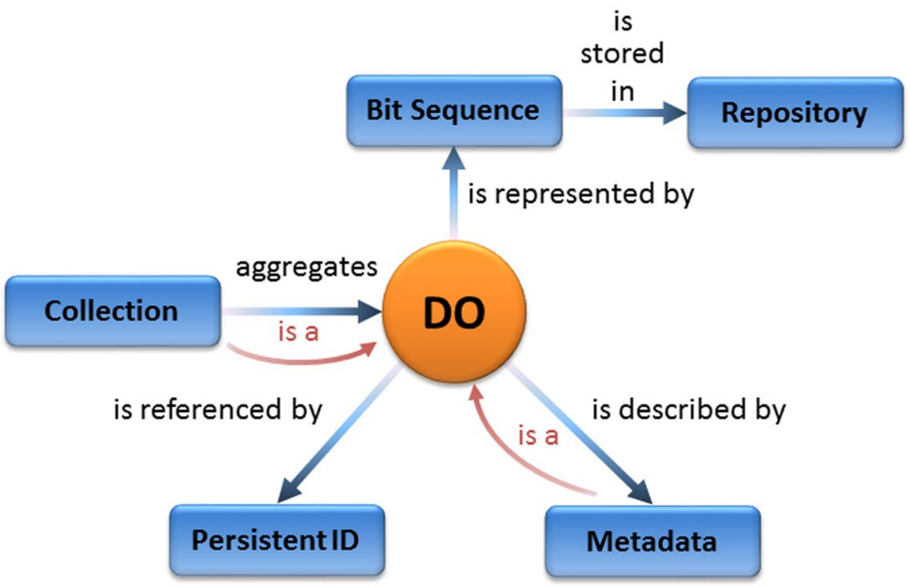


Figure 1 Source: Schwardmann, 2020

The Digital Object (DO) embedded in a structure of other important data elements and concepts.

[2] Schwardmann, U., 2020. Digital Objects – FAIR Digital Objects: Which Services Are Required?. *Data Science Journal*, 19(1), p.15. DOI: http://doi.org/10.5334/dsj-2020-015

How to represent the logical structure of digital objects with the right level of abstraction, however, is in its details still a matter of discussion, it certainly depends on how much of the logical structure is hidden by encapsulation behind a certain layer. And it also will partly depend on the data itself, specific workflows and use cases of data management and reuse. But in any case, the pointer as the most abstract logical representation has a prominent role here, and since data is and must be available across domains and sites, the pointer has to be a reference that is globally unique.

A global reference as URL could be the easiest option, but URLs are unpredictable and unstable references, because they change if the location of the data changes. See also (Klein et el. 2014) for a deeper analysis of this problem and its consequences for scientific reproducibility.

Redirection by means of a catalogue is a common solution for this problem, already known for many centuries from the libraries. This additional level of redirection is the rationale behind persistent identifiers. They are just globally unique strings without any semantics, but each such string has a record in a database that leads to the object, for instance via URL. If it changes, the database record can and has to be changed. These PIDs thus can be seen as the right way to reference data objects. The service to get the path to the object from the identifier string as reference, is called resolution. And since we are dealing with global references, they need to be globally resolvable, meaning that there must be a simple globally organized way that leads to the referred object. Such persistent identifiers are already widely used as global references in several domains of data management and publication and different highly reliable, global providers of PIDs exist since many years. The Handle system (Handle 2019), also used by ICOS and forming the basis of the DOI system, has an inherent, highly scalable global resolution mechanism. Therefore, the PIDs of the Handle system can actually fulfil the role of pointers as logical representation of digital objects.

1.1 Persistent Identifier, Handles and DOIs

There are more clear advantages to use the Handle system as PID technology to describe FDOs. Handles can have a much broader scope than just for DOI, and the policies, which are necessary for publications, are not always flexible enough to fulfil the needs of data management or data sharing between researchers. For data management or data sharing usually digital content related or community specific information, often in a finer granularity and often in a tight connection to the reference, is much more important than bibliographic information. Therefore, there is a need for additional governance structures beyond DOI for Handles to ensure reliable PID services with a much higher flexibility in PID usage and policies.

1.2 Data Types

In addition to the virtualization by reference it is crucial to provide a description of the object that is understandable also by machines to overcome the highly inefficient current way of data handling and to choose and prepare flexible services for digital objects in scientific workflows. And it would be helpful, if these descriptions would be available already at the reference level and thus in the Handle system.

This principle if for example used in the simple characterization of digital objects via MIME types, where the ending of a reference URL gives the necessary information. But for the reusability of data many other and more refined parameters are necessary. Such metadata enhancements of the digital objects are called data types.

The notion of the digital object emphasizes a close coupling between metadata, data and the persistent identifier as pointer. With the abstract structures of the FDO this becomes more explicit. Already in the early RDA working groups "PID Information Types" and "Data Type Registries" the coupling was made even tighter by allowing certain kinds of metadata to become part of the identifier record in the resolution database. Such metadata is called PID information type and build, as shown in Figure 2, a substantial encapsulation of complexity into a generic structure.

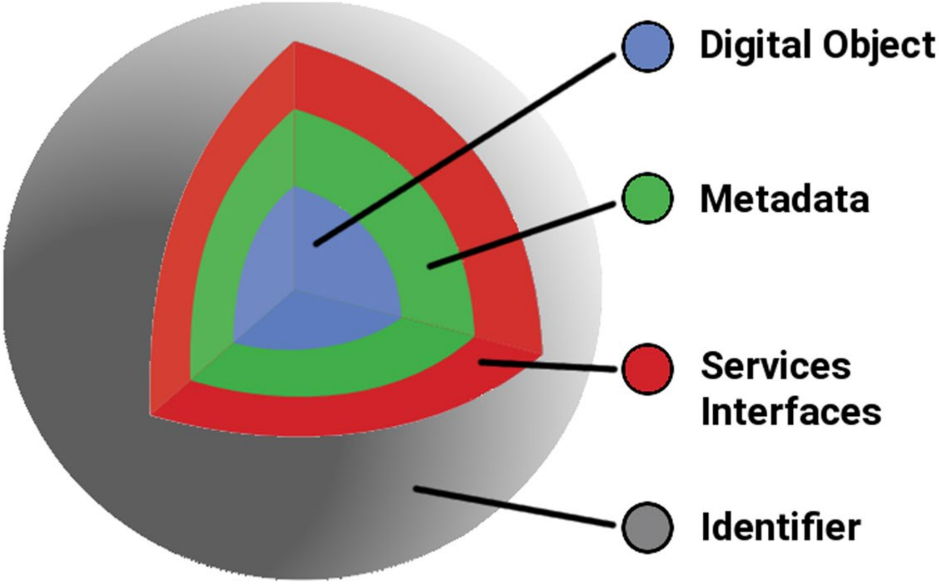


Figure 2 Encapsulation of the digital object, metadata and interfaces for services into a single logical element referenced by a persistent identifier (Source: Schwardmann, 2020)

Encapsulation of the digital object, metadata and interfaces for services into a single logical element referenced by a persistent identifier.

