D6.3
Report on identification and citation service case studies

WORK PACKAGE 6 – INTER-RI DATA IDENTIFICATION AND CITATION SERVICES

LEADING BENEFICIARY: LUND UNIVERSITY

Author(s):	**Beneficiary/Institution**
Margareta Hellström (lead), Maria Johnsson and Alex Vermeulen	LU (Lund University, ICOS)
Dan Lear	MBA (EMBRC)
Markus Fiebig	NILU (ACTRIS)

Accepted by: Alex Vermeulen (WP6 leader) and Zhiming Zhao (Theme 2 leader).

Deliverable type: [REPORT]

Dissemination level: PUBLIC

Deliverable due date: 28.02.2019/M46

Actual Date of Submission: 28.02.2109/M46

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

ABSTRACT

This third and final deliverable of ENVRIplus Work Package 6 (WP6), "Inter RI data identification and citation services", provides a summary of three use cases related to data citation: 1) the publication of marine biodiversity data from peer-reviewed journal to EU data infrastructure; 2) the development of a citation and usage tracking system for greenhouse gas monitoring data; and 3) facilitating quantitatively correct data usage accounting. All of these have a strong bearing on the need to correctly acknowledge the efforts of data producers at multiple levels (research infrastructures, institutes and individuals), as well as support the aggregation of statistics mapping out downstream usage of data products. In addition, we summarise how identification and citation practices fit into the context of current trends in data management, discuss how data usage statistics can be used to estimate the impact of research infrastructures, and provide an update to the citation services landscape inventory first performed in deliverables D6.1 and D6.2.

REPORT REVIEWER(S)

Project internal reviewer(s):

Project internal reviewer(s):	Beneficiary/Institution
Malcolm Atkinson (Theme 2 expert)	UEDIN

DOCUMENT VERSION HISTORY

Date	Version
1.12.2018	Outline of report, writing starts
8.2.2019	Version sent for internal review
26.2.2019	Version sent to Theme 2 and WP 6 leaders
28.2.2019	Final version, submitted by project office

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the lead author (Margareta Hellström margareta.hellstrom@nateko.lu.se)

TERMINOLOGY

The acronyms and specialist terminology used in this report are explained in Appendix A. In addition, the complete and up-to-date ENVRIplus project glossary is provided online here: https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, 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 2015a]

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 harmonize 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.

EXECUTIVE SUMMARY

This third and final deliverable of ENVRIplus Work Package 6 (WP6), "Inter RI data identification and citation services", provides a summary of three use cases related to data citation: 1) the publication of marine biodiversity data from peer-reviewed journal to EU data infrastructure; 2) the development of a citation and usage tracking system for greenhouse gas monitoring data; and 3) facilitating quantitatively correct data usage accounting. The cases illustrate in their own ways the tight connections between Identification & Citation and the other “pillars” of research data management practices, such as Cataloguing, Metadata management, and Provenance. They also point to the importance of designing and implementing efficient means of harvesting accurate and sufficiently detailed information on various types of usage, and then feed these into analysis tools that can produce suitable impact metrics for a range of different purposes: total numbers of downloads and visualisations of data at the level of a research infrastructure, the demand for near real-time data from a given measurement site or campaign, or the average number of citations in scientific literature associated with an individual engineer. Furthermore, the cases highlight the need for RIs to engage with both their respective data producers (especially in the case of distributed or extended collaborations contributing their data to a central database) and data end users, and offer support for and also training in best practices for citing and identifying data.

Intentionally left blank

TABLE OF CONTENTS

ABSTRACT 2

REPORT REVIEWER(S) 2

DOCUMENT VERSION HISTORY 2

DOCUMENT AMENDMENT PROCEDURE 2

TERMINOLOGY 2

PROJECT SUMMARY 3

EXECUTIVE SUMMARY 3

TABLE OF CONTENTS 5

1 ABOUT WORK PACKAGE 6 7

2 IDENTIFICATION & CITATION IN CONTEXT 8

2.1 "Out of cite, out of mind"? 8

2.2 Guiding principles for citation of data in scientific literature 9

2.3 Referencing data in machine-actionable workflows 10

2.4 Estimating "impact" by tracking downstream usage 12

3 CITATION SERVICES 13

3.1 Revisiting the service landscape 13

3.1.1 Project FREYA 14

3.1.2 FORCE11 14

3.1.3 Scholix 14

3.1.4 Make Data Count 14

3.1.5 OpenAIRE 15

3.2 Data publishing workflows: Basic components 16

3.2.1 Making data citable 16

3.2.2 From PIDs to reference lists: creating useful citation strings 16

3.2.3 Monitoring dataset usage statistics 17

4 THE PUBLICATION OF MARINE BIODIVERSITY DATA FROM PEER-REVIEWED JOURNAL TO EU DATA INFRASTRUCTURES 19

4.1 Background 19

4.2 Proposed implementations 19

4.3 Results 21

5 DEVELOPMENT OF A CITATION AND USAGE TRACKING SYSTEM FOR GREENHOUSE GAS MONITORING DATA 23

5.1 Background 23

5.1.1 About ICOS 24

5.1.2 Services offered by the ICOS Carbon Portal data centre 24

5.1.3 The flow of data and metadata in ICOS 24

5.1.4 Strategy for assigning PIDs to ICOS data products 25

5.2 Proposed implementations ... 26
5.3 Results ... 28
5.3.1 Adapting the ICOS data model to support identification and citation ... 28
5.3.2 Resolving ICOS PIDs to dynamic landing pages ... 28
5.3.3 Supporting the citing of ICOS Data ... 29
5.3.4 Collecting bibliometric data usage statistics ... 30
5.3.5 Collecting portal services usage statistics ... 31
6 FACILITATING QUANTITATIVELY CORRECT DATA USAGE ACCOUNTING ... 32
6.1 Background ... 32
6.2 Proposed implementations ... 33
6.3 Results ... 34
7 OUTCOMES AND CONCLUSIONS ... 35
8 IMPACT ON PROJECT ... 36
9 IMPACT ON STAKEHOLDERS ... 37
ACKNOWLEDGEMENTS ... 38
REFERENCES ... 38
APPENDIX A. ACRONYMS AND SPECIAL TERMS ... 42
A.1. Terminology & glossary specific to this deliverable ... 42
A.2. Other technical terms and acronyms used in ENVRIplus deliverables ... 43
A.3. Organisational acronyms ... 45
A.4. ENVRIplus project-related acronyms & terms ... 47
APPENDIX B. CITATION BEST PRACTICES FOR THE ENVRI COMMUNITY ... 49
Technical aspects ... 49
Citation practices ... 50

1 ABOUT WORK PACKAGE 6

The overarching objective of ENVRIplus Work Package 6 (WP6), "Inter RI data identification and citation services", is to improve the efficiency and uptake of data identification and citation in the environmental and Earth science fields by identifying and enabling provision of convenient, effective and interoperable identifier management and citation services. This WP highlights identification and citation practices in environmental RIs, reviews available technologies and develops common services for these operations. In addition, it aims to set up a dialogue between ENVRI community partners and relevant actors and organisations involved in the provisioning of services related to identification and subsequent citation of digital representations of objects from all stages of the research life cycle [ENVRIplus 2015b].

The first WP6 deliverable (D6.1, [Hellström 2017]), summarised the associated technological needs and requirements of the ENVRIplus partners, suggested and outlined a common system design for Identification and Citation, and mapped the landscape of publishers, PID service providers and other actors in the scholarly data management and curation world.

The second WP6 deliverable (D6.2, [Hellström 2018]), summarised the outcomes of efforts to prepare and initiate a dialogue ("negotiations") with publishers, providers of existing data citation systems and other scientific organisations on raising awareness of what environmental and climate research infrastructures view as essential identification and citation-related services required in order to reach the ultimate goal of a "global data citation system".

This final WP6 deliverable is concerned with the development of a number of use case studies. Two of these, the implementation of an on-line, standards-based publication mechanism for marine biological data; and the development of a workflow and guidance for citation tracking models, are outlined in the WP6 Description of Work [ENVRIplus 2015b]:

- Use-case study 1:
    - Develop a full data lifecycle model for biological data linked to the standards developed and promoted through GBIF and the experience of EMSO, EMBRC and SeaDataNet, involving a taxonomy for known and new species, biodiversity with geo-positioning potentially coupled with physical environment data and genomic data. The resources of the UK marine biodiversity data centre (DASSH) and the National Marine Biological Library will provide the infrastructure to ensure data are captured alongside the peer-reviewed publication and made available through the GBIF community.
    - Establish GBIF Integrated Publishing Toolkit for the publication of data from the Marine Biological Records journal.
    - Define workflows for the research community to facilitate the provision of DOIs and the archiving of data in compliance with international standards.
- Use-case study 2:
    - Implement a full and integrated DOI and related metadata system for ICOS as an example of a distributed RI with very heterogeneous data. International cooperation with other infrastructures [working on greenhouse gases] will also be part of this.
    - Test citation tracking models. ICOS has a well-defined internal data workflow and already existing data transfers to global user networks and is therefore an ideal test case.

Not part of the original description of actions of the project, a third use case related to identification and citation of data was developed and managed together with WP9, as outlined in deliverable D9.1 [Chen 2017]. The aims of this use case (implementation case IC_9), Facilitating quantitatively correct data usage accounting, was to investigate how to facilitate quantitatively correct data usage accounting, and suggest ways for ENVRIplus partners to implement this.

2 IDENTIFICATION & CITATION IN CONTEXT

In this Chapter we briefly discuss a couple of important aspects of data citation, including a recapitulation of guiding principles, referencing digital objects in data processing, and the motivations behind estimating the impact of research through usage statistics. This is followed by Chapter 3 in which we revisit the data citation service provider landscape that was first mapped out in deliverable D6.1 [Hellström 2017], and summaries of the use cases in Chapters 4-6.

2.1 "Out of cite, out of mind"?

The practice of citing articles, books and other works, and providing standardized and unambiguous bibliographic references to such sources, is considered an integral part of scientific writing. However, the adaptation of the same (or similar) routines for citing data (and other research resources) has not yet become second nature to researchers. Instead, it is often the case that datasets are referred to as "data from Andersson et al. [personal communication]", or, "the Carousel site data set from 2017", or "field data collected by Pettersson and Lundström in 2017", at best with some contact information given in the form of a footnote. In the cases that data are referred to in a more comprehensive manner, citations may only be provided in-line in the running text or as a footnote, rather than being included in a bibliography section. Furthermore, even within a given domain – such as Earth sciences, or Life sciences – many different citation string formats are being used, sometimes lacking enough details on for example all contributors or their respective roles [Socha 2013].

There are many different reasons for this "double standard" – none of them in the least malicious, but they nevertheless hinder the process towards getting peers, employers and funders to recognize research data production and curation as valid career-building activities.

(1) There is a tendency to think of data as a scientific outcome not requiring as much intellectual effort as e.g. analysing it and writing up the results.

(2) Other resources that are cited, such as papers, colleagues or instruments have an obvious conceptual identity. Data, often being continuously collected, processed into many forms and easily copied, is not very visible and has an elusive identity in the minds of its users. Connected to this, there has been a lack of guidance from e.g. publishers on how to treat and refer to data sets used as a basis for publications. While many journals require authors to make their own outcomes available as supplementary information, there has been less attention paid to ensuring both that complete (raw) datasets are properly and sustainably curated, and that data from sources external to the authors are comprehensively and accurately cited.

(3) The pressures from funders have a general impact on institutional behaviour, but the forms these pressures take is varied and not consistent, and it rarely aligns with international practice adopted by environmental research communities. This aggravates the tendency of researchers to put data citation on the back burner.

(4) It is by no means obvious where the transitions from original data to new data entities occur. When data are aggregated, processed and derivatives created by applying expertise judgement, sophisticated novel algorithms or new representations and visualisations. It may become difficult or cumbersome to individually cite all used resources, especially in the absence of comprehensive provenance metadata attached to the resulting research outputs.

These issues can be understood as cultural differences, and the respective degrees of entrenchment in established working practices vary across different scientific disciplines. The situation is further complicated by inconsistencies in domain- and sub-domain-specific approaches to data sharing, including rules and licensing for downstream data usage. In order to navigate in the presence of these intellectual and socio-political challenges, it is necessary to find pragmatic strategies that work in the new era of “digital economy” and “big data”, while supporting sustainable practices in both research and socio-economic development efforts. However, for the foreseeable future we must accept that the adoption of recommended practices will not move at the same pace in all domains, and the uptake is unlikely to be universal.

Compared to other European research domains, the ENVRI Community research infrastructures have started the process towards integrate support for identification and citation into their data management practices. However, as evidenced by the requirements analysis performed by WP5 [Atkinson 2016], and subsequently confirmed in the work of WP6 (see deliverables D6.1 [Hellström 2017] and D6.2 [Hellström 2018], as well as this report), there remains much to be done.

2.2 Guiding principles for citation of data in scientific literature

In order to aid all involved actors to improve their practices, the CODATA-ICSTI Task Group on Data Citation Standards and Practices [Socha 2013], and then later also the Data Citation Synthesis Group of FORCE11 [FORCE11 2014], have set out to formulate sets of basic principles for data citation, which may be summarized as follows:

- Data is to be considered as legitimate products of research, and citations of data should be managed in the same way as references to e.g. articles. All contributors to a data set should be listed and recognized in a way that ensures they are given scholarly credit as well as normative and legal attribution.
- Claims made in scholarly literature that rely on data should include a citation to the corresponding data source. When citing data, the references should facilitate the access to and reuse of both data and associated metadata, code and other materials used in their production. Importantly, enough provenance and fixity information should be included in the citation to enable end users to retrieve exactly the same subset (time slice or other granular portion) used to support the original claim.
- Persistent and globally unique identifiers should be assigned to all data and then included in citations. The identifiers should be machine-actionable, in a way that ensures that the information resulting from the resolution of the identifier can be directly ingested by scripted workflows. Both the identifiers and the related metadata that describe a data object and its state should persist even after the data object has ceased to exist.
- The formatting and standards for data citation chosen by a given scientific community should reflect its needs and practices, while still supporting interoperability of data citation practices across disciplines and domains.

IN ENVRIplus, efforts have also been undertaken in order to provide guidelines for referencing data and other research-related resources. In Appendix B, we reproduce an excerpt from chapter 8 of the ENVRIplus deliverable D6.1 [Hellström 2017], summarizing the recommended citation best practices for the ENVRI Community. We note, however, that the responsibility for ensuring proper and unambiguous citation of data (and indeed all other components and resources used in research activities) must be shared by all involved actors: data producers, data managers and data scientists involved in curation, data publishers, data users and funding agencies [NSB 2005].

All these principles are highly relevant for the ENVRI Community research infrastructures, especially considering the lifecycles and processes traditionally followed by environmental and Earth science research projects, where both input and output data (as well as associated metadata) are expected to persist for a foreseeable future. Provided they are properly curated and managed, the data remain accessible and reusable also to researchers not directly involved in their production.

However, one very important aspect of collaborative data sharing in environmental sciences falls outside this expectation: the organised production, pooling and sharing of simulation output data at an agreed time for an agreed period. Originally developed in computational fluid dynamics and cosmology, this approach is now being used also in studies of climate, long-term ocean behaviour, and plate tectonic. The large data volumes involved, combined with the continuous refinement of the applied models, makes it impractical to store the model output data beyond a fixed period. During this period, the project participant community can fully exploit the data and all contribute to the scientific output, before the model output data is discarded, leaving behind only relevant metadata. This means that although the data used in the studies can be cited, it can no longer be retrieved, making the work practically irreproducible. This methodology for coordinated collaboration on computationally intensive (and hence energy-demanding and GHG-emitting) projects can be very effective, and should be encouraged in ENVRI-related research activities whenever significant CPU and storage resources need to be leveraged. However, this requires detailed and extensive preparations and planning, including agreements on data, representations and metadata.

2.3 Referencing data in machine-actionable workflows

Data sets are of course not to be seen in isolation; after all, science revolves around the analysis of data, often combined from multiple sources and processed in many consecutive steps with the help of several algorithms and software packages (not forgetting the human judgement steering and selecting those processes). As the volume of accessible relevant data is increasing dramatically, also in the environmental and Earth science domain, data processing needs to become more and more automated or machine-driven [EC2018, Wittenburg 2019]. At the same time, traditional practices of peer-to-peer sharing of data and metadata between colleagues (often in the same research sub-discipline) on request basis are not scalable due to increased demand and amounts of data objects involved. This is leading many communities to expose their data holdings via public catalogues, and to make their data downloadable (with or without access control) via trusted repositories – thus supporting unmediated data reuse both inside and outside of the originating scientific domain [EC 2018].

The FAIR principles [Wilkinson 2016] are a set of 15 broad guidelines and recommendations. FAIR stands for Findable, Accessible, Interoperable and Reusable, and can be applied to both data and metadata, as well as the digital representations of other entities including people, services and other resources. Wilkinson et al. stress that while applying the FAIR principles will also facilitate the work of individuals, the main intent is to support and facilitate machine-actionability of data and metadata resources. Several of the principles (F1, F3, and A1) explicitly mention the importance of assigning globally and unique persistent identifiers (PIDs) to data and related metadata, and describe how these PIDs can subsequently be used to link data and metadata objects together as well as support retrieval of the objects.

The idea of a Global Digital Object Network (GDON)[1] is an emerging concept that brings together making digital resources FAIR, supporting the pervasive application of machine actionability, and use of unique and persistent identifiers. The proponents of GDON argue that Digital Objects – consisting of core bit streams "wrapped" by metadata, typing information that defines allowable operations, and a persistent identifier with associated kernel information (see Figure 2-1) – are

[1] Previously known as Global Digital Object Cloud (GDOC), a term coined by L. Lannom of CNRI [Lannom 2017].

the building blocks of a future FAIR-compliant infrastructure for data management and processing. The multi-layered Digital Object model allows for abstraction, binding and encapsulation mechanisms that effectively decouple the end user (human or machine) from the actual physical resources of storage, computation etc. At the same time, GDON manages which operations are allowed to be performed while respecting related access rights, licenses and other limitations put on the DO use [Wittenburg 2019]. Optimally, the operations would be mediated via a standardised protocol, supporting both basic Create-Read-Update-Delete (CRUD) functionality[2] as well as the capability of initiating more advanced operations. One such protocol is Digital Object Interoperability Protocol (DOIP), developed by the DONA Foundation [DONA 2018].

