

FAIR Research Objects for realising Open Science with the EOSC project RELIANCE

Anne Fouilloux[‡], Elisa Trasatti[§], Federica Foglini^{||}, Alejandro Coca-Castro[¶], Jean laquinta[#]

[‡] Simula Research Laboratory, Oslo, Norway

[§] Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

^{||} Institute of Marine Sciences, National Research Council, Venice, Italy

[¶] The Alan Turing Institute, London, United Kingdom

[#] Information Technology Department, University of Oslo, Oslo, Norway

Corresponding author: Anne Fouilloux (annef@simula.no)

Academic editor: Francisco Andres Rivera Quiroz

Abstract

The numerous benefits of Open Science (OS) and of the four FAIR foundational principles - Findable, Accessible, Interoperable and Reusable - are increasingly valued in academia, although what OS and FAIR entail is still largely misunderstood. In such conditions, putting into practice OS and applying the FAIR principles is challenging and underrated. However, realising OS is perfectly within our grasp provided that an infrastructure supporting the management of the research lifecycle is available. [ROHub \(https://www.rohub.org/\)](https://www.rohub.org/) is a Research Object (RO) management platform implementing three complementary technologies: Research Objects, Data Cubes and Text Mining services. ROHub enables researchers to collaboratively manage, share and preserve their research while they are still working on it (rather than after the work is finished). In this paper, three communities from Earth Sciences, namely Geohazards, Sea Monitoring and Climate Change, demonstrate how ROHub helped them to understand each other and to work openly and, more importantly, how communities of practice play an important role in facilitating reuse and interdisciplinary collaboration. These findings are illustrated with several use cases from these various communities.

Keywords

research object, reproducibility, replicability, reusability, interdisciplinary, open science practices, environmental sciences

Introduction and motivation

Open Science (OS) emphasises collaboration, transparency and sharing of ideas, data, software, workflows and methods (United Nations Educational, Scientific and Cultural

Organization 2021) for ongoing research work and not only at the end when publishing the final results. This approach can significantly speed-up the transfer of knowledge within academia or industry (but also from academia towards industry and vice versa) and, therefore, foster innovation at a more rapid rate. By embracing OS principles, organisations can accelerate innovation and create new knowledge, products, services and solutions that eventually benefit society as a whole. Therefore, OS and innovation strategy are closely linked. The FAIR (Findable, Accessible, Interoperable and Reusable) principles (Wilkinson et al. 2016) are often presented as necessary pre-requisites for realising OS. These principles aim to enable machine-readable data, to ensure that scientific data are easy to find, accessible to all, interoperable with other data sources and that it can be reused in different contexts. To go further and facilitate, in particular, cross-disciplinary research, rich metadata can be added to any data (datasets, software, workflows). This can be done either "manually" (following existing standards) or automatically (discovered, for instance, through a text mining service). In addition, communities of practice play an important role and are key to innovation. They can provide a safe environment for individuals having a common interest, to learn together, collaborate, share knowledge and agree on best practices for their communities. Therefore, applying OS principles not only relies on the use of open data and often, although not necessarily, open source software or tools, but also requires an agreement on the standards and common practices (data formats, coding norms, workflow management systems) to adopt and follow, as well as infrastructure enabling collaboration between practitioners and stakeholders. Sharing while doing is indeed much more challenging than offering open access to any final results. FAIR Digital Objects (FDOs) (Schultes and Wittenburg 2019) and/or FAIR Research Objects (ROs) are often referred to as a way to implement and realise OS. Research Objects are digital artifacts (Bechhofer et al. 2013) that "encapsulate all the components of a research project, including data, software, workflows and documentation into a single package which can be easily stored, shared, reused and reproduced". ROs aim at making research more transparent, reproducible and reusable, by providing means to package and share the various components of a research project in a standardised and machine-readable format.

This paper extends what was presented at the [1st international conference on FAIR digital Objects](#) (Fouilloux 2022, Fouilloux et al. 2022a). In the first part, we introduce ROHub (Garcia-Silva et al. 2019), a Research Object management platform which supports the preservation and lifecycle management of scientific investigations, research campaigns and operational processes. ROHub is described with an emphasis on the co-design with three Earth Science communities. This co-design work led to the definition of different types of ROs: bibliographical, data-centric, workflow-centric and executable ROs. After introducing ROHub, we present different use cases from these communities, highlighting the rationale behind the definition of different ROs. In particular, we discuss executable ROs in-depth by focusing on their metadata and ontologies that enable their re-execution (reproducibility and reusability) i.e. setting up the services and resources (computational environment, input data) for their reuse. Examples of executable ROs with Jupyter notebooks as the main resource are shown

and used to exemplify the need for community of practice to really enhance reusability. This leads to the definition of best practices for writing Jupyter notebooks that significantly improve reusability. Finally, we reflect on the experiences of evolving and reproducing Research Objects using the ROHub approach and what future work remains to be done to fully realise OS.

Method

The [RELIANCE project \(REsearch Lifecycle mAnagement for Earth Science Communities and Copernicus users in EOSC\)](#) delivers a suite of innovative and interconnected services that extend European Open Science Cloud (EOSC)'s capabilities to support the management of the research lifecycle within Earth Science Communities and Copernicus Users. The project has delivered three complementary technologies: Research Objects (ROs), Data Cubes and AI-based Text Mining.

[ROHub \(https://www.rohub.org/\)](https://www.rohub.org/) is a Research Object management platform that implements these three technologies: it has been developed to enable researchers to collaboratively manage, share and preserve their research work. ROHub implements the full RO model and paradigm: resources associated to a particular research work are aggregated into a single digital entity (the Research Object) and metadata relevant for understanding and interpreting the content is represented as semantic metadata that is user and machine readable.

By using ROHub, practitioners can ensure that their research work is well-organised and easily accessible to collaborators, while also being preserved for future use. The fact that ROHub is implementing the RO model and paradigm are especially significant, since this means that the platform is designed to meet the highest standards of data management and sharing. The use of contextual metadata is also a great feature, as it ensures that important contextual information about the research work is well preserved and can be easily understood by both humans and machines. Overall, ROHub is a valuable tool for anyone looking to improve data management, sharing practices and more generally working following Open Science principles.

RO-Crate

RO-Crate (Soiland-Reyes et al. 2022) is a community-driven initiative that provides a standard format to create and share research objects. RO-Crate stands for Research Object Crate, which is a lightweight and extensible metadata container format that enables the description of the components of a Research Object, including data, software, workflows and documentation and their inter-relationships. RO-Crate is based on schema.org annotations in JSON-LD and allows researchers to create a comprehensive description of their research project, which can be easily shared and reused by others.

