DOI: https://doi.org/10.3897/arphapreprints.e108817

FAIR Research Objects for realizing Open Science with RELIANCE EOSC project

Anne Fouilloux, Elisa Trasatti, Federica Foglini, Alejandro Coca-Castro, Jean Iaquinta

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 30/06/2023*. DOI: https://doi.org/10.3897/arphapreprints.e108817

FAIR Research Objects for realizing Open Science with RELIANCE EOSC project

Anne Fouilloux[‡], Elisa Trasatti[§], Federica Foglini[|], Alejandro Coca-Castro[¶], Jean Iaquinta[#]

‡ 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)

Reviewable v 1

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 in practice OS and applying the FAIR principles is challenging and underrated. However, realising OS is perfectly within grasp provided that an infrastructure supporting the management of the research lifecycle is available. RoHub is precisely a Research Object (RO) management platform implementing three complementary technologies: Research Objects, Data Cubes and Artificial Intelligence-based 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, organizations 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 known as necessary pre-requisites for realizing OS. These principles aim to ensure that scientific data is 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, open source software and 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 infrastructures 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 (DOs) and/or FAIR Research Objects (ROs) are often referred to as a way to implement and realize 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 standardized and machine-readable format.

In the first part of this paper, 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: Geohazards, Sea Monitoring and Climate Change. 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 rational 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 examplify the need for community of practice to really enhance reusability. The "Environmental Data Science Book" initiative (see https://edsbook.org) is then described. This leads to the definition of best practices for writing Jupyter notebooks that significantly improve reusability. Finally, we conclude with highlights of strengths and weaknesses of ROs as FAIR Digital Objects, and what future work remains to be done for fully realize OS.

Method

RoHub (https://reliance.rohub.org) is a Research Object management platform 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 FAIR Digital 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-organized 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 (Sefton et al. 2022, Wittner et al. 2023, Carragáin et al. 2019) 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, including Dublin Core, DataCite, and Schema.org, making it easy to integrate with existing data and metadata repositories.

The benefits of using RO-crate and research objects are, among 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 standardized 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 independant to the metadata descriptions, allowing them to include various description standards. ROs could be easily converted into FDOs.

Different types of Research Objects for different purposes

A RO commonly begins its life as an empty "Live RO". ROs aggregate new objects through their whole life-cycle. It means that a RO 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 added as internal or external (linked by reference) resources and can be modified at any time.

One can copy and keep ROs in time through snapshots which reflect their status at a given point in time. Snapshots can have their own Digital Object Identifiers (DOIs) which facilitates tracking the evolution of the research. Eventually, a RO can be published and archived (so called "Archived RO") with a permanent identifier of its own (DOI): it then becomes immutable. New Live ROs can be derived based on an existing Archived RO, for instance by forking it.

To guide researchers, different types of Research Objects can be created:

- **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*[2].
- **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 workflow as a whole.
- **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, each type of RO (except Basic RO) has a template folder structure that we recommend researchers to select. For instance an executable RO 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 Objects is always the same, with mandatory metadata information such as the title, description, authors & collaborators, sketch (featured plots/images), the content of the RO (with different structures depending on the type of ROs). 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 ROs in RoHub are digital objects that are FAIR, and for instance findable in Openaire explore (https://explore.openaire.eu/), including Live ROs.

Use cases

The development of RoHub is the fruit of co-design, and 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 examples 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 meant to be selected when none of the other types of ROs 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 main resource that has permanent identifier*[3] 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 can be archived as RO-Crate in Zenodo (see Fouilloux et al. 2022) which provides 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 on 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 Volcanoes' Supersites. These reports are usually sent to stakeholders whenever geohazard events occur, but keeping them in RoHub allows 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 skakeholders.

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 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 in 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 discuss 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 a RO, users can direclty 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 visualize data as shown on Fig. 4: users can discover the entire datacube collection e.g. 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

At the moment, the concept of FAIR datacubes in RoHub is limited to datacubes available in the RELIANCE ADAM platform. While this limitation could be lifted in the future, there would still be a need for users to create data-centric RO with datasets they generated or derived from datacubes or other datasets.

As part of the RELIANCE project, a collaboration with the Norwegian Infrastructure for Research Data (NIRD) and the Polytechnic University of Madrid (UPM) was established. UPM automatically created data-centric ROs from the datasets already stored in the NIRD archive. 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 NorESM1*[4]: simulations from researchers are usually very little visible 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 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 get 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 ROs types. This is still under discussion within the Earth Science communities involved in the project.