An extensive use of additional fields might slow down the resolution infrastructure, so one has to choose the minimal required elements wisely. So an additional RDA working group on "Kernel Information Types" developed rules and a profile (Weigel 2018) for a set of simple and most frequently needed metadata elements that should be stored together with the PID. The profile can be extended for the needs of scientific communities for instance and the rules are guidelines for these extensions. Currently this powerful technology is not supported by the DOI providers for paper and data publication, but for scientific data management it is available through the more general Handle system.

1.3 Data Type Registries

These data types need some kind of standardization to fulfill a minimal level of interoperability. The classical way along the procedures of international standardization bodies is either too specific or not flexible and fast enough to cover the needs of diverse research and economic areas in this fast growing area of data management.

A more promising approach is to provide community driven, reliable registries that contain reviewed type definitions in machine readable and interpretable form, uniquely referred and disambiguated again by PIDs. The PIDs of the type definitions can be used as keys for the metadata relevant to the Digital Objects as value, either in the PID record or a special metadata record.

Such registries with type definitions are called Data Type Registries (DTRs) and have been a topic for the Research Data Alliance (RDA) also since its first days (Lannom 2015). Two working groups made recommendations that led to a prototypical deployment of a working DTR implementation based on Cordra. Cordra is an open source software for managing digital objects, now available in version 2.0. ePIC is running two instances of Cordra, configured as DTRs on behalf of ePIC, one for production data types and one for the preparation of data types and testing. The type definitions are openly available. To create or change types an account is needed. A distinctive feature of the ePIC DTRs is the ability to define types in a hierarchical manner, such that also complex data types can be easily defined and for instance schemata for the value domain can be derived from the definition (Schwardmann 2016).

Because DTRs enable the disambiguation and correct assignment of types for humans and machines, they build an integral part of the FDO framework. With the correct choice of PID information types, depending on the needs in a scientific community, such FDOs enable fast decisions at the reference level about the relevance of data for certain scientific questions, allow the identification of the location and prepare the automated staging of remote data for the processing in a scientific workflow, for instance with high performance computing, or even the automated decision that a remote computation would need less effort.

Examples here are the detection of duplicates based on checksums, of earlier versions based on ‘was derived from’ relations or of the candidates for format conversion based on mime types and version numbers. A metadata service based on the metadata location given in a PID information type would be another example. Also decisions in workflows can be based on such PID information types as for instance the decision to move the application to the data or the data to the application based on the data size.

Collection representations can be based completely on PID information types, and a wide range of additional services and applications are proposed as part of the collection API and also beyond. It would also be very beneficial for repository interoperability to provide a collection enhancement based on a common agreement, like described by the RDA working group on Research Data Collections (Weigel 2017) to enable more flexibility for structures imposed on digital objects.

These examples show that the elementary service of resolution for retrieving PID information types from the PID record is required, but also services to describe the types in data type registries and to retrieve this information are needed. This additionally asks for interoperability between DTRs and services that monitor this interoperability. And in a next step services are required, that provide a set of information types that can be expected from (a class of) PIDs, so called PID profiles.

1.4 Repositories

Finally, the data services for FDOs itself need to be based on repositories providing reliable access to elementary digital objects. Currently often these repositories are giving some data representation enhanced with data base systems that provide a local layer of data and metadata indexing. Without even a PID registration for the provided data. A FAIR and global data perspective proposes a clear statement to overcome this situation. The FDO has to replace all other kind of representation of data inside repositories and a more generic approach to metadata indexing is also necessary. In some cases, it will be possible to provide adapters around legacy repository architectures, but overall this transformation is a big effort and may take a while. However, this effort is worthwhile to not end up with a fragmented data space with all its interoperability gaps, as we have it today.

2 Unified ICOS metadata and data ontology at ICOS CP

In order to introduce the concepts of the Carbon Portal and to introduce the main concepts needed to understand the Fair Digital Objects and its connection to data Type registries we repeat in this chapter the relevant part of RINGO Deliverable 5.5.

2.1 A brief introduction to ontologies, RDF and OWL

In the last decades, the use of ontologies in information systems has become more and more popular in various fields, such as web technologies, database integration, multi-agent systems, Natural Language Processing, etc.

There are several types of ontologies. The word "ontology" can designate different computer science objects depending on the context. For example, an ontology can be:

- a thesaurus in the field of information retrieval or
- a model represented in OWL in the field of linked-data or
- an XML schema in the context of databases
- etc.

It is important to distinguish these different forms of ontologies to clarify their content, their use and their goal. It is also needed to define precisely the vocabulary derived from the word ontology. In eScience often the terms "controlled vocabulary" and "community standards" are used. Ontologies are a perfect way to formally describe these and make them available in a machine interpretable and interoperable way.

The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. RDF is a recommendation from the W3C for creating meta-data structures that define data on the Web. RDF is used to improve searching and navigation for Semantic Web search engine (Web 3.0 applications).

RDF is composed of Triples: (1) the subject (the web page), (2) a property or predicate (an attribute name) and (3) an object (the actual value of the attribute for the web page).

1. The subject is a resource. Resource is anything that can have a Unique Resource Identifier (URI); this includes all the world's web pages, as well as individual elements of an XML document.
2. The property is a resource that has a name. For example the Dublin Core Metadata Initiative propose to use the name "dc:creator" to represent the author property. Property can be associated to a property type defined in an RDF Schema (RDFS). RDFS defined a RDF vocabulary composed of property type and resource type.
3. The object can be a URI, a literal (a string of character representing a number, a date, a noun etc...) or a blank node.

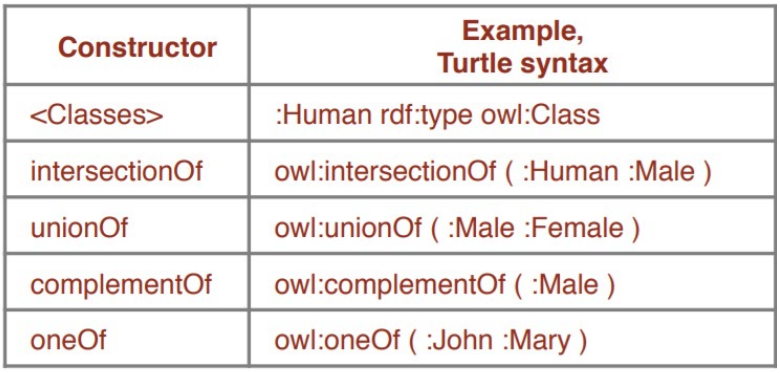
[see http://www.w3.org/RDF/ for more details].

