Ecological Informatics 61 (2021) 101245

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Integrating data and analysis technologies within leading environmental research infrastructures: Challenges and approaches

Robert Huber [a,*], Claudio D'Onofrio [b], Anusuriya Devaraju [c], Jens Klump [d], Henry W. Loescher [e], Stephan Kindermann [f], Siddeswara Guru [c], Mark Grant [c], Beryl Morris [c], Lesley Wyborn [g], Ben Evans [g], Doron Goldfarb [h], Melissa A. Genazzio [e], Xiaoli Ren [i], Barbara Magagna [h], Hannes Thiemann [f], Markus Stocker [j]

[a] MARUM – Center for Marine Environmental Sciences, University of Bremen, Leobener Str. 8, POB 330440, 28359 Bremen, Germany
[b] Dept. Phys. Geography & Ecosystem Science, ICOS Carbon Portal, Lund University, Sölvegatan12, SE-223 62 Lund, Sweden
[c] TERN Australia, University of Queensland, Brisbane, Australia
[d] SIRO, 26 Dick Perry Avenue, Kensington, WA, Australia
[e] Battelle, National Ecological Observatory Network (NEON), Boulder, CO, USA
[f] DKRZ – Deutsches Klimarechenzentrum GmbH, Hamburg, Germany
[g] National Computational Infrastructure (NCI), Australian National University, Canberra, Australia
[h] Environment Agency Austria, Spittelauer Lände 5, 1090 Vienna, Austria
[i] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[j] TIB—Leibniz Information Centre for Science and Technology, Hannover, Germany

ARTICLE INFO

Keywords:
Scientific data analysis
Research infrastructures
Data service providers
Data analysis environments

ABSTRACT

When researchers analyze data, it typically requires significant effort in data preparation to make the data analysis ready. This often involves cleaning, pre-processing, harmonizing, or integrating data from one or multiple sources and placing them into a computational environment in a form suitable for analysis. Research infrastructures and their data repositories host data and make them available to researchers, but rarely offer a computational environment for data analysis. Published data are often persistently identified, but such identifiers resolve onto landing pages that must be (manually) navigated to identify how data are accessed. This navigation is typically challenging or impossible for machines.

This paper surveys existing approaches for improving environmental data access to facilitate more rapid data analyses in computational environments, and thus contribute to a more seamless integration of data and analysis. By analysing current state-of-the-art approaches and solutions being implemented by world-leading environmental research infrastructures, we highlight the existing practices to interface data repositories with computational environments and the challenges moving forward.

We found that while the level of standardization has improved during recent years, it still is challenging for machines to discover and access data based on persistent identifiers. This is problematic in regard to the emerging requirements for FAIR (Findable, Accessible, Interoperable, and Reusable) data, in general, and problematic for seamless integration of data and analysis, in particular. There are a number of promising approaches that would improve the state-of-the-art. A key approach presented here involves software libraries that streamline reading data and metadata into computational environments. We describe this approach in detail for two research infrastructures. We argue that the development and maintenance of specialized libraries for each RI and a range of programming languages used in data analysis does not scale well.

* Corresponding author.
E-mail addresses: rhuber@uni-bremen.de (R. Huber), claudio.donofrio@nateko.lu.se (C. D'Onofrio), adevaraju@marum.de (A. Devaraju), Jens.Klump@csiro.au (J. Klump), hloescher@battelleecology.org (H.W. Loescher), kindermann@dkrz.de (S. Kindermann), s.guru@uq.edu.au (S. Guru), m.grant3@uq.edu.au (M. Grant), beryl.morris@uq.edu.au (B. Morris), lesley.wyborn@anu.edu.au (L. Wyborn), ben.evans@anu.edu.au (B. Evans), doron.goldfarb@umweltbundesamt.at (D. Goldfarb), mgenazzio@battelleecology.org (M.A. Genazzio), renxl@igsnrr.ac.cn (X. Ren), barbara.magagna@umweltbundesamt.at (B. Magagna), thiemann@dkrz.de (H. Thiemann), markus.stocker@tib.eu (M. Stocker).

https://doi.org/10.1016/j.ecoinf.2021.101245
Received 6 November 2020; Received in revised form 21 January 2021; Accepted 27 January 2021
Available online 6 February 2021

Based on this observation, we propose a set of established standards and web practices that, if implemented by environmental research infrastructures, will enable the development of RI and programming language independent software libraries with much reduced effort required for library implementation and maintenance as well as considerably lower learning requirements on users. To catalyse such advancement, we propose a roadmap and key action points for technology harmonization among RIs that we argue will build the foundation for efficient and effective integration of data and analysis.

1. Introduction

One of the great challenges of data-driven research is that data rarely come in a form that is immediately ready for analysis. Across industry and academia it is estimated that in a data-driven project 80% of the effort is spent on data preparation (Press, 2016; Wickham, 2014). As a consequence, a relatively small amount of time is spent on the actual analysis through which the primary value can be realised, and some research may not even be attempted because the necessary data preparation would be too time consuming.

Technological advances over the past decade have provided us with an unprecedented increase in compute resources for research, ushering in a new way of doing science (Hey et al., 2009). However, we will not be able to reap the dividend for research from these developments if we cannot reverse the unfavourable ratio of time spent on data preparation versus data analysis, thus allowing more resources to be shifted towards activities that allow us to extract knowledge from data.

The increasing need for long term, systematic, standardised monitoring required to understand the environment has led to the development of Research Infrastructures (RI). At the same time, technology advancements in storage and compute has provided an opportunity to make any data collected readily available for reuse. RIs are platforms that acquire, curate and publish continuous observation data for research and policy making. Important for RIs is that their data holdings and compute services are accessible and reusable not only for human users but also for machines (Weigel et al., 2020). Preparing the data for machine access paves the way for data reuse in computational environments for data analysis, e.g., in Jupyter Notebooks and High Performance Computing, and thus more actively supports the automated transformation of published data into analysis-ready data.

We conducted a survey of state-of-the-art approaches for integrating data and analysis implemented by world-leading RI in the Earth System and Environmental Sciences. Major RI advances in Earth System and Environmental Sciences studying global challenges such as climate change, geohazards or biodiversity loss have led to an enormous increase in the amount of data available in these domains. This trend was further fueled in recent years by the commissioning of large RIs that enable the permanent observation of the Earth System. The resulting research data managed by these infrastructures collected by individual researchers, groups, or projects are not only voluminous but also extremely heterogeneous, which reflects the multidisciplinarity as well as the large range of methods and technologies used in data acquisition and processing.

Within this scope, we studied automated approaches that improve access to research data published by repositories and facilitate automated transformation of published data into analysis-ready data in computational environments to enable a seamless integration of data and analysis. In our study, we paid particular attention to observational time series and whether such data can be efficiently (i.e., automatically) loaded into data structures for data analysis with Python as programming language and using Jupyter as a commonly used computational environment.

By analysing current approaches and solutions being implemented by world-leading RIs, we highlight the existing practices to interface data repositories with computational environments, underscore the challenges faced at this interface, and identify technology gaps in approaches by individual RIs. The survey also highlights the heterogeneity of existing practices and shows that there is enormous potential for practice harmonization.

A continuing challenge is the significant effort required to develop technologies that match the requirements of the many distinct application programming interfaces (APIs) implemented by data repositories with the many programming languages used by researchers for data analysis. We show that in practice this heterogeneity means that each data repository needs to develop, publish and maintain individual technologies for each (major) programming language used by researchers for data analysis (e.g., R, Python, Julia, Go, MATLAB, just to name a few). Such development and maintenance is inefficient and for many RIs untenable.

Building on this observation, we propose a roadmap for future coordinated development among RIs in the Earth and Environmental Sciences, and potentially in other disciplines, that will see a harmonization in approaches for seamless integration of data and analysis, and inevitably lead to increased efficiency, reduced development and maintenance costs, and lower learning curves for users.

The paper is structured as follows: Section 2 (Survey) presents the conducted survey, with a short description of the surveyed RIs and a description of the activities conducted to understand if and how the RIs support machine discovery of data access, given an identifier (e.g., digital object identifier (DOI)). Building on the survey, sections 3 (Solutions) and 4 (Discussion) present and discuss state-of-the-art solutions. In Section 5 (Roadmap), we suggest that the solutions can inspire a concerted technology harmonization among RIs that would enable the development and maintenance of RI and programming language independent solutions for data-analysis integration. Section 6 (Conclusions) closes this work with final remarks.

