RESEARCH OBJECT AS MECHANISM FOR ENSURING RESEARCH EXPERIMENT REPRODUCIBILITY WITHIN VIRTUAL RESEARCH ENVIRONMENT

MARCIN KRYSTEK[1], CEZARY MAZUREK[1], RAUL PALMA[1], JULIUSZ PUKACKI[1] AND JOSE MANUEL GOMEZ-PEREZ[2]

[1] Poznan Supercomputing and Networking Center
Jana Pawła II 10, 61-139 Poznań, Poland

[2] Expert System
Calle Prof. Waksman 10, 28036 Madrid, Spain

(received: 21 August 2017; revised: 21 September 2017;
accepted: 27 September 2017; published online: 11 October 2017)

Abstract: A Research Object (RO) is defined as a semantically rich aggregation of resources that bundles together essential information relating to experiments and investigations. This information is not limited merely to the data used and the methods employed to produce and analyze such data, but it may also include the people involved in the investigation as well as other important metadata that describe the characteristics, inter-dependencies, context and dynamics of the aggregated resources. As such, a research object can encapsulate scientific knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge within and across relevant communities, and in a way that supports reliability and reproducibility of investigation results. While there are no pre-defined constraints related to the type of resources a research object can contain, the following usually apply in the context of scientific research: data used and results produced; methods employed to produce and analyze data; scientific workflows implementing such methods; provenance and settings; people involved in the investigation; annotations about these resources, which are essential to the understanding and interpretation of the scientific outcomes captured by a research object.

The example research object contains a workflow, input data and results, along with a paper that presents the results and links to the investigators responsible. Annotations on each of the resources (and on the research object itself) provide additional information and characterize, *e.g.* the provenance of the results. Therefore, exploitation of the RO model should be considered as a way to provide additional reliability and reproducibility of the research.

The concept of the RO was introduced to the environment created in the EVER-EST project in the form of Virtual Research Environment (VRE). a group of Earth Scientists, who are observing, analyzing and modeling processes that take place on land and see, was examined against their needs and expectations about the possible improvements in their scientific work.

The results show that scientist expectations are focused on knowledge sharing and reuse, and new forms of scholarly communications beyond pdf articles as supporting tools of knowledge cross-fertilization between their members. The Research Object concept seems a natural answer for these needs. However, the model, in order to be sufficient and usable, must become a part of the working environment and needs to be integrated with the actual tools. Therefore, great efforts have been undertaken to create a generic, technical solution – VRE, which implements the expected functionalities.

In this article we present a concept of the VRE as a tool that takes advantage of the Research Object model in order to integrate and simplify the information exchange, as well as persist, share and discover assets of the reusable research. Moreover, we are presenting example scenarios of the VRE usage in the four different Earth Science domains.

Keywords: research object, virtual research environment, Earth science, semantic web, preservation, evolution

DOI: https://doi.org/10.17466/tq2017/21.4/x

1. Introduction

Usually any research process starts with creating an initial hypothesis. In order to verify this hypothesis scientists conduct an experiment which produces results. The experiment results are then analyzed and conclusions are made. The valuable conclusions are published to disseminate the research results. A process of this kind leads to discovering new knowledge based on which a new hypothesis can be made. In this sense, scientific research can be seen as an incremental process with a specific lifecycle. This process involves many different people in many different stages. Unfortunately, the traditional way of disseminating research results in the form of paper publication is very imperfect, especially in the experimental domains. It is very common that the research process cannot be reproduced by another research team in order to verify the results [1]. This problem is becoming more essential, since on-line datasets and software tools are usually used in the current research in Internet services. Any of them may change or become unavailable at any time. It is also becoming a challenge to coordinate and trace activities in continuously increasing research teams. In response to these challenges the Research Object (RO) model was introduced. The RO is a formal description of the scientific process which includes a set of tasks with sequential constraints (workflow) and resources used to perform the research (data sets, software, people) [2]. The ROHub [3] platform is an implementation of the RO model. It provides a set of functionalities supporting researchers in their daily work. The further evolution of the ROHub platform was possible due to cooperation with researchers focused on the Earth Science domain. Further integration with external tools and services will lead us to define and create the Virtual Research Environment (VRE).

In this article we have briefly presented the concept of the Research Object, the functionality of the ROHub platform, example VRE usage scenarios as well as the VRE architecture, functionality and implementation methodology.