The OWL Web Ontology Language is a standard recommended by the W3C. It is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. The

OWL is intended to provide a language that can be used to describe concepts and relations between them that are inherent in Web documents and applications. OWL language is used for:

1. formalize a domain by defining concepts called classes and properties of those classes,
2. define instances called individuals and assert properties about them,
3. reason about these classes and individuals to the degree permitted by the formal semantics of the OWL language.

One of the most powerful features of OWL is that it can be represented in RDF and thus can be stored and exposed in the same way as the ontology that it describes. This way a complete RDF database and its description as a formal ontology can be made machine readable and interpretable.



Constructor	Example, Turtle syntax
<Classes>	:Human rdf:type owl:Class
intersectionOf	owl:intersectionOf (:Human :Male)
unionOf	owl:unionOf (:Male :Female)
complementOf	owl:complementOf (:Male)
oneOf	owl:oneOf (:John :Mary)

Figure 3.1 Example of some OWL constructors

2.2 The ICOS CP data and metadata handling

2.2.1 Design and philosophy

The philosophy of CP is to treat all data objects equal and preserve the complete integrity of all data objects, so the actual data is never touched or changed up to the bit level. This goes for all data levels, i.e. from raw data, NRT data, final data quality-controlled data up to elaborated data products. CP strives for the maximum granularity of Data Objects.

The metadata that accompanies the data objects is maintained in a versioned so-called RDF triple store, following the Web 3.0, linked open data approach.

2.2.2 Data object handling

Before ingestion CP requires the uploader to calculate the SHA256 checksum of the data object. All ingestion data transport uses standard http(s) put and get methods and can be invoked by for example using the curl program. In the first stage of ingestion the uploader informs through a small metadata packet in JSON format of the object specification and the checksum of the data object together with some minimal provenance metadata that informs on the uploader, the spatial and/or temporal coverage that the data relates to for as far as applicable and depending of the object specification also on other important information like station, measurement level and instrument ID.

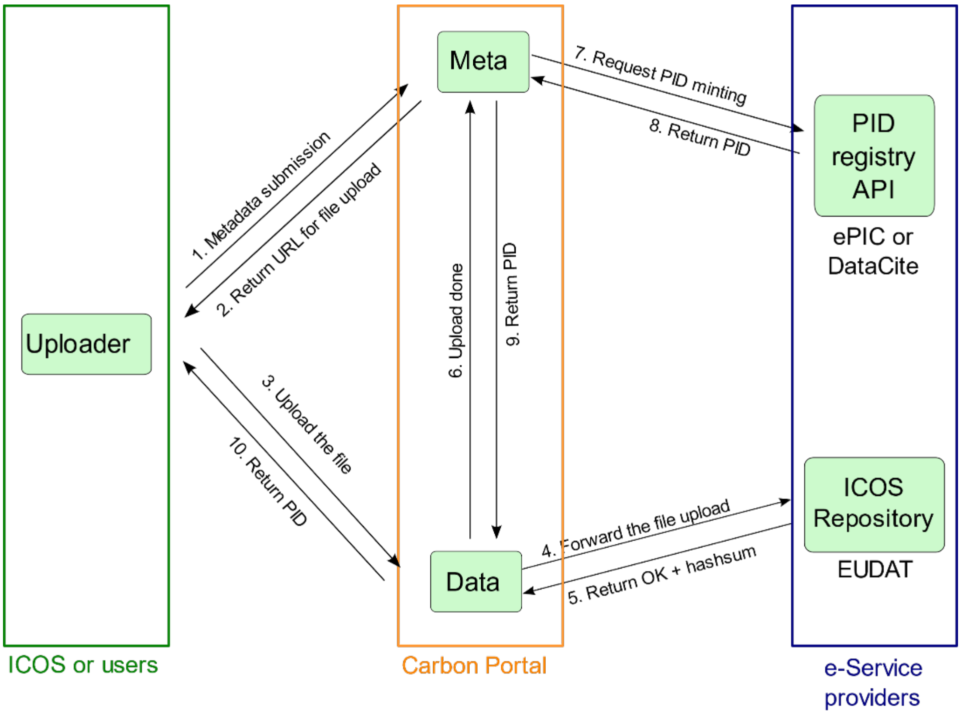


Figure 3.1 Schematic diagram of the data ingestion process at Carbon Portal. In 10 steps the new data is registered, ingested, minted a PID and stored together with the relevant metadata in the ICOS repository and the trusted repository.

Only objects with a known and registered Object Specification type are accepted. After successfully registering in this first step the user can start uploading the data object. While the uploader streams the data to CP, the data is forked and streamed at the same time to the B2SAFE trusted repository.

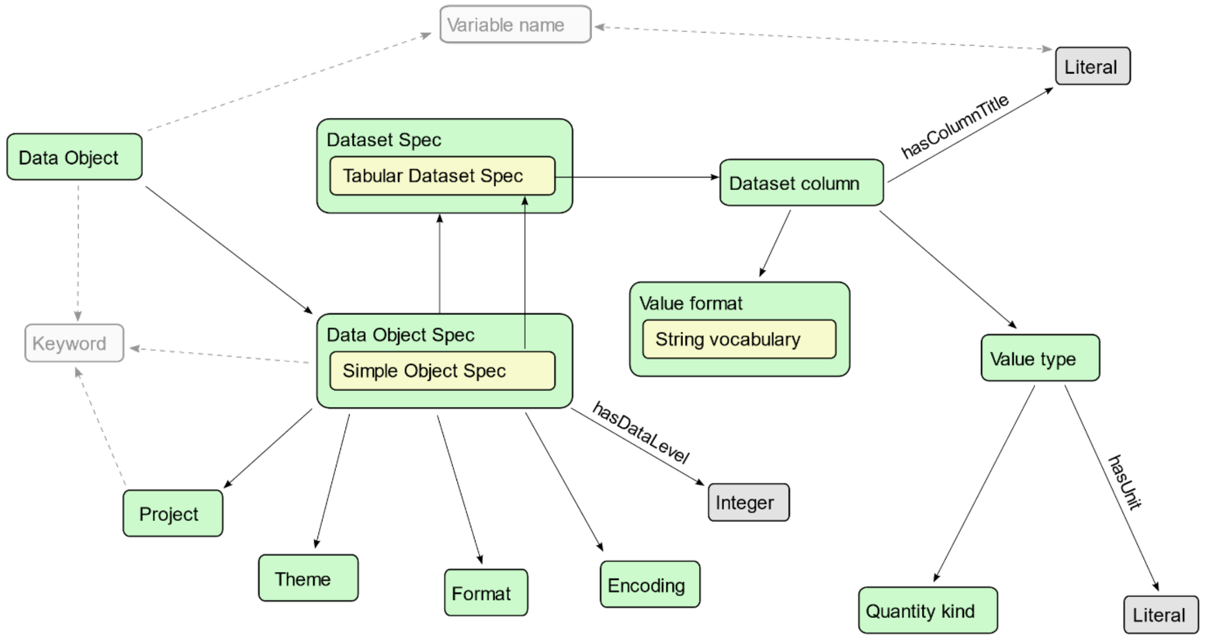


Figure 3.2 Relationship diagram of the ICOS data object specification and the metadata elements that describe the format, value type and quantity kinds of the different variable in an (ICOS) data file and relation to project and theme to which the data belongs.

