Managing, Preserving and Disseminating Research Objects in Earth Science with the ROHub Science Gateway

Raul Palma and Cezary Mazurek
Poznan Supercomputing and Networking Center
Poznan, Poland
{rpalma, mazurek}@man.poznan.pl

Jose Manuel Gomez-Perez and Andrés García
Expert System
Madrid, Spain
{jmgomez, agarcia}@expertsystem.com

***Abstract*— Research Objects (ROs) are semantically enriched information units encapsulating all the materials and methods relevant to a particular scientific investigation, their associated metadata and the context where such resources were produced and came into play. Their purpose is to enhance the sharing, preservation and communication of data-intensive science, facilitating validation, citation and reuse by the community. For such mission, infrastructure and tools for RO governance are critical. ROHub is the platform of reference in the management of ROs and their lifecycle. It enables researchers to preserve their work and make it available to others, as well as to discover and reuse pre-existing scientific knowledge. In this paper, we introduce ROHub to the Science Gateways community and present new capabilities and extensions specific to Earth Sciences, beyond previous efforts in experimental disciplines.**

Keywords—Research Objects, Earth Science

I. INTRODUCTION

Research in data-intensive disciplines is increasingly consuming and generating a variety of digital resources during the course of scientific investigations. This has steadily increased the need for means to systematically capture the lifecycle of scientific investigations, which at the same time provide a single-entry point to all the related resources, including data, publications, computational resources, and the researchers involved in the investigation. In Earth Science, for example, the high-level research and information lifecycle involves tasks such as: 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 [1].

Research Objects (ROs) [2] provide the mechanisms to support researchers in these tasks. Originally conceived to support the scientific endeavour in experimental disciplines like Genomics or Astrophysics, ROs are rapidly being adopted in other fields, with special interest in Earth Sciences. With the necessary extensions and updates, research objects can support also earth scientists to manage their scientific investigations lifecycle, providing structured containers that aggregate all the resources related to a particular experiment/observation, and the means for sharing, validating and disseminating the research work as a single information unit, to be interpreted and reused by the community in the future.

Such capabilities require both an underlying (research object) model and the technological support implementing this model. The former, known as the RO model, specifies the semantic vocabulary and relations for capturing and describing ROs, their provenance and lifecycle. The latter is provided by ROHub, a holistic RO management platform implemented natively on top of the RO model. ROHub supports scientists throughout the research lifecycle to manage and to structure their resources as high-quality ROs, fostering collaboration within and across scientific communities with such ROs at the center.

In the following, we introduce the RO model with a concrete example, followed by a description of ROHub and the recently implemented extensions to both the model and the platform in support of Earth Sciences communities. Finally, we illustrate the usage of research objects and ROHub with a working example and conclude with a discussion on the ongoing work.

II. RESEARCH OBJECTS

A research object can aggregate an arbitrary number of heterogeneous resources, which can be internal or external (linked by reference) to the research object location, such as the data used or the results produced in an experiment study, the (computational) methods employed to produce and analyse that data, and the people involved in the investigation. Additionally, the resources in the research objects can be organised within folders (a special type of resource), to facilitate their inspection. Similarly, the research object can encapsulate any number of annotations associated to these resources (or the research object itself), enabling the understanding and interpretation of the scientific work, such as provenance and evolution information, descriptions of the computational methods, dependency information and settings about the experiment executions.