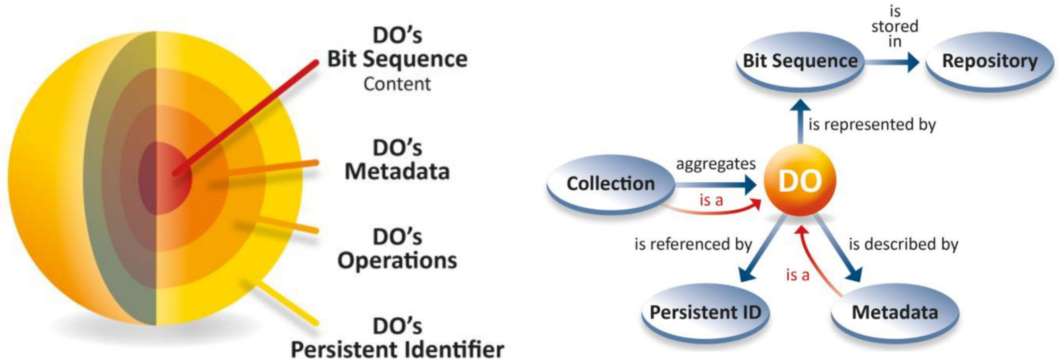


Figure 2-1: *Left: Central concept of the GDON approach, a Digital Object can be considered as a bit sequence encoding content, and described by metadata that enables access and interpretation of the content. The DO is assigned a type which governs which operations can be applied on the bit sequence. The DO's persistent and unique identifier allows it, its associated metadata and type to be located. Right: A Digital Object and its associated elements, as defined by the Core Data Model developed by the RDA Data Foundation & Terminology Group. Illustrations adapted from [Wittenburg 2019].*

The growing scale of data and the risks and climate impacts of moving it, call for virtualised implementations, where the *logical* packaging of the data is used to control and optimise the execution of workflows by delegating work on large cores to locations close to the data. This approach is already being implemented by IS-ENES, and the DRIP software (Dynamic Real-time infrastructure planner) [Wang 2017] goes some way towards implementing it in the SWITCH project [Zhao 2015]. Other ENVRI Community members that have expressed an interest in implementing, or at least supporting interoperability with DOIP-type operations, include ICOS and DiSSCo. These RIs have, together with other European and US research infrastructures and organisations, formed the C2CAMP (Cross Continental Collection Access and Management Pilot) initiative [C2CAMP 2018a], which aims to set up test bench pilots for testing GDON components and their interactions [C2CAMP 2018b].

For the processing of scientific information to be effective, it is vital not only that relevant data and associated metadata can be found, accessed and acted on via scripts and workflows, but also that adequate machine-interpretable provenance information is captured throughout the entire scientific workflow. A first step towards providing this is to include, in the metadata of output datasets, a list of resolvable persistent identifiers of all input data – a somewhat crude, but yet effective form of data citation. More stringent approaches towards full provenance collection and curation are described in e.g. the ENVRIplus WP8 deliverables D8.5 [Magagna 2018] and D8.6 [Goldfarb 2018].

[2] See e.g. https://en.wikipedia.org/wiki/Create,_read,_update_and_delete

2.4 Estimating “impact” by tracking downstream usage

In ENVRIplus, much of the discussions in Theme 2 around data usage accounting and impact assessment have focussed on the question of how to ensure that data producers are given due credit for their work. Indeed, as pointed out by the CODATA-ICSTI Task Group on Data Citation Standards and Practice [Socha 2013],

> “Credit is the universally recognized currency for both quality and quantity of scientific activity. Citing data allows the use of metrics to evaluate use and impact factor of data sets, potentially encouraging *data creators* to make their data available for use by others. This fosters transparency and enables recognition of scholarly effort.”

The European Commission Expert Group on FAIR Data recently released its final report and action plan for making FAIR data an integral part of the European Open Science Cloud (EOSC) [EC 2018]. The report makes a number of recommendations related to impacts, highlighting actions that must be taken to ensure benefits to producers of FAIR data. A central tenet (Recommendations 6 and 26) is the need to change the current metrics used for assessing the career and achievements of individual scholars to also include efforts to collect and “FAIRify” research data. Also, research infrastructures and other data service providers should be incentivised to support FAIR data through efforts to encourage reuse of FAIR outputs (Recommendations 21 and 24).

However, the spectrum of stakeholders that will benefit from the implementation of comprehensive and effective practices for usage tracking is much broader, including also host institutes and organisations, data centres and repositories, curators and other data managers, funding agencies, publishers, end user communities and even society as a whole [Socha 2013].

There are in fact many rationales for both identifying data user categories and quantifying the downstream usage of scientific outputs, such as articles, reports, monographs and now also data:

- Proving the usefulness of a project
- Allowing to give credits to the individuals and institutes behind the output
- Giving feedback to repositories and curators
- Providing feedback to funders about the return of their investments/grants
- Indicating to policymakers and stakeholders what “societal impact” has occurred

One often mentioned reason is the desire to measure the impact of a specific project or RI, for example through quantitative Key Performance Indicators (KPIs). But limiting the analysis to the collection of pure statistics can be very misleading. Estimating, let alone measuring, the overall impact of an RI is complex and multi-faceted. First, the scope of the “impact” must be defined, e.g. scientific domain-wise, societal, regional or something else, followed by the expression of the relevant facets as Key Impact Indicators (KIIs) against which the outcomes can be compared.

Secondly, the potential sources of information should be listed, together with their respective estimated trustworthiness and degree of completeness. Examples of sources related to scientific data usage include:

- Mentions of data sets in media – via full-text searches in media archives or the web
- Number of data publications – from bibliometrics and info collected by RIs
- Citations of data in scientific literature - bibliometrics
- Number and classifications of users: scientists, regional & global projects – from RIs
- Data usage:
    - Downloads – from repositories and portals (some operated by RIs)
    - Visualizations – from portals and similar services (operated by RIs)
    - Computational environments (human- & machine-driven processing) – from RIs

However, as stated in the recently published ICOS impact assessment report [van Belle 2018], simply relying on end users to voluntarily report back to an RI that their data has been used for a scientific purpose will invariably lead to an underestimate of total data use and impact. It is necessary to use data harvesting of different types to get a more complete sense of how much, for what purpose, and by whom RI data products are being used.

A further complicating factor is coupled to how the access to a dataset is enabled technically: some files or databases may be replicated across several data centres located across the world, possibly operated independently of each other. Unless stringent schemas of accounting for e.g. downloads or other access types are applied at all relevant centres, and reported regularly to the primary curating organisation, total usage statistics may be severely mis- or underreported. A special case is represented by the application of so-called "Digital object over cloud" or Named Data Networking (NDN) architectures, where digital objects may be cached in the memory of network nodes in order to optimise access times [Koulouzis 2018]. Here the caching mechanism must ensure that each individual (authorised and legitimate) access of the object is registered and reported back to the data owner or repository holding the original copy. Some end users may also pass on copies of data sets they downloaded to third parties. Unless licensing, usage instructions and relevant referencing information are included, the data may end up being inappropriately used or incorrectly cited.

Machine-initiated (actually script- and/or workflow engine-driven) actions can be tracked and counted by distinguishing programmatic access via (RESTful) API calls combined with the AAI methods + user agent type, but it may be difficult to correctly estimate the total number of derived outputs that originate from a single access of this type. Even more difficult is to count (or account for) human-initiated data usage, as each download does not necessarily equate one instance of use.

Finally, interesting insights into how data are actually being used and for what purposes, can be gained by not only collecting overall statistics and total numbers, but by looking for linkages and correlations between e.g. the cited/used data object itself and its characteristics, where the reference was made and by whom, and in what context the citation occurs. This approach, which builds on similar types of analyses for article citations pioneered by the scientific library community [Bollen 2008], has been taken up by several recent initiatives, including the "PID Graph" concept from Project FREYA (see Chapter 3).

3 CITATION SERVICES

As the needs for convenient and easy management of citations of data and other non-article research resources grow, more and more organisations and providers are entering the scene, offering an ever-increasing palette of services and functionalities. In this Chapter we provide an update of the service provider landscape surveys that we started in WP6 deliverables D6.1 and D6.2, before taking a look at the basic components of data publishing workflows.

3.1 Revisiting the service landscape

Many of the projects and initiatives that were described in the previous WP6 deliverables D6.1 [Hellström 2017] and D6.2 [Hellström 2018] are now making fast progress towards providing mature services, with associated examples of use case implementations in various organisations. In the following section we give a brief update on four of these initiatives: Project FREYA (see also Ch. 2.8 in D6.2), FORCE11 (Ch. 2.1 in D6.2), Scholia (Ch. 7.2 in D6.1 & Ch2.3 in D6.2), and Make Data Count (Ch. 6.3 in D6.1 & Ch. 2.4 in D6.2).

3.1.1 Project FREYA

The Horizon2020 Project FREYA [FREYA 2019] continues its work on focusing on the three following main areas:

1. “PID Graph” which is supposed to connect and integrate the PID systems
2. “PID Forum” which works with community engagements through networks such as Research Data Alliance (RDA)
3. “PID Commons” which focuses on the sustainability of the PID services developed within Project FREYA

During 2018 a couple of deliverable reports from Project FREYA were published, that all describe and explain different PID systems and the current status for the PID systems in a very detailed manner. In their deliverable report “D 4.1 Integration of Mature PID Types” [Lavasa 2018], the different FREYA partners report from different local tests and development of PID systems. This is done in a very concrete way and from the angles of the different disciplinary contexts. Through its Ambassador programme and its connection with European Open Science Cloud, Project FREYA has high ambitions to develop PID service systems that will be useful far beyond the project’s lifetime.

3.1.2 FORCE11

FORCE11 [FORCE 11 2019] continues to act as an important creator of public opinion. The FAIR Data Principles originally crafted by FORCE11 are gaining increasing importance among researchers and their stakeholders. One of the strengths of FORCE11 is to gather people to discuss and work together, and their recent FORCE11 conference held in Montreal in October 2018 brought to the stage a lot of new ideas and innovations around data citation and data metrics [FORCE 11 2018].

3.1.3 Scholix

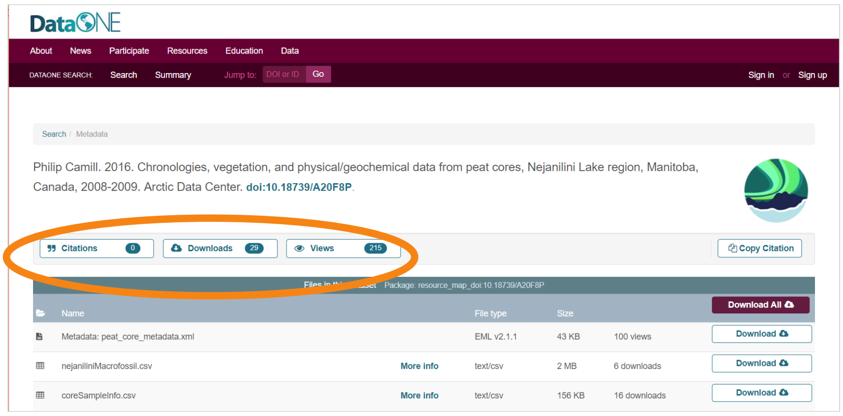
Recently a number of repositories and databases have contributed use cases that implement the technology developed by Scholix. These include the Life Science repository Europe PMC (https://europepmc.org/), the not for profit digital repository Dryad (https://datadryad.org/) and the abstract and citations database Scopus (https://www.scopus.com/) [Scholix 2019, Europe PMC 2018]. As an example of why repositories are participating in the Scholix initiative, Dryad notes that as they are members of DataCite[3], which is one of the Scholix hubs, it is natural to make their data-to-literature links available via Scholix aggregators such as the Data Literature Interlinking (DLI) service. Publishers may query the DLI to find datasets related to their journal articles, and generate a link back to Dryad, thus increasing data re-use, and facilitating research discovery [Dryad 2017]. This solution is very interesting and should be further investigated by the ENVRI Community research infrastructures.

3.1.4 Make Data Count

The project Make Data Count (MDC), described in the previous deliverable reports D6.1 and D6.2, is continuing their work, and has been involved in launching the COUNTER Code of Practice for Research Data Usage Metrics Release 1 [MDC 2018a].

[3] DataCite (https://datacite.org/) is a non-profit organisation that provides a DOI-based data identification service.

The release of this Code of Practice is an important step in the work for consistent and standardised reports of research data usage, and will hopefully help and guide organisations in their future of installing research data usage metrics. The MDC project has also helped two organisations, DataONE and Dash (California Digital Library), to implement metrics for research data usage and citation in their local services [MDC 2018b]. Figures 3-1 and 3-2 below illustrate how data citation metrics are displayed by DataONE and Dash, respectively.

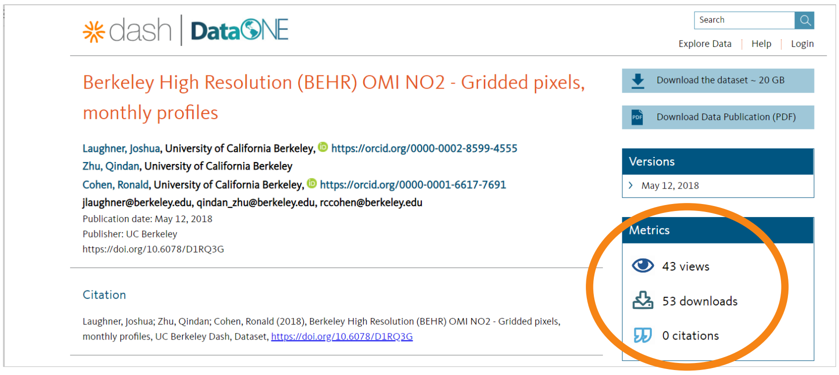


Philip Camill. 2016. Chronologies, vegetation, and physical/geochemical data from peat cores, Nejanilini Lake region, Manitoba, Canada, 2008-2009. Arctic Data Center. doi:10.18739/A20F8P.

Package: resource_map_doi:10.18739/A20F8P

Name		File type	Size		
Metadata: peat_core_metadata.xml		EML v2.1.1	43 KB	100 views	Download
nejaniliniMacrofossil.csv	More info	text/csv	2 MB	6 downloads	Download
coreSampleInfo.csv	More info	text/csv	156 KB	16 downloads	Download

Figure 3-1: *Example of DataONE shows citation metrics of a data set record, see* *https://search.dataone.org/view/doi:10.18739/A20F8P*



Berkeley High Resolution (BEHR) OMI NO2 - Gridded pixels, monthly profiles

Laughner, Joshua, University of California Berkeley, https://orcid.org/0000-0002-8599-4555
Zhu, Qindan, University of California Berkeley
Cohen, Ronald, University of California Berkeley, https://orcid.org/0000-0001-6617-7691
jlaughner@berkeley.edu, qindan_zhu@berkeley.edu, rccohen@berkeley.edu
Publication date: May 12, 2018
Publisher: UC Berkeley
https://doi.org/10.6078/D1RQ3G

Citation

Laughner, Joshua; Zhu, Qindan; Cohen, Ronald (2018), Berkeley High Resolution (BEHR) OMI NO2 - Gridded pixels, monthly profiles, UC Berkeley Dash, Dataset, https://doi.org/10.6078/D1RQ3G

Figure 3-2: *Example of data citation metrics for a data-set record at Dash (California Digital Library), see* *https://oneshare.cdlib.org/stash/dataset/doi:10.6078/D1RQ3G*

3.1.5 OpenAIRE

OpenAIRE which has been developing services for Open Science since 2008 now becomes a more permanent organisation, and will be delivering services to the future portal of EOSC [OpenAIRE 2018]. OpenAIRE also co-ordinates National Open Access Desks (NOADs), a network of local open access offices located in nearly every EU member state. The NOADs play an important role locally when it comes to training and support in Open Science. On the OpenAIRE portal the services are currently organised according to the following themes: Explore, Provide, Connect, Monitor, Develop. Behind each of these themes there are a set of services, tools and guidelines on how to use them [OpenAIRE 2019]. OpenAIRE addresses several target groups through these “themes”: researchers, content providers, research communities and funders. The question is whether OpenAIRE will be able to maintain and develop services and tools for all these target groups on a long-term basis. This is certainly a large task and may be a too high ambition. The strength of OpenAIRE could be to support in training and general guidelines to the target groups, and to act as a co-ordinator in the discussion of service integration between different research communities. This kind of support from OpenAIRE would be very useful in the ENVRI Community and later in the now starting ENVRI-FAIR project.

3.2 Data publishing workflows: Basic components

3.2.1 Making data citable

The data flow from producer to PID provider requires three basic components:

1. **Data producer** = a researcher who has created a dataset which he wants to share
2. **Repository** = a publisher where the researcher may publish their data, for example Zenodo, PANGAEA
3. **PID provider** = a service provider delivering persistent identifiers, for example DataCite, Crossref

When "Data producers" want to share their outputs, they start by searching for a suitable repository where to publish their data. They might need to register at the "repository" in order to publish data there. They also need to prepare the dataset for ingestion into the repository. This preparation may involve registering the provenance information of the dataset as well as additional technical details of the creation of the dataset. The data producers describe their dataset and upload it into the repository, according to the guidelines of the repository.

After the dataset has been ingested, it might undergo some quality control by the repository, by a data curator or a data librarian. At ingestion, the dataset will also be assigned a persistent identifier, according to administrative rules of the repository. Often, as in the case the identifier is a DataCite DOI, the existence of the dataset (and basic metadata about it) will be exposed and disseminated also through the channels of the "PID provider". See Figure 3-3 for a schematic of the process.