When the object specification defines the data format of the file, a check is performed after the complete upload, to check the compliance to the data format and even possibly the validity of the data columns and spatial and temporal coverage as contained in the data file. Any deviation from the definition or prescribed metadata results in refusal of the file and abortion of the ingestion. The successful parsing of the data for text files also results also in the generation of binary CP-internal representations of the data that are used for the visualisation of time series in the data preview.

After upload completion, the checksum of the upload is compared with the registered checksum and when ok, a handle PID is minted for the data object and returned to the user. The metadata from the metadata packet is then added to the metadata repository and enriched with information on the PID, the checksum and other Object Specification dependent metadata. The suffix of the data object PID consists of the first 18 characters of the checksum of the data object and is thus unique for the data object. Later the PID suffix can at any time be compared with the SHA256 checksum of the data object to ensure that the data is up to the bit and exact copy of the original data object.

2.2.3 The metadata system

The metadata database can be queried using an open SparQL endpoint at https://meta.icos-cp.eu/sparql/?query=. The metadata store fully supports data versioning and data collections. It is machine actionable through standard http(s) protocol. The metadata store is fully described by the underlying ontology, that again itself is defined in RDF through the OWL language.

The design of the metadata system is fully configurable to act with a single or multiple portal frontend(s) using a single or multiple metadata stores. This means that for example multiple infrastructures can have their own differently styled data portal and use one single metadata store, or that one infrastructure has one portal that uses several external metadata stores, or that several infrastructures use one common portal that relies on a set of federated metadata stores, one per infrastructure. All completely transparent to the outside user.

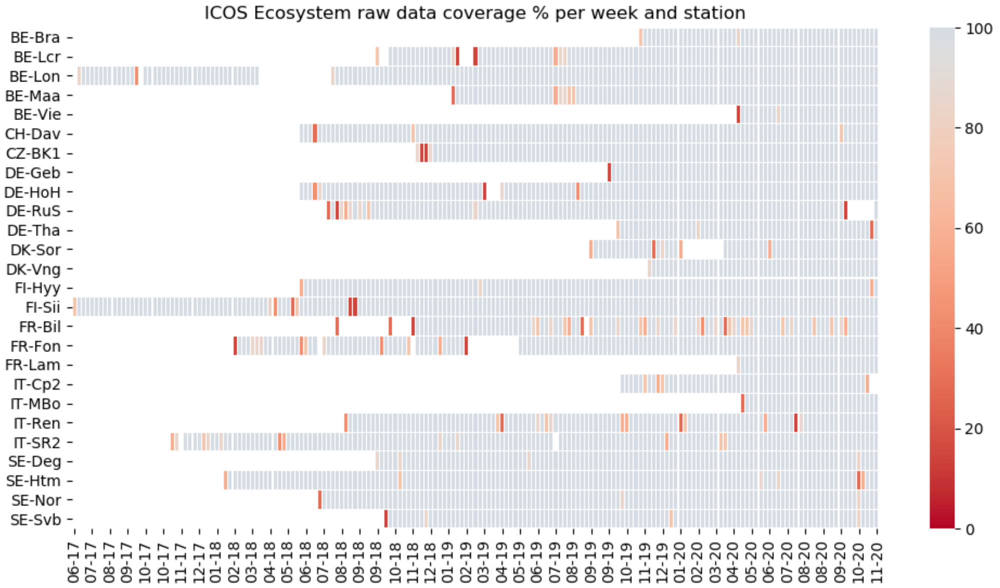
The ICOS CP metadata store is for example shared with the Swedish SITES national infrastructure that has its own dedicated and styled portal.

2.2.4 Data discovery

The main entry point for data discovery for humans is https://data.icos-cp.eu. Here a set of filters can be easily set to filter to the data sets that the user is looking for. The list of data objects that fulfils the set of filters is display dynamically. Changing the filters also dynamically updates the remaining options for the other filters that comply with the other filter settings. Filters can at will be added, removed and applied incrementally. From the results page the user can view the most relevant information on the data object and/or drill down to the data object landing page for all relevant metadata. Most data objects can be previewed, see data visualisation. Most data objects can also be added to the user's data cart for easy download, see data access.

An example of the powerful possibilities of open access to metadata trough SparQL is the overview of time coverage from L0 data for all stations from the heatmap tool[3]. This is a small python program that collects through one single SparQL call the metadata of all available raw data for a domain and then plots per station the availability of data in percentage per week. These heatmaps give a fast overview of the performance in sending the required raw data from the stations through TCs to Carbon Portal. Gaps identify periods with problems in either the measurements or the data transfer.

[3] https://github.com/ICOS-Carbon-Portal/python-tools/tree/master/heatmap



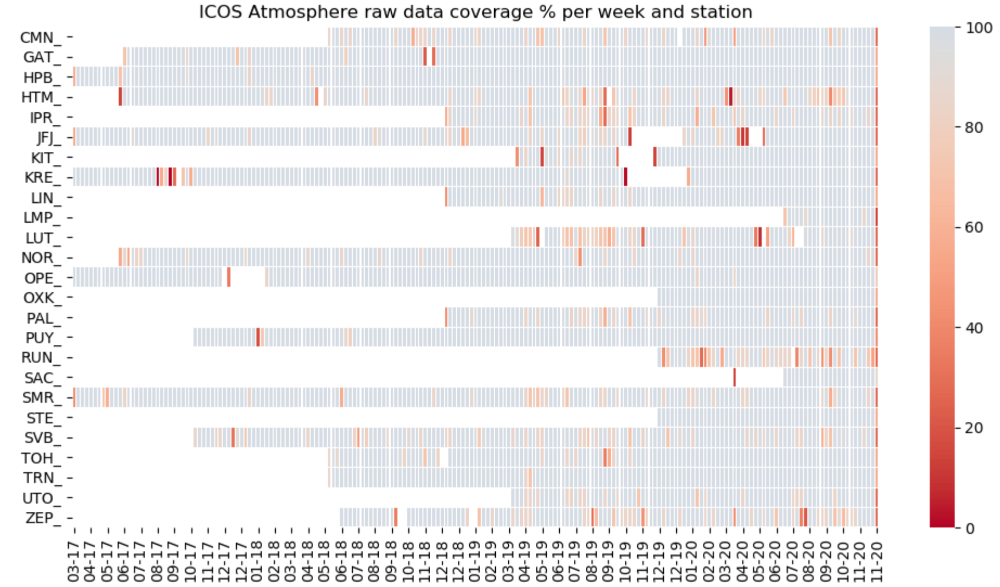


Figure 3.3 Heat maps of data coverage of ICOS raw data for the domains Atmosphere and Ecosystem from March 2017 up to now. Time coverage over each week is indicated from blue (100%) to red (0%).