2. Survey

This section summarizes a systematic review of the approaches implemented by world-leading RIs in Earth System and Environmental Sciences to enable data and metadata access for both humans and machines. We first present the RIs included in the survey. We then detail the survey design and the conducted activities. Finally, we present our findings in a survey evaluation.

2.1. Selected research infrastructures

This section briefly introduces selected RIs and their data and information systems.

PANGAEA (www.pangaea.de) is a data publisher in Earth & Environmental Science, jointly managed by the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI) and the Centre for Marine Environmental Sciences (MARUM) at the University of Bremen. PANGAEA is a trustworthy repository (World Data System, CoreTrustSeal) which provides continued access and long term preservation of more than 400,000 datasets from various sub-disciplines of Environmental Sciences. These datasets have been collected through research infrastructures, projects and programs, and also includes long tail data collected or created by individual researchers. Access to these datasets is enabled through support for numerous community-specific as well as cross-domain standards. All published datasets are also tagged with a persistent identifier (DOI).

The *Terrestrial Ecosystem Research Network (TERN*, https://www.ter

n.org.au/) collects, curates and publishes data on temporal and spatial changes in Australia's terrestrial ecosystems. Established in 2009 with Australian government NCRIS (National Collaborative Research Infrastructure Strategy) grant funding, TERN's Data Discovery Portal gives open access to over 2500 open data collections. TERN's data collections are derived from continental-scale gridded remote sensing, soil and landscape products, plot-based soil and vegetation surveillance monitoring sites, calibration and validation campaigns for remote sensing, and sensors such as phenocams, acoustic monitors and eddy-covariance flux towers. TERN develops standardised ecological monitoring protocols and systems for data collection, storage and management.

AuScope (https://www.auscope.org.au/) is an NCRIS-funded Australian research infrastructure that develops and delivers practical tools to enhance accessibility of geoscience datasets. The AuScope Virtual Research Environment (AVRE) provides a unifying platform for all AuScope Programs' data and analytical needs (Wyborn et al., 2018). Its Scientific Software Solutions Centre (SSSC) provides an environment where scientific software can be published, discovered, shared with collaborators, and described for automated execution (Squire et al., 2018). The process of registering software at the SSSC captures description, license, versioning, and citation relevant information, as well as a machine-readable description of the software environment required to run it. With these elements, software from the SSSC can be chained together into workflows in virtual laboratories. This includes automated data preparation by client applications and management of output data. The elements are identified by persistent and versioned Uniform Resource Identifiers (URI).

The *Commonwealth Scientific and Industrial Research Organization (CSIRO,* https://www.csiro.au) is Australia's national research agency, covering a broad spectrum of science, engineering and medical research domains. Many datasets that originated from CSIRO research are made available through the CSIRO Data Access Portal (DAP, https://data.csiro.au). The repository serves both as a general data repository and as an institutional data archive. The CSIRO DAP was among the first data repositories to offer its metadata for harvesting using Schema.org (Noy and Brickley, 2017) in its user interface to make data landing pages machine readable.

The *National Ecological Observatory Network (NEON,* https://www.neonscience.org) is a Research Infrastructure (RI) established by the US National Science Foundation, with the mission to 'enable ecological forecasting of ecosystem function's response to natural and human-induced forcings such as climate, land use and invasive species across a range of spatial and temporal scales' (Schimel et al., 2011). It is a distributed site-based RI of ecological measurements and observations designed to scale from the site to the region-and-continent over the next 30 years. The observatory includes 81 field sites (including terrestrial and aquatic), airborne remote sensing, and a cyber-infrastructure for data acquisition, storage, analyses, and dissemination. NEON has 181 quality-controlled, open-source data products across a range of biotic and abiotic ecological processes and drivers, that include; biodiversity, biogeochemistry, climate, ecohydrology, invasive species and land use.

The *Chinese Ecosystem Research Network (CERN*, http://www.cern.ac.cn/) was established in 1988 by the Chinese Academy of Sciences, to obtain scientific data of ecosystem changes and to study the changes in structure, functions and processes of different ecosystems in China (Fu et al., 2010). Over the past 30 years, CERN has developed into a national innovative scientific and technological facility, including a synthesis center, a data center, five disciplinary sub-centers, and 44 networking stations. CERN's monitoring and experimental activities produce various data that are processed, integrated, and accessed through ecological stations, sub-centers, data centers, and the synthesis center under standardised procedure. CERN conducts network observation and experimentation across China's diverse ecosystems on a long-term basis, serves as a nexus for national ecological research, promotes data sharing, and creates an educational center and collaborative base for ecological researchers.

The *Integrated European Long-Term Ecosystem, Critical Zone & Socio-Ecological Research Infrastructure (eLTER RI* https://www.lter-europe.net/elter-esfri) comprises a wide range of highly instrumented European sites focusing on terrestrial, fresh- and transitional water ecosystems and also addressing socio-ecological interactions. Currently in the phase of preparing its operational implementation, it enables the in-situ and co-located acquisition and long-term preservation of ecosystem characteristics and Essential Variables ranging from biogeochemistry to biodiversity as well as socio-ecological characteristics. The RI provides and develops e-infrastructure for data managers, developers and scientists, including the Dynamic Ecological Information Management System - Site and Dataset Registry (DEIMS-SDR) (https://deims.org/), the vocabulary service EnvThes (http://vocabs.lter-europe.net/EnvThes/), the eLTER Data Integration Portal (DIP) (http://dip.lter-europe.net) and Central Data Node (CDN) (https://cdn.lter-europe.net/).

The *Integrated Carbon Observation System (ICOS*, https://www.icos-cp.eu/) is a European-wide greenhouse gas research infrastructure. ICOS produces standardised data on greenhouse gas concentrations in the atmosphere, as well as on carbon fluxes between the atmosphere, the earth and oceans. ICOS provides long term, high quality observations that follow the global standards for the best possible quality data on the atmospheric composition for greenhouse gases (GHG), greenhouse gas exchange fluxes measured by eddy covariance and CO_2 partial pressure at water surfaces. ICOS data is based on the measurements from over 140 stations across 12 European countries and is available at the ICOS data portal (https://data.icos-cp.eu/) with open access to data and metadata for download and instant graphical preview. A virtual research environment is provided as well with Jupyter Notebooks to the public, for collaborative research groups and education with direct access to the data.

The *European Network for Earth System Modeling Climate Data Infrastructure (ENES CDI*, https://is.enes.org/) is a Research Infrastructure which aligns and pools national services and resources to support the European climate research community. It is closely integrated into the worldwide Earth System Grid Federation (ESGF, https://esgf.llnl.gov/). Core services, such as data ingestion, hosting and access for climate simulations in the multi-PByte range, are complemented by persistent identifier (PID) services that enable data versioning, data replication, collection building and annotation with external information. High-level collections are associated with DOIs whose persistence is guaranteed by the World Data Centre for Climate (WDCC). Processing services close to the data are stepwise integrated into the infrastructure. This includes JupyterHub installations as well as web service interfaces based on OGC WPS (Open Geospatial Consortium Web Processing Service) standard.ENES provides standardised and quality-controlled, open data collections from various climate modeling activities. Most prominent examples are the collections from Coupled Model Intercomparison Projects (CMIP) and the Coordinated Regional Climate Downscaling Experiments (CORDEX).

NCI Australia (National Computational Infrastructure) (https://nci.org.au/) hosts data collections that are co-located with high-performance supercomputer infrastructure and cloud systems that generate data, process data streams or analyze data. The vast majority of this data has been from the climate, weather, geophysics and environmental sciences. As well as available through filesystem access and vast co-located software library, NCI publicly delivers the geospatial data through interoperable protocols wherever possible, including ISO (International Organization for Standardization) geospatial standards, OGC, and OpenDAP (Open-source Project for a Network Data Access Protocol), plus global federations such as the ESGF NCI has delivered these through a mixture of servers including GeoNetwork (https://geonetwork.nci.org.au) as a data discovery service. While NCI delivers large amounts of data through services such as THREDDS, it has developed its own scalable data services, e.g., through GSKY (https://gsky.nci.org.au). These services increasingly ensure that the computational processing of the data is handled on the server-side (Evans et al., 2015). Each

community on the system can also augment their computational ecosystem on the common data.