RO-Crate enables the inclusion of additional metadata fields and the use of different metadata standards, depending on the requirements of the project. RO-Crate is also designed to be compatible with existing data and metadata standards, making it easy to integrate with repositories.

The benefits of using RO-Crate and Research Objects in general are, amongst others, increased transparency and reproducibility in research, improved data management and sharing and the ability to more easily reuse and build upon existing research. By providing a standardised format for creating and sharing research objects, RO-Crate can facilitate new collaborations, data reuse and knowledge discovery, leading to more efficient and effective scientific research practices.

RO-Crate enables a high degree of interoperability within ROHub. Nevertheless, various disciplines have evolved their own procedures and description standards and the concept of FAIR Digital Objects (FDOs) have emerged. FDOs are independent from the metadata descriptions, allowing them to include various description standards. RO-Crate can be seen as one possible implementation of FDOs if used along with FAIR signposting (Soiland-Reyes et al. 2022a), making ROHub a promising platform for FDOs.

Different types of Research Objects in ROHub for different purposes

An RO in ROHub commonly begins its life as an empty "Live RO". ROs aggregate new objects through their whole life-cycle (Belhajjame et al. 2012). It means that an RO in ROHub is filled incrementally by aggregating according to its typology new relevant resources, such as workflows, datasets, codes, documents that are being created, reused or repurposed. These resources can be added as internal or external (linked by reference) resources and can be modified at any time.

In ROHub, one can copy and keep ROs in time through snapshots which reflect their status at a given point in time: the "original" RO is still available and can continue to evolve. Snapshots can have their own Digital Object Identifiers (DOIs) which facilitate tracking the evolution of the research. Eventually, an RO in ROHub can be published and archived (so called "Archived RO") with a permanent identifier (DOI): it then becomes immutable. In ROHub, new Live ROs can be derived, based on an existing Archived RO, for instance, by forking it. Many ROs cited in this paper have not yet been archived because the associated research work is still on-going and not yet published: ROHub and ROs are supporting FAIR and Open Science practices.

To guide researchers, different types of Research Objects can be created from templates in ROHub:

- **Bibliography-centric:** includes manuals, anonymous interviews, publications, multimedia (video, audio) and/or other material that support research.
- **Data-centric:** refers to datasets which can be indexed, discovered and manipulated. Data cubes are particular data-centric ROs that can be discovered with data cube services such as the [ADAM platform](#) (Mantovani et al. 2020).

- **Executable:** includes the code, data and computational environment along with a description of the research object and, in some cases, a workflow. This type of ROs can be executed via specific services and is often used for scripts and/or Jupyter notebooks.
- **Software-centric:** also known as “Code as a Research Object”. Software-centric ROs include source codes and associated documentation. They often contain sample datasets for running tests.
- **Workflow-centric:** contains workflow specifications, provenance logs generated when executing the workflows, information about the evolution of the workflow (version) and its components/elements and additional annotations for the entire workflow.
- **Basic:** can contain anything and is used when the other types do not fully cover the creator’s need.

To facilitate the understanding and the reuse of the ROs in ROHub, each of these types of ROs (except Basic RO) has a template folder structure that we recommend researchers to select. For instance, an executable RO in ROHub has four folders:

- **“biblio”:** where researchers can aggregate documentations, scientific papers that support the development of the software/tool that is in the tool folder;
- **“input”:** where all the input datasets required for executing or reusing the RO are aggregated;
- **“output”:** where some or all the results generated by executing the RO are aggregated;
- **“tool”:** where the executable tool is aggregated. Typically, one aggregates a Jupyter notebook and/or executed workflows (Galaxy, Snakemake or Common Workflow Language workflows).

In addition to the different types of ROs and associated template structures, researchers can select the type of resources that constitutes the main entity of their RO: for instance, a Jupyter notebook can be selected as the main entity of an executable RO. As shown on Fig. 1, this additional metadata is then visible to everyone (and machine readable) to search and facilitate reuse.

The general overview of any type of Research Object is always the same, with mandatory metadata information such as the title, description, authors and collaborators, sketch (featured plots/images), the content of the RO (with different structures depending on the type of RO). Additional information is displayed on the right panel, such as number of downloads, additional discovered metadata (automatically extracted from the text content of ROs by the RELIANCE text enrichment services), free keywords (added by the end-users) and citation. Regarding the text mining feature, an additional tab called “Enrichment” has been added to provide more comprehensive information. This

additional feature has been requested by end-users. However, it is still under development and information presented is sometime difficult to grasp for newcomers, but it is nonetheless helpful for cross-disciplinary research. The *toolbox* and *share* sections allow end-users to download, snapshot and archive a given RO and/or share it. All the ROs in ROHub are digital objects that are FAIR and, for instance, findable in [Openaire explore](#), including Live ROs.

Use cases

The development of ROHub has been ongoing for several years (Palma et al. 2014, Garcia-Silva et al. 2019) with more recent work undertaken within the [RELIANCE project](#) where co-design has taken place and where it was validated through multidisciplinary and thematic real-life use cases provided by three Earth Science communities: Geohazards, Sea Monitoring and Climate Change communities. Below, we provide use cases for each type of RO. Each use case belongs to either one of the three communities or is interdisciplinary. The main objective of these use cases is to show the added value of ROs for researchers and not to focus on technicalities or concepts. These are selected for illustration only and some types of ROs (such as basic, bibliographical, executable or workflow ROs) could be used in slightly different manners, depending on local/institutional or community practices.

Basic RO to aggregate videos, presentations

Basic ROs are intended for selection when none of the other types of ROs in ROHub is fit for purpose or when a very small amount of resources are to be aggregated. One common usage of Basic ROs is for aggregating videos and presentations delivered during conferences, workshops or other events. For example, the basic RO "[AGU 2022 - Environmental Data Science Book: a community-driven resource showcasing open-source Environmental science](#)" (Book Community et al. 2022) contains a video as the main resource that has a permanent identifier and a few additional resources related to a talk given and recorded during the American Geophysical Union Fall meeting. The recorded talk contains metadata that increases its findability and the RO [is archived as RO-Crate in Zenodo](#) (Fouilloux et al. 2022), thereby providing much richer information (and metadata) than what is commonly recorded as a research outcome by researchers (title and conference abstract only).