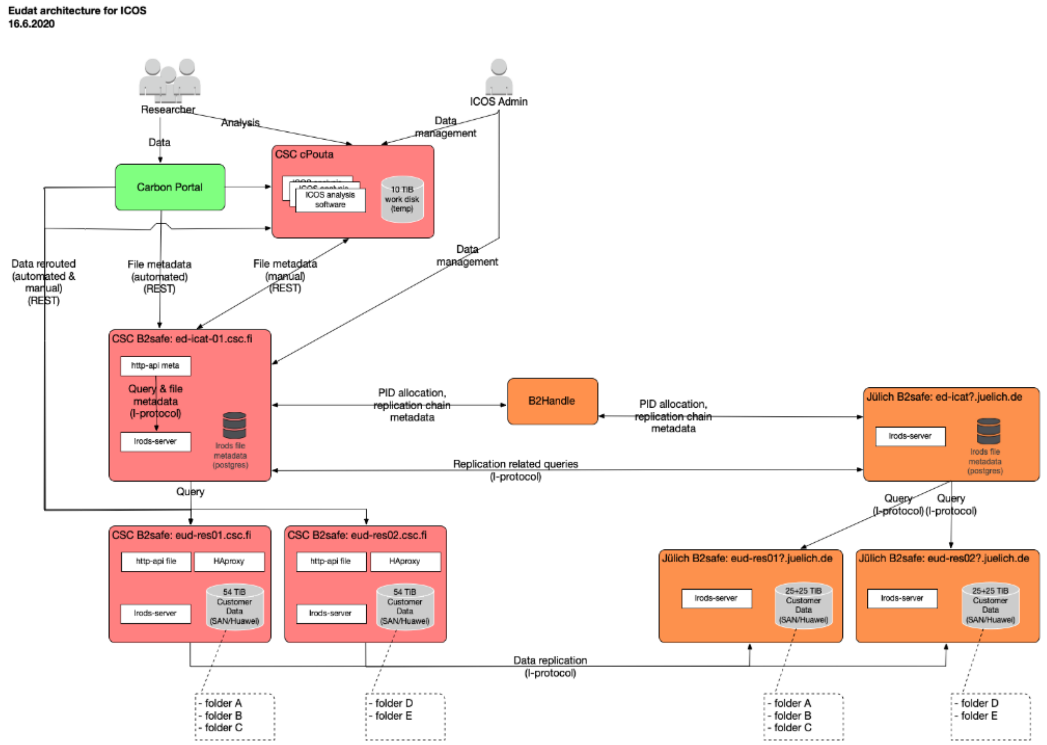
2.2.5 Data access

Data access is provided through the PID (or DOI) of the data objects. Resolving this PID through the Handle or DataCite DOI system leads normally to a landing page that contains a link to the data object(s). In case of non-ICOS data objects this link can point to another data portal due to data license restrictions. Raw data objects are currently also not directly downloadable but require contact with the relevant thematic centre.

The data discovery tool allows to add selected data objects to the user's data cart from where the collected objects can be downloaded in one batch into a single zip archive.

2.2.6 Trusted data repository usage (B2SAFE)

While the data is ingested as depicted in Figure 6.1, simultaneously a copy is streamed to the EUDAT B2SAFE server at CSC in Finland. At the end of the transfer also at their server the checksum of the received file is calculated and compared with the SHA256 checksum at CP. Only when all checksums are ok, the transfer is considered successful and the provenance metadata is finalized in the CP RDF store. In all other cases the uploading client is returned an error and transfer should be retried at a later stage. The EUDAT B2SAFE system will transfer a second copy to the EUDAT B2SAFE server in Jülich later. Both the CSC and the Jülich B2SAFE system are built using redundant data services that also have independent backup systems that allow to restore the data in case of a data storage failure. This adds to the data security already in place through similar systems at ICOS CP, the Thematic Centres and/or data providers and extends beyond the lifetime of these facilities. The ICOS data can be easily identified using the ICOS PIDs and the independently minted Handle PIDs that EUDAT minted for the data files using the B2Handle system. Also, if required, the ICOS data can be exposed through the EUDAT B2SHARE service. At any time ICOS CP can access and retrieve the ICOS data objects stored at B2SAFE through the ICOS PIDs.



Eudat architecture for ICOS
16.6.2020
Researcher
ICOS Admin
Analysis
Data
Data management
CSC cPouta
ICOS analysis software
10 TIB work disk (temp)
Carbon Portal
Data rerouted (automated & manual) (REST)
File metadata (automated) (REST)
File metadata (manual) (REST)
Data management
CSC B2safe: ed-icat-01.csc.fi
http-api meta
Query & file metadata (I-protocol)
Irods-server
Irods file metadata (postgres)
PID allocation, replication chain metadata
B2Handle
PID allocation, replication chain metadata
Jülich B2safe: ed-icat?.juelich.de
Irods-server
Irods file metadata (postgres)
Replication related queries (I-protocol)
Query
Query (I-protocol)
Query (I-protocol)
CSC B2safe: eud-res01.csc.fi
http-api file
HAproxy
Irods-server
54 TIB Customer Data (SAN/Huawei)
CSC B2safe: eud-res02.csc.fi
http-api file
HAproxy
Irods-server
54 TIB Customer Data (SAN/Huawei)
Jülich B2safe: eud-res01?.juelich.de
Irods-server
25+25 TIB Customer Data (SAN/Huawei)
Jülich B2safe: eud-res02?.juelich.de
Irods-server
25+25 TIB Customer Data (SAN/Huawei)
Data replication (I-protocol)
- folder A
- folder B
- folder C
- folder D
- folder E
- folder A
- folder B
- folder C
- folder D
- folder E

Figure 3.4 Architecture of the data transfer from ICOS to the B2SAFE service and replication at the two N2SAFE instances at CSC (Finland) and FZ Jülich (Germany).