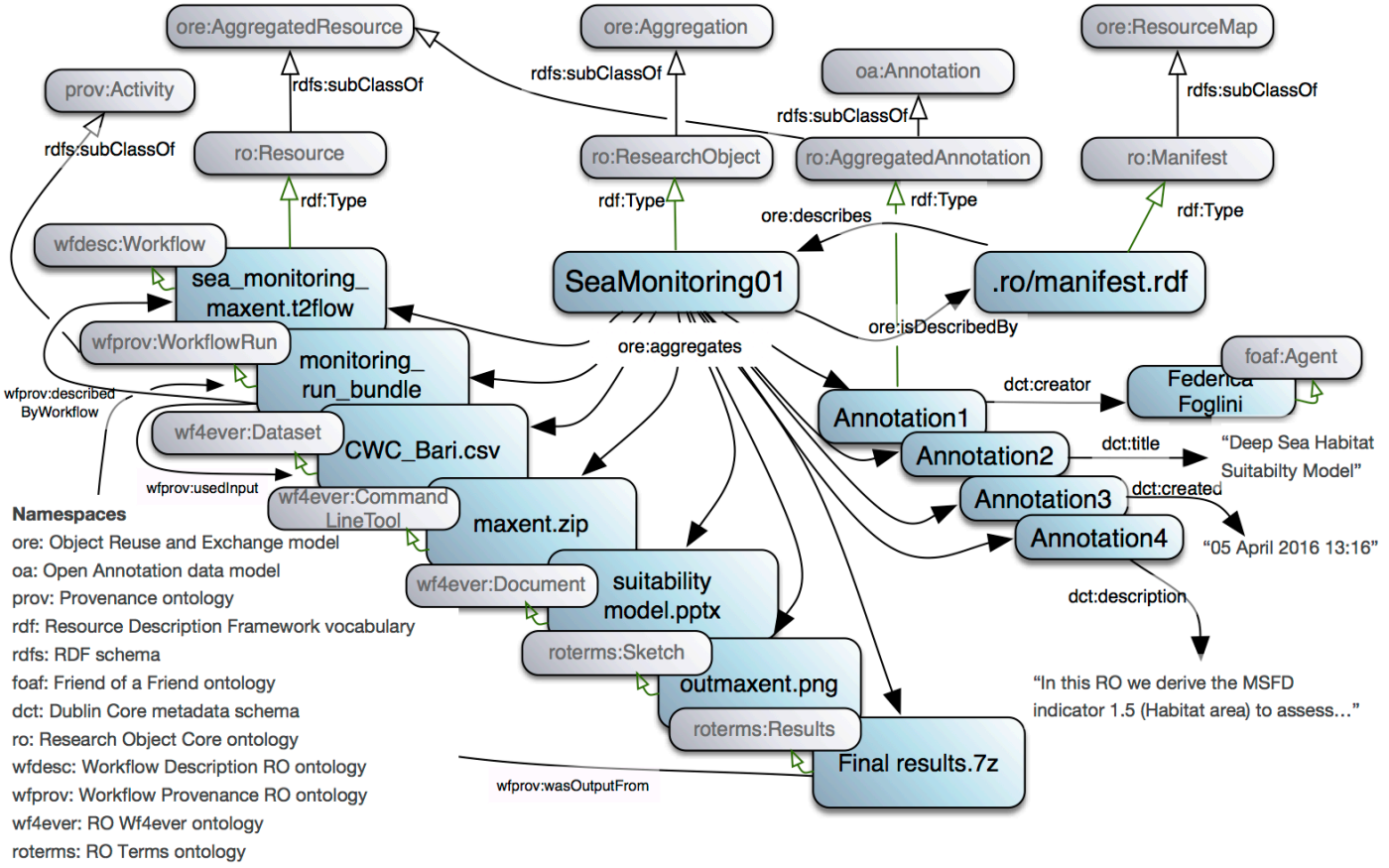

Figure 1 Partial view of an exemplary research object structure

To represent the rich, and potentially complex, structure of a research object, the underlying model was implemented as a suite of lightweight ontologies building upon existing vocabularies: OAI ORE (Object Exchange and Reuse) for specifying aggregation of resources, the Annotation Ontology (AO) to support the annotations, and the PROV Ontology to represent provenance information. A complete specification of the model can be found in [3].

Figure 1 depicts a partial (and simplified) view of a research object structure that illustrates the RO model with a concrete example. This research object is the result of the study elaborated in Section V.

III. ROHUB

ROHub (www.rohub.org) enables scientists to manage and preserve their research work through ROs, to make it available for publishing, to collaborate and to discover new knowledge (see [4,5] for a more detailed description of ROHub and its origin).

A. ROHub Implementation

ROHub comprises both a backend service and a frontend (client) application. The backend provides a set of REST APIs [6] implementing the RO model, which can be used to access ROHub programmatically. The two primary ones are the RO API and the RO Evolution API, which define the formats and links used to: i) create and maintain ROs, the resources aggregated and the associated annotations (metadata); ii) change the lifecycle stage of a RO, create an immutable copy (snapshot or archive) from a working (live) research object and fetch their evolution provenance. The backend also provides APIs for notifications, search, access control and user management, plus a SPARQL endpoint. The frontend exposes RO functionalities to the end-users through a web GUI. This is the main interface for researchers to interact with ROHub.

B. Key Features Overview

Create, manage and share ROs: ROHub provides different methods for creating ROs: from scratch, from a zip file or by importing resources from other repositories. It also supports different access modes for sharing ROs (open, public or private), allowing to specify who can read/write to the RO.

Discover, explore and reuse ROs using a faceted or keyword search interfaces, or using directly the SPARQL endpoint, for discovering ROs that can then be inspected, downloaded, and reused to create new ones.