Workflow-centric RO for reproducible process

Workflow-centric ROs allow to store and share 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 "**5 years CLM-FATES simulation for Nordic site ALP1**" (Fouilloux 2021b) contains a Galaxy workflow 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) (https://workflowhub.eu/) with a Galaxy history where this workflow was executed with test data to examplify 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 rerun 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 e.g. 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 for writing Jupyter notebooks to improve the re-usability beyond the state of the art.

From workflow to executable RO

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 command line or through a Graphical User Interface in a web portal. 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 WorfklowHub (https://workflowhub.eu). 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 to aggregate workflows, executed workflows e.g. 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 (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 in the future this functionality will be made available from Galaxy e.g. a RO-Crate will be automatically generated from a given Galaxy history.

Along the same lines, executable ROs can be used to examplify 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*[5]). This tool is very complex and customizable, 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 possibilities of a given tool. Then end-users can find it complex to go beyond these simple cases: the example runs one day of a fully-coupled climate model and is therefore not realistic for computing climate trends, however it can be reused as-in, for instance to make much longer climate simulation (by changing the duration of the run).

Reproducible Jupyter notebook

Another very popular 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*[6] is an open community making it possible to create sharable, interactive and reproducible environments. Public instances such as mybinder.org (https://mybinder.org/) provide very limited resources and can only be used to run very simple notebooks. EGI provides two Jupyter services: EGI notebook*[7] (based on JupyterHub) and EGI Binder*[8]. For working open, EGI notebook is very useful because it allows to share 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 customized 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*[9] and directly accessed from the Jupyter notebook (thus reducing the amount of data transferred) with larger compute and storage resources available by EGI. In the future, the actual amount of compute and storage resources needed for reproducing a Jupyter notebook could be added as metadata similarly to what is done for adding the computational environment. The only remaining issue with these services is that it is only accessible to European researchers and their collaborators which in effect narrows down Open Science to Europe.

Reuse it and go beyond the state of the art

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 RELIANCE technical team. 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 atmosphere perspective e.g. 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 Copernicus Atmosphere Monitoring Service for different cities was analyzed 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 EGI notebook service but if the RO contains a computational environment (such as requirements.txt or environment.yaml) that is linked to the notebook (this requires to add metadata "Software Requirements" as well as the corresponding computational environment file to the notebook), then EGI Binder will be launched.

Reproducibility is the first and necessary step to build beyond the state-of-the-art (as well as proper licenses such as MIT licenses). Then both communities started to work together and investigated the creation of a combined use case where both point of views e.g. atmosphere 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. Futhermore, data was 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 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 necessary involving all the initial researchers (but still citing them, as RoHub offers fork mechanism). The ability to reuse a Jupyter notebook that has been created by others can be significantly enhanced if best practices were 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 maximizing open infrastructure services to translate research outputs into curated, interactive, shareable 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 (https://the-turing-way.netlify.app/) and Pangeo (https://pangeo.io/), 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 CI/CD tools, FAIR practices through RoHub (https://reliance.rohub.org/) and Binder (https://mybinder.org/) to facilitate fully documented, shareable 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's take a popular notebook example from the EDS book: Fig. 3 summarizes the overall use-case scenario: the executable RO "**Sea ice forecasting using IceNet (Jupyter Notebook) published in the Environmental Data Science book***[1]" (Coca-Castro et al. 2022a) is a Jupyter notebook reproducing the scientific results of the Nature Communications paper titled "**Seasonal Arctic sea ice forecasting with probabilistic deep learning (https://www.nature.com/articles/s41467-021-25257-4)**" (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 e.g. 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" e.g., when creating the original Jupyter notebook.

Several of ROs created and curated by EDS book community have been reused. Overall 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 among ecologists but Python is vastly unknown while it is the other way around for climate modellers. An idea that needs to be explored is the creation of "individual" modular containers for each section of a Jupyter notebook (for instance, download and preparation of input data, data analysis, visualization) 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 were meant to accomodate specific needs of the Earth-Science communities. At the beginning of the RELIANCE project, ROs were still mostly created when the work was finished, e.g. 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. But, 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.