2.2. Design

To obtain a state-of-the-art overview of the current practices on data access offered by the selected RIs, we asked participating infrastructure representatives to fill a questionnaire on technical details and implemented standards of their data infrastructure. Specifically, we asked for persistent identifiers used, the metadata and data formats and standards offered as well as the implemented access protocols and interfaces. Furthermore, we asked the participants to provide example links to both data and metadata resources as well as the required authentication protocols. Finally, we asked the respondents to gather practices and thoughts on the following questions: *Describe how a data scientist can write a script (any language) that based on a DOI/PID loads the identified (meta)data into a data frame (native data structure in your language of choice). Do you or third parties offer special libraries for data access? If a DOI/PID is not sufficient, what information does the data scientist need to load the data of your RI into a data frame?*

Of primary interest are cross-domain practices based on widely used and easily implementable web standards suitable for both human and machine access and processing. We focused on two approaches for disseminating meta(data): embedding metadata within a web page and using content negotiation.

2.2.1. Embedded metadata

A very common method to expose machine readable metadata is to embed metadata in the HTML (Hypertext Markup Language) code of the landing page a PID/DOI resolves to. Traditionally, in the scholarly context this has been achieved using Dublin Core (Weibel and Koch, 2000) within META tags (e.g., title, date, creator, identifier) as recommended by the Dublin Core initiative (Kunze, 1999).

Links (Uniform Resource Locators - URLs) to data objects can be embedded in a landing page's HTML or in the response header following the typed links convention (RFC8228, Nottingham, 2010) which in the scholarly context has been refined by Van de Sompel and Nelson (2015) in their signposting initiative.

During recent years, the use of JSON-LD (JavaScript Object Notation for Linked Data; Sporny et al., 2020) encoded Schema.org metadata, e.g., in the HEAD section of an HTML document, has gained popularity in research related web pages. The use of Schema.org offers two important advantages and therefore has been implemented by a large number of data providers. First, it allows us to describe data sets in detail using the Schema.org/Dataset type. Second, its use improves search engine harvesting and thus visibility and discoverability of described data sets. In Schema.org/Dataset, links (URLs) to data objects can easily be captured using the Schema.org/distribution property.

2.2.2. Content negotiation

To offer web based content in different formats, content negotiation is a common approach to enable access to metadata optimized for both humans and machines. This is done by a client application sending HTTP (Hypertext Transfer Protocol) header requests in which the expected response format is specified using a valid MIME (Multipurpose Internet Mail Extensions) type within the Accept header field. This enables the server to deliver the metadata in the form requested by the client. JSON-LD encoded metadata can be requested using the application/ld + json MIME type, as shown in the following example:

```
GET doi:10.1594/PANGAEA.80968 HTTP/1.1
Accept: application/ld+json
```

Assuming the MIME type is supported, content negotiation could also be used to offer direct access to downloadable data objects such as NetCDF (Network Common Data Form format) files. This mechanism is thus particularly useful if data is offered to users in various alternative formats (Lóscio et al., 2017).

To discover links (URLs) to data objects within the served content generally requires domain knowledge. For example, XML (Extensible Markup Language) encoded ISO19139 (Geographic MetaData XML) uses the 'CI_OnlineResource' element while in XML encoded EML (Environmental Markup Language) this is done using the 'distribution' element. As described above, if Schema.org/Dataset encoded content is accessed through content negotiation, data object links can be discovered via the Schema.org/distribution property. This heterogeneity complicates the discovery for machines of links to data objects in metadata.

Both methods are currently widely used by RIs. Content negotiation is a W3C (World Wide Web Consortium) recommended practice for providing data on the web (Lóscio et al., 2017) and currently gains momentum in particular within the Open Data community through the emerging 'content negotiation by profile' approach (Svensson et al., 2019). JSON (JavaScript Object Notation) encoded metadata, in particular following the schema.org/Dataset specification, has been strongly promoted by major search engines (Guha et al., 2015) and is recommended by Google's dataset search, one of the largest and fastest growing search engines for research data (Brickley et al., 2019).

2.3. Evaluation

To determine if the described technologies are in use by the participating RIs, we tested direct machine access to data given a DOI or another (persistent) identifier used by the RI. The diagram in Fig. 1 shows the overall approach. We used F-UJI (Devaraju and Huber, 2020, https://github.com/pangaea-data-publisher/fuji) - a tool which allows RIs to estimate the FAIR level of a given data set - to perform metadata retrieval from a landing page as well as to test content negotiation for each data set. F-UJI also reports if the links to data objects are listed in metadata. Additionally, we manually inspected each landing page HTML source code in order to validate F-UJI's results. (See Tables 1 and 2.)

Our survey showed that all RIs offer a number of standardised exchange protocols and associated services to enable machine access to their metadata catalogues. However, no common agreement or mainstream practice exists regarding the choice of offered metadata schemas and associated exchange interfaces. The mainstream standards Catalogue Service for the Web (CSW) of the Open Geospatial Consortium (OGC) and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) were mentioned by all RIs except NEON and ICOS, which offer a GraphQL (query language for APIs) API and a SPARQL (SPARQL Protocol And RDF Query Language) endpoint, respectively. Except ICOS, all RIs offer community accepted standard metadata formats, among which ISO19115 (Geographic Information - Metadata), ISO19139, and EML are most frequent. Most RIs offer Schema.org metadata as JSON-LD in order to improve search engine discoverability. Similarly, the choice of data formats is not harmonized among the RIs, which is partly due to the large number of offered data types ranging from time series data to multimedia objects. However, for time series data, the common choice of available formats is limited to text formats, in particular CSV (comma-separated values) and TSV (tab-separated values), or NetCDF, which are offered by all RIs. Besides data set based access, RIs have also started to offer data via APIs such as OGC SOS (Sensor Observation Service) or WFS (Web Feature Service) allowing to retrieve customized data sets whose extent is determined via query parameters. Similarly, OpenDAP is used as query based data interfaces by some of the investigated RIs.

Persistent identifiers are especially important for reusing data, since they serve as reference in scientific citations. Therefore, they usually are the only available information for finding corresponding metadata and to identify and download the data objects, e.g., data files or data streams. All examined RIs routinely identify all their data sets or select data sets with PIDs. DOIs and Handles are the predominantly used

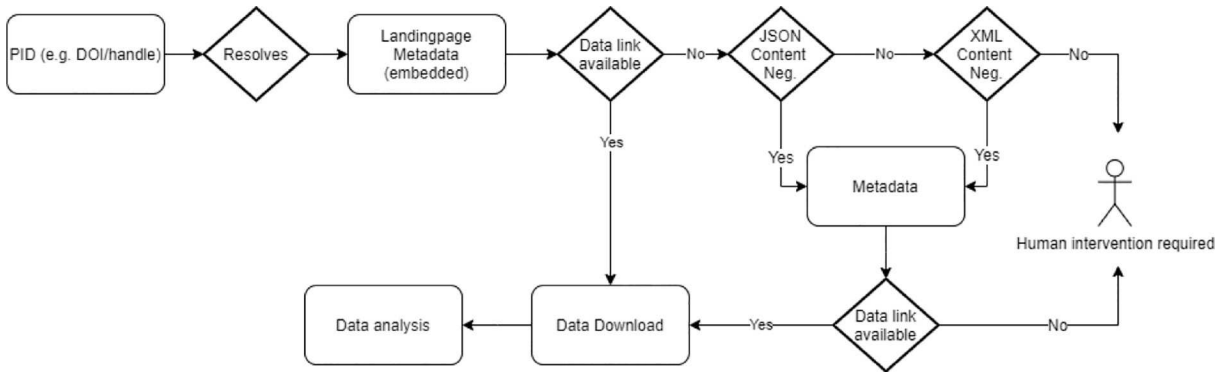

Fig. 1. A schematic overview of HTTP based methods to expose metadata in a machine as well a human friendly manner offering various routes for machine based discovery of links to downloadable data.

Table 1
Overview of most frequently used standards and interfaces for metadata and data access offered by investigated RIs.