Assess RO quality: The RO overview panel shows a progress bar of the RO quality based on a set of basic RO requirements (Figure 2). Further quality information can be found in the quality panel, where ROs can be assessed against

predefined checklist templates for specific domains or community needs.

Manage RO evolution: ROHub allows creating snapshots of the current state of the RO for sharing or release, keeping their versioning information and associated changes. RO evolution can be visualized from the History panel.

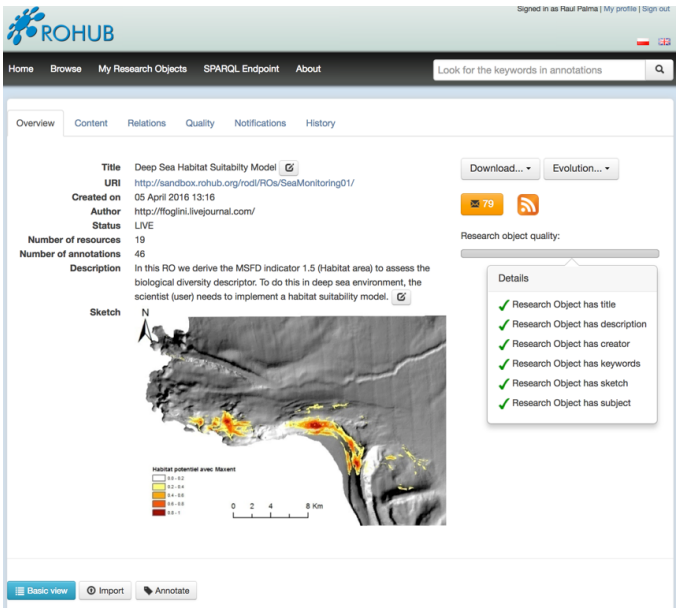

Figure 2 ROHub – RO overview panel

Nested ROs: An RO can aggregate any type of resource, including internal, links to external resources and other ROs. The latter allows aggregating RO bundles [7] that are self-contained ROs serialised as ZIP files and generated by 3rd party tools (e.g., workflow management systems).

Preserve and monitor ROs: Long-term preservation features include RO fixity checking and quality monitoring that generate notifications of changes. RO content and quality changes are shown in the notification panel, and an atom feed is available to get automatic notifications. Additionally, the quality monitoring has an interface that can be reached from the quality panel to visualise the RO quality through time.

Semantic enrichment: An RO can be enriched automatically with structured metadata extracted from its textual content, including the main concepts, domains, lemmas and named entities, in order to facilitate its discovery via the faceted/keyword search interfaces. Such metadata complements the metadata provided explicitly by scientists, offering a richer, machine-readable description of the RO.

DOI and citation: Now a DataCite (www.datacite.org) DOI allocator, ROHub can assign a DOI to the released ROs, enabling citation and stimulating scholarly communication and sharing before actual paper publication. DOI assignment follows RO release after automatically checking that the RO follows DataCite's policies, through the checklist mechanism described above.

IV. Extensions for Earth Science

ROHub is a domain-agnostic platform that has been tested in Experimental Sciences. Currently, we are extending its capabilities to support the specific needs of the Earth Science community as part of EVER-EST project. The analysis of such needs led to the new features in the model and in the platform, including support for:

Geospatial and time information. The most relevant metadata in Earth Sciences, includes geographical location and the time period covered or associated to the RO.

Data access policies to specify more detailed information about the possible use of digital content in publishing, distribution, and consumption of digital media across all sectors and communities.

Intellectual properties rights to specify detailed information on the terms of use for a given resource.

RO Fork functionality to create a new RO from an existing one to start a new line of work or extend a previous one, citing automatically the source RO.

The resulting RO model extensions are publicly available (https://github.com/wf4ever/ro/tree/earth-science) and ROHub is currently being extended to support them and to provide related user interfaces. Such new capabilities include amongst others: access and manipulation of geopositioned ROs through a map interface, definition and enforcement of data access policies and intellectual property rights, and the creation of new ROs by forking existing ones.