Bibliographical RO to preserve reports

An example of Bibliography-centric RO is displayed in Fig. 2. Bibliographical ROs are not meant to replace existing reference managers (such as EndNote, Mendeley, Zotero etc.), but they are fully complementary. In this particular example, INGV (Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy) automatically creates ROs containing timely and up-to-date information about Volcano Supersites. These reports are usually sent to stakeholders whenever geohazard events occur, but keeping them in ROHub allows us

to better preserve this information and make it more widely available. At a later stage, one or several ROs could be created and would contain a collection of bibliographical ROs: for instance, all bibliographical ROs related to a specific supersite could be aggregated together into a new RO. These collections would be helpful for sharing historical information with stakeholders.

Data-centric RO to facilitate the use of open datasets

Data-centric ROs are used to create FAIR datacubes (Mantovani 2021) or to aggregate datasets generated or used by researchers. This type of RO is important because researchers do not always add rich metadata to their datasets which makes them more difficult (and less likely) to be reused.

By default, a data-centric RO would contain the following folders:

- **"biblio"**: papers or documentation that would support the usage/re-usage of the datasets aggregated in the RO;
- **"raw data"**: used if datasets aggregated in data were obtained by transforming raw data. This folder is not always used by researchers even when they have created datasets from other, existing datasets. There is a lack of common practices around data citation and, while researchers are very keen to make their data citable, they sometime omit to cite the original datasets (often large data providers);
- **"data"**: datasets are aggregated here and additional metadata can be added in the metadata folder (see below);
- **"metadata"**: additional metadata information can be added; for instance to help the re-use of the datasets. In practice, this folder is not often used and, when it is used, it mostly points to documentation (which could, for instance, be located in the biblio folder).

[EU FAR - EU Funds by Area Results](#) (Marin et al. 2022) is an example of a data-centric RO. In this RO, the "biblio" folder has not been populated, but two other folders called "presentations" and "maps" were created by the authors: they contain reports and presentations in .pdf format. What can or cannot be put into the "biblio" folder is not always clear and often depends on the scientific discipline and/or individual researcher's interpretation. This is where community of practice is becoming important, especially for cross-disciplinary research. This will be discussed later in the paper.

FAIR datacubes

In the [RELIANCE project](#), the concept of FAIR datacube (Mantovani 2021) has been introduced to facilitate reuse. It makes it possible to bring reusability one step further: by adding a datacube into an RO, users can directly browse and select datacubes available in the [RELIANCE ADAM platform](#) (Mantovani et al. 2020). Then the selected datacube contains rich metadata and users have the possibility to open it from ROHub to visualise data as shown on Fig. 4: users can discover the entire datacube collection (Palma 2023

); for example, select different dates, zoom-in and zoom-out on different geographical areas. Each datacube has a DOI which makes it easy for researchers to reuse and cite the exact dataset used.

FAIR datasets for everyone

The [RELIANCE ADAM platform](#) has been integrated in ROHub which simplifies the creation of datacubes in ROHub: all the metadata are automatically extracted and added to the RO. It is possible to add any types of datacubes in ROHub, but at the moment, all the necessary metadata would need to be created manually which makes it difficult for end-users. While this limitation could be lifted in the future, there would still be a need for users to create data-centric ROs with datasets they generated or derived from datacubes or other datasets and that may not be stored as datacubes (typically vector data).

As part of the RELIANCE project, a collaboration with the Norwegian Infrastructure for Research Data (NIRD) and the Polytechnic University of Madrid (UPM) has been established. UPM automatically created data-centric ROs from the datasets already stored in the NIRD archive (Shammas 2022). For instance, "[NorESM1-M CMIP5 historical \(3.2\) r2 raw output](#)" (Norwegian Climate Centre 2022) is a data-centric RO containing datasets generated by running the [Norwegian Earth System Model NorESM](#): simulations from researchers are usually hardly ever available and exploited, whereas they often contain datasets and information that could well be re-used. Most of the time, the same simulation is carried out again and again, because researchers are not aware of the existence of previous datasets and/or cannot find them easily. Data-centric ROs allow us to increase the visibility and FAIRness of datasets generated by scientists.

Software-centric RO to go beyond software releases

Software-centric ROs were initially created for researchers to share software, for instance, Python packages, such as the "Volcanic and Seismic source Modelling (VSM)" (Trasatti 2022). This type of RO is not used much in ROHub (only six such ROs were created): many researchers and Research Software Engineers are used to obtain a persistent identifier for their repository with Zenodo and do not find any added value there. However, for instance, for private repositories or containers, additional metadata can be added for each individual record in the RO and potentially this could increase the FAIRness of the software. The text enrichment service in ROHub is also a plus, compared to the current citation usage through Github and Zenodo: it can help to show users which software or other ROs are related to a given Software-centric RO. In future releases of ROHub, one would have to decide and eventually reduce the number of RO types. This is still under discussion within the Earth Science communities involved in the project.

Workflow-centric RO for a reproducible process

Workflow-centric ROs allow the storage and sharing of the "process" used by researchers. This can be either an automated workflow using a Workflow Management

System (Galaxy, Cylc, Snakemake, Nextflow etc.) or a simple script or text file detailing the list (and order) of tasks that need to be executed to reproduce the research results. For instance, **Galaxy** (The Galaxy Community et al. 2022) is an **open-source** platform for **FAIR data analysis** that enables users to create complex and fully reproducible workflows, either using a command line or through a Graphical User Interface in a web portal. The Galaxy workflow entitled "[5 years CLM-FATES simulation for Nordic site ALP1](#)" (Fouilloux 2021b) contains a Galaxy workflow that has been exported from Galaxy as a standalone Galaxy workflow text file as well as a link to a shared Galaxy history that illustrates the use of the workflow. This is very similar to [WorkflowHub](#) (Goble et al. 2021) with a Galaxy history where this workflow was executed with test data to exemplify it. Another relevant usage of workflow-centric RO is to use workflow-centric ROs to describe the protocol followed in the research process. This can be very relevant for experimental research. The workflow-centric RO "[Microplastics monitoring methodology in seawaters](#)" (Rapa et al. 2022) describes the research protocol as a sketch, which, of course, improves the reproducibility of the research work, but does not provide a fully automated workflow to re-run and regenerate the actual research outputs.

Executable RO to improve reusability

Executable ROs are very similar to workflow-centric ROs and, actually, many users consider them interchangeably. However, in that case, the workflow is executed on real datasets and not on a sample/test dataset; for example, the actual research outputs can be fully reproducible and reusable. In the section below, examples are provided for Galaxy workflows and interactive Jupyter notebooks. We then discuss the need for best practices when writing Jupyter notebooks to improve their re-usability beyond the state-of-the-art.