RI	PID	Metadata access	Data formats	Data access
PANGAEA	DOI	OAI-PMH, HTTP	HTML, TSV	HTTP
TERN	DOI, PID	OAI-PMH, CSW	NetCDF, CSV, GeoTIFF	HTTP, OpenDAP, WFS, WMS
NEON	DOI, UUID	REST API, GraphQL, HTTP	CSV, HDF5, GeoTIFF	HTTP, REST API
CERN	DOI	HTTP	XLS, CSV, TIFF, SHP, NetCDF	HTTP
eLTER	DOI (external)	REST API, OAI-PMH, CSW	CSV, XLS, NetCDF, GeoTIFF, SHP	HTTP, SOS
ICOS	DOI, handle	HTTP, SPARQL	JSON, CSV, XML, TSV	HTTP
IS-ENES/ ESGF	DOI, handle	HTTP, OAI-PMH	NetCF, GRIB	HTTP, OpenDAP
NCI	DOI	CSW, OpenSearch	NetCDF, HDF5, GeoTIFF, CSV	HTTP, OpenDAP, WMS, WCS
AuScope AVRE	DOI (external)	CSW	XML	HTTP, WMS, WFS

Acronyms used in this table not explained elsewhere in this document: WMS: Web Map Service; WCS: Web Coverage Service; WFS: Web Feature Service; HDF5: Hierarchical Data Format; GeoTIFF: Geographic Tagged Image File Format; UUID: Universally Unique Identifier; SHP: Shapefile format; GRIB: General Regularly-distributed Information in Binary form format.

persistent identifiers.

Some of the above-mentioned mainstream protocols (e.g., OAI-PMH or CSW) provide standard methods to access data sets using identifiers. However, the identifiers listed by these catalogue services often differ in design from persistent identifiers. For example, OAI-PMH identifiers must follow a URI syntax different from RFC2396 (Berners-Lee, 1998), which effectively excludes DOIs in OAI-PMH metadata. This is a challenge for FAIR (meta)data because the principles prescribe the use of persistent identifiers. The same can be said for systems using OpenDAP or OGC data exchange standards, such as SOS or WFS, which do not allow data retrieval by persistent identifier. More important, however, is the fact that no standard or common (and machine actionable) agreement exists for how links to data objects shall be included in metadata. Consequently, none of the investigated RIs provides machine-actionable links to data objects on their landing pages or within provided metadata. As a result, an information gap exists between identifiers and data access. For machines this gap is particularly challenging since they fail to automatically identify the link.

As we highlighted, some of the above-mentioned protocols do not allow retrieval of data based on persistent identifiers. As illustrated in Fig. 1, we, thus, focus on access methods based on the HTTP protocol and evaluate how RIs address the gap between identifiers and data, and which possibilities and practices exist to make data accessible for machine-based data analysis environments using persistent identifiers as an entry point.

PANGAEA and IS-ENES provide metadata embedded in their landing page's HTML expressed as Dublin Core META tags. PANGAEA, CERN and IS-ENES offer their metadata encoded as JSON-LD, following the Schema.org/Dataset convention, embedded in the HTML of their landing pages. Additionally, PANGAEA, TERN, CSIRO, IS-ENES, ICOS and NCI offer JSON encoded metadata via content negotiation. Except ICOS which offers a proprietary JSON encoding, all RIs offer this JSON metadata encoded as standardised Schema.org/Dataset style JSON-LD. NEON also provides JSON-LD schema.org/Dataset encoded metadata embedded in the HTML of its landing page. However, in NEON this HTML is dynamically generated using JavaScript and access to metadata thus relies on a client that interprets JavaScript.

Except PANGAEA, none of the investigated RIs serves JSON metadata containing a standard compliant link to a downloadable data object. ICOS' custom JSON and XML metadata formats do contain data file name and access URL, but are not machine actionable due to a license agreement which needs to be manually accepted. PANGAEA and NEON include sufficient metadata about data objects such as actionable links to data files and encoding information (MIME type) in their Schema.org/Dataset encoded JSON-LD metadata. However, only PANGAEA uses this data object metadata (namely its MIME type) to enable direct download of data files using content negotiation, for example:

```
GET doi:10.1594/PANGAEA.80968 HTTP/1.1
Accept: text/tab-separated-values
```

To summarize, we found that despite the comparably high standardization level among RIs attained during the past years, it still seems to be a challenge to provide the same level of information to both machines and humans. When using a DOI or URL as starting point, a human user is immediately directed to a landing page. From there, for humans it is generally straightforward to discover rich metadata and download the corresponding data. In contrast, it is surprisingly difficult for machines to find interpretable metadata with links to data objects using a DOI as a starting point. Although content negotiation is supported by some RIs, links to downloadable data objects are rarely included in this metadata. Also, these links are rarely embedded in machine-readable form in the HTML code of the landing pages.

Table 2
Evaluation results performed on selected datasets from investigated RIs using the HTTP based methods illustrated in Fig. 1 for machine based discovery of links to downloadable data.

	Evaluated PID	PIDs available	Data link on landing page	Metadata embedded	Content negotiation JSON	Content negotiation XML	Data link in JSON	Data link in XML
PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.896543	Y	Y	Y[a,b]	Y	N	Y	–
TERN	https://doi.org/10.4227/05/5344F1159A1A9	Y	N	–	Y[a]	N	N	N
CSIRO	doi:https://doi.org/10.4225/08/563869A931CFE	Y	–	–	Y[a]	N	N	–
NEON	https://data.neonscience.org/data-products/DP1.00001.001	Y[e]	Y[c]	Y[a,b,c]	N	N	Y[c]	–
CERN	https://dx.doi.org/10.11922/sciencedb.293	Y	N	Y[a]	N	N	–	–
eLTER	https://deims.org/dataset/75a7f938-7c77-11e3-8832-005056ab003f	Y[e]	N	N	N	N	–	–
ICOS	https://hdl.handle.net/11676/8YwZj8CQEj87IuI9P6QkZiKX	Y	N	N	Y[d]	Y[d]	Y[d]	N
IS-ENES/ESGF	doi:10.22033/ESGF/CMIP6.4397	Y	N	Y[a,b]	Y[a]	N	N	–
NCI	doi:10.25914/5eaa30de53244	Y	Y	N	Y[a]	N	N	–

Footnotes:
[a] Schema.org.
[b] Dublin Core.
[c] Content generated by JavaScript.
[d] Proprietary or custom format.
[e] Partly implemented.

3. Solutions

The above described approaches can be used to automate ingesting data and metadata into computational environments. Some of the investigated RIs have proposed solutions by developing specialized software libraries that automate such ingestion. For example, PANGAEA has recently published the Python library `pangaeapy` (Huber et al., 2020) (https://pypi.org/project/pangaeapy/), which allows reading data sets into a native Python object given a DOI. The library leverages PANGAEA's web services to retrieve rich metadata and load the associated tabular data into a Python (`pandas`) data frame object (McKinney, 2012). The data frame data structure is commonly used in the wide range of statistical and data visualisation libraries, e.g., `scipy` or `matplotlib`. The direct availability of data in this data structure streamlines data processing for data scientists. `Pangeapy`'s data structures for metadata follow the PANGAEA data model (Diepenbroek et al., 2017), which is centered around the `Dataset` class structure containing several supportive classes such `Event` and `Parameter` for information about the geographical and methodological context and the observed property, respectively, as well as the `Data` themselves, for the individual measurements and observations. To represent PANGAEA specific metadata, `pangaeapy` implements the `PanDataSet` class, which provides attributes to hold objects for, among other, `PanEvent` and `PanParam` classes. Individual metadata values can be accessed through their corresponding object attributes. The tabular data of a data set is stored in the `data` attribute of `PanDataSet`, which holds a data frame object. In some cases, PANGAEA's tabular data does not contain temporal or geographical information (e.g., latitude and longitude) as this information is instead in the metadata of a PANGAEA data set, e.g., in an associated `Event`. In such cases, `pangaeapy` adds additional data columns to the data frame to hold these data.

Within a Python based data analysis environment such as Jupyter, PANGAEA data and metadata can be loaded with the following statement:

```
pandata = PanDataSet('doi:10.1594/PANGAEA.889516')
```

Similarly to PANGAEA, ICOS provides a Python library (`icoscp`) (https://pypi.org/project/icoscp/) to provide users high-level, performant and easy access to tabular data. In ICOS, persistent identification of digital data objects resolves to a human friendly landing page. The provided metadata includes a filename needed to conventionally download the file. With a single instruction, the `icoscp` library loads this metadata and data into a Python data frame. The following statements can be used to load ICOS data into a data frame by either (1) the local identifier, (2) the Handle, or (3) the landing page URL:

```
icosdata = Dobj('XA_Ifq7BKqS0tkQd4dGVEFnM')
icosdata = Dobj('https://hdl.handle.net/11676/XA_Ifq7BKqS0tkQd4dGVEFnM')
icosdata = Dobj('https://meta.icos-cp.eu/objects/XA_Ifq7BKqS0tkQd4dGVEFnM')
```

As with `pangaeapy`, the returned Python object contains a standard `pandas` data frame and attributes describing the data columns. A column description contains for example: type = "gross primary CO2 production", unit = "µmol m-2 s-1", kind = "particle flux", whereas the metadata for the data set itself contains the citation string, among other information. Currently, `icoscp` is limited to loading time series data (CSV). As a rule of thumb: data sets that can be previewed in the data portal are accessible through the library. Overall, the `pangaeapy` and `icoscp` libraries considerably streamline loading data into computational environments and, thus, support making data analysis ready. This concerns in particular the reading of data into suitable data analysis formats as well as data harmonization and cleansing.