V. Exemplary Use Case

In this section, we introduce an excerpt of one real scenario provided by a virtual research community from EVER-EST project, and then highlight the current limitations in the existing technologies and practices to illustrate how the use of the research object and ROHub can contribute to the preservation, sharing and reuse of research outputs. The RO associated to this scenario (depicted in Figure 1) is available at: http://sandbox.rohub.org/rodl/ROs/SeaMonitoring01/

Sea Monitoring Scenario: A researcher needs to define the habitat extent of the Cold Water Coral in the Bari Canyon and to provide this information to assess the good environmental status related to the descriptor D1 (Biodiversity, Indicator Habitat extent) within the Marine Strategy Framework Directive for the Italian waters. To this scope, the researcher needs a habitat suitability model for the Cold Water Corals. The researcher needs to search high resolution bathymetric data, Cold Water Coral occurrences data and to run a good model to obtain a reliable map of habitat suitability for Cold Water Corals. The researcher needs to release the results to colleagues from different institutions working at the Marine Strategy Framework Directive, to share the model with them, to reuse the model in different locations, and to re-run the model after one year using new data from the same location. For this scenario, it is very important to share data and results

within the community, to reuse the models coming from different scientists working at the same topic, to preserve the results and to publish methodologies and final maps.

Current limitations: Currently there is not a reference site where a scientist can find publications on this specific topic, workflows executing the models, links to the data to be used and results (to mention a few). There are no specific repositories that are used to preserve and reuse all this information. Generally, there is no information about the quality of the models and the methodologies applied and described in the paper. Within the Marine Strategy Framework Directive (http://data.europa.eu/eli/dir/2008/56/oj) there is a big lack of communication and all the relevant information is dispersed in different repositories.

Overcoming the limitations with ROHub: ROHub allows the scientists to encapsulate the data, provenance of workflows executions, results, documentation and other resources related to the particular study, and to effectively preserve, share and reuse these resources through a single information unit. Moreover, ROHub allows the scientists to manage, track and visualize the complete scientific life cycle of the study, to collaborate throughout this process, and to disseminate the associated research object at different stages with colleagues or with the community (see Figure 3), so that other scientists can reuse the models in different locations and using different datasets. For the monitoring purpose the research object gives the possibility to access to all the resources necessary to exactly re-run the same model in the same location at different time giving the opportunity to evaluate the differences in habitat extents applying the same methodologies. Some of the ROHub features used by the end-users in this scenario, in their words, include:

Semantic search: To reuse an existing research object about habitat suitability models the researcher will pose a query with appropriated concepts (e.g. ecology, habitat suitability, habitat extent) in order to easily find the most suitable ones. Semantic search gives the opportunity to retrieve concepts from all documents in the research object and it is really effective to find the research objects containing data and information that you need for your work. Without the semantic search a normal keyword search could be performed, however it was difficult to find the effective concepts.

Checklists: To have information about the research object quality, and to select the research objects that effectively work. It is important to reuse research objects with a running workflow and real link to data. The checklist is a good tool to evaluate a research object without losing time in verifying manually its content.

DOIs: After reusing a workflow, with different data input and modifying the model parameters a new research object will be created with the selected data and the best suitable parameters. This new research object will be released with a DOI that gives the opportunity to be properly cited by other scientists in the community. The use of DOIs encourages researchers to create research objects containing new data and research outcomes and specially to share them with the community, since they enable citation and credit. This mechanism adds incentives for scientists to share their work and stimulate reuse, accelerating the incremental development of science.

Scientific lifecycle management: The researcher can keep track of the evolution of the scientific study, release preliminary results after reaching intermediate milestones in