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 AI derived semantic metadata can be used to deliver more accurate search results and content-based recommendations with so-called "Collaboration Spheres*[10]" 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 is 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 optimize 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 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 realization 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, like 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 or and exploitation of research products, all the while making it clearly visible everybody's actual contributions. Once that 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 came into play and highlight the importance to have space and "venues" to discuss on 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 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 & 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
- 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
- Carragáin EÓ, Goble C, Sefton P, Soiland-Reyes S (2019) A lightweight approach to research object data packaging. A lightweight approach to research object data packaging. Bioinformatics Open Source Conference (BOSC2019). . Zenodo https://doi.org/10.5281/zenodo.3250687
- 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, Iaquinta 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.
- 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, Iaquinta 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, Iaquinta 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, Coca-Castro A, Iaquinta J, Andersson T, Barlow N, Hosking S, Book Community, Environmental Data Science (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
- Fouilloux A, Foglini F, Castellan G, Belgacem M, Iaquinta 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/998dccd6-7192-4d88-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, Boler 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, Droesbeke B, Leo S, Pireddu L, Rodríguez-Navas L, Fernández J, Capella-Gutierrez S, ndez SC, Ménager H, Grüning B, Serrano-Solano B, Ewels P, Coppens F, et al. (2021) Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. In: Zenodo (Ed.) WokflowHub. Zenod https://doi.org/10.5281/zenodo.4605654
- Iaquinta 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
- Mantovani S, Natali S, Folegani M, Cavicchi M, Barboni D, Ferraresi S (2020) The ADAM platform. EGU conferenc 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.
- 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
- Sefton P, Ó Carragáin E, Soiland-Reyes S, Corcho O, Garijo D, Palma R, Coppens F, Goble C, Fernández J, Chard K, Gomez-Perez JM, Crusoe M, Eguinoa I, Juty N,

Holmes K, Clark J, Capella-Gutierrez S, Gray AG, Owen S, Williams A, Tartari G, Bacall F, Thelen T, Ménager H, Rodríguez-Navas L, Walk P, whitehead b, Wilkinson M, Groth P, Bremer E, Castro LJ, Sebby K, Kanitz A, Trisovic A, Kennedy G, Graves M, Koehorst J, Leo S, Portier M, Brack P, Ojsteršek M, Droesbeke B, Niu C, Tanabe K, Miksa T, La Rosa M, Decruw C, Czerniak A, Jay J, Serra S, Siebes R, de Witt S, El Damaty S, Lowe D, Li X, Gundersen S, Radifar M (2022) RO-Crate Metadata Specification 1.1.2. Zenodo https://doi.org/10.5281/zenodo.3406497

- 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
- 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, Tabernero DL, Lariviere D, Nasr E, Afgan E, Zambelli F, Heyl F, Psomopoulos 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://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en
- 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
- Wittner R, Gallo M, Leo S, Soiland-Reyes S (2023) Packing provenance using CPM RO-Crate profile. Zenodo https://doi.org/10.5281/zenodo.7676923

Endnotes

*1 https://edsbook.org/gallery/modelling/polar-modelling-icenet/polar-modelling-icenet.html

*2 https://reliance.adamplatform.eu

*3 https://w3id.org/ro-id/b167006f-4678-4b8a-8e8b-224befdd8038/resources/54d872a8-14ef-4291-ba35-8a7393004530

*4 https://noresm-docs.readthedocs.io/en/latest/

*5 https://www.cesm.ucar.edu/

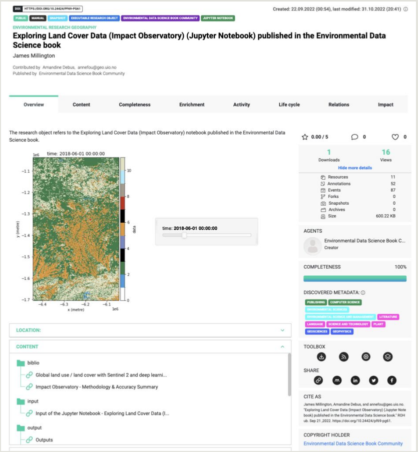
*6 https://jupyter.org/binder

*7 https://notebooks.egi.eu/hub/welcome

*8 https://replay.notebooks.egi.eu/

*9 https://www.egi.eu/service/datahub/

*10 https://reliance.expertcustomers.ai/spheres/index.html



DOI: HTTPS://DOI.ORG/10.24424/PF69-PG61

Created: 22.09.2022 (00:54), last modified: 31.10.2022 (20:41)

PUBLIC MANUAL SNAPSHOT EXECUTABLE RESEARCH OBJECT ENVIRONMENTAL DATA SCIENCE BOOK COMMUNITY JUPYTER NOTEBOOK

ENVIRONMENTAL RESEARCH GEOGRAPHY

Exploring Land Cover Data (Impact Observatory) (Jupyter Notebook) published in the Environmental Data Science book