Reproducible Jupyter Notebook

Another very "common" usage of executable ROs in ROHub is for curating computational notebooks where the main resource is simply a Jupyter notebook. Such Jupyter notebooks are widespread in many scientific disciplines and, in particular, among Earth Scientists. JupyterHub and/or Binder are often used by researchers to highlight the reproducibility of their work or part of it. The [Binder Project](#) (Jupyter et al. 2018) is an open community making it possible to create sharable, interactive and reproducible environments. Public instances, such as [mybinder.org](#), provide very limited resources and can only be used to run very simple notebooks. EGI provides two Jupyter services: [E GI notebook](#), based on JupyterHub) and [EGI Binder](#). For open working, the EGI notebook is very useful because it allows the sharing of data and notebooks while working through a live executable RO. However, the actual computational environment cannot be easily selected (limited number of choices) and one has to (re)install packages on a regular basis. For customised computational environments, it is often preferable to select an EGI Binder because this allows the users to generate a bespoke computational environment associated with the Jupyter notebook. This is done similarly with mybinder.org, but the main advantage here is data can be shared with [EGI datahub](#) and directly accessed from

the Jupyter notebook (thus reducing the amount of data transferred) with larger compute and storage resources made available by EGI. In the future, the actual amount of compute and storage resources needed for reproducing a Jupyter notebook could be specified as metadata similarly to what is done for adding the computational environment. The only remaining issue with these services is that they are only accessible to European researchers and their collaborators which, in effect, narrows down Open Science to Europe.

From workflow to executable RO

Being able to re-execute a complex workflow is very important, for instance, to automate a repetitive pipeline relying on daily weather forecasts or as the basis for deriving new research work. The description of a workflow used in any standard Workflow Management System is often insufficient to understand how to reuse it. Examples and real-life use case workflow execution with inputs and the corresponding generated outputs, links to documentation, papers or tutorials are useful for end-users. An executable RO can be created to gather all the information related to the execution of a computational workflow. When using the Galaxy platform, Galaxy tools and workflows are fully annotated (Serrano-Solano et al. 2022): users normally share their Galaxy histories (and also reference them in their papers) and the workflows themselves can be stored in [WorkflowHub](#). WorkflowHub (Goble et al. 2021) is a registry to describe, share and publish computational workflows: [CWL](#), [RO-Crate](#), [Bioschemas](#) and [GA4GH's TRS API](#) are used in accordance with the **FAIR principles**. WorkflowHub supports most workflow types, including Galaxy, Snakemake and Nextflow workflows. ROHub allows us to aggregate workflows and to execute workflows; for example, the research datasets (inputs and outputs) and any other material relevant for understanding and reusing the corresponding research work. The executable RO titled "[Galaxy workflow and Galaxy histories for air quality analysis](#)" (Iaquinta and Fouilloux 2021) contains a [Galaxy workflow published in WorkflowHub](#) (Fouilloux 2021a) which has been executed in Galaxy: the Galaxy outputs are shared and links to each output were added in the output folder of this RO. This improves the FAIRness of Galaxy histories and this functionality has been added recently in Galaxy; for example, an RO-Crate can be automatically generated from a given Galaxy history (De Geest et al. 2022). In the future, ROHub could be connected to Galaxy and end-users could request to automatically create (and snapshot or archive) an RO in ROHub from a Galaxy history. Overall, a better integration between WorkflowHub and ROHub is desired as it would facilitate cross-disciplinary research.

Along the same lines, executable ROs can be used to exemplify the usage of a given tool: for instance, "[Galaxy CESM Tool Example](#)" (Fouilloux and Iaquinta 2022) shows how to run a climate model called the [Community Earth System Model \(CESM\)](#). This tool is very complex and customisable, with the possibility to define different climate scenarios and providing an example complements potential training material. Indeed, training materials are often short and cannot address all the different capabilities offered by a given tool. Then end-users can find it complex to go beyond these simple cases: the

Galaxy CESM Tool example runs one day of a fully-coupled climate model and is, therefore, not realistic for computing climate trends (given the small amount of computing resources that were available); however, it can be reused "as-is", for instance, to make much longer climate simulation (by changing the duration of the run).

Reuse it and go beyond the state of the art

The integration of EGI notebook and EGI Binder in ROHub significantly increases the re-usability of an executable RO, in particular, Jupyter notebooks. The executable RO "[Changes in air and water quality during the Covid-19 Lockdown in the Venice Lagoon](#)" (Fouilloux et al. 2023) illustrates the strength of ROHub and executable RO with Jupyter notebooks. This RO has been developed by scientists from the Sea Monitoring and Climate Change, based on examples provided by the technical team of the [RELIANCE project](#). The impact of the Covid-19 lockdown in the Venice Lagoon was first investigated from the perspective of the Sea Monitoring community: that led to a first RO named "[Snapshot 2021 study case: Lockdown impacts on the Northern Adriatic Sea at selected site: AcquaAlta Platform Water quality](#)" (Belgacem et al. 2021). In parallel, the Climate Change community investigated the same problem, but from the atmospheric perspective, for example, by investigating the "[Impact of the Covid-19 Lockdown on Air quality over Europe](#)" (Fouilloux et al. 2021): in this case, the study was expanded to Europe and air quality data from the Copernicus Atmosphere Monitoring Service for different cities was analysed to assess the impact of the lockdown on air quality.

Reusing Jupyter notebooks for cross-disciplinary research is often challenging, but this becomes much easier with ROHub. First, the text mining enrichment service can help users to find relevant ROs. Second, the integration with EGI notebook and EGI Binder allows users to replay Jupyter notebooks from ROs (one simply has to right-click on the Jupyter notebook resource to be re-directed to the EGI notebook or Binder service). By default, users are redirected to the EGI notebook service where the user can select one of the available computational environments to execute the notebook. However, if the EGI notebook has been upgraded after the notebook's creation, there is no guarantee that the execution will be successful. To improve the "long-term" reproducibility, users can associate a customised computational environment with the notebook: when the RO contains a computational environment (such as Pip's requirements.txt or Conda's environment.yml) that is linked to the notebook^{*1}, then EGI Binder is launched. The usage of [conda-lock](#) helps to create a more robust computational environment: different computational environments are created for different operating systems and the exact version used for each package is recorded. Users may still encounter issues when reusing "old" Jupyter notebooks and the usage of more long-term solutions — for instance, upgrading a Jupyter notebook to use newer versions of the required packages — is desirable, but out of scope.

Reproducibility is the first and necessary step to build beyond the state-of-the-art (as well as proper licences, such as MIT licences). Then both communities started to work together and investigated the creation of a combined use case where both point of views,

for example, atmospheric air quality and water quality would be investigated over the Venice Lagoon: a new notebook was then derived. All team members described this step as much smoother than usual, thanks to ROHub and its integration with EGI notebook. Furthermore, data were already shared from the two original ROs, therefore, downloading was not an issue either.

