A Research Object-based Toolkit to Support the Earth Science Research Lifecycle

Raul Palma
Poznan Supercomputing and Networking Center
Poznan, Poland
rpalma@man.poznan.pl

Andres Garcia-Silva
Expert System
Madrid, Spain
agarcia@expertsystem.com

Jose Manuel Gomez-Perez
Expert System
Madrid, Spain
jmgomez@expertsystem.com

Marcyn Krystek
Poznan Supercomputing and Networking Center
Poznan, Poland
mkrystek@man.poznan.pl

***Abstract*—Data-intensive science disciplines, like Earth Science, are increasingly producing and consuming a variety of digital resources during the course of a scientific investigation. Instead of having these resources in isolated repositories, scientists are seeking ways for managing and making these resources available from a single place, and at the same time they are also increasingly interested in the adoption of FAIR principles to enhance the visibility and reusability of scientific results. This has called for new methods to improve the access and communication of results. Research Objects are a key building block towards realising this vision. They provide a structured way (a model) to describe scientific resources related to an investigation, along with the context in which they were used and the people involved. But research objects are as useful in practice as the availability of tools supporting their adoption. In this paper, we present a toolkit, tailored for Earth Sciences, comprising a set of services and applications around research objects that support scientists throughout the research lifecycle to manage, share, find and reuse scientific results, and we discuss initial insights into the community adoption.**

Keywords—Research Objects, Knowledge Sharing and Reuse, Earth Science, ROHub

I. Introduction

A research lifecycle specifies a series of sequentially related stages or phases in which information, data and methods are produced or manipulated during the course of a research process [1], and they are usually tailored to specific community needs. In Earth Science, for example, the research and information lifecycle involves tasks like: access to data (e.g., raw data and/or a variety of added value products); sharing results (with colleagues and/or community); execution of data analytic methods and generation of models; validation and dissemination of findings; and collaboration with colleagues [2]. Throughout this process, earth scientists are increasingly producing and consuming a variety of digital resources, as is the case in many other data-intensive science disciplines. For example, they need to work with heterogeneous datasets generated by data providers such as space agencies, specialized organizations and research projects that produce earth observation data.

As a result, these communities, together with a diverse group of stakeholders from academia, industry, funding agencies and publishers are calling for innovative ways to manage their scientific resources to enhance the visibility of research results, encourage reuse, and foster a broader accessibility [3]. Instead of having their resources in isolated repositories, scientists need to systematically capture the lifecycle of scientific investigations and provide a unified entry point to information about the hypotheses investigated, the data consumed and produced during experimentation or observation, the computations carried out, the conclusions that were derived, the researchers involved in the investigation, and the different licensing models over data or software, to name but a few factors. But at the same time such information units should also include a description of the underlying context and relations between these resources, in order to foster the reusability of results. Moreover, scientists should get appropriate credit over such units to encourage sharing and publication.

Research Objects (ROs) enable such a vision, and have the potential to accelerate science and stimulate the uptake of good practices in data-intensive science. An RO [4] is a semantically enriched information unit that encapsulates all the materials and methods relevant to a scientific investigation, the associated annotations and the context where such resources were used and produced. ROs address technical challenges like preservation, reproducibility, and interoperability, and include metadata that make them uniquely identifiable, processable, and machine readable. They encourage the release of scientific resources in addition to text publication, in the sense that data, methods and software can be encapsulated as a citable unit. Thus, ROs also address some of the social aspects in the scientific enterprise [5], by fostering author accreditation of their respective contributions, enabling discussion around the investigation, and ultimately supporting collaboration. As a consequence, ROs are particularly fitted to support FAIR principles [6], a concise and measurable set of guidelines to enhance data reusability, with focus on enhancing the ability of machines to automatically find and use data.

Nevertheless, ROs are as useful in practice as the availability of tools supporting their adoption. With this in mind, we have built an RO-based toolkit, comprising a set of services and applications that support scientists throughout

978-1-5386-9156-4/18/$31.00 ©2018 IEEE
DOI 10.1109/eScience.2018.00020

the research lifecycle to manage, share, find and reuse scientific results. The toolkit has been particularly tailored to support researchers in Earth Science, and is currently being tested and used by different communities. Note that there are some existing tools that leverage the RO paradigm to some extent, which are available for developers and scientists. For example, there are libraries that help developers to generate RO bundles[1] in different programming languages[2], and tools like Latex2RO[3] to create research objects from LaTex papers. In addition, platforms like myexperiment.org and seek4science.org can be used by scientists to share and reuse scientific workflows, datasets, models and simulations. Nevertheless, none of these tools implement the full RO model and specifications, neither they support the specific needs of observational disciplines including Earth Science.