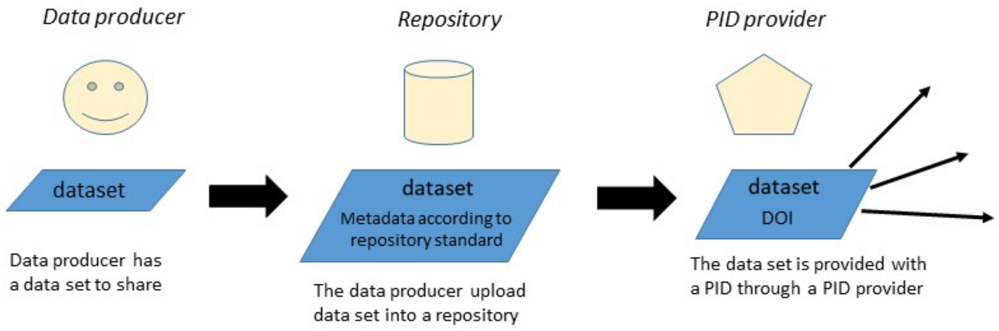

Data producer
Repository
PID provider
dataset
dataset
Metadata according to repository standard
dataset
DOI
Data producer has a data set to share
The data producer upload data set into a repository
The data set is provided with a PID through a PID provider

Figure 3-3 *The three stages of making data citable.*

There are many reasons for a researcher to share a dataset in a repository, not only in order to get credit for research work, it might also be a requirement from the research funder, or it might be for re-usability of the data. Factors such as costs, technology, user access might affect the researcher's choice of a repository for publishing research data. But, also the ways in which the repository will expose and disseminate the research data, might be of importance to the researcher, for example whether citation statistics of the research data will be displayed.

3.2.2 From PIDs to reference lists: creating useful citation strings

DataCite provides a service for creating citations from items, where it provides a simple interface to extract metadata automatically from a DOI and builds a full citation string. The service is a collaboration with Crossref, mEDRA[4], ISTIC[5] and JaLC[6]. The service extracts metadata from the

[4] The multidisciplinary European Registration Agency for DOIs, https://www.medra.org/

[5] China Institute of Science and Technology Information, http://www.doi.org.cn/portal/index.htm

[6] Japan Link Center, https://japanlinkcenter.org/top/english.html

following fields: Title, Creator, Publisher, Year and DOI. The service does not select metadata from the field "Collaborator", which would have been useful in items with resource type "Collection", neither from the field "Contributor". However, the service does support a long list of different citation styles, which is common when citing publications [DataCite 2019a].

3.2.3 Monitoring dataset usage statistics

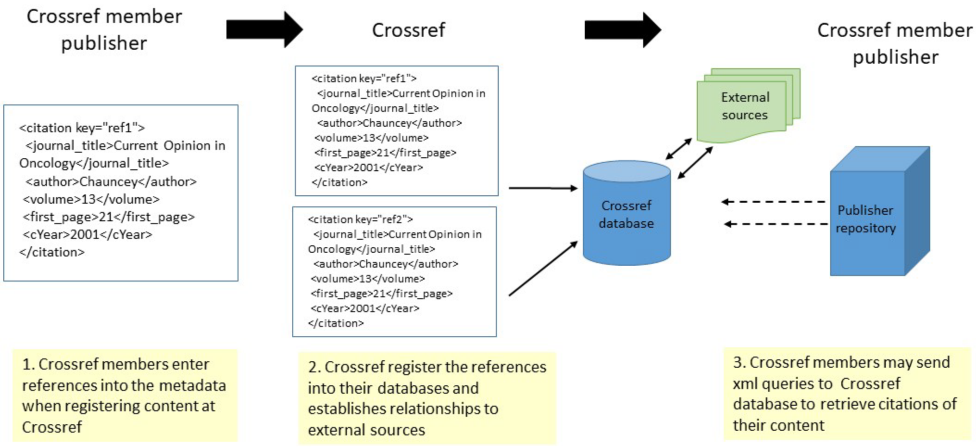
We have earlier described the "Make Data Count" project, which works on developing tools and standards for measuring data citations. A research infrastructure may measure their research data usage in many ways. To go from tracking citations of publication to monitoring of citations of research data citations is quite a big step for a research infrastructure. It requires the research data to be properly described and to be discoverable in systems where data citation metrics are implemented. However, systems for data citation metrics are at this moment still under development and being expanded, and its usage by research infrastructures is still limited. In the following section we will present some services for data citation metrics:

Event Data

DataCite and Crossref have together developed "Event Data" which could be seen as a shared infrastructure containing information about "events" that happen to DOIs outside their own metadata [DataCite 2019b, Crossref 2019, Dasler 2018]. Event Data is collecting and exposing links to DataCite and Crossref DOIs. Linking events are relations between two DOIs, or a DOI and a URL. For DataCite DOIs these links are described in DataCite metadata using the property <relatedIdentifier> with a list of controlled relation types that can be used [DataCite 2019b]. Members of DataCite can retrieve statistics of "events" that have happened to their DOIs via the DataCite REST API, where it is possible to do queries filtered by DOI, DOI prefix, source of the event, and relation type of the event. Usage reports of the "events" are provided according to the "Code of Practice for Research Data Usage Metrics" and the "SUSHI Specification for research data usage metrics". For ENVRIplus research infrastructures "Event Data" seems to be an interesting service for capturing usage statistics of their DOIs in a very specific way. As Event Data is a service under intense development, there is probably opportunity for ENVRI RIs to participate and test this service.

Cited-By

Crossref provides the service "Cited-By" to their members, which is a service where publishers can get information on who are citing their content (articles or other material with DOIs).



```
<citation key="ref1">
  <journal_title>Current Opinion in
Oncology</journal_title>
  <author>Chauncey</author>
 <volume>13</volume>
 <first_page>21</first_page>
 <cYear>2001</cYear>
</citation>
```

```
<citation key="ref1">
  <journal_title>Current Opinion in
Oncology</journal_title>
  <author>Chauncey</author>
 <volume>13</volume>
 <first_page>21</first_page>
 <cYear>2001</cYear>
</citation>
```

```
<citation key="ref2">
  <journal_title>Current Opinion in
Oncology</journal_title>
  <author>Chauncey</author>
 <volume>13</volume>
 <first_page>21</first_page>
 <cYear>2001</cYear>
</citation>
```

1. Crossref members enter references into the metadata when registering content at Crossref

2. Crossref register the references into their databases and establishes relationships to external sources

3. Crossref members may send xml queries to Crossref database to retrieve citations of their content

Figure 3-4. *The steps involved in the Crossref "Cited-By" service.*

The obligation of the publisher is to enter references in the metadata when registering content at Crossref. Crossref is then monitoring citations of the publisher's references, stores this information and gives access to it back to the publisher. The publisher may set-up alerts with Crossref for monitoring citations to their content [Crossref 2018]. Figure 3-4 illustrates the steps involved.

Google Dataset Search

During the autumn of 2018, Google launched a search engine for datasets called "Google Dataset Search" which should enable easy access to datasets to scientists, journalists and many other types of users [Google 2019]. Google encourages publishers, data repositories and other data producers to adapt their data to the new search engine, for example by implementing schema.org mark-up into the metadata flow [Noy 2018]. The Google Dataset Search engine is still in a beta version and in an early phase when it comes to content and coverage from data producers. However, it has potential to be a powerful search engine for datasets, if data producers start to integrate their metadata flows with Google Dataset Search on a broad basis.

Data Citation Index

The Data Citation Index (DCI) database, hosted by Clarivate Analytics, was described in detail in D6.1 [Hellström 2017]and their services are about the same now as in 2017. Clarivate Analytics declares to have a most focused engagement in covering data repositories in the database, currently about 350 repositories are included. Clarivate Analytics are in close partnerships with the major data services; DataCite and the Australian National Data Service (ANDS). This means that metadata from these two data services are included into Data Citation Index [Clarivate 2017, Clarivate 2019a). ANDS has developed a process with Clarivate that enables Research Data Australia records to be transformed into a DCI compliant format [ANDS 2017].

For a publisher or other type of data producer there are many benefits being included in Data Citation Index. First, it is a good way for a data producer to enhance discoverability of their datasets by being included in Data Citation Index, which is a database specialised in covering research data. Second, as Data Citation Index tracks and monitor citations in its records of datasets, the database provides a good way for the data producers to regularly track citations of their own datasets. Clarivate provides a web service API for publishers and data producers to track citation of DOIs and URLs to their own datasets [Clarivate 2019b]. We think that this API service will be of great interest to ENVRIplus RIs.

Dimensions

Dimensions from DigitalScience (https://www.digital-science.com/), is a service that performs searches for mentions of a given persistent identifier in a number of different sources, and then computes various citation-related metrics. Dimensions was originally conceived for tracking usage of articles, books and reports, but is in fact agnostic to the type of resource that is related to the identifier it searches for, as long as the identifier is supported (currently DOI, PubMed IDs and Dimension IDs). Sources that are indexed include Crossref, ArXiV, PubMed and PubMed Central, as well as openly available databases, which as of January 2018 contained over 90 million records.

In addition, Dimensions performs full-text searches (in part of the records) and is also, in collaboration with its sister company Altmetric (https://www.altmetric.com/), able to search e.g. social media, traditional media and funder records [Bode 2018]. The metrics relevant to data objects include "Times cited" (the number of times a publication was cited by other publications, for all years), and "Recent citations" (number of citation received in the last two years). (Other metrics, more relevant to articles, are the "Relative Citation Ratio (RCR)" and "Field Citation Ratio (FCR)".) The metrics can be extracted manually via the Dimensions discovery web application (https://app.dimensions.ai/discover/publication), or – for use on a repository's web site or on landing pages – be retrieved in numerical form via an API and/or displayed as embedded "badges" [Mori 2018]

4 THE PUBLICATION OF MARINE BIODIVERSITY DATA FROM PEER-REVIEWED JOURNAL TO EU DATA INFRASTRUCTURES

This Chapter, contributed by Dan Lear of the Marine Biological Association and EMBRC RI, presents the "use case g" as described in the WP6 Description of Action.

4.1 Background

Significant effort has taken place in the marine biological community to aggregate and standardise access to records of species and habitats. Through national and international initiatives including the UK Marine Environmental Data and Information (MEDIN) partnership, the European Marine Observation and Data network (EMODnet) Biology programme and the Ocean Biogeographic Information System (OBIS), unprecedented volumes of data are now accessible as open, standards-based resources. However it is recognised that some sectors are less engaged with the publication of open data. Traditional publication of biodiversity findings in peer-reviewed journals is a valuable source of species occurrence records that are not routinely captured by aggregators. Data journals, such as the Biodiversity Data Journal (https://bdj.pensoft.net/) are one mechanism to ensure effective data flow. However many, traditional journals routinely publish articles that contain valuable distribution and occurrence data. These data can often describe the first occurrence of a species or range extension which can be associated with impacts of invasive non-native species or changes in distribution relating to climate change or other pressures.

The aim of this use case was to develop a full data lifecycle model for biological data linked to the standards developed and promoted through the Global Biodiversity Information Facility (GBIF).

4.2 Proposed implementations

Through this use case we investigated the implementation of workflow using the Marine Biodiversity Records journal. Published by Bio Med Central, part of the Springer Nature group Marine Biodiversity Records publishes original research documenting the geographical changes in species distribution, including those related to the introduction of novel or alien species. As such, the journal's remit aligns with a clear need to rapidly disseminate such information and ensure the most up-to-date data are included in regional and global data aggregators.

Within the biodiversity community the global standards for the collation of species occurrence data are coordinated through GBIF. The GBIF community have put significant effort into the development of open source software to support the publication and sharing of biodiversity data. The Integrated Publishing Toolkit (https://www.gbif.org/ipt) is a Java-based application which includes support for the allocation of DOIs and is optimised for large datasets, for example 50M species occurrence records, which would include spatial and temporal aspects and details of the sampling methodology employed will occupy somewhere in the region of 3.5GB of storage.

The IPT is built around species data being held in Darwin Core Archive (DwC-A) format (https://github.com/gbif/ipt/wiki/DwCAHowToGuide). DwC-A is a standard developed by the biodiversity informatics community to share species-level occurrence data in comma or tab-delimited text files. The structure of related text files within DwC-A is organised in a 'star' formation to link related extension data files [GBIF 2017], see Figure 4-1. Each DwC-A file will also contain a descriptor file (meta.xml) and associated metadata, often using the Ecological Markup Language (EML) schema.

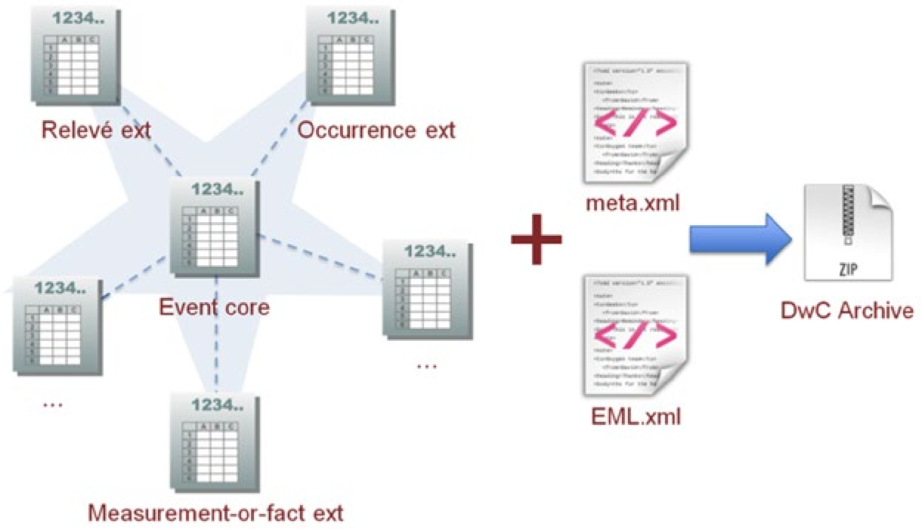


Figure 4-1. *The Darwin Core Archive model that is the basis of the GBIF Integrated Publishing Toolkit.*

Data can be stored within the IPT as either standalone text files or linked to a relational database system via ODBC and extracted on a predefined publication timescale. Version control and the issuing of DOIs is automated within the IPT through direct links to DataCite or EZID. Within the IPT species occurrence data are structured as 'resources'. These resources can cover a wide range of datasets including a single occurrence of a species with a distinct geographic area, a large interdisciplinary dataset covering a whole oceanic region or a time series with a range of species being recorded over several decades. The granularity of the resource is the decision of the data custodian, granting great flexibility in the manner in which the data are published.

The processes of ingestion and publication within the GBIF IPT can be broken down into distinct steps, as illustrated in Figure 4-2 below. In order to ensure interoperability with the global GBIF infrastructure, it was decided that the use-case would implement an instance of the IPT on the servers at the Marine Biological Association (MBA) data centre. It would then be necessary to liaise with the editor and publishers of the Marine Biodiversity Records to ensure the relevant pipeline was in place to ensure data relating to the publishes articles could be captured and ingested within the MBA-hosted IPT.

The GBIF community have developed a number of Microsoft Excel-based templates for the capture of species occurrence records that can be easily ingested by an IPT instance. These templates would be provided to authors via the online submission process for Marine Biodiversity Records. Submission of the completed template was via a dedicated email address and automated ingestion of the standardised templates into the MBA IPT instance.

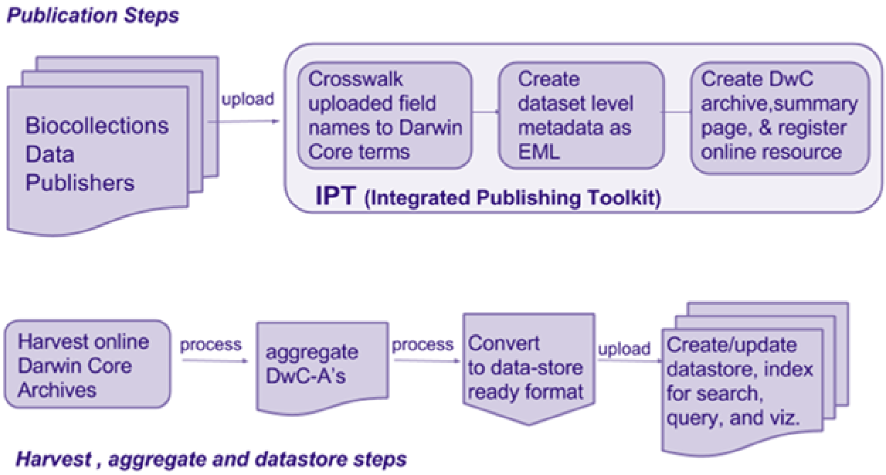

Publication Steps
Biocollections Data Publishers
upload
Crosswalk uploaded field names to Darwin Core terms
Create dataset level metadata as EML
Create DwC archive,summary page, & register online resource
IPT (Integrated Publishing Toolkit)
Harvest online Darwin Core Archives
process
aggregate DwC-A's
process
Convert to data-store ready format
upload
Create/update datastore, index for search, query, and viz.
Harvest , aggregate and datastore steps

Figure 4-2. *The ingestion and publication processes of the GBIF Integrated Publishing Toolkit.*

4.3 Results

The installation and configuration of the IPT application was straightforward. The Java application was deployed within Apache Tomcat, hosted at the MBA data centre. Early discussions took place with the publishers and Editor-in-Chief of the Marine Biodiversity Records journal and an agreement was put in place that the process for the submission of template-based species occurrence data.

The instructions for authors[7] now includes the following text:

> "Marine Biodiversity Records is pleased to offer an exciting new service to authors. In partnership with DASSH, the Archive for Marine Species and Habitats Data, we are now able to provide an innovative data publication service, increasing the visibility of your research and making the data citable through the allocation of a DOI. By completing the simple template here (http://www.dassh.ac.uk/mbrSpeciesRecord.xlsx) and returning the completed template to dassh.enquiries@mba.ac.uk the data will be managed within the archive, made accessible via a unique and persistent DOI and contribute to international data aggregation projects such as OBIS and GBIF."

However, despite the technology side of the project being simple to implement and all the required processes being established, there was extremely limited engagement from publishing authors. The mechanism and tools were available to the authors for an initial period of 12 months before further contact was made with the publishers.