ICOS keeps its metadata following an ontology-based RDF store. The ontology relations are modelled in OWL. The whole metadata store is available through the SparQL endpoint, including the OWL definitions, so that the complete metadata and its relations can be read for machine to machine communication. The ontology models the relations between the person, stations, instruments, measurements, and data processing actions as shown in Figure 3.4. At ingestion the digital objects are based on their data object specification and provenance information in the objects specification JSON package linked to the ontology and stored in the repository and streamed to the trusted repository, while performing the checksum comparison between source and target to ensure the data integrity. The RDF store is versioned, this is to say that every entry in the RDF store is serialized and kept together with the timestamp of the change. This means that the state of the ontology can be restored to its previous state of any point in time.

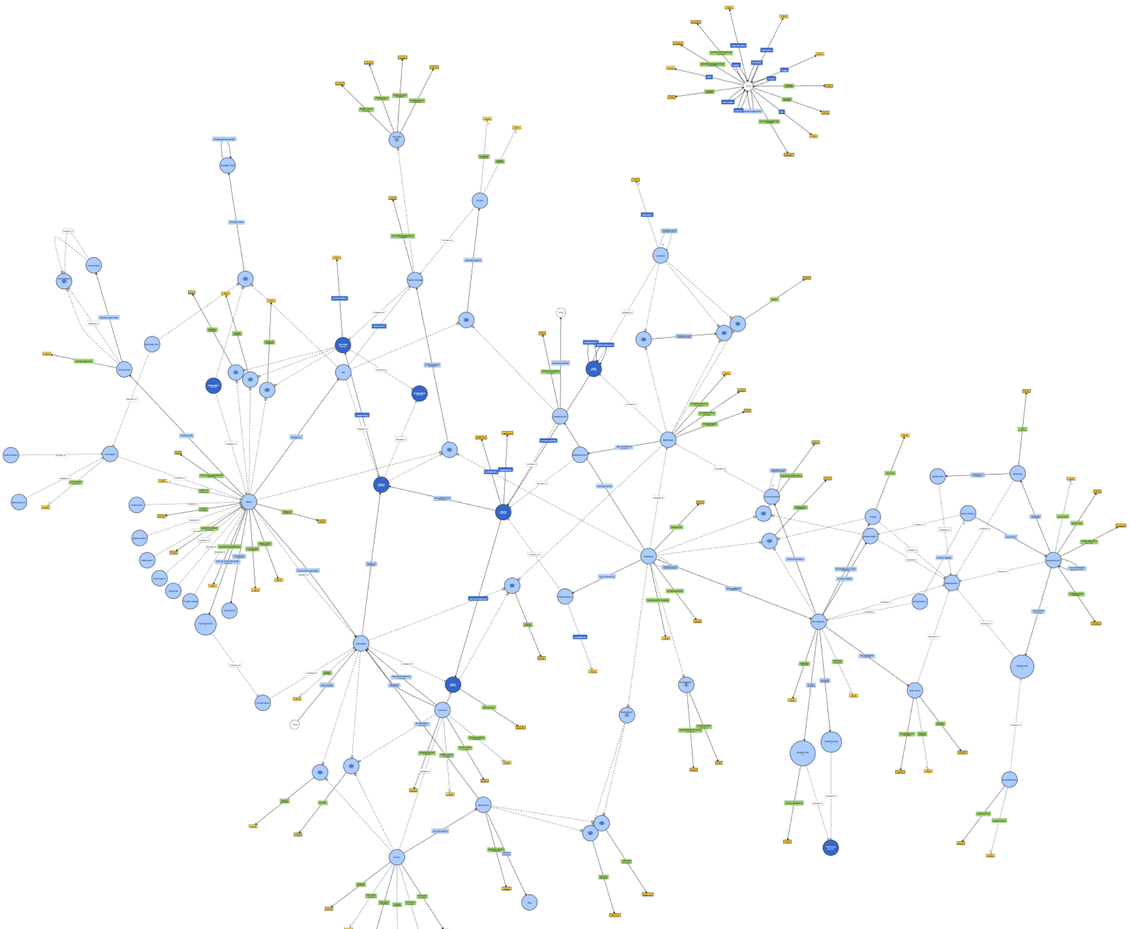


Figure 3.5 Overview of the ICOS CP ontology, modelled in OWL (http://www.visualdataweb.de/webvowl/#iri=http://meta.icos-cp.eu/ontologies/cpmeta/) as of June 2020.

When data is discovered through the portal app or a SparQL query the PID of the data object will resolve through the Handle system into a landing page, which contains either human and/or machine readable metadata that is gathered on the fly by following the relations defined by the ontology. That way always the most accurate and up to date version of the metadata available can be shown that corresponds with the start and end dates of acquisition, submission or processing contained in the metadata belonging to the object.

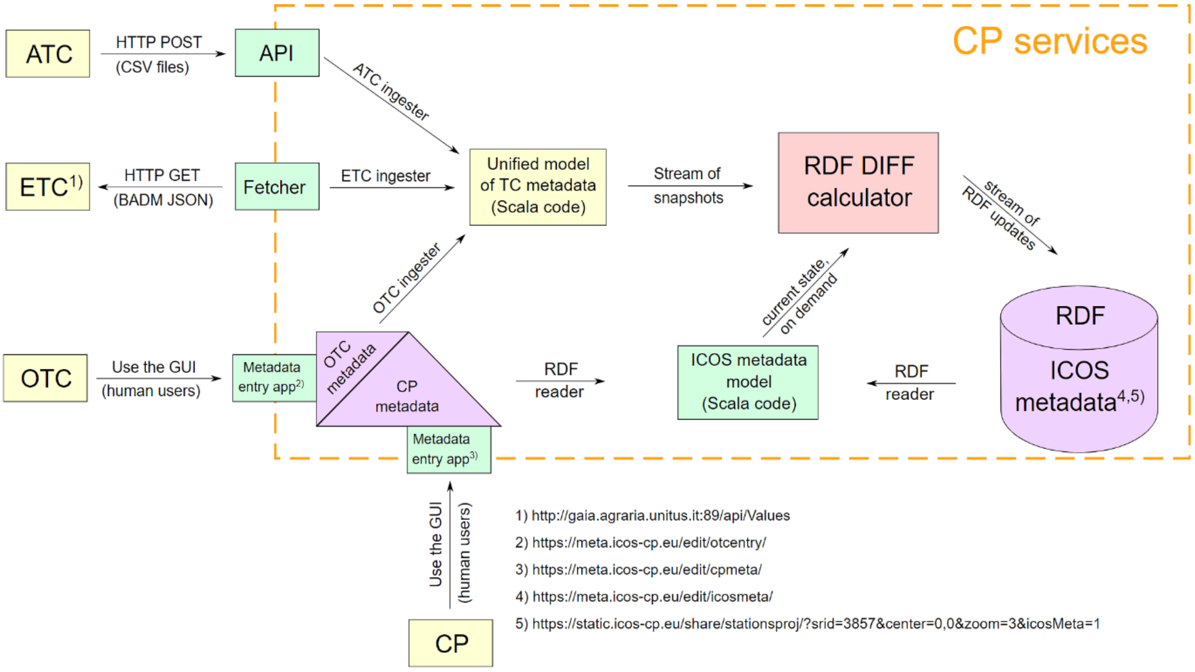
This is practically the most efficient, flexible and consistent way to associate time dependent metadata with the data objects while avoiding duplication of data and metadata and reducing the risk of metadata getting out-of-sync or not properly adjusted after corrections.