The rest of this paper is structured as follows: section II introduces the toolkit, section III presents extensions in the RO model for Earth Science, section IV introduces the individual services and applications, section V discusses initial insights about the community adoption, and section VI highlights the conclusions of this work.

II. Research Object Toolkit for Earth Science

A high-level view of the components comprising the toolkit is depicted in Fig. 1. The core of the toolkit is the RO model (described in Section III), which provides an agreed upon vocabulary with formal semantics for sharing scientific outcomes that are interoperable and machine-readable. Around the RO model the inner circle depicts the features required and the outer circle shows their technical support (described in Section IV). The implementation of the toolkit has been driven and validated by numerous earth scientists from different communities (see section V on Community Adoption), in the context of projects building e-research infrastructure, such as EVER-EST[4] and CoopEUS[5].

The toolkit takes into account that: i) rich and expressive metadata is a key factor for sharing and reuse, ii) scientific results need to be visible and easily discovered, iii) scientists need to receive due credit for their work, and iv) RO management capabilities need to be integrated in existing analytic tools already in use by earth scientists in order to foster adoption. The overall interaction between these components is as follows.

Semantically Rich Metadata

The RO model (see Section III) provides the vocabulary to capture metadata about the individual resources, and to aggregate them in a single unit that includes also metadata but at a high level of abstraction about the structure, content and lifecycle of the associated investigation. While structure and lifecycle metadata can be generated automatically by an RO management system, like ROHub[6], metadata about the RO content (e.g.., text files, papers, slides, etc.) typically requires the scientist input, and thus it is usually scarce. To address this issue, we developed a semantic enrichment service that conducts natural language processing against the RO payload (see Section IV-B). Additionally, we include the functionality to monitor the availability of relevant metadata, aggregated resources, and the overall RO quality, through the use of checklists, defined according to the RO usage scenarios (see Section IV-C).

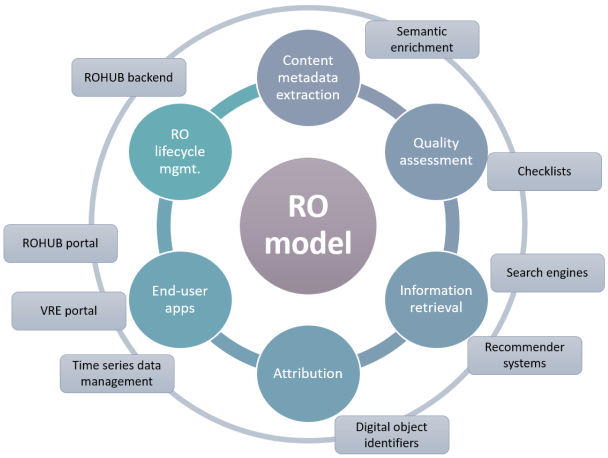

Fig. 1. High-level view of the Research Object based toolkit to support research lifecycle in Earth Science

Visibility and Discoverability

We make sure that ROs are indexable and searchable by search engines and tools that leverage the available metadata (see Section IV-A-2). Furthermore, we developed a recommender system that identifies ROs that are similar (in terms of their content) to other ROs selected by a scientist and provided as input to the system (see Section IV-D).

Attribution

Dynamic accreditation is achieved through an extension of the RO lifecycle with a fork mechanism inspired by software development practices [7], which facilitates reuse while generating an automatic citation of the source RO. Moreover, to enable proper citation of ROs from other publications and sites (e.g., articles, press, etc.), ROHub, now a DataCite[7] member, can assign Digital Object Identifiers (DOI) to ROs upon release of intermediate or final research results (See Section IV-A-1).

Research Object Management

Our goal is to integrate RO management capabilities within the software that facilitates Earth Science allowing scientists to keep the tools they are used to while exploiting the full potential of ROs. To address this issue, our toolkit includes multiple end-user interfaces. On the one hand, ROHub offers a generic RO management portal (see Section IV-A-2) where scientists can create, manage and reuse ROs, and have access to the full set of RO features. On the other

[1] http://w3id.org/bundle
[2] http://www.researchobject.org/specifications/
[3] https://github.com/dgarijo/Latex2RO
[4] http://ever-est.eu/
[5] https://www.neonscience.org/observatory/strategic-development/coopeus-project
[6] http://www.rohub.org/
[7] https://www.datacite.org/

hand, we describe two applications working on top of ROHub's back-end (see Sections IV-E-F): a Virtual Research Environment (VRE) that brings together earth observation datasets and analytical tools, and a time series data management application to more easily query and visualize real-time data on a map.