We initiated discussions with the publishers to gain further insight into the lack of engagement. There were concerns raised around the linking to downloads on 3rd party websites, which was easily addressed by providing a copy of the occurrence templates for inclusion on the online submission portal. Further investigation of the journal submission portal web usage statistics showed a surprisingly limited number of access requests for the submission guidelines. The information regarding data submission was therefore replicated within the revision letter templates that are generated and sent to all authors during the peer review process

The SpringerNature group have a tiered approach to data sharing within their published policy, as summarized in Table 4-1 below [SpringerNature 2019].

[7] https://mbr.biomedcentral.com/submission-guidelines/preparing-your-manuscript/marine-record

Table 4-1. *SpringerNature group's Research Data Policy Types.*

Policy Type	*Policy Summary*
Type 1	Data sharing and data citation is encouraged
Type 2	Data sharing and evidence of data sharing encouraged
Type 3	Data sharing encouraged and statements of data availability required
Type 4	Data sharing, evidence of data sharing and peer review of data required

The Marine Biodiversity records journal is a Type 3 publication, and therefore the submission of data alongside the manuscript cannot be enforced and relies on the willingness and engagement of the authors. Without clear benefits for each author, the submission of data is seen as an extra unnecessary step, despite measures being put in place to ensure the process is not overly onerous or complex.

Additionally the use-case and other work undertaken within WP6 have identified issues relating to the persistence of citations in data aggregated to GBIF. In part, these issues came to light due to a change in status of the MBA Data Centre during the lifetime of the ENVRIplus project. In January 2018, the MBA became the accredited UK Node of OBIS, as such the aggregation of data at the national level and subsequent publication for harvesting by the European OBIS node (EurOBIS) became a primary responsibility. OBIS supports the use of the GBIF IPT as the preferred technology platform for the publication of marine biodiversity data, and as such limited infrastructure changes were required.

However, the OBIS community have undertaken significant work in extending the GBIF 'star' schema approach to publishing data into a form that can capture the complexities of marine biodiversity observations and the associated data that is often collected alongside traditional occurrence records, such as species traits or environmental parameters.

The OBIS-ENV schema (http://bdj.pensoft.net/articles.php?id=10989), see Figure 4-3, provides an alternative structure for the DwC-A format by addition an addition Extended MeasureOrFacts file, linked to either the event or occurrence that is being described [De Pooter 2017]. This additional file allows complex marine biological studies and ocean observation activities to be represented in an "event hierarchy".

Through our work implementing the new OBIS-ENV based approach, we became aware that upon ingestion citations were either being replaced or presented in different formats. This obviously presents problems of provenance and consistent citations across aggregators. There is an ongoing discussion (see e.g. https://github.com/gbif/registry/issues/43#issue-307247880) taking place on these issues between the OBIS and the GBIF community, but at the time of writing it is unclear if the proposed solution has been implemented across all OBIS nodes.

This use-case has been valuable in demonstrating the disconnect between the availability of standard-based technology solutions and their uptake by the targeted community. Given further resources and time we hope to stimulate and increase the interaction through targeted outreach and engagement, and by ensuring the benefits are clearly communicated to the marine biodiversity research community. Such engagement could include additional training, especially for early career scientists to develop the culture of data publication, and in building a vocal and proactive community who recognise and promote the inherent benefits.

By developing a better understanding of the reasons for the lack of uptake we can modify the tools, improve interfaces and remove the barriers, both real and perceived, to ensure data are rapidly integrated and made as widely accessible as possible.

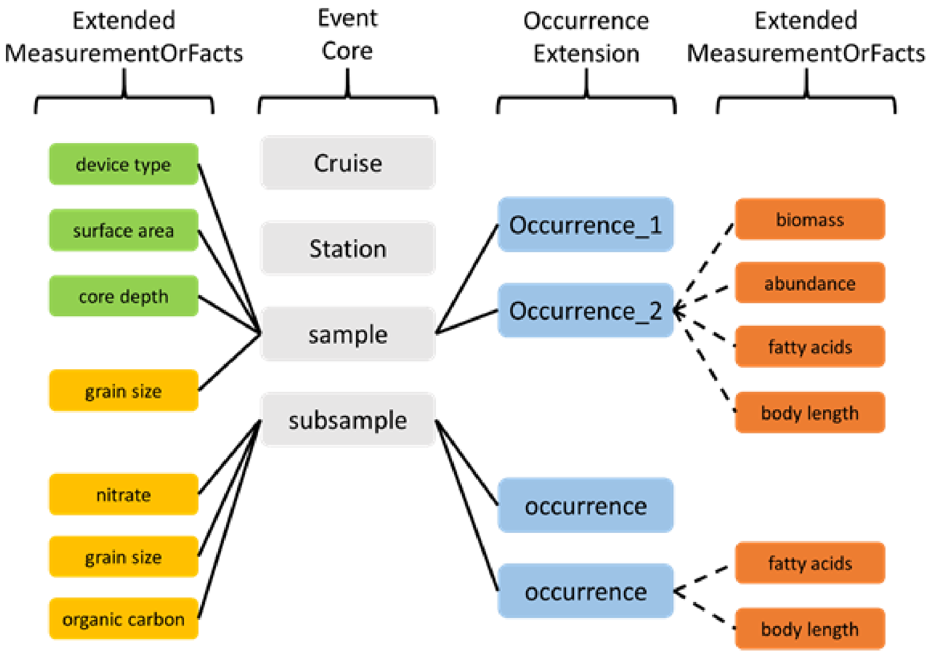


Figure 4-3. *Example of an OBIS-ENV schema with possible additional parameters in the Extended MeasurementOrFacts, see text for details.*

5 DEVELOPMENT OF A CITATION AND USAGE TRACKING SYSTEM FOR GREENHOUSE GAS MONITORING DATA

This Chapter, contributed by Margareta Hellström, Maria Johnsson and Alex Vermeulen of the ICOS Carbon Portal, presents the "use case h", as described in the WP6 Description of Action.

5.1 Background

Already from its conception, the ICOS research infrastructure has been aware of the importance to 1) collect, process and manage ICOS data in a standardised, transparent and trustworthy manner; 2) distribute all ICOS data products in a way following Open Data principles; and 3) track the usage of ICOS-related data and other outputs. Connected to points 2) and 3), the ICOS Data Policy [ICOS ERIC 2016] states that:

> "It is important for ICOS RI and ICOS ERIC and further Data Users to acknowledge the persons and organisations, which have originally generated the ICOS Data or processed the different levels of ICOS Data. For this purpose, a persistent identifier with the information of the Data Providers/authors will be accompanied with every ICOS Data set. ICOS RI and Carbon Portal of the ICOS ERIC will seek the most feasible technical solution for attaching the identifiers to the ICOS Data together with clear information about how to properly acknowledge the ICOS community in the products that make use of ICOS Data. ICOS RI and Carbon Portal will also seek technical solutions to maintain a proper and up-to-date attribution/citation database, tracking the use of the ICOS Data to individual research paper, with a strong traceability even to the individual sites/instruments."

This subchapter aims to provide background on ICOS RI by briefly describing its organisation, services, data and metadata flow and strategies for assigning persistent identifiers.

5.1.1 About ICOS

ICOS, Integrated Carbon Observation System, is a pan-European research infrastructure with a mission to provide standardized, long term, high precision and high quality observations on the carbon cycle and GreenHouse Gas (GHG) budgets and their perturbations. The ICOS observing network consists of over 130 observation stations, each related to one or more of the three domains Atmosphere, Ecosystem and Ocean, and operated by its (currently 12) member countries. The collected data is processed and quality controlled at Thematic Centres (one for each domain), before being openly distributed via the ICOS data centre named Carbon Portal (CP). All ICOS data are meant to be FAIR: easy to find, available for open access, fully traceable and complete with all relevant metadata, and interoperable with other (environmental) data and services.

5.1.2 Services offered by the ICOS Carbon Portal data centre

The ICOS Carbon Portal (https://icos-cp.eu/) is tasked with designing and implementing the set of services that form the backbone of the data management of the whole ICOS RI. The service list includes 1) data ingestion & storage, including the minting of persistent identifiers; 2) staging data from the repository to high-performance and high-throughput computing resources; 3) easy-access cataloguing on top of an ontology-based metadata database (RDF triple store accessible via a SPARQL endpoint); 4) single-sign on AAI (authorization, authentication and identification) for ICOS services; 5) a virtual research environment (VRE) platform for user-initiated data processing (based on Jupyter Notebook running on virtual machine instances); and 6) data discovery, including searching, visualising and downloading of ICOS-related data products including usage tracking. As far as possible, the CP bases all its data management and computing services on Open Source technology, and developed code and relevant documentation is maintained through a GitHub repository (https://github.com/ICOS-Carbon-Portal).

5.1.3 The flow of data and metadata in ICOS

Figure 5-1 below illustrates the overall flow of data to and from ICOS. Observations are made by the measurement stations. The raw data are passed on to the ICOS repository, and then processed by the respective three Thematic Centres (Atmosphere, Ecosystem and Oceanic). Some physical samples are processed by the Central Analytical Laboratories and the Ecosystem Thematic Centre. Near-real time data, as well as finalized & aggregated data products are passed on to the Carbon Portal data centre for long-term storage and curation. End users can locate, view and access ICOS data via the Carbon Portal discovery services.

ICOS central facilities, and external users can also use the Carbon Portal to curate and disseminate their own operational or research outputs, which are then labelled as "elaborated" data. In addition, the Carbon Portal also hosts and curates a number of legacy data products from greenhouse gas- and environmental monitoring projects pre-dating ICOS. The main end user communities are scientific users from inside and outside the ICOS community – including other research infrastructures – as well as ICOS station managers and technicians; but also citizen scientists, students, the general public and policy makers are important target users.

With its distributed architecture (the head office, data portal and thematic centres co-hosted by a total of nine countries) and coverage of three domains, ICOS needs an easy-to-maintain cataloguing system to hold the information about stations, people, instruments, data objects and documents. Metadata on all of these are collected at all levels throughout the data lifecycle, and most of this information is ultimately stored in a common ICOS metadata store, hosted by the CP.

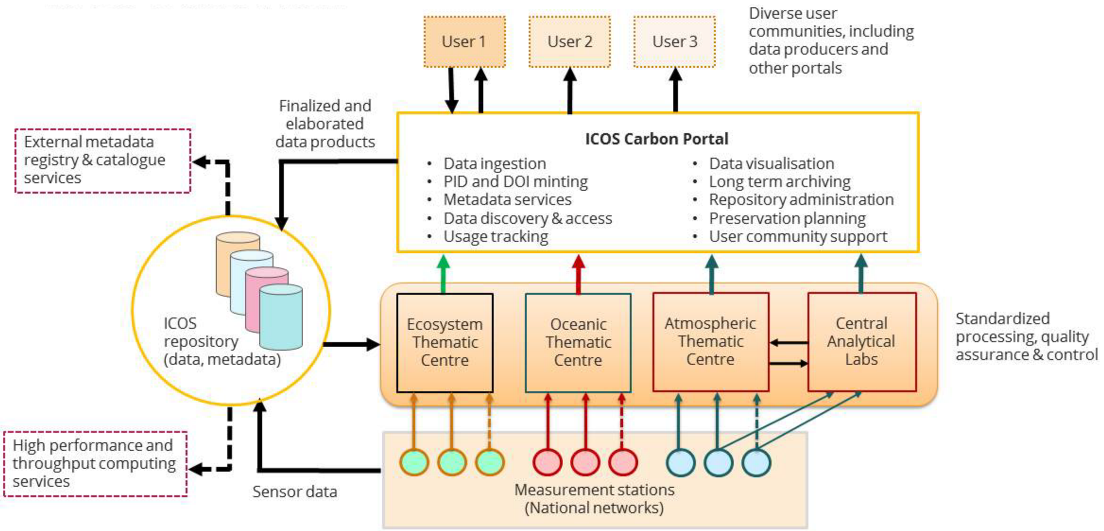

User 1
User 2
User 3
Diverse user communities, including data producers and other portals
Finalized and elaborated data products
External metadata registry & catalogue services
ICOS Carbon Portal
Data ingestion
PID and DOI minting
Metadata services
Data discovery & access
Usage tracking
Data visualisation
Long term archiving
Repository administration
Preservation planning
User community support
ICOS repository (data, metadata)
Ecosystem Thematic Centre
Oceanic Thematic Centre
Atmospheric Thematic Centre
Central Analytical Labs
Standardized processing, quality assurance & control
High performance and throughput computing services
Sensor data
Measurement stations (National networks)

Figure 5-1. *A schematic view of the data flow in ICOS, indicating the roles of the ICOS Central Facilities – the three Thematic Centres, the Central Analytical Labs and the Carbon Portal data centre & repository. Connections to external e-services (portals and computational services) and various end user communities are also indicated.*

The metadata store is based on semantic web and linked data – a design that is both flexible and scalable, while ensuring interoperability and compliance with relevant international standards, including GeoDCAT-AP, ISO-19115 and the European INSPIRE directive.

As there was no pre-existing, 100% applicable, data model, ICOS started to build its own ontology, with Prov-O [W3C 2013] as a major inspiration. The Protegé tool is used to build, edit and manage the ontology components, which are then exported as OWL (Web Ontology Language). The instantiation of the ontologies is implemented in the backend and in the frontend as web application (the ICOS meta service, see https://github.com/ICOS-Carbon-Portal/meta) written in Java and Scala. The ICOS data model is currently being revised and extended [D'Onofrio 2019].

The metadata are stored as time-stamped assertions in RDF (Resource Description Framework) triple format in a triple store database, based on the RDF4J framework. The time-stamping makes it possible to extract the status of a given metadata item at any point in time, so-called "time machine" functionality. ICOS makes all info available via a SPARQL end point for use by humans and machines (ICOS and external parties); see the client interface (with examples) at https://meta.icos-cp.eu/sparqlclient/.

5.1.4 Strategy for assigning PIDs to ICOS data products

The ICOS Carbon Portal is designed to manage and/or distribute data objects of a number of different categories. All data objects are assigned a Handle system-based persistent identifier at the time of ingestion into the Carbon Portal repository. The PID of these individual data objects can be resolved via e.g. the handle.net resolver service, which redirects to the object's landing page hosted by the Carbon Portal. The landing page lists the most relevant metadata of the object, including a direct link to access to the data object.

The Handle identifiers have the form "11676/<suffix>", where "11676" is the ICOS-specific Handle.net prefix, and <suffix> is a globally unique string provided by ICOS, based on the base-64 value of the SHA-256 hash sum of the data object it refers to. (Additional checks for uniqueness are performed before the registration of a new object's identifier, in order to avoid clashes with existing PID records.)

The types of data objects currently curated and managed by the Carbon Portal include:

- ICOS-produced data products
    - Raw observations ("level 0") from Atmosphere, Ecosystem and Ocean measurement station networks
    - Near-real time observational data ("level 1" or "NRT"; aggregated half-hourly or hourly values), automated quality control
    - Finalized observational data ("level 2"; aggregated half-hourly or hourly values, full quality control)
- Elaborated products directly related to ICOS ("level 3"), for example
    - Atmospheric modelling outputs (flux inversions)
    - Spatially interpolated flux fields based on eddy flux or ship board measurements
    - Model or remote sensing input fields
- "Legacy" observational data products, for example
    - Precursors of ICOS
    - InGOS[8]
- Other important externally contributed data sets
    - Global Carbon Project[9] outputs

Objects belonging to the other categories ("Level 2", "Level 3" and the other types) are also registered with DataCite, either as single objects or as collections of data objects. This means that they are also assigned a DOI and that associated metadata are stored in the DataCite catalogue following the DataCite Metadata Schema [DataCite 2017]. This allows the data objects to be found also through searches on the DataCite portal, and also provides full integration with the Citation Formatter service from Crossref and DataCite.

Resolving the DOI, e.g. via the resolving services of Handle.net and the DOI Foundation, results in a redirect to the landing page of the object or object collection, hosted by the Carbon Portal. The DataCite DOI identifiers have the form "10.18160/<suffix>", where "10.18160" is the ICOS-specific DataCite prefix, and <suffix> is a globally unique string. (The strings used for DOIs are also computed starting from the data object hash sum, but are designed to be shorter and easier to read for humans. As in the case of the Handle PIDs mentioned above, additional tests for uniqueness are of course performed before submitting the DOI registration request.)

Any updated versions of a given object, for example dynamic, growing time series or corrected data sets, are assigned a completely new PID. In order to provide unbroken provenance chains, the metadata record of the old version is updated with a link to the superseding object, and vice versa. This strategy is applied to data objects of all types described above.

5.2 Proposed implementations

ICOS needs to collect information about the use of its data and services, in order to assess its scientific and societal impact. This is requested by funding agencies and other stakeholders that are financing the operations and construction of the ICOS components. Because funding comes from funders, institutes and universities in the member countries, it is important to be able to break down the statistics to applicable granularity, such as measurement station and country.

[8] Integrated non-CO2 Creenhouse gas Observing System, https://www.ingos-infrastructure.eu/

[9] https://www.globalcarbonproject.org/

In order to obtain statistics of the usage of its data products and services that are as complete as possible, ICOS needs to be able to collect information across a multitude of sources and platforms. Data download statistics will be gathered and applied to the records of all contributors by the Carbon Portal. Similarly, all searches and visualization requests to the Carbon Portal are logged. (See 5.3.1 below.)

In addition to data usage tracking, ICOS will also need to periodically carry out harvesting of "traditional" bibliometric data sources, to extract information on articles describing research made using ICOS data, technical reports describing measurement protocols and instrumentation development done at ICOS measurement stations and Thematic Centres, and other mentions of ICOS in media, see 5.3.2 below for an example.

However, when data objects are cited on the web or in literature it is still an open issue how to harvest this use and how to attribute the contributors. There are different solutions for a data publisher such as the ICOS Carbon Portal to track and monitor data citations of its datasets. Firstly, the data publisher should have a clear overview of the export data flow from its data portal. To which external websites or repositories is data transferred from the original data portal? And what happens to the data at external websites or repositories? Secondly, there are several ways of monitoring how data is used outside the original data publisher, i.e., citations, downloads etc.