2. Research Object

A Research Object (RO) is defined as a semantically rich aggregation of resources that bundles together essential information relating to experiments and investigations. This information is not limited merely to the data used and the methods employed to produce and analyze that data, but it may also include the people involved in the investigation as well as other important metadata that describe the characteristics, inter-dependencies, context and dynamics of the aggregated resources. As such, a research object can encapsulate scientific knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge within and across relevant communities, and in a way that supports reliability and reproducibility of investigation results [4]. While there are no pre-defined constraints related to the type of resources a Research Object can contain, in the context of scientific research the following usually apply:

- Data used and results produced
- Methods employed to produce and analyze data
- Scientific workflows implementing such methods
- Provenance and settings
- People involved in the investigation
- Annotations about these resources, which are essential to the understanding and interpretation of the scientific outcomes captured by a research object

The Research Object contains a workflow, input data and results, along with a paper that presents the results and links to the investigators responsible. Annotations on each of the resources (and on the research object itself) provide additional information and characterize, e.g. the provenance of the results (the results were obtained by executing the workflow on the input data).

Scientific workflows represent a key technology paradigm in the scientific community as they allow scientists to delineate the steps of a complex analysis, record the steps of computational experiments and expose this to peers using workflow design, execution and sharing tools and platforms. A scientific workflow can be defined as a series of structured activities and computations that occur in scientific problem-solving. From a computational perspective, such a workflow could be defined as a directed acyclic graph whose nodes correspond to analysis operations and whose edges specify the flow of data between those operations. In any case, the usefulness of workflows goes beyond the mere description and execution of a set of computations since they play an important role as an executable artifact for sharing, exchanging and reusing scientific in-silico methods, as demonstrated by existing workflow repositories, such as myExperiment [5] and crowdLabs [6]. Their high scholarly value lies in the fact that:

- They allow the assessment of the reproducibility of results;
- They can be reused by the same or by a different scientist;
- They can be repurposed for other goals than those for which they were originally built;

- They can validate the method that led to a new scientific insight;
- They can serve as live-tutorials, exposing how to take advantage of the existing data infrastructure.

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 the 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.

Research Objects (ROs) provide the mechanisms to support researchers in these tasks. Originally conceived to support the scientific endeavor in experimental disciplines like Genomics or Astrophysics, ROs are being rapidly 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 the lifecycle of their scientific investigations, 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 [7] 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 [2].

3. ROHub

ROHub [2] (www.rohub.org) enables scientists to manage and preserve their research work through ROs [8, 9], to make it available for publishing, to collaborate and to discover new knowledge. ROHub comprises both a backend service and a frontend (client) application. The backend provides a set of REST APIs [10] 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.

The ROHub portal is a web client application providing a comprehensive user interface for the management and preservation of Research Objects (ROs). It integrates and provides access to different research object services, including:

- **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 specifying 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. The RO evolution can be visualized from the History panel.
- **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 visualize 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 the DataCite policies, through the checklist mechanism described above.

ROHub has been designed as service providing functionality and tools for the users. In general it is domain agnostic. This means that the RO model is not tightly associated to any domain and may describe any data and any scientific method. The current experience in the RO model usage in the Earth Science domain shows that all user scenarios can be successfully accomplished and all the required data and metadata could be aggregated in the RO. However, there are some additional semantic models describing specific metadata like geospatial data or time sequences. In order to follow the standards and support better integration with external tools used by the scientists, the RO model has been extended and new models have been included. Also, the ROHub portal user interface provides most of the functionality required by the users. The missing parts apply mostly to tools specific for the Earth Science domain like maps. The EVER-EST [11] users have requested also for an additional set of tools and services for collaboration, real time information exchange, e-learning and integration with computational resources. In order to meet all these requirements and preserve the domain agnostic character of the ROHub platform, a concept of the Virtual Research Environment (VRE) has been introduced. The VRE will integrate the functionality related to RO management and all services and tools required to effectively conduct research in the Earth Science domain.

4. Virtual Research Environment motivation

In this paragraph we describe the VRE usage scenario introduced by one of the EVER-EST [11] partners. It introduces the main challenges and problems of a daily scientific work and emphasizes how the usage of the VRE and the RO model can effectively overcome them.

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 the 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 on the same topic, to preserve the results and to publish methodologies and final maps.

Currently there is no 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 just 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.

The VRE platform allows overcoming these limitations by taking advantage of the Research Object concept. Each science investigation is represented by a new instance of the RO. The RO itself contains data, formal description of the undertaken tasks (workflow), results, documents and other resources describing the scientific process. Such an RO can be then maintained, shared or reused if required. VRE users are taking advantage of the ROHub and VRE integration and getting the same functionality covered by the dedicated user interface. Other VRE users interested in the same or similar domain may simply reuse an existing RO by customizing its description and adapting it to a new localization or different data sets. In order to monitor environment changes, the RO gives a possibility to access all the resources required for repeating a whole scientific process in the automatic manner at any time. In this way changes in the studied environment can be easily detected.