III. Research Object Model Extensions

The RO model defines a vocabulary, with formal semantics, for the representation and description of ROs. It is implemented as a suite of ontologies building upon existing standards and well-known vocabularies, such as OAI ORE (Object Exchange and Reuse)[8], the Annotation Ontology[9], and the PROV Ontology[10]. The suite includes the core ontology (for describing the aggregation of resources and their annotations), the evolution ontology (for describing the RO lifecycle), a vocabulary with useful terms, and two ontologies to describe scientific workflows and their execution provenance. A complete specification of the model can be found in GitHub[11] and detailed in [8].

Initially, the RO model was developed in the context of experimental disciplines like genomics and astrophysics (see [8]), where scientific workflows play a central role to enable reproducibility. However, even though that is also a relevant aspect for Earth Science communities, these are more focused on observations, e.g. involving the analysis of time series satellite data, rather than experimentation. Accordingly, we carried out a gap analysis to identify the required extensions to be implemented in the model. For this task, i) we asked members of our pilot communities to complete a detailed questionnaire; ii) we carried out two hackathons with members of our pilot communities where they received training and starting playing with ROs.

After the processing and analysis of the input collected, we identified the following key categories of information that were missing in the model:

1) *Geospatial,* including the coordinates of the region relevant for the RO and the observation it represents.
2) *Time-period,* e.g., time span covered in the observation.
3) *Intellectual property rights,* including copyright holder and starting year, type of license and attribution.
4) *Data access policy,* i.e. the access level and policies under which the RO can be accessed.
5) *General metadata,* including the main scientific discipline of the RO, the size and format of the resources it aggregates, the date when the RO was released, its digital object identifier (DOI), the status according to the research lifecycle, and its target community.

Additionally, concepts and properties related to executable resources have been extended to consider not only scientific workflows but also other types of processes, such as web services, scripts, command line tools and dedicated software, frequently used in Earth Sciences. Earth scientists also requested new types of ROs, to classify them according to the kind of resources they aggregate and their main focus. So, in addition to the existing workflow-centric RO, we extended the RO types to characterize data-centric, code-centric and service-centric, as well as discussion and bibliographic ROs. Finally, the RO lifecycle was extended with a new status (forked) to characterize an RO branch.

Some of the required changes were considered relevant for the general RO community and were incorporated in the corresponding ontologies, after re-aligning the different branches and updating them to the latest version of the base vocabularies[12]. Other updates that were too specific to the Earth Science domain were implemented in a new ontology extension with the metadata elicited in our analysis[13].

IV. Services and Applications

As previously introduced in section II we have identified different features complementing the RO model that have been realised through different services and applications. In the following we introduce each of these components.

A. ROHub

The RO management platform ROHub [9,10] enables scientists to manage and preserve their research work via ROs, to make it available for publishing, to collaborate and to discover new knowledge.

Built entirely around the RO concept and inspired by sustainable software management principles, ROHub is the only existing platform implementing natively the full RO model and paradigm.

ROHub comprises a backend service, a reference web client application (ROHub portal), and integrates multiple added-value RO services, as described below.

1) ROHub backend

ROHub backend service exposes a set of Restful APIs implementing the RO model to support programmatically access to the provided functionalities [11]. The two main APIs are: i) the **RO API** that enables the storage and retrieval of ROs and their aggregated resources, as well as annotating them; ii) the **RO evolution API** that enables the management of the ROs lifecycle. The backend also exposes the following APIs: OpenSearch with geospatial extensions, notification, user management, and access control. The full specification of all APIs is publicly available at GitHub[14].

The functionalities implemented to support the Earth Science needs include:

- Extended RO lifecycle support (as described in the model), including the capability to generate forks from an RO, inspired by Open Source Software development practices[15]. Forking an RO means to create a copy of an existing RO, e.g., to test new ideas without affecting the original one, or to start a new research process based on it.

[8] http://openarchives.org/ore
[9] https://www.w3.org/ns/oa
[10] https://www.w3.org/TR/prov-o/
[11] http://wf4ever.github.io/ro/
[12] https://github.com/ResearchObject/specifications/issues/13
[13] https://github.com/wf4ever/ro/tree/earth-science
[14] https://github.com/rohub/apis/wiki/RO-Services-and-APIs
[15] https://help.github.com/articles/fork-a-repo/

This is a key mechanism to foster reuse, which addresses proper accreditation by generating an automatic citation of the source RO.

- OpenSearch[16] API with geospatial extensions[17]. OpenSearch is the de facto standard used in Earth Science to search across data repositories, so it was necessary to expose such API, and especially its geospatial extensions, to facilitate integration in existing tools. This also enhanced findability and accessibility to ROs, particularly in Earth Science.
- DOI generation for RO. Now, a DataCite member, ROHub can assign DOI to snapshots or final releases of a RO. DOI are an important tool to encourage scientists for the adoption of RO since they can see the benefits of releasing results, even intermediate, that will be properly credited. DOI contribute to the findability of research data and methods, since they are persistent and searchable through a public DOI registry. Moreover, they are dereferenceable, meaning that, with a single click, the user is redirected to a landing page with the main RO metadata.