Community of practice with the Environmental Data Science Book

The previous example (Fouilloux et al. 2023) showed the strength of executable ROs including Jupyter notebooks. The example was relatively simple and, most importantly, all the initial partners and researchers were involved in the creation of the final executable RO. However, when working open, one aims at allowing anyone to derive new ROs (here Jupyter Notebook) without necessarily involving all the initial researchers (but still citing them, as ROHub offers with a fork mechanism). The ability to reuse a Jupyter notebook that has been created by others can be significantly enhanced if best practices are defined and adopted by the communities. This is the role of the community of practice.

[The Environmental Data Science Book](#) (EDS Book, Coca-Castro and Environmental Data Science Community (2022)) is a pan-european community-driven resource hosted on GitHub and powered by Jupyter Book. The resource leverages executable ROs in ROHub with Jupyter notebooks as the main resource, cloud resources and technical implementations of the FAIR principles to support the publication of datasets, innovative research and open-source tools in environmental science. The EDS book does not aim at replacing academic journals. It is a pedagogical opportunity maximising open infrastructure services to translate research outputs into curated, interactive, sharable and reproducible executable notebooks which benefit from a collaborative and transparent open review process. Building upon existing global open science communities, such as [the Turing Way](#) and [Pangeo](#), the EDS book provides clear guidelines for writing modular and reusable Jupyter notebooks, for submission and reviewing, templates for creating and scheduling notebooks using GitHub Actions Continuous Integration/Continuous Development (CI/CD) tools, FAIR practices through ROHub and [Binder](#) to facilitate fully documented, sharable and reproducible notebooks.

The quality of the published content is achieved by an open review policy supported by GitHub related technologies. Beyond the reproducibility that is ensured at the publication stage, the EDS book facilitates reuse. Let us take a popular notebook example from the EDS book: Fig. 3 summarises the overall use-case scenario: the executable RO "[Sea ice forecasting using IceNet \(Jupyter Notebook\)](#)" published [online](#) in the Environmental Data Science book (Coca-Castro et al. 2022a) is a Jupyter notebook reproducing the scientific results described in the corresponding publication (Andersson et al. 2021). The rendered version of the IceNet Jupyter notebook is made available in the EDS book which allows everyone to read it: some of the code cells are "hidden" (can be unwrapped by end-users while reading the Jupyter notebook) to highlight the most important sections of the Jupyter notebook, still keeping it fully reproducible. Fouilloux et al. (2022) shows a use-

case scenario where this RO has been re-used, for example, forked (Coca-Castro et al. 2022b) and modified with a new list of authors and the original authors as contributors. This clearly speeds up re-usage and creation of derivative work is facilitated because best practices (and light review process) were adopted "by design", for example, when creating the original Jupyter notebook.

Several of the ROs created and curated by the EDS book community have been reused. Overall, the feedback from the environmental science community is very positive; however, the need for understanding a specific programming language (Python, Julia, R) remains. This is clearly a barrier for inter-disciplinary research because researchers do not usually know many programming languages and each scientific discipline often makes use of a particular programming language. For instance, R is widely used amongst ecologists, whereas Python is not as well-known in that community. On the other hand, the situation is reversed for climate modellers. An idea that needs to be explored is the creation of "individual" modular containers, for example, canonical workflow building blocks (Soiland-Reyes et al. 2022a), for each section of a Jupyter notebook (for instance, download and preparation of input data, data analysis, visualisation) that could be incorporated in web portals, such as Galaxy (The Galaxy Community et al. 2022). The different tools could then be reused from a Graphical User Interface (and still from the command lines for those who are familiar with command lines) to create and execute fully annotated and reproducible workflows.

Discussion

While other platforms exist, such as [WorkflowHub](#), [Aperture Neuro](#) or [BioCompute Objects](#), none of them was meant to accommodate specific needs of the Earth-Science communities. At the beginning of the RELIANCE project, ROs were still mostly created when the work was finished, for example, to aggregate results produced within a research project and for publication purposes only, since some journal editors started to make it mandatory to provide supplementary material additions to published papers. Then, at best, having ROs when starting a project and/or reusing existing ROs to create derivative works was seen as "useful" by researchers. However, when ROHub began to integrate EOSC services, such as EGI datahub, EGI notebook or EGI Binder, ROs became "live FAIR digital objects" that evolve at the same pace as the research work and with little additional effort from researchers. Gradually, it became "convenient", since it was very straightforward to make data and documents available for co-workers with a single location (instead of having copies) and to share Jupyter notebooks (including not only the source code, but also outputs), so that they could get feedback on the implemented methods, interpretation of results, alternative approaches etc. There are still several features in ROHub that are not fully exploited. For instance, most ROs in ROHub have permalinks, but are never archived or snapshots created: ROs in ROHub are indexed on OpenAIRE and most end-users do not understand the potential advantages of archiving ROs and, more importantly, creating snapshots. More training

(self-paced material or online videos etc.) with concrete and real-life use cases exemplifying the advantages of each of the ROHub features would be helpful.

The text mining services also, as they improved over time, based on users feedback, now bring more information about the Research Objects, since they can access not only purely text documents (papers etc.), but also other metadata and what is a novelty: the source code itself within Jupyter notebooks. This makes it possible to discover ROs potentially relevant to researchers who would not have looked into them, based on "ordinary" keywords only. In addition, the derived semantic metadata can be used to deliver more accurate search results and content-based recommendations with so-called "[Collaboration Spheres](#)" (Rico et al. 2017) and/or "Similar ROs" where users can find other authors developing other ROs of interest, even if the title or original keywords that each of them used to describe their work had *a priori* nothing in common. The text mining service and its automatic metadata discovery are very promising to increase and facilitate inter-disciplinary research collaboration.

The number of ROs increases steadily with more than 3000 ROs and 150 users (March 2023): the vast majority of these ROs (about 2000) are bibliographical resources and basic ROs that contain reports, videos or other resources that would not be easily findable otherwise. Data-centric ROs are mostly datacubes which can be easily explained by the possibility to discover datacubes with the [ADAM platform](#): for data providers, this is clearly a way to advertise their data platform and track the usage of datasets. Collecting statistics and tracking reuse of data-centric ROs could be a way for data providers to optimise their platform and develop a more user-centric roadmap. Executable ROs are becoming more and more popular since the EGI notebooks and EGI Binder have been integrated into ROHub: these EOSC services seamlessly allow us to reproduce and reuse Jupyter notebooks that can require significant computational and storage resources.