Having data available in data frames is an important first step in overall data processing and analysis. As a showcase for PANGAEA and ICOS data processing and analysis, we used `pangaeapy` and `icoscp` to synthesise data from both RIs in a common computational environment, for which we used Jupyter. We chose two complementary data sets (Diverres et al., 2020; Knust and Rohardt, 2018) with data collected during ship-based physical oceanography and carbon dioxide measurements (Pfeil et al., 2013). Together, the data represent two close transects across the Atlantic ocean, measured during two ship expeditions that occurred within a 2 months' time frame.

With the libraries, we load the data sets directly into data frames using the respective persistent identifier as follows:

```
icosdata = Dobj('https://hdl.handle.net/11676/xgu4rfCmqvXb4w1wGGD6mYsB')
icosdata_frame = icosdata.get()
pandata = PanDataSet('https://doi.org/10.1594/PANGAEA.889516')
pandata_frame = pandata.data
```

We used the `pandas` built-in `plot` method and `matplotlib` to plot the sea surface temperature data against their measurement timestamp of both data sets. As Fig. 2 shows, the data sets are temporally closely connected.

The showcase highlights (code in Fig. 2) that only the data are uniform while metadata remains heterogeneous, with RI-specific syntax and semantics. Indeed, the columns of interest are labelled `Temp` and `Temp [degC]` in PANGAEA and ICOS data, respectively. Both are water temperature observations in degree Celsius. This can be inferred from the metadata, but is implicit and not (easily) correctly interpreted by machines. The depth (sensor depth in the water) for the measurement is unknown, which highlights a lack of harmonized annotation practices for accurately describing observable properties. Unfortunately, resolving this requires manual intervention. In the future, this semantic problem must be addressed more systematically, e.g. with a unified parameter nomenclature or a semantic interoperability framework (Magagna et al., 2020).

We used the `cartopy` (Met Office, 2020) Python module to show the geographical variation of measured temperatures along both transects. The module provides advanced cartographical plotting features to add a geographical context to `matplotlib` plotting results using a variety of geographical projections. The plotted map (Fig. 3) nicely shows the geographical complementarity of both datasets as well as the expected longitudinal variation of observed sea surface temperatures.

In addition to data analytics, both libraries support proper data citation, the `citation` property in each data structure, which prints the data citation including the preferred persistent identifier (Fig. 4). (See Fig. 4.)

The presented chart and map plotting examples nicely show the advantage of libraries that streamline the data ingestion and harmonization tasks and thus contribute to ensuring that data scientists can focus more on data analysis. Although the example given here is very Python-centric, we note that the proposed approach can be implemented using other languages. For example, the R library `pangaear` (Chamberlain et al., 2016) offers functionality comparable to `pangaeapy`.

4. Discussion

The implementation by RIs of harmonized and standardised access to data and metadata described here shows some clear trends. For example, we observe that cross-community metadata standards are gaining momentum in environmental RIs. Here, embedding metadata in landing pages following the Schema.org convention for structured data on the web (JSON-LD) plays a very important role. This is an interesting development, since the main focus in recent years has been on community specific formats. For example, there is a remarkable diversification of OGC and ISO standards in the numerous community profiles and extensions to the ISO19115 metadata standard (see, e.g., Brodeur et al., 2019). These are highly specialized formats in contrast to the generic and easy to implement Schema.org.

An analysis of Google Trends shows a decreasing interest in ISO19115 after about 2011, and at the same time an increasing interest in Schema.org (Fig. 4). In 2013, the Schema.org/Dataset type was included, which enabled data providers to richly describe their data assets and major search engines to harvest them.

Three years later, Wilkinson et al. (2016) published the FAIR data principles, which describe "concise, domain-independent, high-level principles that can be applied to a wide range of scholarly outputs" while recognizing the importance of discipline specific requirements. Since then, the FAIR principles have had overwhelming success within the scientific community and are endorsed by major scientific stakeholders including publishers, funders and policy makers (see, e.g., Stall et al. (2019) for initiatives in Earth and environmental sciences).

Both the FAIR data principles as well as search engine optimisation (SEO) approaches have similar requirements for domain agnostic provision of metadata and have a comparably high standard with respect to detail and completeness. As Schema.org serves two purposes (SEO and FAIR), it is now used by a rapidly increasing number of data providers to enable FAIR metadata and data provision.

The use of persistent identifiers is another prerequisite for FAIR data, and their advantages have been described in detail by Philipson (2019).

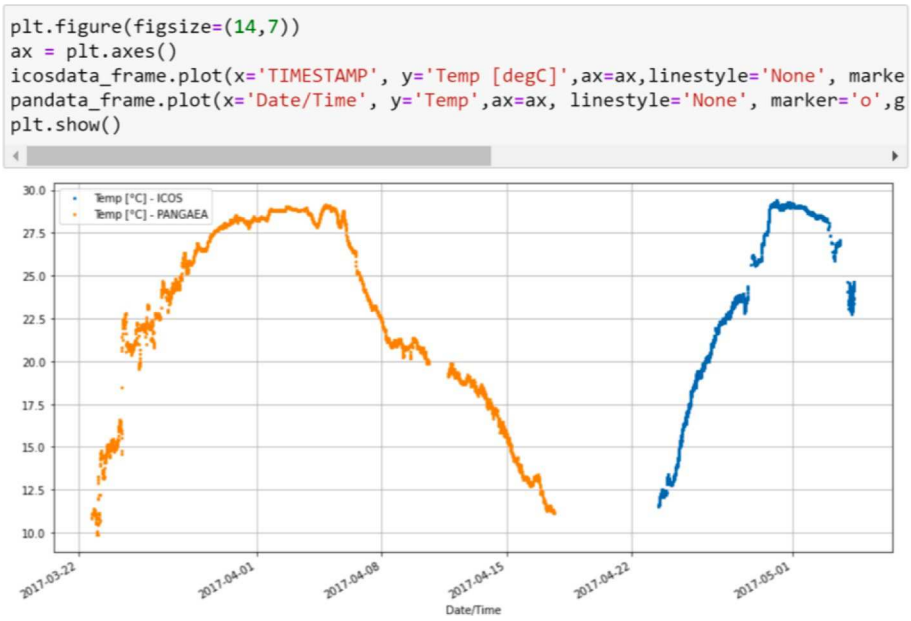

Fig. 2. Time series of water temperature in degree Celsius for two ships crossing the Atlantic ocean. On the left in orange, the data set published by PANGAEA, with observations from Europe to South America. On the right in blue, the data set published by ICOS, with data collected from Europe to Brazil. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

```
proj=ccrs.cartopy.crs.Miller()
plt.figure(dpi=150)
ax = plt.axes(projection=proj)
ax.set_extent([-60, 20, 40, -40])
ax.stock_img()
land_50m = feat.NaturalEarthFeature('physical', 'land', '50m',edgecolor='grey',f
ax.add_feature(land_50m)
ax.coastlines()
ax.scatter(pandata_frame['Longitude'],pandata_frame['Latitude'],c=pandata_frame[
ax.scatter(icosdata_frame['Longitude'],icosdata_frame['Latitude'],c=icosdata_fra
plt.show()
```

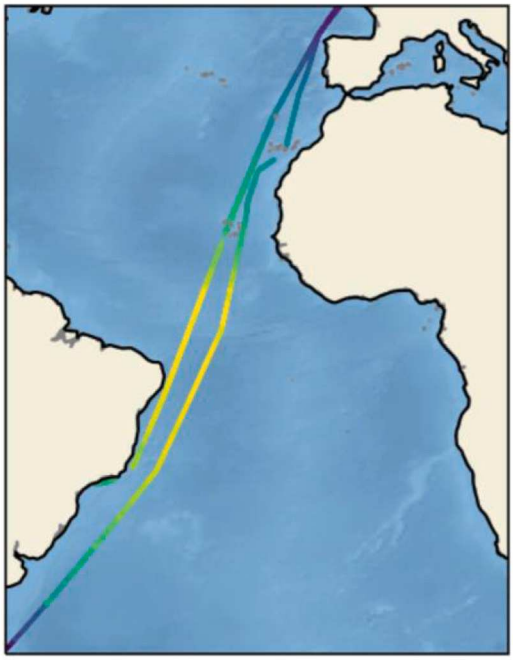

Fig. 3. The PANGAEA and ICOS data sets plotted with geolocation and colour gradients (dark blue, minimum, to yellow, maximum) to represent the sampled water temperature. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

```
for references in [icosdata.citation, pandata.citation]:
    print(references)
```

```
Diverres, D., Lefèvre, N., ICOS RI, 2020. ICOS OTC Release, FR-SOOP-France-Brazil , 2017-04-23-2017-05-04, https://hdl.handl
e.net/11676/xgu4rfCmqvXb4w1wGGD6mYsB
Knust, Rainer; Rohardt, Gerd (2018): Continuous thermosalinograph oceanography along POLARSTERN cruise track PS105 (ANT-XXXI
I/4). Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA, https://doi.org/10.159
4/PANGAEA.889516
```

Fig. 4. Example of how to obtain the citation strings for both data sets.

