From workflows to Research Objects: an architecture for preserving the semantics of science

Kevin Page[1], Raúl Palma[2], Piotr Hołubowicz[2], Graham Klyne[3], Stian Soiland-Reyes[4], Don Cruickshank[5], Rafael González Cabero[6], Esteban García Cuesta[7], David De Roure[1], Jun Zhao[3], José Manuel Gómez-Pérez[7]

[1]Oxford e-Research Centre and [3]Department of Zoology, University of Oxford, UK
[2]Poznan Supercomputing and Networking Centre, Poland [7]iSOCO, Madrid, Spain
[4]School of Computer Science, University of Manchester, UK
[5]Electronics and Computer Science, University of Southampton, UK
[6]Facultad de Informítica, Universidad Politícnica de Madrid, Spain

Abstract Research Objects (ROs) provide a flexible model to collate and describe the semantic context of science. In this position paper we describe how ROs can also provide a foundation for interoperability within RESTful architecture design, enabling the development of new services and clients alongside compatible enhancements to existing software including myExperiment. To illustrate this we introduce an infrastructure, known as the Wf4Ever Toolkit, providing services and clients to encapsulate, preserve, and re-use ROs.

1 Introduction

While a workflow can describe an experiment, to aggregate the wider digital context of scientific processes and their conduct – input and output data, method, software, actors, analysis, dissemination, sharing, re-use, and the links and relationships between these gathered resources – we need Research Objects (ROs) [2] and a rich array of tools that, through ROs, can support these needs. To achieve this goal, the Workflow4Ever (Wf4Ever) project[8] is building infrastructure to enable the capture, preservation and re-use of ROs, not simply as a semantic encoding, but as building blocks for interoperability and co-ordination in a distributed architecture of services and their APIs.

2 An RO-centric architecture

Adopting the practices outlined by Page *et al* [5], our RO-centric architecture builds interoperability through *models*, *APIs*, and *services*:

Models, particularly the *RO model* [1] for aggregation and annotation, are the linchpin of architecture interoperability embodied by APIs. Specialised models supplement RO for *in silico* workflow definition (*wfdesc*), workflow provenance (*wfprov*) and RO evolution (*roevo*, incorporating versioning and lifecycle).

APIs. Application Programming Interfaces follow a strict definition of REST [4] with "follow your nose" navigation of resources for transitions in application

[8] The EU Wf4Ever project (270129) is funded under EU FP7 (ICT-2009.4.1).

state and caching considerations to the fore. The RO model and its associates are, in their RDF incarnation, the primary representation of many resources exposed by the APIs and thereby provide a linked data interface. Beyond this use of REST and linked data our architecture is substantially characterized by the choice of supplemental resources and representations that are provided. A good example of this relationship between model and API, and its centrality to the architecture, can be seen **RO Storage and Retrieval (ROSR) API**[9] used by many of the services and clients below.

Services implement shared provision of functionality accessed through one or more APIs (each may be implemented by one or more service). The next section outlines a number of implemented services that demonstrate the interoperability afforded by adoption of an RO-centric architecture.

3 The Wf4Ever Toolkit

The business of capturing, enhancing, and preserving the scientific process is undertaken by the **Wf4Ever Toolkit**[10], an RO-centric architecture with services categorised within three layers and four functional sub-groupings (figure 1; in the following text services and APIs in **bold**, conceptual models sans serif).

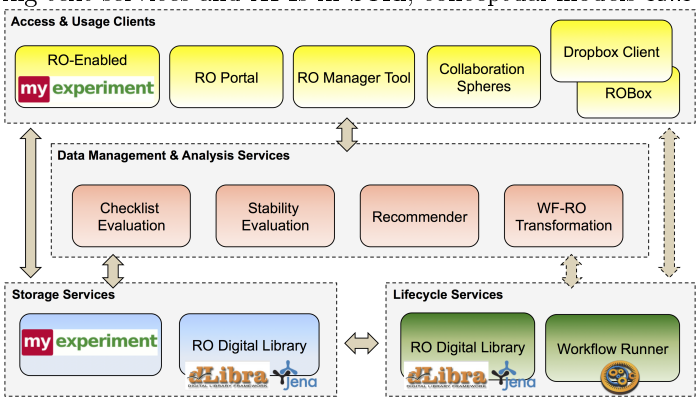

Figure 1. The Wf4Ever Toolkit and architecture

Storage services. The Toolkit makes use of two *storage* services constructed around the core RO model. The **RO Digital Library (RODL)** provides foundational capabilities within the Toolkit and a reference implementation of the **RO Storage and Retrieval (ROSR) API**. Extending the dLibra Digital Library software, RODL provides core functionality including *(i) Retrieval:* serving ROs to clients in multiple representations via content-negotiation including self-contained zip archives and RDF manifest descriptions, provision of indexing and query interfaces (including SPARQL) for ROs and their metadata; *(ii) Storage:* creating, editing and deleting internal and external resources and annotations

[9] http://www.wf4ever-project.org/wiki/display/docs/RO+SRS+interface+6

[10] Implementation repositories at *https://github.com/wf4ever* : *rosrs*, *workflow-runner*, *Stability-service-API*, *epnio*, *wf-ro*, *Collaboration-spheres* and *portal*.

that, alongside semantic metadata, comprise RO structures, minting identifiers (URIs) as required; *(iii) Maintenance:* adding and removing RO resources whilst ensuring consistency with metadata and manifests, and managing user identities and accounts and their relationships to ROs.