Conclusions

ROHub has played a central role in the early adoption of the Open Science and FAIR principles by several Earth Sciences communities dealing with Geohazards, Sea Monitoring and Climate Change. It provided an easy-to-use and accessible infrastructure where different types of FAIR Research Objects could be created by scientists and shared with their colleagues or with the rest of the world. The way ROHub itself was used has significantly evolved between the beginning of the RELIANCE project towards its end. This demonstrated a change of mindset and the realisation that the products of research could be much more than mere communications and that collaborative work promotes creativity, innovation and cross-skilling (Open Science) that can significantly improve the quality of research outputs.

In the near future with more compute and storage resources made available (GPUs, HPCs etc.) and with, for instance, "collaborative" Jupyter notebooks (where several contributors will be able to work simultaneously on the same piece of code, as is already

done on text documents), exploiting platforms like ROHub will be a no-brainer to save time and energy from original ideas, to advance science, to involve more actors in the research process and/or exploitation of research products, all the while making clearly visible everybody's actual contributions. Once that is understood, researchers will be able to contribute more "casually" to the discussion on Open Science principles and how to apply these principles to their own discipline and in their respective communities. This is where community of practice comes into play and highlights the importance to have space and "venues" to discuss these best practices.

Acknowledgements

The RELIANCE (REsearch Lifecycle mAnagemeNt for Earth Science Communities and CopErnicus users in EOSC) project has received funding from the European Union's Horizon 2020 INFRAEOSC programme under grant agreement No 101017501. Alejandro Coca-Castro's work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/W006022/1, particularly the "Environment and Sustainability" theme within that grant and the Alan Turing Institute.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Andersson TR, Hosking JS, Pérez-Ortiz M, et al. (2021) Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nat Commun* 12: 5124. <https://doi.org/10.1038/s41467-021-25257-4>
- Bechhofer S, Buchan I, Roure DD, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C (2013) Why linked data is not enough for scientists. In this paper we make the case for a scientific data publication model on top of linked data and introduce the notion of Research Objects 29 <https://doi.org/10.1016/j.future.2011.08.004>.
- Belgacem M, Chiggiato J, Bastianini M (2021) Snapshot 2021 study case: Lockdown impacts on the Northern Adriatic Sea at selected site: AcquaAlta Platform Water quality. ROHub URL: <https://w3id.org/ro-id/0869e396-3733-4aff-8fb2-94c8937b28aa>
- Belhajjame K, Corcho O, Garijo D, et al. (2012) Workflow-centric research objects: First class citizens in scholarly discourse. In: *SePublica2012* (Ed.) Proc 2nd Work Semant Publ., 903. Second International Conference on the Future of Scholarly Communication and Scientific Publishing, 2012. URL: <http://ceur-ws.org/Vol-903/paper-01.pdf>
- Book Community EDS, Coca-Castro A, Iaquinta J, Andersson T, Barlow N, Hosking .S, Fouilloux A (2022) AGU 2022 - Environmental Data Science Book: a community-driven resource showcasing open-source Environmental science - archive. Simula Research Laboratory <https://doi.org/10.24424/ch0e-b129>

- Coca-Castro A, Environmental Data Science Community (2022) Environmental Data Science Book: A community-driven online resource to showcase and support a collaborative, reproducible and transparent Environmental Data Science. 0.0.1. Zenodo. Release date: 2022-1-29. URL: <https://doi.org/10.5281/zenodo.5918932>
- Coca-Castro A, Andersson T, Barlow N (2022a) Sea ice forecasting using IceNet (Jupyter Notebook) published in the Environmental Data Science book. ROHub URL: <https://w3id.org/ro-id/ac327c3a-5264-40a2-8c6e-1e8d7c4b37ef>
- Coca-Castro A, Fouilloux A, Iaquina J (2022b) Sea ice forecasting using IceNet (Jupyter Notebook) forked from the Environmental Data Science book. <https://w3id.org/ro-id/18269477-c1b8-4aa8-9b0e-372c7bb6b65c>. Accessed on: 2023-1-17.
- De Geest P, Coppens F, Soiland-Reyes S, Eguinoa I, Leo S (2022) Enhancing RDM in Galaxy by integrating RO-Crate. Research Ideas and Outcomes 8 <https://doi.org/10.3897/rio.8.e95164>
- Fouilloux A (2021a) *Investigation of lockdown effect on air quality between January 2019 to May 2021*. 1. WorkflowHub. Release date: 2021-12-20. URL: <https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.251.1>
- Fouilloux A (2021b) 5 years CLM-FATES simulation for Nordic site ALP1. ROHub URL: <https://w3id.org/ro-id/4086e5b7-284e-4551-a156-e3453ddcee58>
- Fouilloux A, Iaquina J, Mantovani S (2021) Impact of the Covid-19 Lockdown on Air quality over Europe. ROHub URL: <https://w3id.org/ro-id/53aa90bf-c593-4e6d-923f-d4711ac4b0e1>
- Fouilloux A (2022) FDO Conference 2022: FAIR Research Objects for realizing Open Science with RELIANCE EOSC project. URL: <https://doi.org/10.24424/nz65-v565>
- Fouilloux A, Iaquina J (2022) Galaxy CESM Tool Example. ROHub. Release date: 2022-6-12. URL: <https://w3id.org/ro-id/f99a5c78-6e3d-44a6-a283-3a61e46e249b>.
- Fouilloux A, Foglini F, Trasatti E (2022a) FAIR Research Objects for realizing Open Science with RELIANCE EOSC project. Research Ideas and Outcomes 8 <https://doi.org/10.3897/rio.8.e93940>
- Fouilloux A, Coca-Castro A, Iaquina J, Andersson T, Barlow N, Hosking S, Book Community, Environmental Data Science (2022b) AGU 2022 - Environmental Data Science Book: a community-driven resource showcasing open-source Environmental science - archive. Simula Research Laboratory <https://doi.org/10.24424/ch0e-b129>
- Fouilloux A, Foglini F, Castellani G, Belgacem M, Iaquina J, Mantovani S (2023) Changes in air and water quality during the Covid-19 Lockdown in the Venice Lagoon. 1.0. ROHub. Release date: 2023-1-08. URL: <https://w3id.org/ro-id/998dccc6-7192-4d88-af39-6018c71e6bdf>.
- Garcia-Silva A, Gomez-Perez JM, Palma R, Krystek M, Mantovani S, Foglini F, Grande V, Leo FD, Salvi S, Trasatti E, Romaniello V, Albani M, Silvagni C, Leone R, Marelli F, Albani S, Lazzarini M, Napier H, Glaves H, Aldridge T, Meertens C, Bolter F, Loescher H, Laney C, Genazzio M (2019) Enabling FAIR research in Earth Science through research objects. Enabling FAIR research in Earth Science through research objects, Future Generation Computer Systems 98 <https://doi.org/10.1016/j.future.2019.03.046>.
- Goble C, Soiland-Reyes S, Bacall F, Owen S, Williams A, Eguinoa I, Driesbeke B, Leo S, Pireddu L, Rodríguez-Navas L, Fernández JM, Capella-Gutiérrez S, Ménager H, Grüning B, Serrano-Solano B, Ewels P, Coppens F (2021) Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. In: Zenodo (Ed.) WorkflowHub. Zenodo <https://doi.org/10.5281/zenodo.4605654>