It also allows to model complex relations between metadata elements that are hard to efficiently model in traditional relational database management systems. A disadvantage of the flexibility and complex relations is that querying the ontology for even quite simple questions like: ’give me the list of all L2 data objects' can become a time-consuming operation due to the required traversal of a complex tree of a large amount of RDF triples. At Carbon Portal we invested quite some effort in designing a caching and query optimisation system using 'magic' indexing that makes sure that the queries that are used in our apps and data portal all execute within say 50 milliseconds.

2.3 Unified metadata gathering

As discussed in the chapter 2 the ICOS Thematic Centres and Calibration Labs perform independent data processing and offer services to the national networks. However, all data, from raw (L0) to Near Real Time (L1) and final quality controlled and calibrated data (L2) is ingested at Carbon Portal at the time of the generation of the data. Also, the networks maintain the metadata describing the contributors, the measurement systems and observations through the IT systems present at the Thematic Centres. The TC's also offer software clients that allow the measurement performing persons to enter provenance data and flag the measurement data as part of the Quality Control, to mark periods with problems in instrumentation for example. Also, the TC's add quality information to the data by for example performing analyses of instrument precision, conditions like low turbulence that impact the data quality, calibration uncertainties, drift and spike detection. In general, this QC information is added to the time series data files as separate columns with estimates of the different kinds of uncertainty per parameter or as columns with information in the form of data quality flags.

As part of the RINGO project we defined a comprehensive unified metadata flow from Thematic Centers to the Carbon Portal. The design criteria of this system were to integrate as much as possible the operational (legacy) database systems at the TCs with the data portal, thereby preserving the investments in the robust and proven QA/QC and database systems at the TCs and combining these with the benefits of a linked open data system with connected data licence check, usage tracking and dynamic machine operable data and metadata based on a versioned RDF triple store.



1) http://gaia.agraria.unitus.it:89/api/Values
2) https://meta.icos-cp.eu/edit/otcentry/
3) https://meta.icos-cp.eu/edit/cpmeta/
4) https://meta.icos-cp.eu/edit/icosmeta/
5) https://static.icos-cp.eu/share/stationsproj/?srid=3857&center=0,0&zoom=3&icosMeta=1

Figure 3.6 Diagram of the implemented metadata transfer scheme between Thematic Centers and Carbon Portal. TCs either post or make available the agreed tables with information wrt persons, roles, stations and instrumentation and the information is

routinely synced with the metadata ontology at CP by comparing current and available information and storing the differences. OTC adds the metadata directly into the CP RDF database.

All relevant metadata is synced with the CP repository using the methods shown in Figure 3.5. The Atmospheric TC provides the information as CSV tables that are transferred at regular intervals (daily) to a receiving service at the CP end. The Ecosystem TC provides a service that is polled by the CP at regular intervals (hourly). The Ocean TC uses the CP RDF store and metadata entry GUI to maintain the metadata for the Ocean domain. All three metadata exchange mechanisms rely on standard internet protocols and have been implemented, TCs are free to switch between these free choices to further harmonise the operations across the domains and reduce the complexity of the ICOS metadata handling.

The data from those three sources is then converted to a unified model of the TC metadata and compared with the current metadata in the CP final metadata store. Any difference is then translated in an RDF update statement that is logged in the CP metadata repository, just like any other metadata update.

In the first stage TCs provide the relevant station and person (plus role) metadata to make the attribution and dynamic citation generation work. This part is now working for Ocean and Ecosystem. As the corresponding Atmosphere metadata synchronisation only started in June 2020 the integration of this part was completed by the end of Summer 2020. In the next stage instrument information will be integrated and merged with the ICOS ontology as an important pre-condition to be able to complete the provenance information of the observational data sets.

The roles acknowledged in the thematic centres and stations, mapped to the roles that are differentiated in Carbon Portal and DataCite and the weights that determine the order in the citation string at CP and DataCite are detailed in Table 3.1 and Table 3.2 below. Not all roles result in a mention in the author list in the citation string. But all roles will result in a entry in the contributor list for the relevant data objects. Ocean and Ecosystem personnel are acknowledged in the citation string as ICOS RI. In Atmosphere one can specify roles for the Thematic Centres that will become part of the citation string. Also other Thematic Centres could adopt rolesfor their contributors.

Table 3.1 Station roles mapped to Carbon Portal and Datacite. Weights determine the order in the citation, higher means higher weight, no value means the role is listed in the contributor info, but not added as author to the citation string.

Station roles					Order		
Ecosystem	**Atmosphere**	**Ocean**	**Carbon Portal**	**DataCite**	**ETC**	**ATC**	**OTC**
Affiliated	Other		Other	RelatedPerson			
Scientist		Researcher	Researcher	Researcher	3		2
Scientist Flux			Researcher	Researcher	3		
Scientist Ancillary			Researcher	Researcher	3		
Principal Investigator	Principal Investigator	Principal Investigator	Principal Investigator	Contact Person	4	1	3
Co-PI	Deputy PI		Principal Investigator	Contact Person	4	1	
Manager	Station Supervising PI		Administrator	Supervisor	5	2	
	Species PI		Principal Investigator	Contact Person		1	
Technician	Instrument responsible	Engineer	Engineer	DataCollector	1		1
Technician Flux			Engineer	DataCollector	1		
Data manager	Data controller		Data manager	DataManager	2		
	Tank configurator		Engineer	ProjectMember			
Thematic Center Roles							
as ICOS ETC	Data manager	as ICOS OTC	Data manager	DataManager		0	
	ATC staff member		Engineer	Other		0	
	Checker		Data manager	DataCurator		0	
	Data Analist		Data manager	DataManager		0	
	Calib. centre responsible		Researcher	Researcher		0	
	Supervisor		Administrator	Supervisor		0	

Table 3.2 Mapping of the DataCite to the Carbon Portal contributor roles. Only roles indicated with a start are added to the author list and included in the citation string.