2) ROHub portal

The ROHub portal is the generic front-end for ROs that provides an advanced, life cycle management-oriented, tool exposing the full set of RO management capabilities to scientists. It is intended for users who are already familiar with ROs, or who would like to analyse and manage them at a finer grain of detail. Hence, it provides great flexibility and access to all possible operations at a granular level.

The portal is built on top of the ROHub backend, and integrates and provides access to different RO added-value services, like notification, transformation of workflows into ROs, quality and stability assessment, metadata enrichment, rating and exploratory search (see below).

One particular aspect that has been a priority in the portal is to ensure that ROs can be properly indexed and searched via search engines like Google, in order to increase findability and visibility of the ROs. To this end we i) have implemented server-side rendering mechanisms[18] to improve the indexability (and loading time) of the RO landing page (Fig. 2), which exposes its most relevant metadata; ii) are currently investigating and testing the meta tags to include, leveraging the available metadata, to help the engines find and understand these pages[19].

B. Semantic Enrichment Service

To alleviate the scarceness of metadata about the content of RO, and to structure them beyond plain text, we developed a service to automatically enrich ROs with semantic metadata extracted from their aggregated human-generated content [12]. Such metadata enhances both human and machine readability, and thus contributes to RO discoverability and interoperability. The resulting annotations are structured as semantic markup based on a knowledge graph [13] and included as annotations following the RO model. The enrichment process comprises three main stages: the extraction of text from RO resources, the semantic analysis of such text, and the actual generation of semantic metadata.

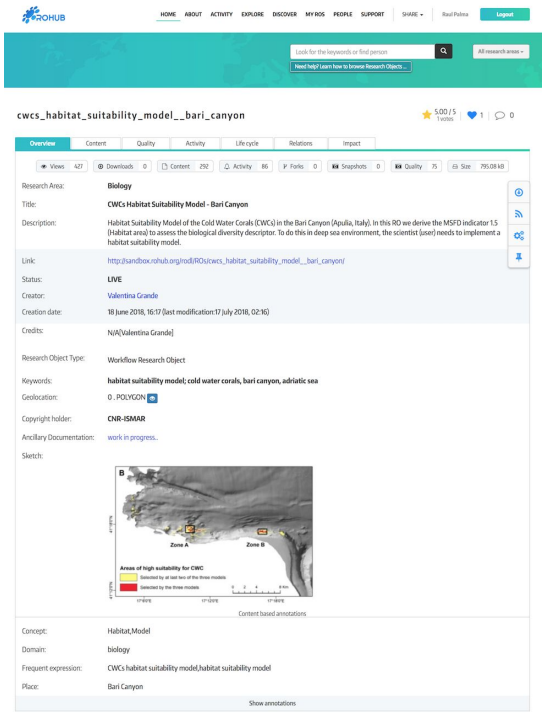

Fig. 2. ROHub portal

The first phase gathers all the text available within RO resources and human annotations. Next the text is semantically analysed using Cogito[20] system, which can identify and produce the following annotations: i) main concepts (most frequently mentioned in a document); ii) main domains (fields of knowledge in which the main concepts are commonly used); iii) main lemmas (canonical form of the most frequent words in the text); iv) main compound terms (Most frequent noun phrases); v) main named entities (most frequently mentioned People, Organizations and Places). Finally, the annotations produced are added as RO metadata, following the annotation ontology and the Content-Desc vocabulary[21].

The service has been integrated with ROHub as a nightly daemon to enrich daily the new and modified ROs, but it can also be called on demand from the ROHub portal.

C. Checklist services

ROs with high quality metadata are more likely to be reused than those with low quality. Moreover, RO quality can change in the long run, e.g., when some input file (e.g., a dataset file) becomes unavailable, degrading the overall quality of the RO and introducing decay. Inspired by wet lab practices, checklists [14] were proposed as a key tool to assess the quality of ROs through their lifecycle [15]. The

[16] http://www.opensearch.org/
[17] http://www.opensearch.org/Specifications/OpenSearch/Extensions/Geo/1.0/Draft_2
[18] https://angular.io/guide/universal
[19] https://support.google.com/webmasters/answer/35769
[20] http://www.expertsystem.com/cogito
[21] https://w3id.org/contentdesc

checklists are made up of statements that specify the required metadata and resources an RO must include and have access to, according to its type and intended purpose. Checklists allow to calculate quality metrics about ROs, including their completeness, stability and reliability [15]. These metrics can be visualized in ROHub portal, which displays the overall RO quality, and provides access to an interactive chart called the RO monitoring tool to see the all the quality metrics over time. Additionally, ROHub portal enables users to evaluate checklists over an RO on demand, e.g., to see how current changes are improving the quality.