5. Virtual Research Environment

The methodology that we followed for the analysis of research objects in the Earth Science communities and the elicitation of requirements had 3 main pillars:

- A **Research Object Requirements Questionnaire** that contained questions related to the goals, content, metadata and requirements of RO users. The questionnaire was given to each of the 4 Virtual Research Communities (VRC) and their answers were analyzed to derive the requirements. The questions were based on the existing RO requirements from other scientific communities and the goal was to establish a) which of these requirements were applicable in Earth Science as well and b) what additional requirements may be needed.
- A **Research Object Hackathon** where users from the 4 VRCs received comprehensive training on research objects and had the chance to create their own objects using the existing tools and models. This exercise helped clarify and distill the requirements derived from the questionnaire and helped identify potential issues and challenges with respect to their implementation.
- A **Research Object Survey** that is addressed to the broader Earth Science community (i.e. not merely the four VRCs of this project) and aims at a more comprehensive understanding of its needs with respect to ROs and scientific knowledge sharing and preservation.

Based on the analysis of the requirements and workshop results example VRE usage scenarios were defined [12]. 4 use cases from different Earth Science domains are described below:

5.1. Sea monitoring

The Sea Monitoring VRC focuses on finding new ways to measure the quality of the maritime environment and it is quite wide and heterogeneous, consisting of multi-disciplinary scientists such as biologists, geologists, oceanographers and GIS experts, as well as agencies and authorities. This scenario can be divided into three main parts: a) collecting data about species and environment variables; b) creating a statistical model and its validation; c) using a statistical model to predict the habitat of species and support the maritime economy.

Given the above, a research object will encapsulate the following resources: bathymetry data, vector data and hydrodynamic models; analytic software; metadata – creation time, authors, statistic parameters, variables, dependencies, access rights; documents – publications and presentations

5.2. Natural Hazards

This scenario focuses on modeling the impact of 3 key hazards – surface water flooding, land instability and high winds – on people, their communities and key assets such as road, rail and utility networks. The partners share scientific expertise, data and knowledge on hydrological modeling, meteorology, engineering geology, GIS and data delivery and modeling of socio-economic impacts.

Let us consider a hazard impact modeling workflow that describes all the elements necessary to build further hazard impact models to be shared with partners located at other organizations across the UK as an example of a research object. The purpose of such a research object would be sharing modeling workflows to enable partners to plan how they might develop their own models for different hazards. This type of approach would help maintain the consistency of the approach and ensure that all impact models use the most up to date versions of generic datasets. Primarily, the research object will be used simply to share a workflow comprising a set of disparate processes developed by individual organizations. These processes would be run locally by the host organization that would describe the process and store this description as part of the research object. Any generic processes common to all hazards could be stored within the research object as scripts for information and re-used locally. Currently, each process that makes up the workflow is run manually by the host organization. This often involves complex scientific modeling. The outputs from this modeling e.g. hazard footprint, could be stored in the research object and re-used by partners for subsequent impact modeling Whilst a complete modeling workflow could not be fully automated (due to the use of complex scientific modeling), some processes could be packaged and therefore automated.

Such a research object will include standards and definitions, modeling methods, descriptions of the natural hazard impact, data sets and software.

5.3. Land Monitoring

Land Monitoring is a transversal issue (common to a number of users ranging from scientists to institutions) that can refer to the monitoring of urban, build-up and natural environments to identify certain features and anomalies or changes over areas of interest as well as of natural resources to observe their condition and exploitation.

The use case to be considered in the project involves the ingestion of satellite images acquired on land areas (with the support of info coming from social sensing sources and other geotagged information) to automatically detect changes and send an alarm to the user. A second alert will be triggered if the detected change is identified as relevant with respect to a predefined list of changes (e.g. new building construction, road interruption, etc.) to be populated following the user requirements.

In this example the RO will contain: datasets – satellite images and geotagged data (videos, images, social media); software; documents – algorithm descriptions, verification procedures and reports, change detection method descriptions; metadata – creation dates, people, places, purpose and content description.

5.4. GeoHazards Supersites

An example of a research object for the Supersite VRC involves the measurement of ground displacement and velocity due to volcanic activity. Ground deformation mapping is a typical use case for this VRC and it may be carried out by different researchers on different volcanoes or even on the same volcano. It normally consists of two consecutive workflows, one to perform the analysis of a multitemporal InSAR image dataset to calculate ground displacement time series, and another to validate the results by comparison with other data or results.

In this scenario the RO will contain: computational workflow; documentation – scientific articles, bibliography, instructions; input and output data and reports; software.