Contributor roles		Contributor roles
Datacite	**Author**	**Carbon Portal**
ContactPerson	*	Administrator
DataCollector	*	Data manager
DataCurator		Engineer
DataManager	*	Principal Investigator
Distributor		Researcher
Editor		Other
HostingInstitution		
Producer		
ProjectLeader		
ProjectManager		
ProjectMember		
RegistrationAgency		
RegistrationAuthority		
RelatedPerson		
Researcher	*	Researcher
ResearchGroup		
RightsHolder		
Sponsor		
Supervisor	*	Administrator
WorkPackageLeader		
Other		

3 Developing ICOS' readiness to adopt FDO and DTR

Consistent use of persistent identifiers throughout the whole data lifecycle and documenting the workflow in the data object metadata, as required for the FAIR data management and subsequent successful implementation of FDO, requires also that the ICOS Thematic Centers use unique and stable identifiers in their own data processing chains and exchange the information with the CP using those identifiers. As described in chapter two we developed in RINGO the use of stable and unique identifiers and the metadata harmonisation over the whole of ICOS and implemented this for persons, stations and organisations. Next to the strong and highly granular PID'ding of data objects and the linked open data approach of ICOS where all data objects are assigned a rigidly defined data object type this has made ICOS ready to map its data and metadata to the FDO approach and registering the data object types using DTR once the standard has reached maturity in the coming years.

Next steps will be to implement the metadata and persistent identification harmonisation also for instruments, documentation like protocols, and software. CP has been very active in the RDA working group for developing the PID definition for instruments, which led to the formulation of a standard that will fit in the FDO principles and is published in Stocker et al, 2018.

ABBREVIATIONS and ACRONYMS

AS	Atmosphere Station
ATC	Atmosphere Thematic Centre
B2FIND	EUDAT CDI service to search for data object metadata
B2SAFE	EUDAT CDI service to store data
CC4BY	Creative Commons - Attribution 4.0 International data license
CDI	Common Data services Infrastructure
CO_2	Carbon Dioxide
CP	Carbon Portal
CTS	CoreTrustSeal
DOI	Digital Object Identifier system
DSA-WDS	Data Seal of Approval-World Data System
DTR	Data Type Registry
ENVRI	ENVironmental Research Infrastructure
ERIC	European Research Infrastructure Consortium
ES	Ecosystem Station
ETC	Ecosystem thematic Centre
FAIR	Findable, Accessible, Interoperable, Reusable
FDO	FAIR Digital Object
GCOS	Global Climate Observing System
GTOS	Terrestrial Observing System
GEO	Group on Earth Observations
GHG	GreenHouse Gas
GCP	Global Carbon Project
ICOS	Integrated Carbon Observing System
IW	Internal Working (data, Level 1)
JSON	JavaScript Object Notation – a lightweight (self-descriptive) data-interchange format
MSA	Monitoring Station Assembly
NEON	National Ecological Observatory Network (USA)
netCDF	network Common Data Form
NOAA	National Oceanic and Atmospheric Administration (USA federal agency)
NRT	Near Real Time
ObsPack	Observation Package
OTC	Ocean Thematic Centre
OWL	Web Ontology Language
RDA	Research Data Alliance
RDF	Resource Description Framework
RDL	Registration, Disclaimer and Licensing
RI	Research Infrastructure
SHA	Secure Hash Algorithms
SHA-256	Hash algorithm using a fixes 256 bits hash size
SOOP	Ships of Opportunity
TC	Thematic Centre
TR	Trusted Repository
PID	Persistent Identifier
SOCAT	Surface Ocean CO_2 Atlas

STILT	Stochastic Time-Inverted Lagrangian Transport (atmospheric transport model)
SparQL	recursive acronym for SPARQL Protocol and RDF Query Language
UNFCCC	United Nations Framework for the Convention on Climate Change
URI	Universal Resource Identifier
VOS	Voluntary Observing Ship
WDCGG	World Data Centre for Greenhouse Gases
WIGOS	WMO Integrated Global Observation System
XML	eXtended Markup Language

References

Berg-Cross, G, Ritz, R and Wittenburg, P. 2015. Core Term Definitions. In: Data Foundation and Terminology Work Group Products. DOI: https://doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF

ePIC. 2019. ePIC Persistent Identifiers for eResearch. Available at http://dtr.pidconsortium.net/

European Commission: Directorate-General for Research and Innovation. 2018. Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. DOI: https://doi.org/10.2777/1524

GEDE. 2019. Group of European Data Experts. Available at https://rd-alliance.org/group/gede-group-european-data-experts-rda/wiki/gede-digital-object-topic-group

Handle. 2019. The Handle System. Available at http://www.handle.net

Kahn, R and Wilensky, R. 2006. A framework for distributed digital object services. Int. J. on Digital Libraries, 6: 115–123. DOI: https://doi.org/10.1007/s00799-005-0128-x

Klein, M, Van de Sompel, H, Sanderson, R, Shankar, H, Balakireva, L, Zhou, K and Tobin, R. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE, 9(12): e115253. DOI: https://doi.org/10.1371/journal.pone.0115253

Lannom, L, Broeder, D and Manepalli, G. 2015. RDA Data Type Registries Working Group Output. DOI: https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458

RDA. 2019. Research Data Alliance. Available at https://rd-alliance.org

RDA Europe. 2019. Research Data Alliance – Europe. Available at https://www.rd-alliance.org/rda-europe

Schultes, E and Wittenburg, P. 2018. FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In: Manolopoulos, Y and Stupnikov, S (eds.), Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, 1003. Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-23584-0

Schwardmann, U. 2016. Automated schema extraction for PID information types. 2016 IEEE International Conference on Big Data. PID:21.11101/0000-0002-A987-7. DOI: https://doi.org/10.1109/BigData.2016.7840957

Stocker, M., Darroch, L., Krahl, R., Habermann, T., Devaraju, A., Schwardmann, U., D'Onofrio, C. and Häggström, I., 2020. Persistent Identification of Instruments. Data Science Journal, 19(1), p.18. DOI: http://doi.org/10.5334/dsj-2020-018

Weigel, T, Almas, B, Baumgardt, F, Zastrow, T, Schwardmann, U, Hellström, M, Quinteros, J and Fleischer, D. 2017. Recommendation on Research Data Collections, RDA. DOI: https://doi.org/10.15497/RDA00022

Weigel, T, Plale, B, Parsons, M, Zhou, G, Luo, Y, Schwardmann, U, Quick, R, Hellström, M and Kurakawa, K. 2018. RDA Recommendation on PID Kernel Information (Version 1), RDA. DOI: https://doi.org/10.15497/RDA00031

Weigel, T, Schwardmann, U, Klump, J, Bendoukha, S and Quick, R. 2020. Making data and workflows findable for machines. Data Intelligence, 2: 40–46. DOI: https://doi.org/10.1162/dint_a_00026

Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3. DOI: https://doi.org/10.1038/sdata.2016.18