ROHub – A Digital Library for Sharing and Preserving Research Objects

Raul Palma, Cezary Mazurek
Piotr Hołubowicz[i]
Poznan Supercomputing and Networking Center
Poznan, Poland
(+48) 618582161
[rpalma,mazurek]@man.poznan.pl

Oscar Corcho
Ontology Engineering Group,
Universidad Politécnica de Madrid
Madrid, Spain
(+34) 913366605
ocorcho@fi.upm.es

José Manuel Gómez-Pérez
iSOCO
Madrid, Spain

(+34) 913349797
jmgomez@isoco.com

ABSTRACT

ROHub is a digital library system, enhanced with Semantic Web technologies, which supports the storage, lifecycle management, sharing and preservation of research objects - semantic aggregations of related scientific resources, their annotations and research context. ROHub includes a set of features to help scientists throughout the research lifecycle to create and maintain high-quality research objects that can be interpreted and reproduced in the future, including quality assessment, evolution management, navigation through provenance information and monitoring features. It provides a set of RESTful APIs along with a Web Interface for users and developers. A demo installation is available at: www.rohub.org.

General Terms

Infrastructure, specialist content types.

Keywords

Methods, preservation, semantic, aggregation, research objects

1. INTRODUCTION

Digital Library systems collect, manage and preserve digital content, with a measurable quality and according to codified policies [1]. These systems have been traditionally focused on the preservation of data and content of rather static nature, i.e., documents, images, datasets. However, research in data-intensive science, conducted in increasingly digital environments, has led to the emergence of new types of content and artefacts [2], such as computational methods that also have a dynamic dimension (i.e. they are executable). For instance, scientific workflows are executable descriptions of scientific procedures that define sequences of computational steps in automated data analysis.

Hence, in order to share and preserve research findings, we need to consider not only the data used and produced, but also the methods employed, and the research context in which these artefacts were conceived. Moreover, in order to enable the reusability and reproducibility of the associated investigations, we need to provide access to all these related artefacts, their research context, as well as information about the usage and provenance of these resources. Similarly, in order to capture the dynamic aspects of these resources, we need information about their evolution and, in the case of computational methods, about their executions.

Research objects (ROs) provide a container for all these associated artefacts. They are aggregating objects that bundle together experimental resources that are essential to a computational scientific study or investigation, along with semantic annotations on the bundle or the resources needed for the understanding and interpretation of the scientific outcomes. The RO model [3] provides the means for capturing and describing such objects, their provenance and lifecycle, facilitating the reusability and reproducibility of the associated experiments. The model consists of the core RO ontology[1], which provides the basic structure for the description of aggregated resources and annotations on those resources, and extensions for describing evolution aspects and experiments involving scientific workflows. Hence, ROs can help scientists in sharing research findings, but scientists also need the appropriate technological support enabling them to create, manage, publish and preserve these objects.

2. ROHub

ROHub is a digital library system supporting the storage, lifecycle management, sharing and preservation of research findings via ROs. It includes different features to help scientists throughout the research lifecycle: (i) to create and maintain ROs compliant with predefined quality requirements so that they can be interpreted and reproduced in the future; (ii) to collaborate along this process; (iii) to publish and search these objects and their associated metadata; (iv) to manage their evolution; and (v) to monitor and preserve them supporting their accessibility and reusability.

2.1 Interfaces

ROHub provides a set of REST APIs[2], the two primary ones being the RO API and the RO Evolution API. The RO API defines the formats and links used to create and maintain ROs in the digital library. It is aligned with the RO model, hence recognizing concepts such as aggregations, annotations and folders. The RO ontology is used to specify relations between different resources. ROHub supports content negotiation for metadata, including formats like RDF/XML, Turtle and TriG. The RO Evolution API defines the formats and links used to change the lifecycle stage of a RO, to create an immutable snapshot or archive from a mutable Live RO, as well as to retrieve their evolution provenance. The API follows the RO evolution model [3]. ROHub also provides a SPARQL endpoint, a Notification API, a Solr REST API, and a

[1] See http://wf4ever.github.io/ro/ and http://researchobject.org/

[2] APIs documentations available at: http://www.wf4ever-project.org/wiki/display/docs/Wf4Ever+service+APIs

User Management API, in addition to a Web interface, which exposes all functionalities to the users. The latter is the main interface for scientists and researchers to interact with ROHub.

2.2 Implementation

ROHub realizes the backbone services and interfaces of a software architecture for the preservation of ROs [4]. Internally, it has a modular structure that comprises access components, long-term preservation components and the controller that manages the flow of data. ROs are stored in the access repository once created, and periodically the new and/or modified ROs are pushed to the long-term preservation repository.

The access components are the storage backend and the semantic metadata triplestore. The storage backend can be based on dLibra[3], which provides file storage and retrieval functionalities, including file versioning and consistency checking, or it can use a built-in module for storing ROs directly in the filesystem.

The semantic metadata are additionally parsed and stored in a triplestore backed by Jena TDB[4]. The use of a triplestore offers a standard query mechanism for clients and provides a flexible mechanism for storing metadata about any component of a RO that is identifiable via a URI.