ICOS Carbon Portal is therefore engaging with citation indexers and publisher organisations in order to identify practical – preferably automated, or requiring minimal manual operations – sustainable solutions for citation statistics tracking. Three services are under review at the time of writing: "Event Data" from DataCite, "Data Citation Index (DCI)" hosted by Clarivate Analytics, and "Dimensions" from DigitalScience. Most likely, ICOS will use a combination of all three services, as they provide complementary types of information. (See chapter 3.2 for an overview of current services for tracking data citation and data usage).

The "Event Data" service is an interesting option, as ICOS is already associated with DataCite through its generation of DOIs. Even if there might be some development work for ICOS through adjustment of queries to the DataCite API, it would be a good opportunity for knowledge exchange with DataCite. The team behind the "Event Data" service is also putting a lot of effort into the standards for data-usage reports through its cooperation with "Code of Practice for Research Data Usage Metrics" [MDC 2018a], which is most interesting to ICOS.

The second service "Data Citation Index (DCI)" could also be a good choice. DCI is steadily expanding its coverage of data repositories, now about 350 are included. Being included as a data repository in DCI means that the registered datasets will be regularly tracked and monitored when it comes to citations of the datasets. And the DCI database is also offering several ways of displaying and visualising citation reports, based on the long experience from their other databases within Web of Science. It will be worthwhile to investigate this further both for ICOS and for the RIs within the ENVRI Community.

The third option would be to use the "Dimensions" service, which offers different ways to extract citation statistics for individual data sets or articles via their Metrics API. Of greatest interest to ICOS data producers, data users and national funders are the "Times cited" and "Recent citations" metrics. The output from the API queries can easily be incorporated into the ICOS metadata catalogue and/or embedded "on the fly" e.g. on data object landing pages – either as numbers or as a Dimension "badge". Use of the Dimensions API and "badge" services is free for non-commercial users like ICOS.

5.3 Results

ICOS’ efforts towards integrating the use of persistent identifiers, support for citation, and collection & visualisation of data usage statistics into the daily research data management are still in progress. In this section, a number of outcomes realised thus far are briefly described.

5.3.1 Adapting the ICOS data model to support identification and citation

One important issue that has emerged during the work is the need to fully integrate all identification and citation-related aspects into the ICOS data model (discussed above in 5.1.3). For identification and citation purposes, it is important that the ICOS data model contains entries corresponding to all relevant entries in e.g. the Dublin Core [DCMI 2012] and DataCite [DataCite 2017] metadata schemas.

This includes Author/Creator, Publication year, Publisher, Identifier, as well as (if applicable) other details associated with the curation and management of the digital object. In addition, information on the ICOS object classification (e.g., data Level, which Thematic Centre processed it, originating measurement station, time interval covered etc.) is often needed in order to compose a descriptive Title. At present, work on streamlining – and as far as possible, automating – the process of collecting relevant underlying metadata from the three domains’ measurement stations, Thematic Centres and other actors involved in the production of ICOS data sets is in progress, with special emphasis placed on ensuring that updates and changes are recorded promptly and accurately.

5.3.2 Resolving ICOS PIDs to dynamic landing pages

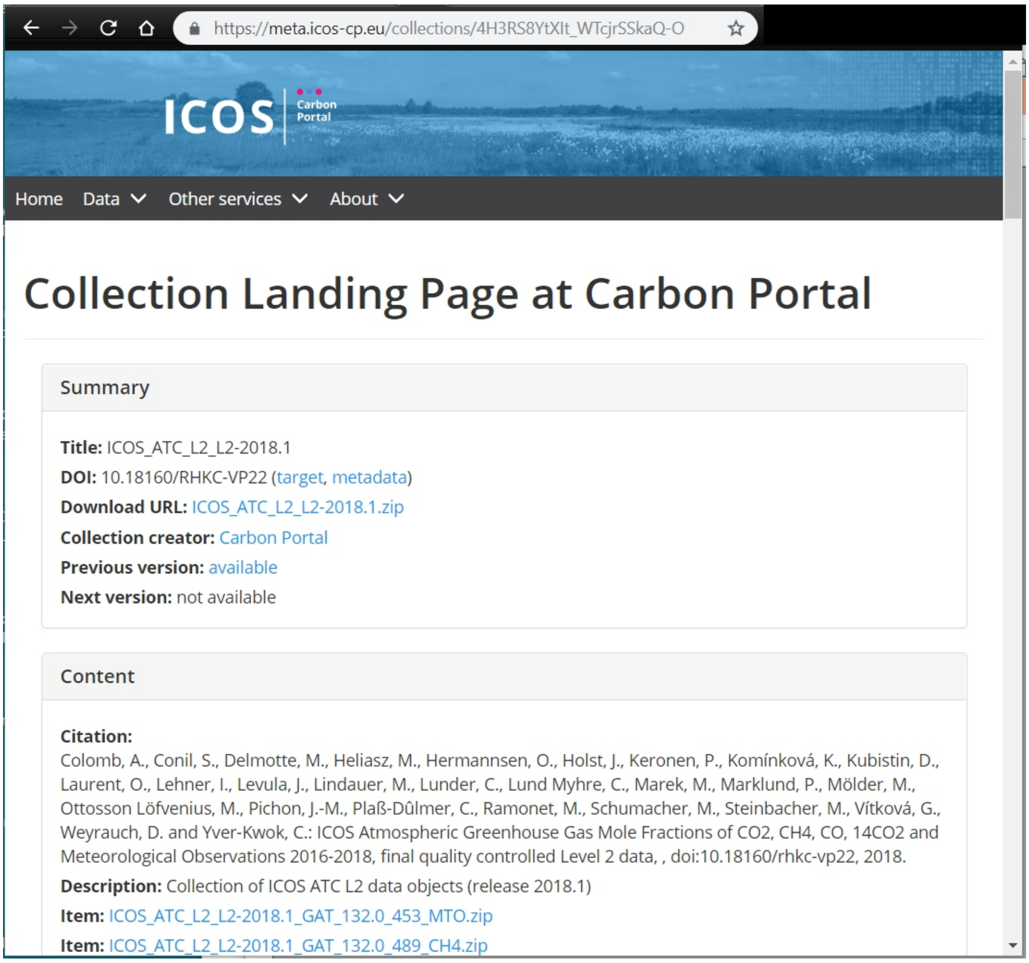
The Carbon Portal is, thanks to its linked data approach, able to serve a landing page for any digital object described in the metadata store – including data sets, but also stations, data type specifications and concepts. All landing pages are created dynamically, i.e. at the moment that their URL is accessed. This means that the displayed information always reflects the current, most up-to-date information. The format, and what information is shown, will vary between the type of objects, with the richest content provided for data objects.

As illustrated in Figure 5-2 below, which shows a screen shot of the upper part of a landing page for a Level 2 “collection” object, Summary and Content metadata are included. On landing pages for L0, L1, L2, and L3 data objects, other metadata categories are also presented, giving details about e.g. Acquisition, Production, Submission and Spatial Coverage.

Content negotiation is supported, allowing end users to specify the desired format of the response by specifying the “Accept:” parameter in the HTTP Request Header. Supported formats are “text/html” (output is HTML, default for web browsers), “application/json” (output in JSON), “application/xml; charset=UTF-8” (output is RDF) and “text/plain; charset=UTF-8” (output is plain text). The three latter options are mainly useful to machine-driven processing of landing pages.

It should be noted that it is currently not possible to directly access the contents of an ICOS data object by resolving its Handle PID, since this always redirects to the landing page, but the landing page contains the link to the actual data object (Download URL: in Figure 5-2). Accessing the object triggers the data licence check, which by default requires a manual acknowledgement by (human) users that they accept the ICOS data policy and data license.

Only after completion of this step, and the subsequent download, is the download counter updated. (When accessing objects through the API of the ICOS data service, the data licence check can be bypassed by passing an API token. Alternatively, users can register at the Carbon Portal and choose to include the acknowledgment of the license into their account profile., or when a user is logged in, and has acknowledged agreement to the licence in her user profile.)


https://meta.icos-cp.eu/collections/4H3RS8YtXlt_WTcjrSSkaQ-O
ICOS
Carbon Portal
Home
Data
Other services
About

Collection Landing Page at Carbon Portal

Summary

Title: ICOS_ATC_L2_L2-2018.1
DOI: 10.18160/RHKC-VP22 (target, metadata)
Download URL: ICOS_ATC_L2_L2-2018.1.zip
Collection creator: Carbon Portal
Previous version: available
Next version: not available

Content

Citation:
Colomb, A., Conil, S., Delmotte, M., Heliasz, M., Hermannsen, O., Holst, J., Keronen, P., Komínková, K., Kubistin, D., Laurent, O., Lehner, I., Levula, J., Lindauer, M., Lunder, C., Lund Myhre, C., Marek, M., Marklund, P., Mölder, M., Ottosson Löfvenius, M., Pichon, J.-M., Plaß-Dûlmer, C., Ramonet, M., Schumacher, M., Steinbacher, M., Vítková, G., Weyrauch, D. and Yver-Kwok, C.: ICOS Atmospheric Greenhouse Gas Mole Fractions of CO2, CH4, CO, 14CO2 and Meteorological Observations 2016-2018, final quality controlled Level 2 data, , doi:10.18160/rhkc-vp22, 2018.
Description: Collection of ICOS ATC L2 data objects (release 2018.1)
Item: ICOS_ATC_L2_L2-2018.1_GAT_132.0_453_MTO.zip
Item: ICOS_ATC_L2_L2-2018.1_GAT_132.0_489_CH4.zip

Figure 5-2. *Screenshot of the ICOS landing page for the data object ICOS_ATC_L2_L2-2018.1 - Collection of ICOS ATC L2 data objects (release 2018.1). The page exposes basic metadata about the data set itself, its contents, and a recommended citation string – the latter includes the data set's DataCite DOI number, https://doi.org/10.18160/RHKC-VP22.*

5.3.3 Supporting the citing of ICOS Data

ICOS data is distributed using a Creative Commons Attribution 4.0 International (CC BY 4.0) license[10], which users have to accept before they can access the data. The CC BY license requires end users to give appropriate credit (i.e. citing data when it is used), provide a link to the license, and indicate if changes were made when re-distributing the data. When citing ICOS data, at a bare minimum the data object's persistent digital identifier should be given in a machine-actionable form (as a HTTP URL, for example http://hdl.handle.net/11676/6PrNhZelwXKHLqO41QRsbheu or https://doi.org/10.18160/RHKC-VP22); this minimal form is sufficient for inclusion in provenance records, but for use in scientific literature much more information (contributors, data set name etc.) is of course required.

[10] https://creativecommons.org/licenses/by/4.0/

As a service to end users, ICOS provides suggested citation strings on the respective data object landing pages. Examples include:

- "ICOS RI, 2019. ICOS ATC NRT Meteo growing time series, Hohenpeißenberg (93.0 m), 2018-06-01-2018-07-16. https://hdl.handle.net/11676/AdviZyTQSYy8eZHwr4ngktga" (a Level 1 data set)
- "ICOS RI, 2018. ICOS ATC CH4 Release, Hohenpeißenberg (93.0 m), 2017-02-15-2018-05-31. https://hdl.handle.net/11676/-ffoiHjX5NDN0Vq_fKuVmas0" (an individual level 2 data set)
- "Colomb, A., Conil, S., Delmotte, M., Heliasz, M., Hermannsen, O., Holst, J., Keronen, P., Komínková, K., Kubistin, D., Laurent, O., Lehner, I., Levula, J., Lindauer, M., Lunder, C., Lund Myhre, C., Marek, M., Marklund, P., Mölder, M., Ottosson Löfvenius, M., Pichon, J.-M., Plaß-Dûlmer, C., Ramonet, M., Schumacher, M., Steinbacher, M., Vítková, G., Weyrauch, D. and Yver-Kwok, C.: ICOS Atmospheric Greenhouse Gas Mole Fractions of CO2, CH4, CO, 14CO2 and Meteorological Observations 2016-2018, final quality controlled Level 2 data, , doi:10.18160/rhkc-vp22, 2018." (a collection of Level 2 data sets)

Further means of assisting users with data citation are being planned, including providing the possibility to download bibliographic metadata in e.g. BibTeX format via links on the landing pages, as well as passing on detailed citation information together with downloaded data objects. (Currently only the persistent identifier is included.)

5.3.4 Collecting bibliometric data usage statistics

As discussed under 5.2 above, ICOS does not yet have a system in place for harvesting data citation metrics information from e.g. DataCite, CrossRef or Dimensions. Implementing this would be trivial in the case of usage of the simple REST interface of Dimensions, but as none of the objects in the Carbon Portal have citations yet, this is not an urgent issue. However, ICOS CP is maintaining a BibBase-powered database of publications that related to ICOS, e.g. because they are 1) based in part on data collected at ICOS stations (including pre-ICOS data sets); 2) cover technological or methodological developments made by ICOS (e.g. in collaboration with the Thematic Centres); or 3) mention ICOS in a research-related context. Input to the list, which is available at https://www.icos-cp.eu/references/, comes from a combination of manual submissions from ICOS personnel and researchers, and on-line searches in publication databases.

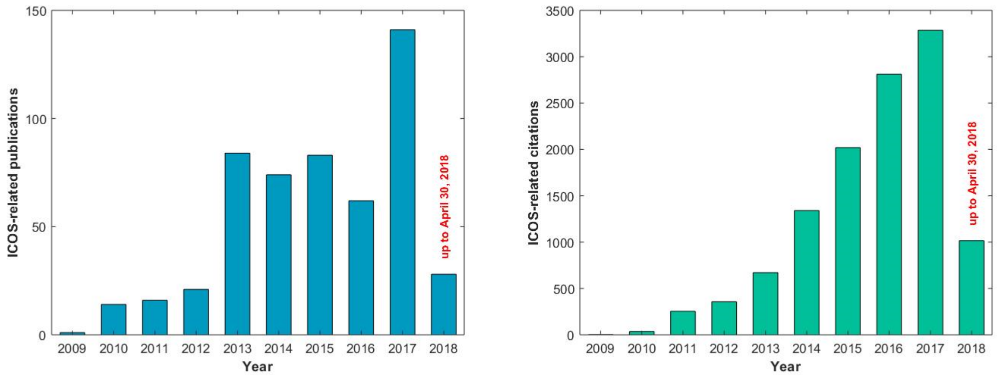


Figure 5-3. *Image showing annual totals of (left) the number of ICOS-related publications and (right) the number of citations of these. (Note the data cut-off of April 30 for 2018!).*

Periodically, statistics are gathered on the number of citations in the scientific literature of the publications in the database. During the most recent such survey, conducted at the end of April 2018, ICOS-related publications published during the interval 2009 to 2018 (cut-off April 30) were looked at. This time period was chosen to reflect the lifespan of the ICOS RI, from the start of the ESFRI-supported preparatory phase up to the beginning of the operational phase. At that time, a total of 524 publications (mainly journal articles) were included, with the total number of 11793 citations – corresponding to an average of 22.5 citations per publication. In addition to looking at the number of citations, other aspects such as the main scientific discipline of the journals, as well as their impact factors, were also analysed [van Belle 2018]. Figure 5-3 illustrates the distribution in time of the number of publications and citations.

5.3.5 Collecting portal services usage statistics

An important contribution to the overall usage and impact of ICOS are the activities at the Carbon Portal web site. These include visits by humans, personally interacting with the portal's discovery services, as well as machine-operated workflows interacting programmatically via the APIs offered by the CP. The discovery services include a search interface, various visualisation tools, and a shopping cart system for downloads. (See also 5.1.2 for more information on CP services.)

Each time a Carbon Portal-related resource locator is activated, all metadata belonging to the data object is recorded into the CP MongoDB database – including the date and time, user agent, IP number, and of course the complete URL including the page address and any attached parameters. The approximate geolocation of the agent (browser or script) associated with the IP address is automatically resolved (via an external service for geolocalisation) and stored in the same record in the dedicated usage database. Currently, no differentiation is made between internal use by ICOS staff and other users.

This database allows the extraction of a number of different statistics measures, which can then be analysed to answer a large range of questions related to e.g. which ICOS data objects (individual ones or product categories) are most often accessed, the temporal and geographical distributions of the requests, and how users interact with the web site itself, including the sequence of page visits, time spent, etc.

The collation and visualisation of the usage statistics is available via https://data.icos-cp.eu/stats/. The page and the underlying analytics functionality are continuously being revised and updated in order to respond to requests from e.g. ICOS organisational components (the Head Office, the Thematic Centres, and the principal investigators of the measurement stations) and stakeholders (funders, national coordinators, and large organisations such as GEOSS and ESFRI).

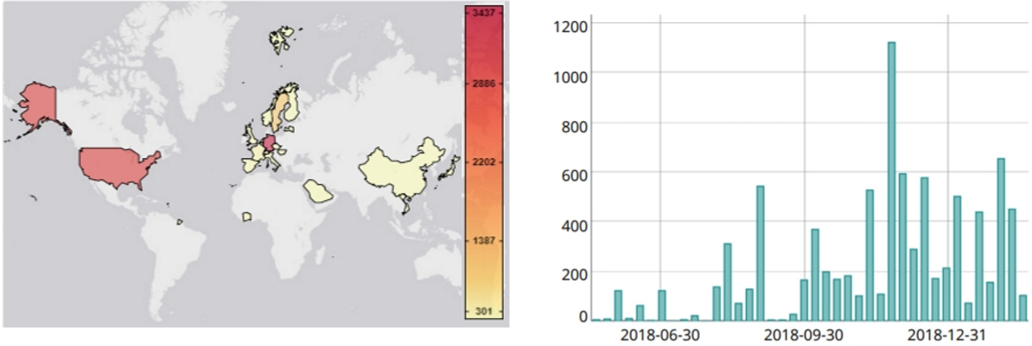


Figure 5-4. *Image showing geographical (left) and temporal (right) distribution of downloads of all Level 2 data sets (CO2, CH4, CO, 14C and meteorological variables) from the Atmospheric Thematic Centre. (Based on screenshots of https://data.icos-cp.eu/stats/ made on February 19, 2019.)*

A recent addition is the capture of information related to the visualisation of data, including which data sets are most often visualised, which measurements or derived variables are most popular to graph, and which types of visualisations (time series, scatter plots or 2d maps) are users selecting. Figure 5-4 shows snapshots of some of the usage metrics available at the time of writing, including the spatial and temporal distribution of all downloads up to date of Level 2 data sets (CO_2, CH_4, CO, ^{14}C and meteorological variables) from the Atmospheric Thematic Centre.