James Millington

Contributed by Amandine Debus, annefou@geo.uio.no
Published by Environmental Data Science Book Community

Overview Content Completeness Enrichment Activity Life cycle Relations Impact

The research object refers to the Exploring Land Cover Data (Impact Observatory) notebook published in the Environmental Data Science book.

time: **2018-06-01 00:00:00**

LOCATION:

CONTENT

- biblio
  - Global land use / land cover with Sentinel 2 and deep learni...
  - Impact Observatory - Methodology & Accuracy Summary
- input
  - Input of the Jupyter Notebook - Exploring Land Cover Data (I...
- output
  - Outputs

☆ 0.00 / 5 0 ♡ 0

1	16
Downloads	Views

Hide more details

Resources	11
Annotations	52
Events	87
Forks	0
Snapshots	0
Archives	0
Size	600.22 KB

AGENTS

Environmental Data Science Book C...
Creator

COMPLETENESS 100%

DISCOVERED METADATA:

PUBLISHING COMPUTER SCIENCE ENVIRONMENTAL SCIENCES ENVIRONMENTAL SCIENCE AND MANAGEMENT LITERATURE LANGUAGE SCIENCE AND TECHNOLOGY PLANT GEOSCIENCES GEOPHYSICS

TOOLBOX

SHARE

CITE AS

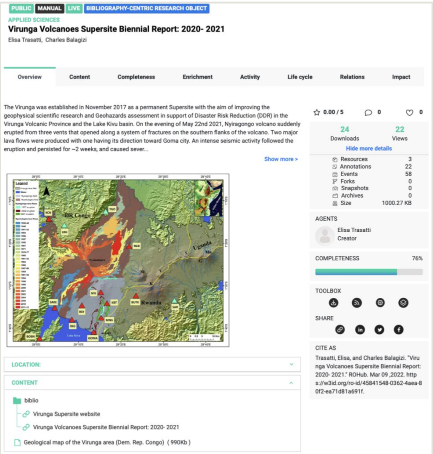
James Millington, Amandine Debus, and annefou@geo.uio.no. "Exploring Land Cover Data (Impact Observatory) (Jupyter Notebook) published in the Environmental Data Science book." ROHub. Sep 21 ,2022. https://doi.org/10.24424/pf69-pg61.

COPYRIGHT HOLDER

Environmental Data Science Book Community

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 BIBLIOGRAPHY-CENTRIC RESEARCH OBJECT

APPLIED SCIENCES

Virunga Volcanoes Supersite Biennial Report: 2020- 2021

Elisa Trasatti, Charles Balagizi

The Virunga was established in November 2017 as a permanent Supersite with the aim of improving the geophysical scientific research and Geohazards assessment in support of Disaster Risk Reduction (DDR) in the Virunga Volcanic Province and the Lake Kivu basin. On the evening of May 22nd 2021, Nyiragongo volcano suddenly erupted from three vents that opened along a system of fractures on the southern flanks 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 caused sever...

☆ 0.00 / 5 · 0 · ♡ 0

24	22
Downloads	Views

Resources	3
Annotations	22
Events	58
Forks	0
Snapshots	0
Archives	0
Size	1000.27 KB

AGENTS

Elisa Trasatti
Creator

COMPLETENESS 76%

TOOLBOX

SHARE

CITE AS

Trasatti, Elisa, and Charles Balagizi. "Virunga Volcanoes Supersite Biennial Report: 2020- 2021." ROHub. Mar 09 ,2022. https://w3id.org/ro-id/45841548-0362-4aea-80f2-ea71d81a691f.

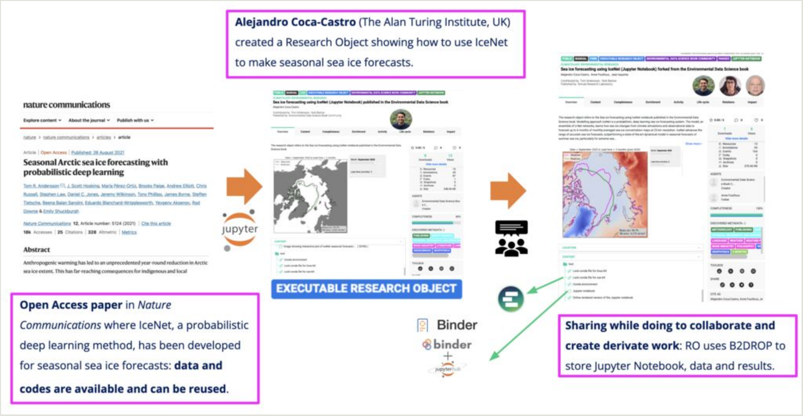
LOCATION:

CONTENT

- biblio
  - Virunga Supersite website
  - Virunga Volcanoes Supersite Biennial Report: 2020- 2021
- Geological map of the Virunga area (Dem. Rep. Congo) (990Kb)

Figure 2.