The environmental domain, in particular, has been using PIDs for a variety of data resources and products. In recent years, the reliability of PID systems has been a cause for concern, resulting in DOIs becoming the tool of choice in most communities, as this system has proven to be the most trusted and consistent (Klump and Huber, 2017), a trend we also observe among RIs investigated in this study. Persistent identifiers assigned by trustworthy data archives are essential to make data citable and thus enable the data-based reproducibility of research results. They represent the link between analysis results, published research data and publications based on them and are thus the bridge between publishers, data and computational environments.

While being frequently used by the investigated RIs, some of the above mentioned catalogue services, namely OAI-PMH and CSW, unfortunately have significant disadvantages with respect to the above described FAIR data practises. As they often require using internal identifiers, OAI-PMH, CSW and other catalogue services are less useful for metadata retrieval within interdisciplinary data science applications. Furthermore, the use of these protocols requires additional knowledge such as the web location of the individual service endpoint. Except for OpenSearch, no standard and widely accepted method exists to expose machine readable links to metadata search or catalogue services within a data set web page. We therefore recommend to offer domain-agnostic machine readable metadata, preferrably Schema.org/Dataset JSON-LD, in closest connection with a human readable data set web page. As discussed above, this can be achieved by implementing commonly used web technologies such as HTTP based content negotiation or embedding metadata within the HTML code of the landing page. Both methods already are used by investigated RIs.

In principle, the harmonized use of persistent identifiers and common approaches for exposing cross-disciplinary metadata should enable

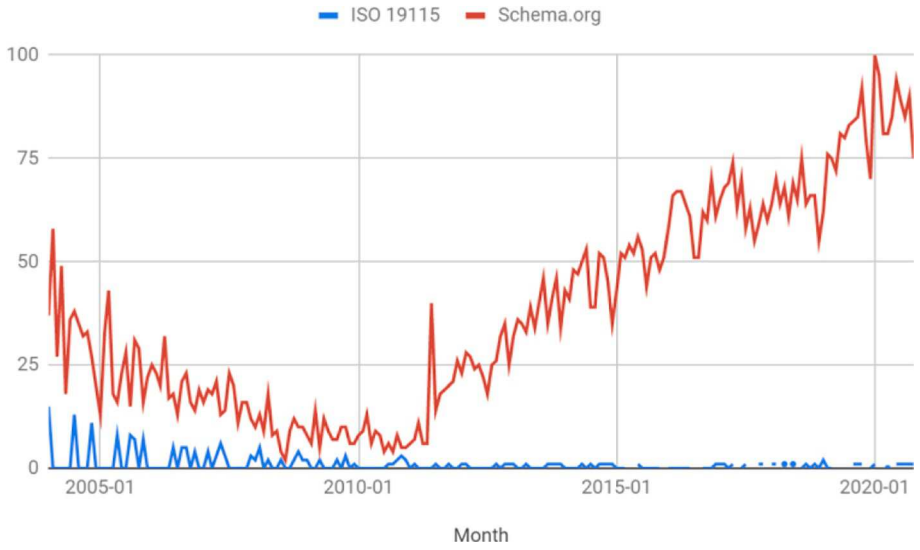

Fig. 4. Google trend analysis, showing in red increased interest in 'Schema.org' and in blue a decrease for ISO19115. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data to be transferred to computational environments using a uniform approach (Weigel et al., 2020). The technology components required to directly access data given a DOI and negotiate content between data provider and consumer does exist. While necessary for a solution, these components are not sufficient for our ultimate goal of having the data in a native data structure of the computational environment. Essential is also the explicit linking of data in metadata (FAIR sub principle F3: "Metadata clearly and explicitly include the identifier of the data they describe") and we observed that such links are often not available or not identifiable by machines. Moreover, data are available in many different formats, which considerably complicates a unified approach since loading data into computational environments relies on additional information. Such additional information needs to be encoded in programming code. A solution is presented by specialized software libraries such as `pangaeapy` or `icoscp`. Specifically for an RI and programming language, implement the required software logic and with this simplify the steps from download to the transformation of metadata and data into a uniform data structure for analysis.

Although the approach to use specialized libraries is promising, our use case revealed some open issues. For example, it takes time to get familiar with multiple libraries. It would therefore be desirable if the libraries (interface) were uniformly designed. Although `icoscp` and `pangaeapy` are very similar, a common library framework for accessing and ingesting data from a variety of RIs would be very helpful and would improve the overall accessibility and reusability of their data sets. Furthermore, a common data access and ingestion framework would ease adoption and development of comparable libraries and tools for other RIs. Another open issue is that the libraries presented here are made available for only one programming language, i.e., Python, and therefore cannot support the large number of data scientists using alternative languages such as Julia or R. A common framework should - to the extent this is possible - be defined independent from programming languages.

We therefore favour and propose a more standards based solution, which builds on standardization and best practices we discussed in this study. As mentioned above, a distinct trend towards the use of Schema.org JSON-LD encoded metadata in combination with persistent identifiers can be observed among all investigated RIs. A common strategy of FAIR implementation in RIs should build on the high level of standardization we observed but include the definition of some common best practises for, e.g., Schema.org implementations, agreement on the link used to point to data in metadata, and a minimum set of metadata elements required by data scientists. These best practises should build on accepted FAIR principles for data objects, including the FAIR sub principle F3. Inclusion of data identifier in the metadata of data to enable its machine based access is also recommended by existing work on the definition of FAIR metrics in the EOSC (European Open Science Cloud) context (Genova et al., 2020) such as the Research Data Alliance (RDA) FAIR data maturity model (RDA FAIR Data Maturity Model Working Group, 2020) or the FAIRsFAIR project (Devaraju et al., 2020). FAIR assessment tools such as F-UJI can help during the journey towards FAIR compliance by testing the inclusion of data identifiers in metadata.

While the technologies and approaches would allow for harmonized access to metadata for use within data analytics, access to data still is problematic due to the rather low degree of standardization and the lack of agreements on data syntax and semantics among the RIs. Additional efforts are therefore necessary to reach such agreement. Since RIs cover a wide range of scientific disciplines, interdisciplinary formats are the most promising. Candidate formats include established scientific binary formats such as NetCDF, HDF or the emerging interdisciplinary text based formats 'CSV on the Web' (Tennison, 2016) or 'frictionless data' (Fowler et al., 2018). The latter two are especially interesting for data scientists as they are based on flat, two dimensional comma separated tables that are particularly easy to load using languages such as Python or R. In contrast, the handling of multidimensional data in NetCDF or HDF requires considerable knowledge about the dimensional structure of the contained data. In summary, the choice of suitable data formats is relatively small and need not be limited to a single format.