6 FACILITATING QUANTITATIVELY CORRECT DATA USAGE ACCOUNTING

This Chapter, contributed by Markus Fiebig of the ACTRIS RI and the Norwegian Institute for Air Quality Monitoring (NILU), presents the "implementation case 9" which was developed and carried out in collaboration between WP6 and WP9.

6.1 Background

In a world of open and free data sharing, producers and providers of environmental data, i.e. all entities involved in the data production chain, face the challenge of having to find sources of continued funding for their efforts since "selling" their data is not an option. In finding funding, they need to justify to funding agencies and users the relevance of their observations and contributions to data production. One way of showing this relevance and to give merit to data producers and providers is by documenting the use of their data products in a quantitative way.

An existing analogy to such a use-based merit system are scientific journal publications, where authors receive merit based on the number of "uses" of the article, i.e. based on the number of citations. Journals are selected by authors and institutions based on the aggregation of those citation scores in recent years, i.e. how visible the result becomes by using a given publication channel. However, aggregating scores, i.e., citation numbers accumulated across repositories, may be difficult to compose if data is stored at different granularities in the different archives.

Of course, the benefit from public investments into collection of geoscientific data needs to be maximised. This is the reason for funding agencies to push towards re-use of data for multiple purposes, including re-distribution and commercial use. The re-use of data for other than the original purposes is expected to facilitate new services, thus generating economic growth. This vision requires increasingly more open data policies, and data sharing free of charge.

On the other hand, this vision depends critically on being accepted by data providers and producers. Otherwise, these groups could hinder realisation of the vision by delaying their participation. Consequently, a system for giving data producers and providers an incentive by quantitative merit for the use of their data products is urgently needed. The objective here is to facilitate a basic, yet quantitatively correct accounting of data use in an open data world. Provision and support of good tools and automation as well as adoption of endorsed standards are also key factors, orthogonal to measurement of use, which we consider here.

By analogy to scientific articles, persistent identification of data by Digital Object Identifiers (DOIs) would be a crucial element of such a service for quantitative accounting of data use. However, at least 4 challenges exist:

- **DOI granularity:** Data DOIs can be issued with very different granularity. This would make usage numbers based on a fine granularity biased in comparison to data identified with a coarser granularity.
- **Data collections:** DOIs can refer to a user defined collection of other datasets, which themselves may be identified by DOIs. The data collection approach makes data very

convenient to cite. However, the contribution of different data producers to such a collection can vary significantly. In this case, data use accounting without quantitatively resolving the share of contributions of individual data producers to a collection would appear rather unfair. Thus, quantitative accounting of data use needs to resolve the contribution of individual data producers.

- **Accounting mechanism:** Indexing agencies will or have been setting up services for counting of use events involving (DOI) identified data. From here, services need to be implemented that break down data use events into the contribution of single data producers, and with a fixed granularity allowing comparisons between data producers.
- **Nature of data use events:** Scientific data can be used in many different ways, e.g. illustration for outreach purposes, trend analysis, constraining models of environmental processes, event analysis, just to mention a few, and data can be accessed once or multiple times for the same use case event. The above list of data uses is non-exclusive and will evolve over time. A consensus needs to be established which of these use events should count towards data use, and weighting factors need to be agreed upon in case some types of data use are regarded as more valuable than others. A list of data use types counted towards use accounting, including weighting factors, would typically be agreed on and continuously updated by a cross-domain working group consisting of experts on data production, data management, and data indexing.

6.2 Proposed implementations

In order to meet the challenges listed above, and to work towards implementing accounting services for data use, the following tasks were defined in the early project phase:

Data identification with homogeneous granularity in primary archives: For making data use numbers comparable, data need to be identified, here by DOIs, with homogeneous granularity. These DOIs, in this context called primary DOIs, would be used as reference for setting up data use accounting. The ambition in the early project phase was to achieve homogeneous granularity, and thus comparable data use metrics, across repositories and RIs. The community of atmospheric RIs in ENVRIplus, i.e. ACTRIS, IAGOS, and ICOS, were identified as the reference group for this task, and an initial agreement between these RIs was reached on this approach.

Transparent data accounting when using data collections: When identifying data in larger studies, e.g. global climatologies of atmospheric parameters, using primary DOIs requires quoting hundreds of DOIs, which would be rather inconvenient. The DOI specification provides for coining DOIs for user specified data collections, which are ideally suited to identify data used in larger studies. However, when using primary DOIs as reference for accounting data use, collection DOIs need to contain references at least to the primary DOI identifiers of all data contained in the collection to facilitate correct accounting. Ideally, the references would also include further provenance information in order to identify and acknowledge contributors to the data product used. This task involves working with frameworks such as the Research Data Alliance (RDA, http://rd-alliance.org/) on setting up a recommendation for identifying data collections that meets this requirement.

Performing correct accounting of data use: For scientific publications, accounting of use is performed by the indexing agencies. If they offer a similar service for data, it needs to be assured that references to collection identifiers are resolved to the primary identifiers to ensure correct accounting of data use. The task involves a dialogue with the indexing agencies to implement this policy.

Defining the nature and relative value of data use events: A standing working group consisting of experts on data production, data management, and data indexing needs to be established to define and maintain a list of data use events considered in accounting of data use, including their relative weight.

6.3 Results

The results of this implementation case are summarised by task below:

Data identification with homogeneous granularity in primary archive: During the implementation, it turned out that the goal of achieving homogeneous granularity of primary data identifiers across atmospheric RIs was too ambitious. Data products are simply too different in nature between repositories, sometimes even within a single RI. However, primary data identifier granularity should be homogeneous at least within one repository, preferably comparable also among repositories of a single RI.

For example, a granularity of one year's worth of data per instrument can be used as default within an RI dealing with continuous time series, ensuring that data use metrics are comparable at least within one repository or across one RI. For domain specific data repositories, this level of comparability in data use metrics will be sufficient for many or most use cases. This approach has been or will be implemented in a number of ENVRIplus RIs.

Transparent data accounting when using data collections: For this task, ENVRIplus was represented by ICOS in the relevant RDA working group on research data collections[11]. This work resulted in a fully finished recommendation for issuing and handling persistent identifiers for data collections that meet the requirement of referencing back to the primary identifiers of the data contained in the collection [Weigel 2017].

The recommendation covers, among others, these aspects:

- Requirement for PID for the collection
- Requirement for PID for objects in the collection.
- Information beyond object's lifetime
- No assumption on collection lifetime
- Collections may not contain sub-collections. These must be resolved.
- Objects may be part of several collections.
- Objects may be stored in different places.
- Provenance information for collection

Moreover, it was discovered that the metadata schema used by DataCite [DataCite 2017] already contains a provision for including references to primary identifiers in the catalogue entries for collections. This means that this task could essentially be accomplished within the project duration. Further work on this task would involve setting up a reference service for coining DOIs for data collections that meets the above mentioned RDA recommendation.

Performing correct accounting of data use: A dialogue with DataCite as an indexing agency collecting use events involving data DOIs revealed that indexing agencies show little willingness to resolve references to primary identifiers contained in collection DOIs when accounting for data use. From the indexing agencies perspective, this approach makes sense due to the issue of heterogeneous granularity of primary data identifiers across or even within domains. As a result, the task needed to be modified. The service of calculating metrics for data use is moved from the

[11] https://www.rd-alliance.org/groups/research-data-collections-wg.html

indexing agency to the primary data repository. Based on its own primary DOIs with homogeneous granularity, the primary repository can access data use events stored at the indexing agency, resolve references in collection DOIs, and thus calculate data use metrics comparable across the repository. A prerequisite for this approach would be machine-to-machine access to the indexing agencies data holdings by the primary data archives. A dialogue about this is ongoing and needs to be continued beyond the duration of ENVRIplus.

7 OUTCOMES AND CONCLUSIONS

This report is the final deliverable from Work Package 6 of ENVRIplus. Its main scope is to summarise the outcomes of two rather different use cases described in the Description of Action. Many differences can be seen between these two cases as one mainly represents the specific domain of marine biodiversity science (Chapter 4), while the other relates to a broad collaboration between environmental researchers in the atmosphere, ecosystem and ocean domains (Chapter 5). The complexity and overall scope of the issues to be investigated also appear to be quite different at first. But after looking into the results and outcomes, it becomes evident that despite the differences, there are many commonalities – and several of these are connected to the more general issue covered by the third case (outlined in Chapter 6).

The cases illustrate the tight connections between Identification & Citation and the other "pillars" of research data management practices, such as Cataloguing, Metadata management, and Provenance. They also point to the importance of designing and implementing efficient means of harvesting accurate and sufficiently detailed information on various types of usage, and then feeding these into analysis tools that can produce suitable impact metrics for a range of different purposes: total numbers of downloads and visualisations of data at the level of a research infrastructure, the demand for near real-time data from a given measurement site or campaign, or the average number of citations in scientific literature associated with an individual researcher. All cases described here highlight the need for RIs to engage with both their respective data producers (especially in the case of distributed or extended collaborations contributing their data to a central database) and data end users, and offer support for and also training in best practices for citing and identifying data.

In fact, all the activities of WP6 have pointed to the existence of more commonalities than differences between the ENVRIplus partners, regardless of their domain, size or level of maturity. While the requirement analysis performed in collaboration with WP5 did indicate that at the beginning of the project most of ENVRIplus RIs had not yet implemented detailed strategies for including Identification & Citation-related aspects in their data management or technological implementation, this situation is already showing improvements. This is partly due to both the advances made by the beneficiaries taking part in WP6, and the increased engagement from all partners in WP6-related discussions at the ENVRI Week meetings.

The interest in implementing Identification & Citation best practices, such as those listed in D6.1 (and also included in Appendix B) has received an addtional boost from the recent developments in the urge of fulfilling the FAIR principle for services and data intended for use in the European Open Science Cloud. The WP6 use cases reported here show how ENVRI partners can relatively easily incorporate many of these best practices in their data management, including assigning a globally unique PID that can be expressed as an unambiguous URL to all datasets intended to be citable (best practice A), ensuring that these PID URLs resolve to a landing page that is readable by both humans and machines (B & C), and RIs actively promoting data citation to users and others by provision of documentation and common citation formats (J).

In connection with their interest in assigning persistent identifiers to their data, some ENVRIplus partners have also discovered limitations or a lack of desired functionalities in many of the available Identification & Citation-related services. Issues include corruption, or even loss, of original author and creator information during harvesting of metadata from repositories to search portals, and difficulties to include individuals with non-author roles into automatically created citation strings. Many of the experienced frustrations came to light during the WP6-initated dialogue ("negotiation") with publishers, PID service providers and indexers covered in D6.2.

Finally, during the project it became possible to update and augment the way that Identification & Citation is represented in the ENVRI Reference Model [ENVRI RM V2.1 2016]. Although some of the model's "viewpoints" remain to be finalised, the concepts of PID server and PID manager are now fully integrated in the science, information and computational viewpoints, and in the accompanying OIL-E ontology. The theoretical basis offered by the Reference Model should facilitate the understanding of how Identification & Citation fits into the conceptual architectures of research infrastructures and similar collaborative entities, and it may also help in designing a global system for data citation.

8 IMPACT ON PROJECT

As outlined in the previous Chapter, the results and outcomes of the three use cases described in this report have significant importance for the ENVRIplus partners and for the ENVRI Community as a whole, regardless of their level of maturity. Here we present case-by-case summaries of the impacts at project level:

Case 1: Publication of marine biodiversity data. This use case highlights many of the issues and concerns that surfaced during the dialogue with providers of identification & citation services that was reported on in deliverable D6.2, including the importance of having well-defined and effective communication channels between 1) data producers and publishers; 2) publishers and PID service providers; and 3) data producers and their end user communities. The successful dialogue that was set up between the different marine sub-domain actors and publishers, and which led to the implementation of augmented publishing workflows, sets a good example for the other ENVRIplus sub-domains. The problems that were encountered, e.g. with poor engagement from individual researchers in using the new services, are likely to be common across the entire ENVRI landscape, indicating that all RIs will have to put efforts into support (including tool and workflow developments) as well as training of their respective end user communities. In addition, it will be important to periodically survey the needs and practices of data users, not forgetting that researchers from new, "non-traditional" disciplines may be starting to use data from ENVRI Community members!

Case 2: citation and usage tracking system for greenhouse gas monitoring data. ICOS design and implementation of a complete PID-centric cataloguing and curation framework based on Open Source software has proven very successful. It may serve as a model for other ENVRI research infrastructures that are in the process of setting up their own data repositories and portals, as the model allows for federated data sharing because of the underlying Linked Open Data concept, seamless streaming of data objects to a trusted repository, data licence check at data access and usage tracking. The ability to serve dynamic landing pages for all ICOS data objects, containing links both to the data itself and to discovery services such as visualisations, as well as a rich selection of metadata, empowers end users to make informed decisions about both content and quality of ICOS data products. At the same time, the fact that ICOS is still in the process of selecting which data citation tracking services to use, serves as a reminder to other ENVRIplus partners that the landscape of service providers is quite complex, and the smorgasbord of functionalities and tools is under constant development.

Case 3: quantitatively correct data usage accounting. This use case has provided the ENVRIplus community with a workable and viable approach towards quantitative accounting of the use of data produced by the ENVRIplus RIs. The specification of the tasks involved in the approach has been completed during the project. The implementation of the tasks is at the pilot application level and also involves negotiations with external partners – work that will be continued in the framework of the ENVRI-FAIR project.

9 IMPACT ON STAKEHOLDERS

The impacts of the use cases on stakeholders are closely tied to the issue of how to assess and quantify the impacts that research infrastructures have. However, relying to a high degree on data citation or access statistics when evaluating the usage or significance of any research collaboration, let alone a research infrastructure, is fraught with problems. Firstly, a data set may be the result of highly specialised expert work, have undergone strict quality control and assurance processes, and score very high on evaluations of FAIRness, but still only be accessed a couple of times by end users after it has been made available. And a few or only one user can still make the difference. Secondly, a data object could be accessed and downloaded many times, but be mentioned very sparingly in the literature – perhaps due to its end users not knowing how to cite data properly, rather than not being fit for its intended use, or because users, like the general public, just consume and do not publish results at all.

With these potentially limiting factors in mind, some of the most important impacts on stakeholders are listed here, ordered by use case:

Case 1: Publication of marine biodiversity data. This use case has highlighted the significant cultural issues regarding the acceptance and adherence to the FAIR data principles. It is clear that despite continued and sustained engagement with the publishers, unless there are either tangible benefits or mandated processes, authors are unwilling to undertake steps they perceive as unnecessary alongside the traditional process of manuscript submission. Further dialogue with authors in the marine biodiversity discipline is recommended to further investigate the cultural barriers to provision of data. Mandates from funders and publishers are beginning to show some improvements in data availability. However, without clear engagement and positive outcomes within the community there is significant risk of data continuing to be not linked to the resulting publication.

Case 2: citation and usage tracking system for greenhouse gas monitoring data. ICOS work on extracting, curating and disseminating statistics on the usage of its data products, including searches, visualisations and downloads, as well as mentions in scientific literature, provides input information to several of the key performance and impact indicators that are relevant to ICOS stakeholders. Examples of ICOS indicators with direct relevance to WP6 and the scope of this deliverable are “Number of ICOS-related articles published”, “Popularity of ICOS data”, “ICOS-related articles used outside the scientific domain” and “Application of ICOS data in globally leading models". Other ENVRI Community members are likely to benefit from using these, or similar, measures of impact and may therefore benefit from applying the ICOS approach.

Case 3: quantitatively correct data usage accounting: This use case provides RIs with the means of giving data producers an incentive for making their data available following the FAIR principles of open data sharing. Data providers will be rewarded by correct usage statistics for the data they provide, which will motivate them to make their research data openly available, thus working towards implementing the EOSC vision.

ACKNOWLEDGEMENTS

The authors would like to thank all the people who participated in the ENVRIplus Work Package 6 for all their valuable contributions and inputs, with especial thanks to our reviewers. MH also acknowledges stimulating discussions with Peter Wittenburg, Carlo Maria Zwölf and the other members of the Digital Objects, Persistent Identifiers and Citation Practices focus groups of GEDE (Group of European Data Experts in RDA Europe).

REFERENCES

[ANDS 2017] ANDS: Establishing a harvest from your data source to the Data Citation Index. https://www.ands.org.au/online-services/research-data-australia/data-citation-index/establishing-a-data-citation-index-harvest, accessed on February 4, 2019.

[Atkinson 2016] M. Atkinson, A. Hardisty, R. Filgueira, C. Alexandru, A. Vermeulen, K. Jeffery, T. Loubrieu, L. Candela, B. Magagna, P. Martin, Y. Chen and M. Hellström: A consistent characterisation of existing and planned RIs. ENVRIplus Deliverable 5.1, submitted on April 30, 2016. Available at http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf.

[Bode 2018] C. Bode, C. Herzog, D. Hook and R. McGrath: A guide to the Dimensions Data approach. Dimensions Report, January 2018. https://doi.org/10.6084/m9.figshare.5783094.

[Bollen 2008] J. Bollen, H. Van de Sompel and M.A. Rodriguez: Towards Usage-based Impact Metrics: - First Results from the MESUR Project. In Proceedings of the Joint Conference on Digital Libraries 2008 (JCDL08), June 16, 2008, Pittsburgh, Pennsylvania, USA. arXiv:0804.3791 [cs:DL], available via https://arxiv.org/abs/0804.3791.

[C2CAMP 2018a] The C2CAMP consortium: Interlinking Digital Repositories. Digital Objects as Foundational Entities in the Global Data World. Information flyer, May 5, 2018 http://doi.org/10.23728/b2share.e16da3527a314fcfa1f5984d0b00ca31.