- laquinta J, Fouilloux A (2021) Galaxy workflow and Galaxy histories for air quality analysis. ROHub URL: <https://w3id.org/ro-id/67edd16f-c3d8-4879-a3a5-2d223e7dce6d>
- Jupyter P, Bussonnier M, Forde J, Freeman J, Granger B, Head T, Holdgraf C, Kelley K, Nalvarte G, Osheroff A, Pacer M, Panda Y, Perez F, Ragan-Kelley B, Willing C (2018) Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. Proceedings of the Python in Science Conference <https://doi.org/10.25080/majora-4af1f417-011>
- Mantovani S, Natali S, Folegani M, Cavicchi M, Barboni D, Ferraresi S (2020) The ADAM platform. EGU conference <https://doi.org/10.5194/egusphere-egu2020-17707>
- Mantovani S (2021) D4.4 FAIR Data Cubes -Release I. Zenodo <https://doi.org/10.5281/zenodo.5153210>
- Marin AM, Chis A, Bogdan C, Glăvan E, et al. (2022) EU FAR - EU Funds by Area Results. ROHub URL: <https://w3id.org/ro-id/941ecf90-82ba-4c56-961b-2f727da5df78>
- Norwegian Climate Centre (2022) NorESM1-M CMIP5 historical (3.2) r2 raw output. ROHub URL: <https://w3id.org/ro-id/707ef635-5f5a-448f-8ae4-1371fd367d77>
- Palma R, Holubowicz P, Corcho O, Gómez-Pérez JM, Mazurek C (2014) ROHub — A Digital Library of Research Objects Supporting Scientists Towards Reproducible Science. Communications in Computer and Information Science77-82. https://doi.org/10.1007/978-3-319-12024-9_9
- Palma R (2023) CAMS European air quality forecasts. URL: <https://w3id.org/ro-id/3939a208-64fc-4800-8b29-6a97676c7508/resources/bbb78b1a-20d6-47db-9e82-20d544bd809c>
- Rapa M, Cecchi T, Poletto D, Castellan G (2022) Microplastics monitoring methodology in seawaters. ROHub URL: <https://w3id.org/ro-id/a1fe8d87-7ca4-4846-ab84-869b9d8a2b57>
- Rico M, Gómez-Pérez JM, Gonzalez R, Garrido A, Corcho O (2017) Collaboration Spheres: a Visual Metaphor to Share and Reuse Research Objects. arXiv <https://doi.org/10.48550/arXiv.1710.05604>
- Schultes E, Wittenburg P (2019) FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. 20th International Conference, DAMDID/RCDL 2018. https://doi.org/10.1007/978-3-030-23584-0_1
- Serrano-Solano B, Fouilloux A, Eguinoa I, Kalaš M, Grüning B, Coppens F (2022) Galaxy: A Decade of Realising CWFR Concepts. Data Intelligence 4 (2): 358-371. https://doi.org/10.1162/dint_a_00136
- Shammass GH (2022) Massive ROs Creator. URL: <https://github.com/oeg-upm/Massive-ROs-Creator>
- Soiland-Reyes S, Bayarri G, Andrio P, Long R, Lowe D, Niewielska A, Hospital A, Groth P (2022a) Making Canonical Workflow Building Blocks Interoperable across Workflow Languages. Data Intelligence 4 (2): 342-357. https://doi.org/10.1162/dint_a_00135
- Soiland-Reyes S, Sefton P, Castro LJ, Coppens F, Garijo D, Leo S, Portier M, Groth P (2022b) Creating lightweight FAIR Digital Objects with RO-Crate. Research Ideas and Outcomes 8 <https://doi.org/10.3897/rio.8.e93937>
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022c) Packaging research artefacts with RO-Crate. Data Science 5 (2): 97-138. <https://doi.org/10.3233/ds-210053>
- The Galaxy Community, Afgan E, Nekrutenko A, Grüning BA, Blankenberg D, Goecks J, Schatz MC, Ostrovsky AE, Mahmoud A, Lonie AJ, Syme A, Fouilloux A, Bretaudeau A,

Nekrutenko A, Kumar A, Eschenlauer AC, DeSanto AD, Guerler A, Serrano-Solano B, Batut B, Grüning BA, Langhorst BW, Carr B, Raubenolt BA, Hyde CJ, Bromhead CJ, Barnett CB, Royaux C, Gallardo C, Blankenberg D, Fornika DJ, Baker D, Bouvier D, Clements D, de Lima Morais DA, Taberner DL, Lariviere D, Nasr E, Afgan E, Zambelli F, Heyl F, Psonopoulos F, Coppens F, Price GR, Cuccuru G, Corguillé GL, Von Kuster G, Akbulut GG, Rasche H, Hotz H, Eguinoa I, Makunin I, Ranawaka IJ, Taylor JP, Joshi J, Hillman-Jackson J, Goecks J, Chilton JM, Kamali K, Suderman K, Poterlowicz K, Yvan LB, Lopez-Delisle L, Sargent L, Bassetti ME, Tangaro MA, van den Beek M, Čech M, Bernt M, Fahrner M, Tekman M, Föll MC, Schatz MC, Crusoe MR, Roncoroni M, Kucher N, Coraor N, Stoler N, Rhodes N, Soranzo N, Pinter N, Goonasekera NA, Moreno PA, Videm P, Melanie P, Mandreoli P, Jagtap PD, Gu Q, Weber RJM, Lazarus R, Vorderman RHP, Hiltemann S, Golitsynskiy S, Garg S, Bray SA, Gladman SL, Leo S, Mehta SP, Griffin TJ, Jalili V, Yves V, Wen V, Nagampalli VK, Bacon WA, de Koning W, Maier W, Briggs PJ (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research* 50 (W1). <https://doi.org/10.1093/nar/gkac247>

- Trasatti E (2022) Volcanic and Seismic source Modelling (VSM). The Python toolkit for modelling geodetic data. ROHub <https://doi.org/10.24424/t83f-5t97>.
- United Nations Educational, Scientific and Cultural Organization (2021) UNESCO recommendation on open science. SC-PCB-SPP/2021/OS/UROS. URL: <https://identifiers.org/ark:/48223/pf0000379949>
- Wilkinson M, Dumontier M, Aalbersberg I, Appleton G (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3: 160018. <https://doi.org/10.1038/sdata.2016.18>

Endnotes

- *1 This requires the addition of metadata "Software Requirements" as well as the corresponding computational environment file to the notebook.