Based on the input collected from our pilot communities, regarding the RO types and their expectations about them, we created corresponding checklists to assess Earth Science (ES) ROs of types: basic, workflow-centric, data-centric, research-product and bibliographic. The basic checklist defines the minimum metadata elements that should be present in an ES RO, while the others extend it, focusing on the type of resource the RO has at its core (workflow, data, etc.). These checklists[22] are available via ROHub portal to assess a loaded RO (Quality tab in Fig. 2).

D. *Recommender service*

To facilitate even more the discoverability of ROs, we implemented a content-based recommender service. The service takes as input user interests, expressed as a collection of ROs, and matches them against other ROs based on their content, exploiting the metadata generated by the RO semantic enrichment process [12]. The service leverages the RO social dimension via forms of interaction among researchers like RO co-authoring and citation.

The user interface, which can be accessed via ROHub and VRE portals, follows a visual metaphor based on concentric spheres, designed to facilitate RO sharing and reuse through goal driven exploration of potentially large collections of ROs. The usability and user satisfaction about this interface have been assessed in the past [16], and the overall recommender approach to facilitate RO discovery and reuse has also been recently validated [12].

E. *VRE*

The Virtual Research Environment (VRE), developed as part of EVER-EST project provides different communities of earth scientists with Virtual Research Community (VRC) portals offering custom services and tools targeted to ease the work in their community specific tasks. To support collaborative research across institutional and discipline boundaries, these portals use the RO concept and paradigm to draw together research data, models, analysis tools and workflows as well as to manage and preserve the full research cycle. However, these interfaces abstract the RO vocabulary and details from the user, providing custom-built access to the core RO management capabilities provided by ROHub in a simple and transparent manner. To achieve this, the portals communicate with ROHub backend via a custom middleware API, which translates high-level user operations into multiple ROHub backend API calls. Currently there are four VRC portals - Land Monitoring[23], Natural Hazards[24], GeoHazards Supersites[25] and Sea Monitoring[26] (Fig. 3).

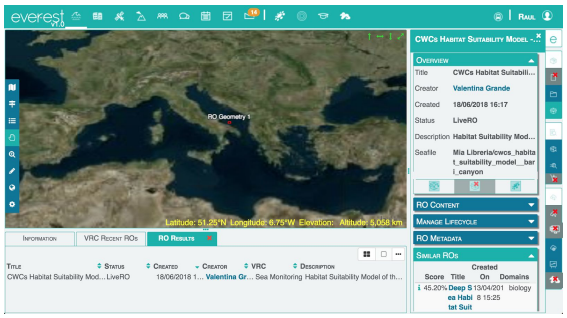

Fig. 3. Sea Monitoring VRC portal

The central role in the design of these portals is played by a 3D virtual globe, the most natural playground for an earth scientist to perform his activity. It provides interactive tools to manage the full research cycle and enables direct interaction and visualization with research data. Next to the virtual globe there is a toolbar that collects and provides access to features related to ROs and other tools that are commonly used by Earth scientists, such as search in OpenSearch catalogues and access cloud services (workflow execution, virtual machines, web processing services).

F. *Time series data management application*

Our toolkit also includes an interactive web-based prototype application[27] (Fig. 4) that integrates time series from UNAVCO[28] and National Ecological Observatory Network (NEON)[29] sensors, and produces workflow-centric ROs. The UNAVCO stations record GPS positions while sensors in NEON towers provide multiple types of data, e.g., wind speed, humidity, etc., at different time resolutions. Users can plot and download time series data by selecting the station, sensor type, and time range.

The time series are accessible from REST services; however, as UNAVCO and NEON provide data in different formats, a Kepler[30] workflow was developed to perform the REST queries and convert the results into GeoCSV[31].

Once the time series have been selected from one or more sensors, an RO may be created in ROHub to encapsulate the data and process used to create it (GeoCSV file and Kepler workflow with parameters used). The communication with ROHub is done via the backend API.