[C2CAMP 2018b] The C2CAMP consortium: Wiki of the Cross Continental Collection Access and Management Pilot (C2CAMP), https://github.com/c2camp/core/wiki. Accessed February 18, 2019.

[Chen 2017] Y. Chen, B. Grenier, M. Hellström, A. Vermeulen, M. Stocker, R. Huber, B. Magagna, I. Häggström, M. Fiebig, P. Martin, D. Vitale, G. Judeau, T. Carval, T. Loubrieu, A. Nieva, K. Jeffery, L. Candela and J. Heikkinen: Service deployment in computing and internal e-Infrastructures. ENVRIplus Deliverable 9.1, submitted on August 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D9.1-Service-deployment-in-computing-and-internal-e-Infrastructures.pdf.

[Clarivate 2017] Clarivate Analytics: Recommended practices to promote scholarly data citation and tracking. https://clarivate.com/wp-content/uploads/2017/12/WOS_Whitepaper_DCI_web.pdf, accessed on February 4, 2019.

[Clarivate 2019a] The repository Selection Process – Clarivate Analytics. https://clarivate.com/products/web-of-science/repository-selection-process/?utm_source=false&utm_medium=false&utm_campaign=false, accessed on February 4, 2019.

[Clarivate 2019b] E-mail conversation with Clarivate support team in February 2019.

[Crossref 2018a] Crossref Cited By. https://www.crossref.org/services/cited-by/, accessed on February 4, 2019.

[Crossref 2018b] Crossref Cited By Factsheet. https://www.crossref.org/pdfs/about-cited-by.pdf, accessed on February 4, 2019.

[Crossref 2019] Crossref Event Data. https://www.eventdata.crossref.org/guide/, accessed on February 4, 2019.

[Dasler 2018] R. Dasler and H. Cousijn: Are your data being used? Event Data has the answer! DataCite Blog 8 October 2018. https://blog.datacite.org/are-your-data-being-used-event-data-has-the-answer/, accessed on February 4, 2019.

[DataCite 2017] DataCite Metadata Working Group. DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.1. DataCite e.V., 2017. http://doi.org/10.5438/0015.

[DataCite 2019a] DataCite: Citation Formatter. https://datacite.org/citation.html, accessed on February 4, 2019.

[DataCite 2019b] DataCite Event Data Guide. https://support.datacite.org/docs/eventdata-guide, accessed on February 4, 2019.

[DataONE 2018] DataONE Implements New Usage and Citation Metrics to Make Your Data Count, blog post 24 October 2018. https://www.dataone.org/news/new-usage-metrics, accessed on February 4, 2019.

[DCMI 2012] Dublin Core Metadata Initiative: Dublin Core Metadata Element Set, Version 1.1, June 14, 2012. http://dublincore.org/documents/2012/06/14/dces/, accessed on February 19, 2019.

[De Pooter 2017] D. De Pooter et al. Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. Biodiversity Data Journal 5: e10989. https://doi.org/10.3897/BDJ.5.e10989.

[D'Onofrio 2019] C. D'Onofrio and the ICOS Data Lifecycle Working Group: ICOS Metadata Specification. Work in progress, accessed February 1, 2019.

[DONA 2018] The DONA Foundation Digital Object Interface Protocol Specification version 2.0, November 12, 2018 Available at https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf.

[Dryad 2017] Dryad News & Views. December, 2017. https://blog.datadryad.org/2017/12/18/improvements-in-data-article-linking/, accessed on February 4, 2019.

[EC 2018] European Commission: Turning FAIR into reality. Final report and action plan from the European Commission Expert Group on FAIR Data. November 2018. https://doi.org/10.2777/1524.

[ENVRI RM V2.1 2016] ENVRI Reference Model V2.1, November 9 2016. https://wiki.envri.eu/download/attachments/8553250/EC-091116-1403.pdf. Accessed 2017-01-10. Also available in wiki format at https://wiki.envri.eu/display/EC/ENVRI+Reference+Model.

[ENVRIplus 2015a] ENVRIplus project description, public part. ENVRIplus Grant Agreement, Annex 1, part B. Horizon 2020 project no. 654182. Associated with document Ref. Ares(2015)1488547. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/ENVRIplus_PartB_public.pdf.

[ENVRIplus 2015b] ENVRIplus Description of Work (DoW), public part. ENVRIplus Grant Agreement, Annex 1, part A. Horizon 2020 project no. 654182. Associated with document Ref. Ares(2015)1488547. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/ENVRIplus_DoW_public.pdf.

[Europe PMC 2018] Europe PMC Tech Blog: Integrating Literature and Data. June, 2018. https://europepmc.github.io/techblog/literature_data_integration/2018/06/04/integrating-literature-and-data.html, accessed on February 4, 2019.

[Fenner 2016] M. Fenner, M. Crosas, J.S. Grethe, D. Kennedy, H. Hermjakob, P. Rocca-Serra, R. Berjon, S. Karcher, M. Martone, M and T. Clark: A data citation roadmap for scholarly data repositories. bioRxiv preprint published 2016-12-28. https://doi.org/10.1101/097196.

[FORCE11 2014] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. M. Martone (ed.) San Diego CA: FORCE11, 2014. https://doi.org/10.25490/a97f-egyk.

[FORCE11 2018] FORCE 2018 Montreal Canada, Conference October 11 & 12 2018. Presentations and posters available at: https://zenodo.org/communities/force2018, accessed on February 4, 2019.

[FORCE11 2019] FORCE 11. https://www.force11.org/, accessed on February 4, 2019

[FREYA 2019] FREYA Connected Open Identifiers for Discovery, Access and Use of Research Resources. https://www.project-freya.eu/en, accessed on February 4, 2019

[GBIF 2017] GBIF: Darwin Core Archives – How-to Guide, version 2.0, released on 9 May 2011, (contributed by D. Remsen, K. Braak, M. Döring, T. Robertson), Copenhagen: Global Biodiversity Information Facility. https://github.com/gbif/ipt/wiki/DwCAHowToGuide, accessed on February 11, 2019

[Goldfarb 2018] D. Goldfarb, B. Magagna, S. Kindermann, M. Stocker, D. Lear, K. Paxmann, C.-F. Enell, R. Slapak, P. Martin, S. Koulouzis and C. Pichot: Data provenance and tracing for environmental sciences: Prototype and deployment. ENVRIplus deliverable D8.6, submitted on October 31, 2018. Available via http://www.envriplus.eu/wp-content/uploads/2015/08/D8.6-Data-provenance-and-tracing-for-environmental-sciences-Prototype-and-deployment.pdf.

[Google 2019] Google Dataset Search Beta. https://toolbox.google.com/datasetsearch, accessed on February 4, 2019.

[Hellström 2017] M. Hellström, M. Lassi, A. Vermeulen, R. Huber, M. Stocker, F. Toussaint, M. Atkinson and M. Fiebig: A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations. ENVRIplus Deliverable D6.1, submitted on January 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/ D6.1-A-system-design-for-data-identifier-and-citation-services-for-environmental-RIs.pdf.

[Hellström 2018] M. Hellström, M. Johnsson, F. Toussaint, S. Kindermann, D. Lear, R. Huber, M. Stocker, I. Häggström and C.-F. Enell: A report on negotiations with publishers, providers of existing data citation systems and other scientific organisations on implementing a global data citation system. ENVRIplus Deliverable D6.2, submitted on April 30, 2018. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D6.2-A-report-on-negotiations-with-publishers-providers-of-existing-data-citation-systems-and-other-scientific-organisations-on-implementing-a-global-data-citation-system.pdf.

[ICOS ERIC 2015] ICOS ERIC Head Office: ICOS data policy, approved by the General Assembly December 10, 2015. Available via https://www.icos-ri.eu/fetch/9b7ae35b-3461-47db-95db-9660ac5bd0de.

[Koulouzis 2018] S. Koulouzis, R.Mousa, A. Karakannas, C. de Laat, Z.Zhao: Information Centric Networking for Sharing and Accessing Digital Objects with Persistent Identifiers on Data Infrastructures, in the proceedings of the 3rd International Workshop on Distributed Big Data Management, in the context of IEEE CCGrid 2018, Washington, US. https://doi.org/10.1109/CCGRID.2018.00098.

[Lannom 2017] L. Lannom: C2CAMP. Presentation at the Managing Digital Objects in an Expanding Science Ecosystem workshop, Bethesda, MD, USA on November 15, 2017. Available via https://www.rd-alliance.org/sites/default/files/CENDI-15.Nov .17-Lannom-Final-2.pdf.

[Lavasa 2018] A. Lavasa et al: Integration of Mature PID Types. FREYA Connected Open Identifiers for Discovery, Access and Use of Research Resources. Project FREYA deliverable D4.1. https://doi.org/10.5281/zenodo.2414839.

[Magagna 2018] B. Magagna, D. Goldfarb, P. Martin, F. Toussaint, S. Kindermann, M. Atkinson, K.G. Jeffery, M. Hellström, M. Fiebig, A. Nieva de la Hidalga and A. Spinuso,: Data provenance and tracing for environmental sciences: system design. ENVRIPLUS project deliverable D8.5, submitted on 30 April, 2018. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D8.5-Data-provenance-and-tracing-for-environmental-sciences-system-design.pdf.

[MDC 2018a] COUNTER Code of Practice for Research Data Usage Metrics Release 1, blog post September 2018. https://makedatacount.org/2018/09/13/counter-code-of-practice-for-research-data-usage-metrics-release-1/, accessed on February 4, 2019.

[MDC 2018b] DataONE Implements New Usage and Citation Metrics to Make Your Data Count, blog post October 2018. https://makedatacount.org/2018/10/24/dataone-implements-new-usage-and-citation-metrics-to-make-your-data-count/, accessed on February 4, 2019

[Mori 2018] A. Mori and M. Taylor: Dimensions Metrics API. Reference & Getting Started, January 2018. https://doi.org/10.6084/m9.figshare.5783694.v2.

[Noy 2018] N. Noy: Making it easier to discover datasets. Blog, 5 September 2018. https://www.blog.google/products/search/making-it-easier-discover-datasets/, accessed on February 4, 2019.

[NSB 2005] US National Science Board. Long-lived digital data collections: Enabling research and education in the 21st century. ReportNSB0540 from the National Science Foundation, September 2005. Available at http://www.nsf.gov/pubs/2005/nsb0540/.

[OpenAIRE 2018] OpenAIRE becomes a fully fledged organisation. October 29, 2018. https://www.openaire.eu/openaire-organisation-in-the-making, accessed on February 4, 2019.

[OpenAIRE 2019] OpenAIRE. https://www.openaire.eu/, accessed on February 4, 2019.

[Scholix 2019] Scholix Implementors. http://www.scholix.org/implementors, accessed on February 4, 2019.

[Socha 2013] Y.M. Socha, ed. and the CODATA-ICSTI Task Group on Data Citation Standards and Practices: “Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data”. *Data Science Journal* vol. 12, 13 Sept 2013. https://doi.org/10.2481/dsj.OSOM13-043.

[SpringerNature, 2019] Research Data Policy Types. https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096, accessed on February 11, 2019.

[van Belle 2018] J. van Belle, J. van Barneveld-Biesma, V. Bastiaanssen, A. Buitenhuis, L. Saes and G. van Veen: ICOS Impact Assessment Report. Technopolis group, August 2018. Available at https://www.icos-ri.eu/sites/default/files/2018-10/ICOS_Impact_Assessment_2018.pdf.

[W3C 2013] World Wide Web Consortium: PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013, https://www.w3.org/TR/prov-o/. Accessed on February 12, 2019.

[Wang 2017] J. Wang, A. Taal, P. Martin, Y. Hu, H. Zhou, J. Pang, C. de Laat, Z. Zhao: Planning Virtual Infrastructures for Time Critical Applications with Multiple Deadline Constraints, International journal of Future Generation Computer System, volume 75, page 365-375, 2017. https://doi.org/10.1016/j.future.2017.02.001.

[Weigel 2017] T. Weigel et al: Recommendation on Research Data Collections. Research Data Alliance. https://github.com/RDACollectionsWG/specification/blob/master/Recommendation%20package/rda-collections-recommendation.pdf, accessed on February 11, 2019.

[Wilkinson 2016] M. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018 (2016). http://doi.org/10.1038/sdata.2016.18.

[Wittenburg 2019] P. Wittenburg, G. Strawn, B. Mons, L. Boninho and E. Schultes: Digital objects as drivers towards convergence in data infrastructures. http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11.

[Zhao 2015] Z. Zhao, P. Martin, J. Wang, A. Taal, A. Jones, I. Taylor, V. Stankovski, I. Garcia Vega, G. Suciu., A. Ulisses, C. de Laat: Developing and operating time critical applications in clouds: the state of the art and the SWITCH approach, In the proceedings of HOLACONF - Cloud Forward: From Distributed to Complete Computing, Pisa, Italy, 2015. Elsevier, Procedia Computer Science, Vol 68, 17-28. https://doi.org/10.1016/j.procs.2015.09.220.

APPENDIX A. ACRONYMS AND SPECIAL TERMS

This appendix is based on the official ENVRI terminology and glossary, as available at the ENVRI community wiki site (see https://wiki.envri.eu/pages/viewpage.action?pageId=14452608).

A.1. Terminology & glossary specific to this deliverable

AAI: Authentication, Authorisation and Identification – used in the context of user access to services

ANDS: Australian National Data Service

C2CAMP: Cross-Continental Collection Access and Management Pilot

CrossRef: Non-profit membership organization making research outputs easy to find, cite, link, and assess.

CRUD: Create, Read, Update, Delete

DataCite: Global non-profit organisation that provides persistent identifiers (DOIs) for research data.

Data collection: A number of datasets grouped together as one entity.

DCAT-AP: Application profile of the Data Catalogue (DCAT) developed for European data portals

DCI: Data Citation Index

DLI: Data Literature Interlinking

DLM: Data-Level Metrics, an aggregation and publication service developed by Scholix

DO: Digital Object.

DOI: Digital Object Identifier.

DOIP: Digital Object Interoperability Protocol

Dynamic data: Refers to datasets that may change over time, e.g. because new data has been added, updates or changes of data have been made.

DwC-A: Darwin Core Archive

EML: Ecological Markup Language

EZID: Service from the California Digital Library allowing to create and manage long-term globally unique IDs for data and other sources.

FAIR: Findable, Accessible, Interoperable, Reusable

FORCE11: international community for scholarly communication.

Fragment dataset: A specific subset of a larger dataset.

FREYA: European research project, funded by Horizon 2020. Follow-up of the THOR project.

GDON: Global Digital Object Network

GeoDCAT-AP: an extension of DCAT-AP for describing geospatial datasets, dataset series, and services

GHG: Greenhouse Gas

Handles: Short for the Handle System, a type of persistent identifier.

HDF5: Hierarchical Data Format (HDF) is a set of file formats (HDF4, HDF5) designed to store and organize large amounts of data.

InGOS: Integrated non-CO_2 Greenhouse gas Observing System

IPT: Integrated Publishing Toolkit, set up by GBIF

ISO-19115: A schema required for describing geographic information and services, developed by ISO (International Standards Organisation)

ISTIC: China Institute of Science and Technology Information

JaLC: Japan Link Center

KII: Key Impact Indicator

KPI: Key Performance Indicator

MEDIN: UK Marine Environmental Data and Information

mEDRA: Multilingual European Registration Agency for DOI persistent identifiers for any form of intellectual property on a digital network.

Metadata 2000: Initiative to define common standards for metadata interoperability.

MDC: Make Data Count

NOAD: National Open Access Desk

NDN: Named Data Networking

OBIS: Ocean Biogeographic Information System

ORCID: Non-profit organization providing unique identifiers for researchers.

OWL: Web Ontology Language

PID: Persistent digital identifier.

Pidapalooza: Series of international events dedicated to technologies and services around persistent identifiers.

Prov-O: Provenance Ontology

Query store: Instead of storing many duplicates of subsets of data it is possible to create specific queries in order to identify and obtain certain subsets of data. The queries may be stored in a query store, enabling re-use.

RDF: Resource Description Framework

REST: Representational State Transfer

Scholix: Scholarly link exchange is a high level interoperability framework for exchanging information about the links between scholarly literature and data, as well as between datasets.

SHA-256: Secure Hash Algorithm (256-bit)

SPARQL: A recursive acronym for SPARQL Protocol and RDF Query Language

SQL: Structured Query Language, a domain-specific language used in programming and designed for managing data held in a relational database management system.

THOR: European research project, funded by Horizon 2020. Precursor to FREYA.

URL: Uniform Resource Locator, a location-based uniform resource identifier.

URN: Uniform Resource Name, a type of persistent identifier.

WoRMS: The World Registry of Marine Species.

XML: Extensible Markup Language.

A.2. Other technical terms and acronyms used in ENVRIplus deliverables

API: Application Program Interface, is a set of routines, protocols, and tools for building software applications

Biodiversity: is the variety of different types of life found on earth

Biodiversity metrics: measurements of the number of species and how they are distributed

CERIF: Common European Research Information Format

CIARD RING: A global directory of information services and datasets in agriculture

D4Science: is an organisation offering a Hybrid Data Infrastructure service and a number of Virtual Research Environments

Data stream: is a sequence of digitally encoded coherent signals used to transmit or receive information that is in the process of being transmitted

Data pipeline: In computing, a pipeline is a set of data processing elements connected in series, where the output of one element is the input of the next one.

DCAT: is a resource description format vocabulary designed to facilitate interoperability between data catalogues

DIRAC: Distributed Infrastructure with Remote Agent Control. High-Throughput computing platform operated by EGI.

EduGAIN: is an international inter-federation service interconnecting research and education identity federations

E-infrastructure: can be defined as networked tools, data and resources that support a community of researchers, broadly including all those who participate in and benefit from research

FIM4R: Federated Identity Management for Research collaborations

gCube: is an open-source software toolkit used for building and operating Hybrid Data Infrastructures enabling the dynamic deployment of Virtual Research Environments by favouring the realisation of reuse oriented policies

HPC: High Performance Computing

HTC: High Throughput Computing

IoT: The Internet of Things - is a scenario in which objects, animals or people are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.