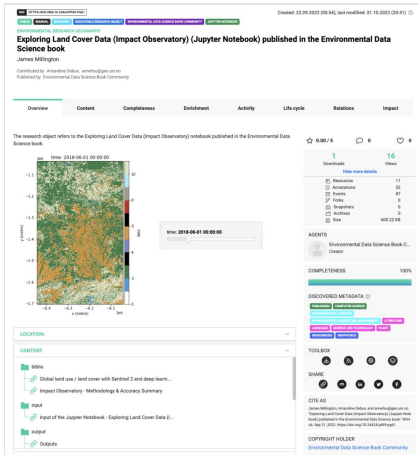


Figure 1.

Example of executable Research Object with a Jupyter notebook as a main resource. DOI: <https://doi.org/10.24424/pf69-pg61>.

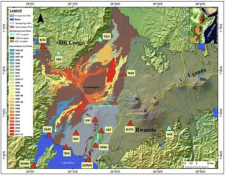
PUBLIC **MANUAL** **LIVE** **BIBLIOGRAPHIC-CONTENT RESEARCH OBJECT**

APPLIED SCIENCES
Virunga Volcanoes Supersite Biennial Report: 2020- 2021
 Elita Tresselt, Charles Bologni

Overview Content Completeness Enrichment Activity Life cycle Relations Impact

The Virunga was established in November 2017 as a permanent Supersite with the aim of improving the geographical scientific research and GeoHazards assessment in support of Disaster Risk Reduction (DQR) in the Virunga Volcanoes Province and the Lake Kivu basin. On the evening of May 23rd 2021, Nyiragongo volcano suddenly erupted from three vents that opened along a system of fractures on the northern flank of the volcano. Two major lava flows were produced with one having its direction toward Goma city. An intense seismic activity followed the eruption and persisted for ~2 weeks, and resumed later.

Show more >



24 Downloads 22 Views

Hide stats details

- Resources 3
- Annotations 32
- Events 58
- Forks 0
- Snapshots 0
- Addresses 0
- Size 1003.27 KB

AGENTS

- Elita Tresselt
- Creator

COMPLETENESS 76%

TOOLBOX

SHARE

CITE AS
 Tresselt, Elita, and Charles Bologni. "Virunga Volcanoes Supersite Biennial Report: 2020-2021". *Virunga*. Mar 09, 2022. <https://w3id.org/ro-id/45841548-0362-4aea-80f2-ea71d81a691f>.

LOCATION

CONTENT

- bbis
- Virunga Supersite website
- Virunga Volcanoes Supersite Biennial Report: 2020-2021
- Geological map of the Virunga area (Dem. Rep. Congol.) (19066)

Figure 2.

Bibliographical Research Object entitled "Virunga Volcanoes Supersite Biennial Report: 2020- 2021" and containing detailed report by INGV from the Virunga Volcano Supersite. This RO has a permanent identifier: <https://w3id.org/ro-id/45841548-0362-4aea-80f2-ea71d81a691f>.

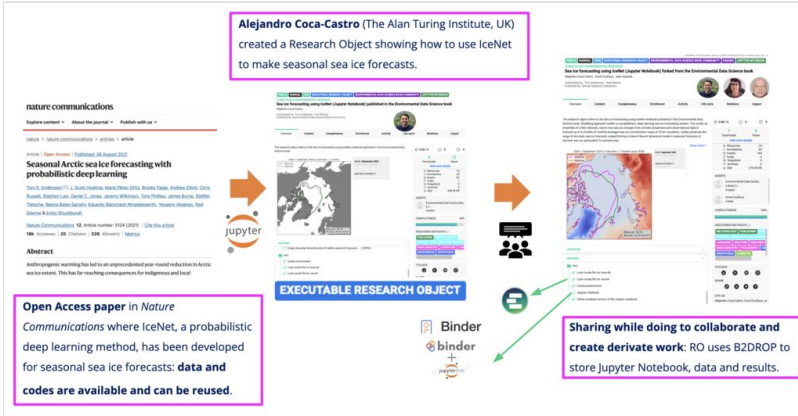


Figure 3.

Use case scenario: an executable RO (Coca-Castro et al. 2022a) created from an Open Access paper (Andersson et al. 2021) is reused to create derivative work in a collaborative way thanks to EOSC services, such as EGI datahub, EGI notebooks, EGI Binder and ROHub.

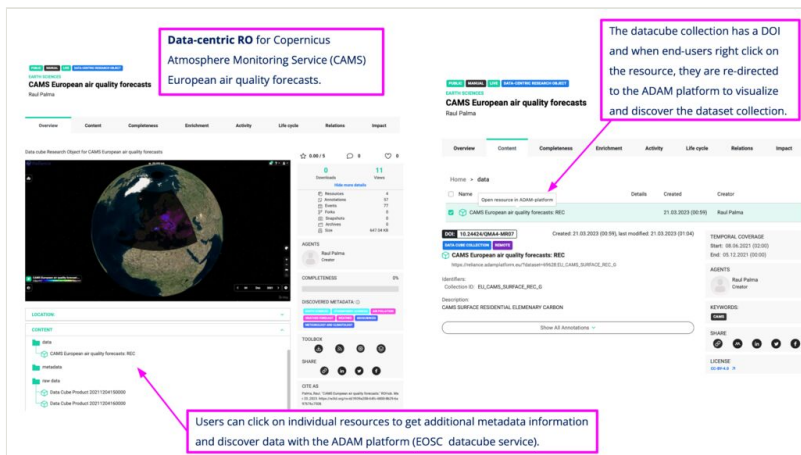


Figure 4.

Data-centric Research Object with datacube collection from the Copernicus Atmosphere Monitoring Service (CAMS) European air quality forecasts (Palma 2023). The figure on the left shows the [data-centric RO](#) and that on the right an example of [datacube discovery with the ADAM platform](#) (to visualise the datacube, users need to register and login to the ADAM platform).