As mentioned earlier, another important aspect is dynamic data retrieval APIs such as OGC SOS, which represent an alternative to offering data-access via machine-actionable persistent identifiers. SOS has for example been used as data infrastructure for national air quality data provisions to the European Environment Agency (Kotsev et al., 2015) and as a platform to provide public real-time access and harvesting of marine data to be archived in PANGAEA (Huber et al., 2012). It is used by RIs such as eLTER to publish site-based environmental observation and measurement data and has also been extended or used as a basis for additional services in diverse application scenarios such as data quality assurance (Devaraju et al., 2015; Goldfarb et al., 2020). Open Source libraries such as the R-package `sos4R` (https://github.com/52North/sos4R) allow programmatic access to data comparable to the above described PANGAEA and ICOS libraries, with the difference that for `sos4R` an API endpoint and the correct parameters must be provided instead of a persistent identifier. Publishing data only via API and not via

regular file-based persistently identified snapshots raises concerns regarding, in particular, reproducibility, since dynamic data can change. Pioneering initiatives such as the RDA Working Group on Dynamic Data Citation (WGDC) (Rauber et al., 2016) have provided detailed guidelines how this gap can be addressed, primarily by assigning PIDs to the queries used to retrieve data sets.

Though not in the scope of this work, we emphasize that integrating data and analysis must also encompass semantic harmonization of terms used in (meta)data. This should allow for less ambiguous and machine-actionable descriptions of complex observable properties beyond what is currently possible or practiced with, e.g., Schema.org/measuredProperty. The RDA InteroperAble Descriptions of Observable Property Terminology (I-ADOPT) Working Group (WG) addresses this problem and is developing a semantic interoperability framework (Magagna et al., 2020).

The last two decades have seen a remarkable change in the use of computational environments. In the early days, community-tailored Virtual Research Environments (VREs, also termed Virtual Research Laboratories or Science Gateways; see, e.g., Barker et al., 2019), provided pre-canned workflows for scientific analysis required by a research community. However, VREs proved to be inflexible as they rely on adaptation to meet the frequently changing requirements (Voss and Procter, 2009). Moreover, custom systems generally require considerable maintenance effort, which consequently results in sustainability challenges for many VREs and broader adoption beyond the community that built them for their specific needs (particularly those that tightly coupled data sets to specific tools, e.g., Candela et al., 2013). In recent years, however, generic, domain-agnostic solutions have become more common. They rely more on programmable data handling and analysis. In this context, the rapid spread of computational notebooks, especially Jupyter, has created an incredible dynamic in the context of data sciences (Perkel, 2018). For example, AuScope has expanded its VREs to enable users with varying skills to specifically target their needs and either access a range of online data and software services to now create their own workflows in their own environment. Both data and software are accessible via standardised interfaces and are being utilised by individual researchers who commonly use computational notebooks to mix and match data, software and tools to create their own exploratory workflows (Wyborn et al., 2018).

For users who utilize analysis tools, be it advanced community specific VREs or desktop based Jupyter notebooks, a common RI approach to expose metadata and data as described above would be very advantageous. Environmental RIs play a major role worldwide for a large number of users from research, industry and policy. The growing number of such facilities and the increasing quality of the measurement methods used have led to a sharp increase in the amount of available research data. This explains the increasing importance of data science, specifically also in environmental sciences (Raban and Gordon, 2020). Both growing data and advanced analytics are essential elements in the production of knowledge required to address urgent societal problems such as climate change, loss of biodiversity or natural disasters. Addressing the problem discussed here is an important responsibility of the e-infrastructures managing the data. These infrastructures must support efficient and effective data-driven, interdisciplinary research. Streamlining data flow into the analysis tools used is an important sub task, towards which this work contributes important insight.

5. Roadmap

The discussed approaches for data and metadata provision would enable a significant fraction of RI-collected data to be more easily integrated into a computational platform of choice. However, the state-of-the-art of creating, publishing and maintaining software libraries specialized for each RI and programming language popular in data analysis does not scale well, is inefficient and ultimately not sustainable.

We argue for a concerted technology harmonization effort among the RIs so that their data assets can be seamlessly integrated into arbitrary computational platforms and programming languages with minimum effort required for development and maintenance of libraries. The requirement of minimum effort relies on developing generic approaches and implementing these in generic libraries, which itself relies on technology harmonization among RIs for which we need a concerted effort.

We thus suggest a roadmap and actions that the RIs involved here now intend to implement and others could follow as well. The plan involves the following measures, which we categorize into immediate (I), short (S) and medium term (M):

- Persistently identify data sets
- Adopt open and globally implementable communication protocols to exchange data and metadata, where possible HTTP
- Resolve persistent identifiers to landing pages that are human and machine readable, where possible HTTP-based resolution
- (S) Offer metadata in a domain-agnostic format, preferably JSON-LD following the schema.org/Dataset specification
- (M) Harmonize RI-relevant metadata (e.g., observed properties, methods, etc.).
- (S) Ensure that data set metadata include an explicit (i.e., machine actionable) link to the corresponding data object so that given a persistent identifier machines can access data without human intervention
- (M) Ensure that accessed data are offered in a web and data science friendly data format to (e.g., 'CSV on the Web' or 'frictionless data')
- (M) Offer metadata and data through content negotiation.

Although the presented work focuses on RIs, we note that the proposed solutions and roadmap are applicable also to environmental data infrastructures not directly associated with RIs. The proposed measures are relatively straightforward to implement and, if widely applied, could contribute to significantly improving the reusability, and thus FAIRness, of environmental data.

6. Conclusions

We presented how some of the world's largest environmental research infrastructures (RIs) make their data available to the scientific community. We found that while access to persistently identified data and metadata is generally straightforward for humans, this is not true for machines. For most RIs, it is thus still a challenge to make analysis-ready data available in computational environments. Moreover, in recent years the FAIR principles have defined further requirements for data providers, including reproducibility and citation of data used in analysis. We argue that it is important to address these challenges in order to efficiently and effectively support data-intensive research in popular modern data analysis languages such as Python and R. We present and discuss current approaches implemented by RIs to address the challenge of seamless integration of data and analysis. Our analysis shows that the state-of-the-art approach is for the RI to provide specialized software libraries in a data analysis language of choice that streamline loading persistently identified data and metadata into a data structure native to the language. This relies on established domain-independent standards and web-based practices that in principle allow for the design of uniform, programming language and RI independent approaches, which could enable seamless integration of arbitrary data and analysis conducted in virtual research environments. For this to become viable in practice, we need a concerted technology harmonization effort among the RIs, for which we proposed a roadmap that the RIs involved here now intend to implement.

Declaration of competing interest

None.

Acknowledgements

This manuscript is in part the product of environmental research infrastructures and related projects working towards the harmonization of data over the past decade. This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreements No. 824068 (ENVRI-FAIR project) and No. 831558 (FAIR-sFAIR project). NEON is a project sponsored by the National Science Foundation (NSF) and managed under cooperative support agreement (EF-1029808) to Battelle. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our shareholders or sponsors. This research would not have taken place without decades of collegial interactions and community engagement with our peers and mentors. AuScope, TERN and NCI are supported by the National Collaborative Research Infrastructure Strategy (NCRIS), an Australian Government Initiative.

References

Barker, M., et al., 2019. The global impact of science gateways, virtual research environments and virtual laboratories. Futur. Gener. Comput. Syst. 95, 240–248. https://doi.org/10.1016/j.future.2018.12.026.

Berners-Lee, T., 1998. Uniform Resource Identifiers (URI): generic syntax. Internet Eng. Task Force RFC 2396. https://www.ietf.org/rfc/rfc2396.txt.

Brickley, D., et al., 2019. Google Dataset Search: building a search engine for datasets in an open Web ecosystem. In: The World Wide Web Conference, Vol. WWW'19, pp. 1365–1375. https://doi.org/10.1145/3308558.3313685.

Brodeur, J., et al., 2019. Geographic information metadata—an outlook from the international standardization perspective. ISPRS Int. J. Geo-Inf. 2019 (8), 280. https://doi.org/10.3390/ijgi8060280.

Candela, L., et al., 2013. Virtual research environments: an overview and a research agenda. Data Sci. J. Vol. 12, GRDI75–GRDI81. https://doi.org/10.2481/dsj.GRDI-013.

Chamberlain, S., Woo, K., MacDonald, A., Zimmerman, N., Simpson, G., 2016. Pangaear: Client for the 'pangaea' Database. Available at: https://CRAN.R-project.org/package=pangaear.

Devaraju, A., Huber, R., 2020. F-UJI – an automated FAIR data assessment tool (version v1.0.0). Zenodo 2, 10. https://doi.org/10.5281/zenodo.4063720.

Devaraju, A., et al., 2015. Q-SOS—A sensor observation service for accessing quality descriptions of environmental data. ISPRS Int. J. Geo-Inf. 4, 1346–1365. https://doi.org/10.3390/ijgi4031346.

Devaraju, A., et al., 2020. FAIRsFAIR Data Object Assessment Metrics (Version 0.4). Zenodo. https://doi.org/10.5281/zenodo.4081213.