[22] https://github.com/wf4ever/ro/tree/earth-science/checklists

[23] http://vre.ever-est.eu/landmonitoring/
[24] http://vre.ever-est.eu/naturalhazards/
[25] http://vre.ever-est.eu/supersites/
[26] http://vre.ever-est.eu/seamonitoring/
[27] https://firemap.sdsc.edu/savi/map.html
[28] https://www.unavco.org/
[29] https://www.neonscience.org/
[30] https://kepler-project.org/
[31] https://giswiki.hsr.ch/GeoCSV

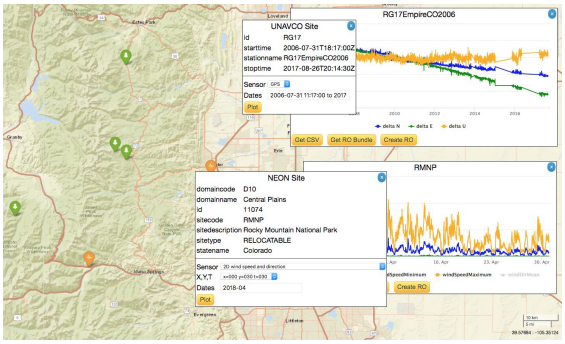

Fig. 4. Time series data management application

V. Community Adoption

Although we are still in the early stages of building a large and established community in Earth Science leveraging the RO concept and technologies in a normal basis, we have a solid infrastructure and we count with a considerable international community of early adopters mainly distributed over Europe and the USA but also with some participation from Australia. These adopters are already producing and exploiting high-quality ROs in their communities. And in order to keep growing our user base, we have i) taken a number of steps to encourage the usage of the applications; ii) implemented different mechanisms to monitor and measure the usage of the infrastructure and the benefits obtained.

A. Featured ROs

After gaining a good understanding of the RO concept and the toolkit applications, key members of our pilot communities created a set of high-quality representative ROs, referred as Golden Exemplar ROs (GEROs), for each of their areas in order to raise awareness and to encourage the usage of the infrastructure. These ROs demonstrate the feasibility and utility of ROs to manage and share data, models and results of the daily work in the particular community, but also more generally in Earth Science. Currently, there are 18 GEROs[32].

Additionally, we have produced in collaboration with our pilot communities over 500 bibliographic ROs[33] (AGROs -
Automatically Generated ROs), aggregating grey literature and reports, which are crucial for their respective institutions but that were not easily accessible (or completely inaccessible) in the past. These AGROs not only allow such important documents to be accessible but most importantly discoverable through a rich set of metadata, including some automatically extracted from the aggregated resources. Moreover, they allow us to build a critical mass of content for others to start using it.

B. Key performance indicators

We defined a set of key performance indicators (KPIs) to: i) assess the success in the adoption of the RO toolkit; ii) identify and analyse usage trends; iii) estimate the extent to which this work is contributing to improve the discoverability, accessibility and reuse of scientific resources in Earth Science. For each KPI, we defined a target for the six-month period Apr-Sep 2018, which were defined with the feedback of key community members regarding their experiences and expectations about ROs and their daily work. So, starting from April 2018, we are collecting KPIs values monthly via ROHub, and compare these values against the targets to get insights and to take necessary actions. The values collected for the last month (June 2018) are depicted in Table 1.

TABLE I. KPIs: values for June 2018 and targets for September 2018

KPI			*Target*	*Measure*
# of ROs in Earth Science		GEROs	8	18
		AGROs	500	501
		Globally	1000	720
# of resources managed by ROs in Earth Science		Total	10000	3288
Average quality of ROs in Earth Science		GEROs	95%	100%
		AGROs	90%	100%
		Released	85%	77%
Impact of ROs in Earth Science	Views	GEROs	100%	100%
		AGROs	40%	100%
	Downloads	GEROs	80%	50%
		AGROs	25%	1,8%
	Forks	Total	25%	0.42%

As we can see from table, we have already reached or exceeded several targets, including number of GEROs and AGROs, average quality of GEROs and AGROs, and percentage of ROs views. Reaching so early the targets in the number of golden and automatic ROs is a good indicator related to community adoption. But more importantly, having already over 700 ROs with over 3200 resources is a clear improvement regarding the discoverability and accessibility to the resources of these communities. As mentioned before, many of the resources in the AGROs are only until now discoverable, through a rich set of metadata, and accessible for reusing. In fact, as we can observe the generated ROs are of very high quality. And given that quality assessment considers conditions like availability and accessibility of the aggregated resources, completeness and machine readability of associated metadata, it further shows the benefits gained by the adoption of ROs among these communities. Even more, all these ROs have been already viewed at least once, i.e., they have been discovered and visualized, which reinforces our contribution to the discoverability of resources.