ICT: Information and Communications technology

IG: Interest Group, open-ended topic group, for example in the Research Data Alliance

IPR: Intellectual Property Rights

KOS: Knowledge Organization Systems - is a generic term used in Knowledge organization about authority lists, classification systems, thesauri, topic maps, ontologies etc.

LOD: Linked open data is linked data that is open content

LOV: Linked Open Vocabularies

Metadata: is data that describes other data. Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier

NGI: National Grid Initiative

NMI: National Metrological Institutes

NREN: National Research and Education Network

NRT: Near Real Time - refers to the time delay introduced, by automated data processing or network transmission, between the occurrence of an event and the use of the processed data (For example, a near-real-time display depicts an event or situation as it existed at the current time minus the processing time, as nearly the time of the live event)

ODP: 1) Open Distributed Processing (for the ENVRI Reference Model); 2) Online Data Processing

OIL-E: The Open Information Linking model for Environmental science - is a semantic linking framework

Ontology: (In computer science and information science) an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse

QoE: Quality of user experience

Over dispersion: a statistical characteristic of data such that the data have more clusters than compared to what might be expected if the data were distributed randomly in proportion to the time/space available.

NetCDF: a file format.

OceanSITES: a worldwide system of long-term, open-ocean reference stations measuring dozens of variables and monitoring the full depth of the ocean from air-sea interactions down to the seafloor

OOI: Ocean Observatories Initiative

RDA: Resource Description and Access, a standard for descriptive cataloguing. See also A.3. (Organisational acronyms) below.

RM: Reference Model - is an abstract framework or domain-specific ontology consisting of an interlinked set of clearly defined concepts produced by an expert or body of experts in order to encourage clear communication

SensorML: The primary focus of the Sensor Model Language is to provide a robust and semantically-tied means of defining processes and processing components associated with the measurement and post-measurement transformation of observations

Semantics: is the study of meaning

Syntax: In computer science, the syntax of a computer language is the set of rules that defines the combinations of symbols that are considered to be a correctly structured document or fragment in that language

SLA: Service Level Agreement

SME: Small and medium-sized enterprise

UV: Unmanned vehicles

VL: Virtual Laboratory

VRE: Virtual Research Environments, web based package tailored to a specific community

WG: Working Group, time-limited topic group, for example in the Research Data Alliance

A.3. Organisational acronyms

ACTRIS: Aerosols, Clouds, and Trace gases Research Infrastructure network. ENVRIplus partner.

AnaEE: Analysis and Experimentation on Ecosystems. European research infrastructure, ENVRIplus partner.

AQUACOSM: EU network of mesocosms facilities for research on marine and freshwater ecosystems open for global collaboration

CDI: Collaborative Data Infrastructure. European e-service provider organisation,

CEA: Commissariat à l'Energie Atomique et aux Energies Alternatives. French research agency, ENVRIplus participant.

CINECA: Consorzio Interuniversitario. Italian non-profit research consortium, ENVRIplus participant.

CNR: Consiglio Nazionale delle Richerche. Italian national research council, ENVRIplus participant.

CNRS: Centre National de la Recherche Scientifique. French research organisation, ENVRIplus participant.

CODATA: Committee on Data for Science and Technology.

ConnectinGEO: Coordinating an Observation Network of Networks EnCompassing saTellite and IN-situ to fill the Gaps in European Observations

COOPEUS: Strengthening the cooperation between the US and the EU in the field of environmental research infrastructures. Project funded under EU FP7, continued as COOP+ under Horizon 2020.

COPERNICUS: previously known as GMES (Global Monitoring for Environment and Security), is the European Programme for the establishment of a European capacity for Earth Observation

CREEM: Centre for Research into Ecological and Environmental Modelling, operated by University of St Andrews (USTAN).

CSC: Center for Science (Tieteen tietotekniikan keskus Oy). Finnish national high-performance computing centre, ENVRIplus participant.

CU: Cardiff University. UK university, ENVRIplus participant.

DANUBIUS: The international center for Adavanced studies on river-sea systems

DASSH: Data archive for seabed species (a UK marine biology resource centre)

DiSSCo: Distributed Systems of Scientific Collections

DKRZ: German Climate Computation Center (Deutsches Klimarechenzentrum GmBH). German research organisation, ENVRIplus participant.

EAA: Umweltbundesamt GmbH - Environment Agency Austria. Austrian governmental agency, ENVRIplus participant.

EEA: European Environment Agency

EGI: Stichting European Grid Initiative. European research foundation, ENVRIplus participant.

EISCAT: EISCAT Scientific Association. European research organisation, ENVRIplus participant.

EISCAT_3D: Multi-static phased array radar system. Operated by EISCAT Scientific Association, ENVRIplus partner.

EMBL: European Molecular Biology Laboratory. European research organisation, ENVRIplus participant.

EMBRC: European Marine Biological Resource Centre. A research infrastructure and consortium of research organisations interested in marine biology. ENVRIplus partner.

EMODNET: The European Marine Observation and Data Network

EMRP: European Metrology Research Programme

EMSC: European-Mediterranean Seismological Centre. European non-governmental organisation, ENVRIplus participant.

EMSO: European Multidisciplinary Seafloor and Water Column Observatory. European research infrastructure, ENVRIplus partner.

EOSC: European Open Science Cloud. Initiative from the European Commission.

EPOS: The European Plate Observing System. European research infrastructure, ENVRIplus partner.

ERIS: Environmental Research Infrastructure Strategy 2030

ESONET VI: European Seafloor Observatory NETwork. European research infrastructure, ENVRIplus partner.

ETHZ: Eidgenössische Technische Hochschule Zürich. Swiss technical university, ENVRIplus participant.

EUDAT: H2020 project on Research Data Services, Expertise & Technology Solutions (previously funded by FP7). Continues as the Collaborative Data Infrastructure (CDI).

EUFAR: European Facility for Airborne Research

EURO-ARGO: European research infrastructure, ENVRIplus partner.

EUROCHAMP2020: European atmospheric simulation chambers

EUROFLEETS: New operational steps towards an alliance of European research fleets. ENVRIplus partner.

EUROGOOS: European Global Ocean Survey System. International non-profit association, ENVRIplus participant.

EuroSITES: European Ocean Observatory Network

FixO3: Fix point open ocean observatories (survey programme). European research infrastructure, ENVRIplus partner.

FMI: Finnish Meteorological Institute (Ilmatieteen Laitos). Finnish research and service agency, ENVRIplus participant.

FZJ: Research Centre Jülich (Forschungszentrum Jülich GmbH). German research centre, ENVRIplus participant.

GBIF: Global Biodiversity Information Facility

GEO: The Group on Earth Observations.

GEOMAR: Helmholtz Zentrum für Ozeanforschung Kiel. German research institution, ENVRIplus participant.

GEOSS: Global Earth Observation System of Systems, coordinated by GEO (The Group on Earth Observations)

GMES: Global Monitoring for Environment and Security, previous name for COPERNICUS.

GROOM: Gliders for research ocean observation and management

HELIX Nebula: partnership between big science and big business in Europe that is charting the course towards the sustainable provision of cloud computing - the Science Cloud

IAGOS: In-service Aircraft for a Global Observing System. European research infrastructure, ENVRIplus partner.

ICOS: Integrated Carbon Observation System. European research infrastructure, ENVRIplus partner.

ICSU: The International Council for Science

IFREMER: Institute Français de Recherche Pour l'Exploitation de la Mer. French research organisation, ENVRIplus participant.

INGV: Istituto Nazionale di Geofisica e Vulcanologia. Italian research institute, ENVRIplus participant.

INRA: Institut National de la Recherche Agronomique. French research institute, ENVRIplus participant.

INSPIRE: Integrated Sustainable Pan-European Infrastructure for Researchers in Europe

INTERACT: International Network for Terrestrial Research and Monitoring in the Arctic. European research infrastructure, ENVRIplus partner.

IPBES: Intergovernmental Platform on Biodiversity & Ecosystem Services

IS-ENES: Infrastructure for the European Network for Earth System Modelling. European research infrastructure, ENVRIplus partner.

JERICO: Towards a joint European research infrastructure network for coastal observatories. European research project, ENVRIplus partner.

LifeWatch: European e-Science infrastructure for biodiversity and ecosystem research. ENVRIplus partner.

LTER: The Long-term Ecological Research Network. International research organisation.

LTER-Europe: European Long-term Ecosystem Research network of 21 national LTER networks. ENVRIplus partner.

LU: Lund University (Lunds universitet). Swedish university, ENVRIplus participant.

MARUM: Centre for Marine Environmental Sciences at University of Bremen (UniHB).

MBA: Marine Biological Association of the United Kingdom. UK research organisation, ENVRIplus participant.

NERC: Natural Environment Research Council. UK research council, ENVRIplus participant.

NILU: Norwegian Institute of Air Research (Norsk Institutt for Luftforskning). Norwegian research institute, ENVRIplus participant.

OASIS: Advancing Open Standards for the Information Society (non-profit consortium)

PANGAEA: Information system and data publisher for geoscientific and environmental data, operated by MARUM and UniHB. German data repository, ENVRIplus participant.

PLOCAN: Oceanic Platform of the Canary Islands (Consorcio Para el Diseno, Construccion, Equipamiento y Explotacion de la Plataforma Oceanica de Canarias). Spanish research organisation, ENVRIplus participant.

RCN: Research Council of Norway (Norges Forskningsrad). Norwegian national research council, ENVRIplus participant.

RDA: Research Data Alliance. International organisation working to promote collaboration on the management of research data. See also A.2 (Other technical terms and acronyms) above.

SCAPE: SCAlable Preservation Environments. European research project, financed under FP7.

SeaDataNet: Pan-European infrastructure for ocean & marine data management. European research infrastructure, ENVRIplus partner.

SIOS: Svalbard Integrated Arctic Earth Observing System. European research infrastructure, ENVRIplus partner.

UCPH: University of Copenhagen (Københavns Universitet). Danish university, ENVRIplus participant.

UEDIN: University of Edinburgh. UK university, ENVRIplus participant.

UGOT: University of Gothenburg (Göteborgs Universitet). Swedish university, ENVRIplus participant.

UHEL: University of Helsinki (Helsingin Yliopisto). Finnish university, ENVRIplus participant.

UiT: University of Tromso (Universitetet i Tromsø). Norwegian university, ENVRIplus participant.

UniHB: University of Bremen (Universität Bremen). German university, ENVRIplus participant.

UNILE: University of Salento (Universitá del Salento). Italian university, ENVRIplus participant.

UNITUS: University of Tuscia (Universitá Degli Studi della Tuscia). Italian university, ENVRIplus participant.

USTAN: The University Court of the University of St. Andrews. UK university, ENVRIplus participant.

UvA: University of Amsterdam (Universiteit van Amsterdam). Dutch university, ENVRIplus participant.

A.4. ENVRIplus project-related acronyms & terms

AC: Active Collab (ENVRIplus Project Management System)

BEERi: Board of European Environmental Research Infrastructures - is an internal advisory board representing the needs of environmental Research Infrastructures

CA: Consortium Agreement - Legal contract between the ENVRIplus beneficiaries

DL: Deliverable / Deadline

DoA: Description of Action

DoW: Description of Work

EB: Executive Board - supervisory body for the execution of the Project

EC: European Commission - is the executive body of the European Union responsible for proposing legislation, implementing decisions, upholding the EU treaties and managing the day-to-day business of the EU

EINFRA-1-2014: H2020 Call for e-infrastructures (Managing, preserving and computing with big research data), funding source for ENVRIplus

ENV SWG: the Strategic Working Group on Environment of ESFRI

ENVRI: FP7 project on Implementation of common solutions for a cluster of ESFRI infrastructures in the field of environmental Sciences. Precursor of ENVRIplus.

ENVRIplus: Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe.

ESFRI: the European Strategy Forum on Research Infrastructures

GA: 1) Grant Agreement - Contract between Coordinator and Commission; 2) General Assembly - GA is the ultimate decision-making body of the consortium

H2020: Horizon 2020, European level research funding scheme

I3: Integrated Infrastructures Initiative (I3) combines several activities essential to reinforce research infrastructures and to provide an integrated service at the European level

INFRADEV-4: Sub-call topic of the H2020 INFRADEV call for Implementation and operation of cross-cutting services and solutions for clusters of ESFRI and other relevant research infrastructure initiatives

PM: Person Month

RI: Research Infrastructure. RIs are facilities, resources and related services used by the scientific community to conduct top-level research in their respective fields, ranging from social sciences to astronomy, genomics to nanotechnologies.

VCP: Virtual Community Platform

WP: Work Package

APPENDIX B. CITATION BEST PRACTICES FOR THE ENVRI COMMUNITY

Note: This appendix corresponds to Chapter 8.3.2 of ENVRIplus deliverable D6.1 [Hellström 2017]. The text is included here for the convenience of the readers.

The ENVRIplus partners should strive to follow the following recommendations for data citation, based on the review of data citation best practices and recommendations from relevant organisations including [Fenner 2016], [FORCE11 2014a], [Socha 2013]:

Technical aspects:

A. All datasets intended for citation have a globally unique PID that can be expressed as an unambiguous URL
B. A PID expressed as a URL resolves to a landing page for a dataset
C. The landing page of a dataset is both human-readable and machine-readable (and preferably machine-actionable) and contains the dataset's PID
D. PIDs for datasets support multiple levels of granularity (including fine-grained subsets as well as collections)
E. Datasets are described with rich metadata (to track provenance information and to create meaningful citations and (including the identifier of the dataset))
F. Metadata are accessible even if a dataset is no longer accessible
G. RIs provide a robust resolver and registry for resolving PIDs and for data discovery
H. Metadata protocols and standards are used, that ensure interoperability with related stakeholders, e.g. cataloguing and indexing services
I. Data are published with a clearly defined data usage license

Citation practices:

J. RIs actively promote data citation (to users, publishers and other stakeholders in their research community (e.g. by providing documentation and how-tos) and by providing common citation formats to users)
K. Citation methods are flexible to support each community while still ensuring interoperability across communities

The citation best practices for RIs are outlined below.

Technical aspects

A. All datasets intended for citation have a globally unique PID that can be expressed as an unambiguous URL

Based on the current and evolving practices and technological requirements of each RI, the choice of PID systems may differ across ENVRIplus partners. It is important to choose a PID system that facilitates interoperability.

B. A PID expressed as a URL resolves to a landing page for a dataset

When resolved through the respective handle system, PIDs will resolve into a URI that points to a landing page that either produces a human-readable or machine-readable summary of the relevant metadata of the data object and a link to the data object itself.

C. The landing page of a dataset is both human-readable and machine-readable

As stated, a dataset that is intended for citation is given a PID, and a corresponding landing page providing information about the dataset. The landing pages should be human-readable and machine-readable, and preferably machine-actionable.

D. PIDs for datasets support multiple levels of granularity (including fine-grained subsets as well as collections)

RIs should support data citation on multiple levels of granularity that suit the characteristics of their data and the needs of the community. Examples of levels of granularity include data points, subsets, datasets, and collection of datasets. Different PID systems may be used for different levels of granularity, providing they are interoperable.

E. Datasets are described with rich metadata on the landing pages

Metadata should include provenance information as well as other curation metadata about the dataset following at a minimum the metadata scheme of the PID system. Provenance information includes a list of all persons who have contributed to the dataset. In cases where there is a long list of contributors, these can be listed on the landing page of each data set instead of in the citation. Links to documentation that provides more information about the dataset are encouraged (see curation deliverable). Note that documentation and other research objects also can be given PIDs and associated landing pages.

F. Metadata are accessible even if a dataset is no longer accessible

Datasets might be deleted or replaced with a later version, in which case a new PID should be minted for the new version of the dataset. Thus, the previous versions of the dataset keep their PIDs and landing pages, and the new version is seen as a completely new dataset. Linking between the landing pages of the previous versions and the current version is encouraged to provide provenance information and facilitate tracking of the dataset usage through its versions. Note that the PID should still be resolvable to a landing page of the type tomb stone when a dataset is no longer accessible – this provides important information even though the dataset is no longer available.

G. RIs provide a robust resolver and registry for resolving PIDs and for data discovery

Citations and PIDs should be as persistent as the objects they cite, or actually even more persistent (see point F above regarding tomb stones for no longer accessible datasets). RIs should provide a robust resolver and registry, which could be handled through mirrored servers or a distributed storage solution.

H. Metadata protocols and standards are used, that ensure interoperability with related stakeholders, e.g. cataloguing and indexing services

RIs should choose metadata standards and protocols that ensure interoperability across RIs and services provided by other stakeholders, such as cataloguing and indexing services. A landing page should be able to generate the metadata in all relevant controlled vocabularies (see Cataloguing deliverable).

I. Data are published with a clearly defined data usage license

Data usage license information should be readily available to provide prospective users, preferably in both human-readable and machine-readable form.

Citation practices

J. RIs actively promote data citation

ENVRIplus partners actively promote data citation practices to potential data users and other stakeholders in their research community. This could include providing documentation and how-to guides to data citation, and by providing common formats to users. RIs can facilitate and promote proper citation by including pre-created citation text snippets on the landing pages of all data sets they are curating. Citations can be automatically generated in a desired format using e.g. the DataCite DOI Citation Formatter API, which supports over 5,000 citation styles [DataCite 2019a].

K. Citation practices are flexible to support each community, while still ensuring interoperability across communities

This entails promoting and developing citation practices as well as choosing and implementing technical solutions that suit the specific RIs, while taking into account the ENVRIplus community as a whole. RIs should interact with their user base — in OAIS terminology, called the “designated community” — to investigate its data citation practices. In many cases, scientific end users are following old, pre-PID habits by routinely referring to data sets in the running texts of articles (e.g. “we used the ABC dataset provided by Andersson et al.” in a paper’s “Materials and Methods” section). This behaviour should be actively discouraged, instead encouraging proper data citation practices by highlighting good examples of proper data citation using PIDs (DOIs) in the context of conference presentations or workshops., and by following the data citation best practices recommended here.