Diepenbroek, M., et al., 2017. Terminology supported archiving and publication of environmental science data in PANGAEA. J. Biotechnol. 261, 177–186. https://doi.org/10.1016/j.jbiotec.2017.07.016.

Diverres, D., Lefèvre, N., ICOS, R.I., 2020. ICOS OTC Release, FR-SOOP-France-Brazil, 2017-04-23–2017-05-04. https://hdl.handle.net/11676/xgu4rfCmqvXb4w1wGGD6mYsB.

Evans, B., et al., 2015. The NCI high performance computing and high performance data platform to support the analysis of petascale environmental data collections. In: Denzer, R., Argent, R.M., Schimak, G., Hřebíček, J. (Eds.), Environmental Software Systems. Infrastructures, Services and Applications. ISESS 2015. IFIP Advances in Information and Communication Technology, Vol. 448. Springer, Cham. https://doi.org/10.1007/978-3-319-15994-2_58.

Fowler, D., et al., 2018. Frictionless data: making research data quality visible. Int. J. Digit. Curation 12 (2), 274–285. https://doi.org/10.2218/ijdc.v12i2.577.

Fu, B.J., et al., 2010. Chinese ecosystem research network: progress and perspectives. Ecol. Complex. 7 (2), 225–233. https://doi.org/10.1016/j.ecocom.2010.02.007.

Genova, F., et al., 2020. EOSC FAIR metrics – second draft for consultation (version second draft). Zenodo. https://doi.org/10.5281/zenodo.4106116.

Goldfarb, D., et al., 2020. Providing a user-friendly outlier analysis service implemented as open REST API. EGU Gen. Assem. https://doi.org/10.5194/egusphere-egu2020-1490. Online, 4–8 May 2020, EGU2020-14903.

Guha, R.V., et al., 2015. Schema.org: evolution of structured data on the web. Commun. ACM 59 (2), 44–51. https://doi.org/10.1145/2857274.2857276.

Hey, T., et al., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft. Microsoft Research, Redmond, WA. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf.

Huber, R., et al., 2012. Real time access and long term archiving concepts for HYPOX observatory data. In: EGU General Assembly Conference Abstracts, p. 2842. https://ui.adsabs.harvard.edu/abs/2012EGUGA..14.2842H/abstract.

Huber, R., et al., 2020. pangaeapy – a Python module to access and analyse PANGAEA data. Zenodo. https://doi.org/10.5281/zenodo.4013941.

Klump, J., Huber, R., 2017. 20 years of persistent identifiers – which systems are Here to stay? Data Sci. J. 16, 9. https://doi.org/10.5334/dsj-2017-009.

Knust, R., Rohardt, G., 2018. Continuous Thermosalinograph Oceanography Along POLARSTERN Cruise Track PS105 (ANT-XXXII/4). Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA. https://doi.org/10.1594/PANGAEA.889516.

Kotsev, A., et al., 2015. Building bridges: experiences and lessons learned from the implementation of INSPIRE and e-reporting of air quality data in Europe. Earth Sci. Inf. 8 (2), 353–365. https://doi.org/10.1007/s12145-014-0160-8.

Kunze, J., 1999. Encoding dublin core metadata in HTML. In: Dublin Core Metadata Initiative Memo. https://www.hjp.at/doc/rfc/rfc2731.html.

Lóscio, B.F., et al., 2017. Data on the web best practices. W3C Recomm. 31 (1). World Wide Web Consortium. https://www.w3.org/TR/dwbp/.

Magagna, B., et al., 2020. Towards an interoperability framework for observable property terminologies. EGU Gen. Assem. https://doi.org/10.5194/egusphere-egu2020-19895. Online, 4–8 May 2020, EGU2020-19895.

McKinney, W., 2012. pandas: a foundational Python library for data analysis and statistics. In: Python for High Performance and Scientific Computing, pp. 1–9. https://www.oreilly.com/library/view/python-for-data/9781449323592/.

Met Office, 2020. "Cartopy: A Cartographic Python Library with a Matplotlib Interface". 0.18.0. https://scitools.org.uk/cartopy.

Nottingham, M., 2010. Web linking. Internet Eng. Task Force RFC 5988. https://tools.ietf.org/html/rfc5988.

Noy, N., Brickley, D., 2017. Facilitating the discovery of public datasets. Tech Blog 24, 1. Google. http://ai.googleblog.com/2017/01/facilitating-discovery-of-public.html.

Perkel, Jeffrey M., 2018. Why Jupyter is data scientists' computational notebook of choice. Nature 563, 145–146. https://doi.org/10.1038/d41586-018-07196-1.

Pfeil, B., et al., 2013. A uniform, quality controlled Surface Ocean CO2 Atlas (SOCAT). Earth Syst. Sci. Data 5, 125–143. https://doi.org/10.5194/essd-5-125-2013.

Philipson, J., 2019. Identifying PIDs playing FAIR. Data Science 2 (1–2), 229–244. https://doi.org/10.3233/DS-190024.

Press, G., 2016. Cleaning Big Data: most time-consuming, least enjoyable data science task, survey says. Forbes 23, 3. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/.

Raban, D.R., Gordon, A., 2020. The evolution of data science and big data research: a bibliometric analysis. Scientometrics 122, 1563–1581. https://doi.org/10.1007/s11192-020-03371-2.

Rauber, A., et al., 2016. Identification of reproducible subsets for data citation, sharing and re-use. Bull. IEEE Tech. Comm. Digit. Libr. Spec. Issue Data Citation 12 (1), 6–15. https://bulletin.jcdl.org/Bulletin/v12n1/.

RDA FAIR Data Maturity Model Working Group, 2020. FAIR Data Maturity Model: specification and guidelines. Res. Data Alliance. https://doi.org/10.15497/RDA00050.

Schimel, D., Keller, M., Berukoff, S., Kao, R., Loescher, H.W., Powell, H., Kampe, T., Moore, D., Gram, W., 2011. NEON Science Strategy; Enabling Continental-Scale Ecological Forecasting. Pub. NEON Inc., Boulder CO, p. 55.

Sporny, M., et al., 2020. JSON-LD 1.1 - "A JSON-based serialization for linked data". W3C Recomm. 16. World Wide Web Consortium. 6. https://www.w3.org/TR/json-ld/.

Squire, G., et al., 2018. Scientific software solution centre for discovering, sharing and reusing research software. In: American Geophysical Union Fall Meeting 2018. American Geophysical Union, Washington, D.C.. https://agu.confex.com/agu/fm18/meetingapp.cgi/Paper/459873

Stall, S., et al., 2019. Make scientific data FAIR. Nature 570 (7759), 27–29. https://doi.org/10.1038/d41586-019-01720-7.

Svensson, L.G., et al., 2019. Content negotiation by profile. In: W3C Working Draft. World Wide Web Consortium, 26 11. https://www.w3.org/TR/dx-prof-conneg/.

Tennison, J., 2016. CSV on the web: a primer. W3C Working Group Note. World Wide Web Consortium, 25 2. https://www.w3.org/TR/tabular-data-primer/.

Van de Sompel, H., Nelson, M.L., 2015. Reminiscing about 15 years of interoperability efforts. D-Lib Mag. 21 (11/12) https://doi.org/10.1045/november2015-vandesompel.

Voss, A., Procter, R., 2009. Virtual research environments in scholarly work and communications. Libr. Hi Tech 27 (2), 174–190. https://doi.org/10.1108/07377830910968146.

Weibel, S.L., Koch, T., 2000. The Dublin core metadata initiative. D-Lib Mag. 6 (12), 1082–9873. https://doi.org/10.1045/december2000-weibel.

Weigel, Tobias, et al., 2020. Making data and workflows findable for machines. Data Intell. 2 (1–2), 40–46. https://doi.org/10.1162/dint_a_00026.

Wickham, H., 2014. Tidy data. J. Stat. Softw. 59 (10) https://doi.org/10.18637/jss.v059.i10, 23 pp.

Wilkinson, M.D., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.

Wyborn, L.A., Friedrich, C., Rawling, T., Wu, M., Squire, G., Klump, J.F., Fraser, R., 2018. Building a multipurpose Geoscience Virtual Research Environment to cater for multiple use cases, a range of scales and diverse skill sets. In: American Geophysical Union, Fall Meeting 2018 abstract #IN33E-0887. https://ui.adsabs.harvard.edu/abs/2018AGUFMIN33E0887W/abstract.