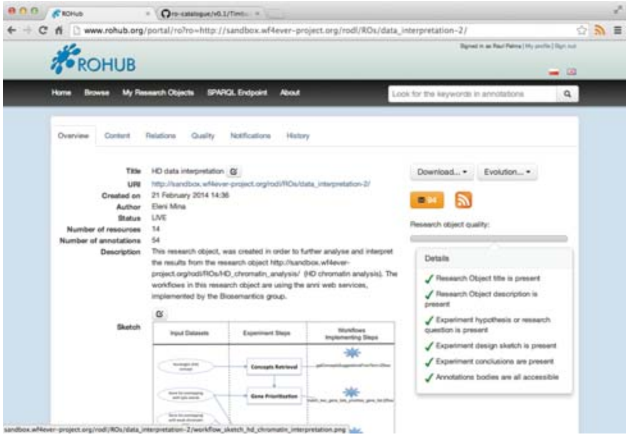
The long-term preservation component is built on dArceo[5], which stores ROs, including resources and annotations. Additionally, ROHub provides fixity checking and monitors the RO quality through time against a predefined set of requirements. If a change is detected, notifications are generated as Atom feeds.

2.3 Main functionalities

Create, manage and share ROs There are different methods for creating ROs in ROHub: (i) from scratch, adding resources progressively; (ii) by importing a pack of resources from other systems (currently myExperiment); (iii) from a ZIP file aggregating files and folders; (iv) by uploading local ROs from the command line using RO Manager Tool[6]. Resources can be added and annotated from the content panel that also shows the folder structure. ROHub provides different access modes to share the ROs: open, public or private. In the open mode, anyone with an account can visualise and edit the RO. In the public mode, everyone can visualise the RO, but only users with correct permissions can edit it. In private mode, only users with correct permissions can visualize and/or edit the RO. ROHub provides a keyword search box and a faceted search interface to find ROs, and a SPARQL endpoint to query RO metadata.

Assessing RO quality Users can visualise a progress bar on the RO overview panel (see Fig. 1), which shows the quality evaluation based on set of predefined basic RO requirements. When clicked, users can visualise further information about the RO compliance. Users can also get more information about the quality of the RO from the Quality panel, where they can choose from different templates to use as the basis for evaluating the RO.

Managing RO evolution From the RO overview panel, users can also create a snapshot (or release) of the current state of their RO, at any point in time, for sharing the current outcomes with colleagues, get feedback, send it to review, or to cite them.


ROHub
ro-catalogue/v0.1/Tim...
www.rohub.org/portal/ro?ro=http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/
Signed in as Raul Palma | My profile | Sign out
ROHUB
Home
Browse
My Research Objects
SPARQL Endpoint
About
Look for the keywords in annotations
Overview
Content
Relations
Quality
Notifications
History
Download...
Evolution...
Research object quality:
Sketch
Input Datasets
Experiment Steps
Workflows Implementing Steps
Concepts Retrieval
Gene Prioritization
sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/workflow_sketch_hd_chromatin_interpretation.png

Title HD data interpretation
URI http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/
Created on 21 February 2014 14:36
Author Eleni Mina
Status LIVE
Number of resources 14
Number of annotations 54
Description This research object, was created in order to further analyse and interpret the results from the research object http://sandbox.wf4ever-project.org/rodl/ROs/HD_chromatin_analysis/ (HD chromatin analysis). The workflows in this research object are using the anni web services, implemented by the Biosemantics group.

Details
- ✓ Research Object title is present
- ✓ Research Object description is present
- ✓ Experiment hypothesis or research question is present
- ✓ Experiment design sketch is present
- ✓ Experiment conclusions are present
- ✓ Annotations bodies are all accessible

Figure 1 ROHub - RO overview panel

Similarly, when the research has concluded, they can release and preserve the outcomes for future references. ROHub keeps the versioning history of these snapshots, and calculates the changes from the previous one. Users can visualise the evolution of the RO from the History panel, and navigate through the RO snapshots.

Navigation of execution runs Scientists can aggregate any type of resources, including links to external resources and RO bundles, which are structured ZIP files representing self-contained ROs that facilitate their transfer and integration with 3rd party tools. Taverna, for example, can export provenance of workflow runs as RO Bundles. In ROHub, bundles are unpacked into nested ROs, exposing their full content and annotations. Hence, scientists can navigate through the inputs, outputs and intermediate values of the run, something potentially useful for future reproducibility.

Monitoring ROs ROHub includes monitoring features, such as fixity checking and RO quality, which generate notifications when changes are detected. This can help to detect and prevent, for instance, workflow decay, occurring when an external resource or service used by a workflow becomes unavailable or is behaving differently. Users can visualise changes in the RO, regarding the content and quality monitoring in the notification panel and they can subscribe to the atom feed to get automatic notifications.

3. ACKNOWLEDGMENTS

This work was supported by Wf4Ever EU project (FP7-270129). We thank Filip Wisniewski, Graham Klyne, Stian Soiland-Reyes and Wf4Ever users for their contributions and feedback.

4. REFERENCES

[1] Candela, L. et al. The DELOS Digital Library Reference Model Foundations for Digital Libraries. DELOS, Italy, Dec 2007

[2] De Roure, D. et al. Towards the preservation of scientific workflows. In Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011). Nov 2011

[3] Belhajjame, K. et al. Workflow-centric research objects: First class citizens in scholarly discourse. In ESWC2012 Workshop on Semantic Publication (SePublica2012) May 2012

[4] Page, K. et al. From workflows to Research Objects: an architecture for preserving the semantics of science. In ISWC Workshop on Linked Science. Nov 2012

[3] http://dlab.psnc.pl/dlibra/

[4] http://jena.apache.org/

[5] http://dlab.psnc.pl/darceo/

[6] https://github.com/wf4ever/ro-manager

[i] Present address: Google, CA, USA. piotrhol@google.com