The use cases developed together with actual users and deeper analysis of their requirements are the keystone in VRE conceptualization. The VRE is defined as a set of integrated services and tools, which are exposed to the user by a common interface. Using the VRE, the user is able to accomplish a whole research procedure starting from defining the initial assumptions, through model building, model verification and ending with result publishing. This process can be realized due to the integration of the following services [13]:

- ROHub – is the centre point of the VRE. ROHub provides the RO management functionality. All activities accomplished by the user in the VRE, description of the experiments together with its results are materialized as parts of the Research Object. This allows building and preserving a clear research path and enables its verification and reproduction in the future.
- Computational platform – shared in the PaaS model (Platform as a Service). The computational platform allows users to define and instantiate virtual machines according to their needs. User VMs are dedicated to perform all automatic steps defined in the user workflow, which is a part of the Research Object.
- Workflow manager – a set of tools for designing and executing task sequences. It is responsible for synchronization, coordination and automation of the execution process.
- Collaboration services – a set of tools for direct user communication and cooperation. The expected interaction models are: text messages,

audio-video teleconferences, document sharing and editing, notes, blogs and WIKI pages.
- E-Learning services – services for creating documents with the source code, equations and visualizations. The user will be able to use these services for data transforming and cleaning, numerical transformations, statistical modeling and machine learning.
- Cloud data storage – allows users to store and share files required in the daily scientific work. It is not a data source for a computational platform but personal user workspace.
- Identity provider – is a service responsible for managing the user identity. It can authenticate the user and authorize his/her access to all services integrated in the VRE.
- VRE portal – a graphical user interface implemented in the form of an Internet portal. It is the main access point to the VRE which integrates all the above services, tools and functionalities. As the VRE architecture is service oriented, some of the services may have their own, dedicated interfaces, offering a different approach to the content presentation and more sophisticated functionalities. A good example of this is ROHub, which has its own Internet portal. However, the VRE portal was designed and implemented based on the user requirements, hence, it supports domain users in their daily scientific work. It provides users with a common interface for all processes in a single place. Users from other domains can be easily supported due to the modular structure of the VRE portal implementation and simple adaptation to the new requirements.

The VRE is a tool for the daily work of a scientist. Due to the integration of multiple services, especially a computational platform, the user may accomplish many tasks in a single place without using different tools, portals and services. The VRE significantly simplifies scientific work by eliminating the need of relocation of data and results as well as creating and managing the computational platform. The Research Object model, which is the basis of the VRE, ensures that all user activities (like creation of a new document, creation of a new model) will be stored in the form allowing recreating the process for verification and reproducibility in a different context. Thus, the VRE not only simplifies the daily work but essentially increases the value of the scientific investigation by increasing the visibility of scientific results and methods, allowing their better verification and adaptation to new different applications.

6. Summary

In this article we have briefly presented the concept of the Research Object – a formal and structural description of a scientific experiment. We have also described the ROHub platform which implements the RO concept in the form of an Internet portal and service. Although, this platform was designed and implemented for a scientist from experimental domains, it proves its usability also in the Earth Science domain. The next step in the RO concept development is the VRE – an environment that integrates services and functionalities in order to create a user-friendly working environment for the scientist, dedicated to the Earth Science domain. Creating the VRE around the ROHub platform improves the quality of models created by users due to verification mechanisms, sharing, exploring, semantically annotating and automating most of the process. The VRE is a tool for modern scientists, who use sets of data in their daily work large and execute complicated computational models, ensuring the reproducibility of an experiment and reuse of a scientific method.

The VRE platform is still under construction. The actual work focuses on further service integration and new functionalities. Releasing the final version of the VRE is planned at the end of 2018.

References

[1] [Online] available at: http://www.reuters.com/article/2012/03/28/us-science-cancer-idUSBRE82R12P20120328

[2] Belhajjame K *et al.* 2012 *Workflow-centric research objects: First class citizens in scholarly discourse* In Proceedings of SePublica 112

[3] [Online] available at: http://www.rohub.org

[4] Belhajjame K *et al.* 2015 *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* doi: doi:10.1016/j.websem.2015.01.003

[5] [Online] available at: www.myexperiment.org

[6] [Online] available at: crowdlab.com

[7] Belhajjame K *et al.* 2012 *Workflow-centric research objects:first class citizens in scholarly discourse* in: Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web

[8] Palma R, Corcho O, Gomez-Perez J Mazurek C 2014 *ROHub – a digital library of research objects supporting scientists towards reproducible science*, *CCIS* in: Presutti V, *et al.* (eds.) **457** 77

[9] Palma R, Corcho R, Hołubowicz P, Pérez S, Page K, Mazurek C 2013 *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) 8

[10] [Online] available at: https://github.com/wf4ever/apis/wiki/Wf4Ever-Services-and-APIs

[11] [Online] available at: http://ever-est.eu/

[12] Gomez-Perez J M, Alexopoulos P, Garcia N, Palma R *D4.1 Workflows and Research Objects in Earth Science Concepts and Definitions* [Online] available at: http://ever-est.eu/wp-content/uploads/EVER-EST_DEL_WP4-D4.1.pdf

[13] Gonçalves P *et al.* *VRE Architecture and Interfaces Definition* [Online] available at: http://ever-est.eu/wp-content/uploads/EVER-EST_DEL_WP5-D5.1.pdf