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



Alejandro Coca-Castro (The Alan Turing Institute, UK) created a Research Object showing how to use IceNet to make seasonal sea ice forecasts.

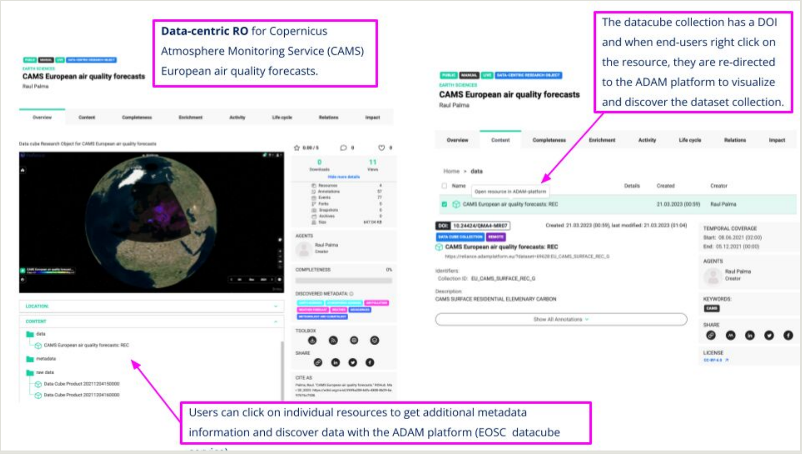
Open Access paper in *Nature Communications* where IceNet, a probabilistic deep learning method, has been developed for seasonal sea ice forecasts: **data and codes are available and can be reused.**

Sharing while doing to collaborate and create derivate work: RO uses B2DROP to store Jupyter Notebook, data and results.

nature communications
Seasonal Arctic sea ice forecasting with probabilistic deep learning
Abstract
jupyter
EXECUTABLE RESEARCH OBJECT
Binder
binder
jupyterhub

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.



EARTH SCIENCES
CAMS European air quality forecasts
Raul Palma
Overview
Content
Completeness
Enrichment
Activity
Life cycle
Relations
Impact
Data cube Research Object for CAMS European air quality forecasts
0.00 / 5
0
Downloads
11
Views
Hide more details
Resources 4
AGENTS
Raul Palma
Creator
COMPLETENESS 0%
DISCOVERED METADATA
TOOLBOX
SHARE
CITE AS
LOCATION
CONTENT
data
CAMS European air quality forecasts: REC
metadata
raw data
Data Cube Product 20211204150000
Data Cube Product 20211204160000
Home > data
Name
Details
Created
Creator
Open resource in ADAM platform
CAMS European air quality forecasts: REC
21.03.2023 (00:59)
Raul Palma
DOI 10.24424/QMA4-MR07
Created: 21.03.2023 (00:59), last modified: 21.03.2023 (01:04)
CAMS European air quality forecasts: REC
Identifiers
Collection ID: EU_CAMS_SURFACE_REC_G
Description
CAMS SURFACE RESIDENTIAL ELEMENARY CARBON
Show All Annotations
TEMPORAL COVERAGE
Start: 08.06.2021 (02:00)
End: 05.12.2021 (00:00)
AGENTS
Raul Palma
Creator
KEYWORDS
CAMS
SHARE
LICENSE

Data-centric RO for Copernicus Atmosphere Monitoring Service (CAMS) European air quality forecasts.

The datacube collection has a DOI and when end-users right click on the resource, they are re-directed to the ADAM platform to visualize and discover the dataset collection.

Users can click on individual resources to get additional metadata information and discover data with the ADAM platform (EOSC datacube

Figure 4.

Data-centric Research Object with datacube collection from the Copernicus Atmosphere Monitoring Service (CAMS) European air quality forecasts. The figure on the left shows the data-centric RO and that on the right an example of datacube discovery with the ADAM platform.