A few KPIs are still below the targets like the total number of ROs and resources; however, at the moment of writing we are preparing a second campaign for generation of additional ROs, both automatically and via inclusion of additional community members, and we are confident to reach the goals. Nevertheless, indicators of reuse (RO download and forks) are still far from the target. As a result, we have increased our efforts, analysing the causes and taking measures whenever possible, to raise such values. Our analysis, supported by discussions with our communities, shows that addressing this issue requires both a change in the mindset of scientists and proper tooling support. Scientists should be encouraged to increase sharing by reusing or repurposing existing results instead of carrying out their research from scratch. And they need to carry out

[32] Currently can be seen at https://tinyurl.com/y9cnuc3q
[33] https://tinyurl.com/yden9cyb

such tasks simply and intuitively, with a low technical entry barrier, providing proper accreditation to the reused work. Although such mechanisms are already available in ROHub (e.g. release of ROs with DOIs, RO fork and automatic citation to the source), our analysis indicates some lack of awareness about such functionalities among user scientists. Also in this line, ROHub recently implemented additional social components enabling users to rate and favourite (like) a RO, providing an indicator of the impact and popularity of the RO. While the amount of data is still limited, we observe a trend indicating a correlation between RO reuse and popularity. Hence, we are currently also making more awareness of such functionalities. Follow up work in this regard includes building scientists' reputation and ranking, based on the impact (rates, likes, views), reuse (downloads, forks), and quality of their ROs, which we believe will encourage them more to share and reuse.

C. *Web analytics*

We started tracking the ROHub portal with Google Analytics since March 1st 2018. Note that these statistics refer only to visitors of ROHub portal, and not about the overall number of users of the whole infrastructure. In fact, earth scientists use typically more the VRC portals than ROHub portal to carry out their daily work.

Some interesting insights we found involve the number of users visiting per day, where we can observe multiple peaks (see Fig. 5). After checking the dates, we noticed that many of the peaks coincide with the dates of dissemination or demonstration events, indicating interest from the target communities, e.g., GeoVol (Latin American workshop on volcanology) 7th-9th March, or EGU (European Geosciences Union) 9th-12th April; however, other peaks are not related directly to events, but to normal activities of our pilot communities. It is worth noticing that we have peaks of 20 users visiting per hour, and with an average session duration of 8 minutes and 39 seconds, we can estimate around 3 concurrent users[34].

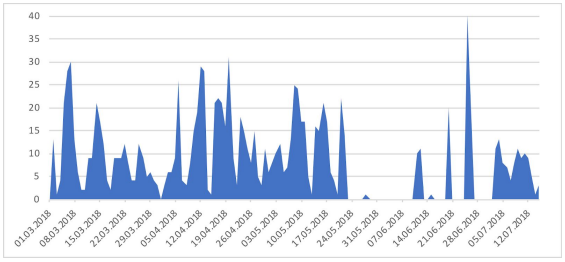

Fig. 5. ROHub web traffic: users per day March - Mid July 2018

Another interesting discovery is related to the users per country. In the first position we have USA with about 20% of the share. Despite the fact that we have engaged some communities there, it was an interesting discovery. Next is France, which started using heavily the portal during the last month. Since we don't have communities directly involved there, it was a little surprising, and thus we are analysing with our pilot communities the engagement of scientists there. In the third place is Poland (where ROHub is developed), followed by Italy (where two communities are located), Spain (where one community and one key technical partner are located) and UK (where another community is located).

[34]Using formula proposed in https://tinyurl.com/y74xtm3z

VI. Conclusions

This paper describes a toolkit of services and applications built around the RO concept and paradigm to support research in Earth Science, enabling the adoption of FAIR principles by these communities. At the centre of the toolkit is the RO model, which provides the vocabulary to represent and describe scientific investigations and their lifecycle as ROs. As part of this work, we extended this model to cover specific needs of Earth scientists, distilled from different surveys and workshops. Around the model, we have built and connected a number of services and applications enabling the adoption of ROs by the Earth Science communities, including:

1) *ROHub* that comprises the core backend service used by all other components (extended with capabilities for DOI assignment, fork and automatic citation, and geospatial search), and a generic Web portal application for RO management;

2) *semantic enrichment service* that extends ROs with metadata extracted from their aggregated content;

3) *checklist services* to assess the availability of core metadata, aggregated resources and the overall RO quality;

4) *recommender service* that enhances RO discoverability based on their content; and

5) two *Web applications tailored for Earth scientists*: a VRE that brings together core RO features, earth observation datasets and analytical tools, and a time series data management application to more easily query and visualize real-time data on a map

The paper also provides insights into the community adoption of the presented toolkit, including a discussion of the steps and measures taken to encourage and assess the usage of the infrastructure. The results obtained so far, although still far from establishing a global RO community in Earth Science, are encouraging and show the potential of the infrastructure. The challenge for the future is to enlarge the user community and leverage the experience gained to encourage other research communities to make the transition to a research environment powered by ROs.

Acknowledgment

We gratefully acknowledge EU Horizon 2020 for research infrastructures under grant EVER-EST-674907. We also acknowledge MEEO for developing the VRC portals, the San Diego Supercomputer Center, UNAVCO and NEON for the time-series data management application, and to the EVER-EST pilot communities: CNR-ISMAR, Supersites - INGV, Natural Hazards Partnership UK and SatCen, for their valuable input and commitment.