The second, **myExperiment** [3], is a well established platform for the social sharing of workflows. As a Storage Service we consider myExperiment "Packs" a form of proto-RO – conversely ROs can be considered "Packs version 2". These can be imported into the RO Digital Library, undergoing conversion to ROs through a web based tool, while myExperiment's workflows and associated social networks give input to the Recommender Service. The user interface of myExperiment also performs a role as a client within the architecture below.

Lifecycle services support the dynamic nature of workflow-centric RO resources over time. *Publication* and *Archival* are preservation states for which an RO must be preserved: the **RODL** service allows creation of a duplicate with a new state (live, snapshot, or archived) through the **RO Evolution API**. Using the roevo ontology it also captures *versioning* metadata of the subsequent relationships between these ROs, the history of which can be retrieved through the API. The **Workflow Runner** service is built upon Taverna workflow server, providing remote workflow *execution* with each run captured as an RO. Numerous workflow runs generate many ROs, so their lifespan may be short – should an RO require longer-term preservation it is exchanged between the Runner and RODL services using the **ROSR API** as implemented by both components.

Data Management and Analysis services provide an extension layer that augments Storage and Lifecycle capabilities: that generate, maintain and provide access to added-value data derived from, or related to, RO resources. Four services are implemented within the current Wf4Ever Toolkit: *(i)* **Checklist Evaluation** performs an assessment of an RO passed by its URI (and retrieved from either a local disk or from a remote service implementing the ROSR API) against a minimum information model[11] for purposes of completeness, repeatability, executability, etc.; *(ii)* **Stability Evaluation** is a derivative service, monitoring the Checklist Evaluation over time as a measure of whether an RO can maintain its original purpose whilst constituent resources change or become unavailable; *(iii)* the **Recommender** service uses keyword, content-based, collaborative filtering, and social network approaches, returning recommended users, ROs, and their aggregated resources; *(iv)* **WF-RO Transformation** is a service that, given a Taverna *t2flow* workflow bundle, generates or updates an RO encapsulating that workflow, extracting workflow description using the wfdesc and roevo ontologies respectively, and storing these resources in the RODL (or other service implementing the ROSR API).

Access and Usage Clients allow users to interact with ROs and the services enabled by them. Some, such as **Collaboration Spheres** [6] and **ROBox**, provide an interface to specific functionality (recommendations and Dropbox compatibility respectively) while others like the **RO Portal** provide a general environment for exploring and modifying ROs, and through RO-centric inter-

[11] http://purl.org/net/mim/ns

change, value-added functions from the architecture. Two other Toolbox clients highlight the interoperability advantages afforded by an RO-centric architecture. **myExperiment** has been extended to support the RO model, to query and retrieve ROs from **ROSR** compatible services and display them alongside Packs. Workflow "heritage" demonstrates this architecture in use: a "parent" workflow is downloaded by a user from myExperiment; who derives "offspring" workflows; encapsulated in ROs through WF-RO and RODL; then queried by myExperiment from RODL using roevo; this relationship - between the parent workflow and the derivative ROs - is displayed in myExperiment as an indicator of reuse.

RO Manager [7] is a command line tool for manipulating ROs on the filesystem of a scientist's workstation and synchronising them (via **ROSR**) with those preserved in RODL. It is the gateway between the local working practices of the user and the network services provided by the Toolkit, presenting storage, lifecycle, and extension functions in a local user interface, while ensuring and utilising compatibility with the RO model and Toolkit APIs.

4 Conclusions and further work

In this paper we have shown how the RO model provides a suitable basis for interoperability within a service-based architecture for management of scientific workflows, and outlined a number of implemented services from the Wf4Ever Toolkit that support scientific preservation via this mechanism of interchange – the RO model, its extensions, the REST APIs built around it, and the services and clients that implement and use these APIs. The architecture and Toolkit are a work in progress, co-evolving with user requirements as first-hand experience is gathered. With foundation services in place and our approach validated we will build upon the flexibility of the RO model and the simplified client development process afforded by our APIs; to further enhance and streamline the scientist's experience both through targeted applications that assist a specific task, and the continued RO-enabling of myExperiment with Wf4Ever services.

References

1. K. Belhajjame et al. Workflow-centric research objects: First class citizens in scholarly discourse. In *Proc. Workshop on the Semantic Publishing*, pages 1–12, 2012.
2. D. De Roure, K. Belhajjame, , et al. Towards the preservation of scientific workflows. In *Proc. 8th Intl. Conference on Preservation of Digital Objects*, 2011.
3. D. De Roure et al. The design and realisation of the virtual research environment for social sharing of workflows. *FGCS*, 25(5):561–567, 2009.
4. R. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, 2000.
5. K. Page et al. Rest and linked data: a match made for domain driven development? In *Proc. 2nd Intl. Workshop on RESTful Design*, pages 22–25, 2011.
6. C. Ruiz, G. Álvaro, et al. A framework and implementation for secure knowledge management in large communities. In *Proc. 11th Intl. Conference on Knowledge Management and Knowledge Technologies*, page 19, 2011.
7. J. Zhao, G. Kylne, et al. RO-Manager: A Tool for Creating and Manipulating Research Objects to Support Reproducibility and Reuse in Sciences. In *Proc. 2nd Intl. Workshop on Linked Science*, 2012.