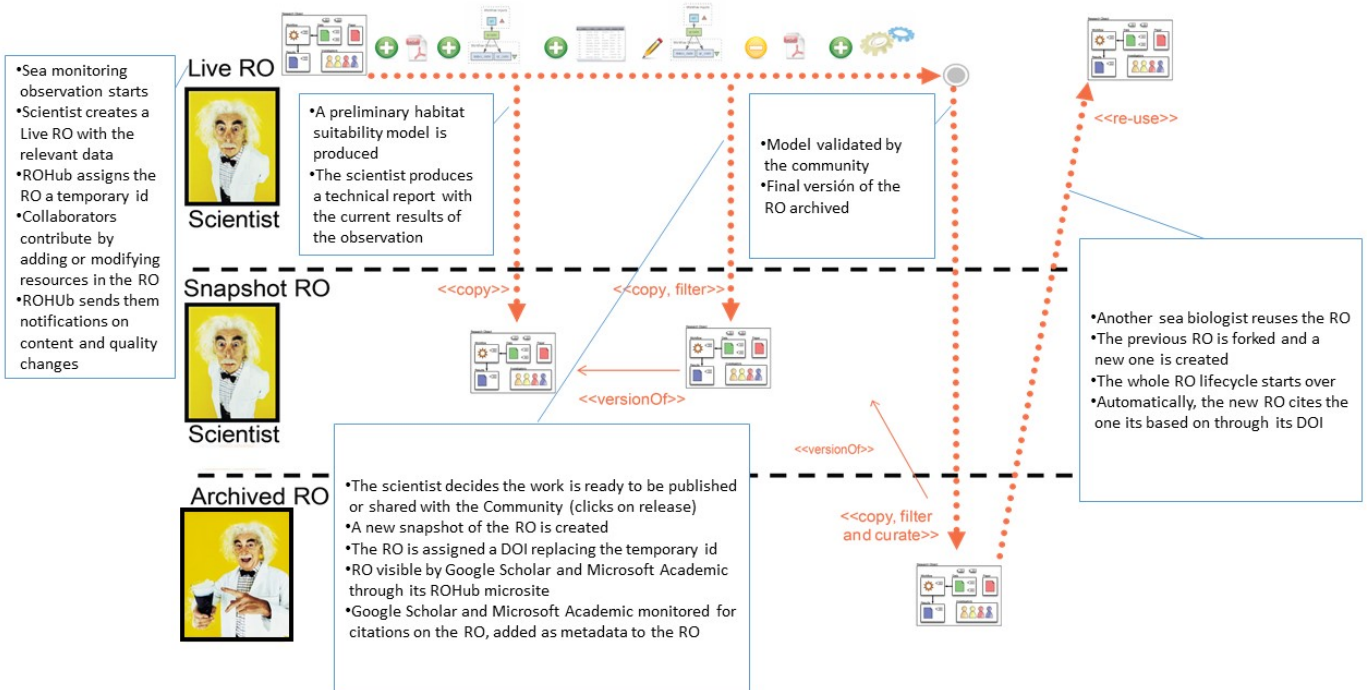

Figure 3 Research object lifecycle example

order to share the results with other colleagues and to keep a record of a particular state in the study. Such intermediate releases can then be compared to analyze the changes, or they can be used to start alternative lines of work.

Notifications: The researcher can receive notifications regarding changes in research object content but also about changes in the quality assessment. Updates on quality downfalls can be particularly useful (e.g., one of the services used is no longer available) in order to take corrective actions. Similarly, team collaborators can be notified about research object editing activity to keep track of the progress in the study, or to know when their input is required.

VI. CONCLUSION

The adoption of the Research Object paradigm by the scientific enterprise can accelerate science through a better management of the scientific information. Benefits of this approach can have an immediate impact on the validation, sharing, preservation and (eventually) reuse of scientific outcomes. However, appropriate tools and infrastructure need to be in place in order to provide the necessary functionalities to manage ROs throughout their entire lifecycle across the different scientific communities. ROHub is the first and main scientific gateway to provide holistic support for the management, sharing and communication of scientific knowledge in the form of ROs. In this paper, we recap on its main features and report the recently implemented and ongoing extensions that enable a variety of scientific communities, and specifically earth scientists, to adopt ROs in their daily work. It is still early to measure the impact that this will have in terms of increased scientific productivity and scholarly communication and citation across the different scientific areas. In addition to further refinement of the methods and tools produced, future work involves piloting the approach in our scientific communities and beyond in order to collect data, e.g. biblio and altmetrics, number of ROs, number of users, etc. that allow assessing such impact.

ACKNOWLEDGMENT

This work is supported by the EVER-EST EU project (HORIZON2020-674907). Special thanks to Federica Foglini, from CNR-ISMAR, whose RO in sea monitoring has been used as an example across the paper.

REFERENCES

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

[2] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garc´ıa-Cuesta, J.M. Gomez-Perez, G. Klyne, K. Page, M. Roos, J.E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C.A. Goble. Workflow-centric research objects: First class citizens in scholarly discourse. In Proceedings of SePublica2012, pages 112, 2012.

[3] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. Gómez-Pérez, S. Bechhofer, G. Klyne, C. Goble, Using a suite of ontologies for preserving workflow-centric research objects, in Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2015. doi:10.1016/j.websem.2015.01.003

[4] R. Palma, O. Corcho, J. Gomez-Perez and C. Mazurek, ROHub – a digital library of research objects supporting scientists towards reproducible science, in: Presutti, V., et al. (eds.) SemWebEval 2014. CCIS, vol. 457, pp. 77–82, Springer, Heidelberg (2014), Crete, Greece, May 2014.

[5] 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.

[6] 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.

[7] Research Object Bundle 1.0 specification. November 2014. https://researchobject.github.io/specifications/bundle/