References

[1] C. Humphrey, "e-Science and the Life Cycle of Research." 2006. Retrieved from ERA Education and Research Archive Website. doi: 10.7939/R3NR4V.URL:https://era.library.ualberta.ca/items/3334684b-fa6a-4c9d-a74b-559fecd42f9f

[2] EVER-EST project, D3.1 - Use Cases Description and User Needs Document. Project deliverable. 2016. Retrieved from EVER-EST

website:https://ever-est.eu/wp-content/uploads/EVER-EST_DEL_WP3-D3.1.pdf

[3] P. E. Bourne, T. W. Clark, R. Dale, A. de Waard, I. Herman, E. H. Hovy, D. Shotton, Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331), Dagstuhl Manifestos 1 (1) (2012) 41–60. doi:10.4230/DagMan.1.1.41. URL http://drops.dagstuhl.de/opus/volltexte/2012/3445

[4] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garcia-Cuesta, J. Gomez-Perez, G. Klyne, K. Page, M. Roos, J. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. D. Roure, C. Goble, Workflow-centric research objects: A first class citizen in the scholarly discourse, in: 2nd Workshop on Semantic Publishing (SePublica), CEUR Workshop Proceedings, Aachen, 2012, pp. 1–12. URL http://ceur-ws.org/Vol-903/paper-01.pdf

[5] A.-L. Barabási, Network theory–the emergence of the creative enterprise, Science 308 (5722) (2005) 639–641. doi:10.1126/science.1112554. URL http://science.sciencemag.org/content/308/5722/639

[6] M. Wilkinson, et al, The fair guiding principles for scientific data management and stewardship, Nature Scientific Data (160018). 2016. URL http://www.nature.com/articles/sdata201618

[7] G. Robles, J. M. González-Barahona, A comprehensive study of software forks: Dates, reasons and outcomes, in: I. Hammouda, B. Lundell, T. Mikkonen, W. Scacchi (Eds.), Open Source Systems: Long-Term Sustainability, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 1–14.

[8] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. Gomez-Perez, S. Bechhofer, G. Klyne, C. Goble, Using a suite of ontologies for preserving workflow-centric research objects, Web Semantics: Science, Services and Agents on the World Wide Web 32 (2015) 16 – 42. doi:10.1016/j.websem.2015.01.003. URL http://www.sciencedirect.com/science/article/pii/S1570826815000049

[9] R. Palma, P. Hołubowicz, O. Corcho, J. Gomez-Perez, C. Mazurek, ROHub—a digital library of research objects supporting scientists towards reproducible science, in: Semantic Web Evaluation Challenge, Springer, 2014, pp. 77–82.

[10] R. Palma, O. Corcho, P. Hołubowicz, S. Pérez, K. Page, C. Mazurek, Digital libraries for the preservation of research methods and associated artefacts, in Proc. 1st International Workshop on the Digital Preservation of Research Methods and Artefacts (DPRMA 2013) at Joint Conference on Digital Libraries (JCDL 2013). pp.8-15. Indianapolis, Indiana, USA, July 2013.

[11] R. Palma, P. Hołubowicz, K. Page, S. Soiland-Reyes, G. Klyne, C. Mazurek. A Suite of APIs for the Management of Research Objects, Proceedings of the Developers Workshop, ISWC. October 2014.

[12] J. M. Gomez-Perez, R. Palma, A. Garcia-Silva, Towards a human-machine scientific partnership based on semantically rich research objects, in: IEEE 13th International Conference on e-Science (e-Science), 2017, pp. 266–275. doi:10.1109/eScience.2017.40.

[13] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna, Semantic annotation for knowledge management: Requirements and a survey of the state of the art, Web Semantics: Science, Services and Agents on the WWW 4 (1) (2006) 14 – 28. doi:10.1016/j.websem.2005.10.002.URL:http://www.sciencedirect.com/science/article/pii/S1570826805000338

[14] B. M. Hales, P. J. Pronovost, The checklist—a tool for error management and performance improvement, Journal of critical care 21 (3) (2006) 231–235.

[15] J. M. Gómez-Pérez, E. García-Cuesta, A. Garrido, J. E. Ruiz, J. Zhao, G. Klyne, When history matters-assessing reliability for the reuse of scientific workflows, in: International Semantic Web Conference, Springer, 2013, pp. 81–97.

[16] M. Rico, J. M. Gómez-Pérez, R. Gonzalez, A. Garrido, Ó. Corcho, Collaboration spheres: a visual metaphor to share and reuse research objects, CoRR abs/1710.05604. arXiv:1710.05604. URL http://arxiv.org/abs/1710.05604