Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

STREP FP7-ICT-2007-6 270192

Objective: ICT-2009.4.1b — "Advanced preservation scenarios"

D2.2v2 Design, implementation and deployment of workflow lifecycle management components - Phase II

Deliverable Co-ordinator: **XX**

Deliverable Co-ordinating Institution: **University of Manchester**

Other Authors: **XX**

This deliverable describes the second phase of delivery of workflow lifecycle management components. It includes a description of the Research Object Model, which facilitates interoperation between components; an initial Research Object Storage and Retrieval Service; RO Manager command line tool; and a definition of a model for workflow abstraction.

Document Identifier:	Wf4Ever/2013/D2.2v2/0.1	Date due:	July 31, 2013
Class Deliverable:	Wf4Ever FP7-ICT-2007-6 270192	Submission date:	June 1, 2013
Project start date	December 1, 2010	Version:	0.1
Project duration:	3 years	State:	Draft
		Distribution:	Public

Wf4Ever Consortium

This document is part of the Wf4Ever research project funded by the IST Programme of the Commission of the European Communities by the grant number FP7-ICT-2007-6 270192. The following partners are involved in the project:

Intelligent Software Components S.A. (ISOCO) – Coordinator Edificio Testa, Avda. del Partenón 16-18, 1^o, 7^a Campo de las Naciones, 28042 Madrid Spain Contact person: Jose Manuel Gómez Pérez E-mail address: jmgomez@isoco.com	**University of Manchester (UNIMAN)** School of Computer Science Oxford Road, Manchester M13 9PL United Kingdom Contact person: Carole Goble E-mail address: carole.goble@manchester.ac.uk
Universidad Politécnica de Madrid (UPM) Departamento de Inteligencia Artificial, Facultad de Informática. 28660 Boadilla del Monte. Madrid Spain Contact person: Oscar Corcho E-mail address: ocorcho@fi.upm.es	**Instytut Chemii Bioorganicznej PAN - Poznan Supercomputing and Netowrking Center (PSNC)** Network Services Department Ul Z. Noskowskiego 12-14 61704 Poznań Poland Contact person: Raul Palma E-mail address: rpalma@man.poznan.pl
University of Oxford (OXF) Department of Zoology South Parks Road, Oxford OX1 3PS United Kingdom Contact person: Jun Zhao, David De Roure E-mail address: jun.zhao@zoo.ox.ac.uk david.deroure@oerc.ox.ac.uk	**Instituto de Astrofísica de Andalucía (IAA)** Dpto. Astronomía Extragaláctica. Glorieta de la Astronomía s/n, 18008 Granada Spain Contact person: Lourdes Verdes-Montenegro E-mail address: lourdes@iaa.es
Leiden University Medical Centre (LUMC) Department of Human Genetics Albinusdreef 2, 2333 ZA Leiden The Netherlands Contact person: Marco Roos E-mail address: M.Roos1@uva.nl	

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed to the writing of this document or its parts:

- iSOCO
- OXF
- PSNC
- UNIMAN
- UPM

Change Log

Version	Date	Amended by	Changes
0.1	01-06-2013	Khalid Belhajjame	Initial outline
0.2	09-06-2013	Khalid Belhajjame	Initial draft of Section 2 on the RO model

Executive Summary

This deliverable describes the second phase of delivery of workflow lifecycle management components. These components are focused around the Wf4Ever Research Object Model (RO Model), which provides descriptions of workflow-centric ROs – aggregations of content. This model is used to structure and describe ROs which are then stored and manipulated by the components of the Wf4Ever Toolkit.

The RO Model provides a framework for describing aggregations of content along with annotations of the aggregated resources, a vocabulary for describing workflows, and a vocabulary for describing provenance. The model underwent few changes in the last year in the light of user comments. We provide here a summary of the new version of the RO model. We also present the components developed for creating and managing Research Objects: the Research Object Storage and Retrieval API (implemented as part of the Research Object Digital Library (RODL)) and a command line tool – the RO Manager. These components and services are also discussed in D1.2v3 (Wf4Ever Sandbox – Phase II), D1.3v1 (Wf4Ever Architecture – Phase II) and D1.4v1 (Reference Wf4Ever Implementation – Phase II).

One of the main development in the last year consists in incorporating research objects within the myExperiment environment to allow scientists who already use myExperiment to create, share and reuse research objects. We discuss the efforts that went into this task, and report on an activity that we conducted to convert all existing Taverna T2 workflows into ROs.

We present advanced management functions that we developed for abstracting and indexing workflows, with the aim of supporting the discovery and reuse of workflows. We present an ontology that we developed for abstracting workflows in terms of motifs that characterize data manipulation and transformation patterns, which we term motifs. We also report on a solution that we developed for indexing workflows based on the services (processes) that they use.

This deliverable should be read in tandem with D1.3v2 (Wf4Ever Architecture – Phase II), D1.4v2 (Reference Wf4Ever Implementation – Phase II), D1.2v3 (Wf4Ever Sandbox – Phase III), D3.2v2 (Design, implementation and deployment of Workflow Evolution, Sharing and Collaboration components – Phase II) and D4.2v2 (Design, implementation and deployment of Workflow Integrity and Authenticity Maintenance components – Phase II) in order to provide a complete picture of the state of the Wf4Ever Phase II components.

Contents

List of Tables

List of Figures

1 Introduction

2 The Research Object Model

The RO model consists of a family of ontologies organized into core and extensions, which we will present in this section.

2.1 RO core ontology

The Core RO Ontology provides the minimum terms that are essential to the specification of research objects. Specifically, it caters for two essential requirements by providing a container structure that can be used by the scientists to bundle the resources and material relevant for their investigation, and by enabling annotations of such a container, its resources, as well as the relationships between resources thereby making the research object interpretable and reusable.

To cater for the specification of aggregation structures, we built the Research Object Core Ontology upon the popular ORE vocabulary. ORE defines standards for the description and exchange of aggregations of Web resources. Figure 1 illustrates the main terms that constitute the Research Object Core Ontology, which we describe in what follows.

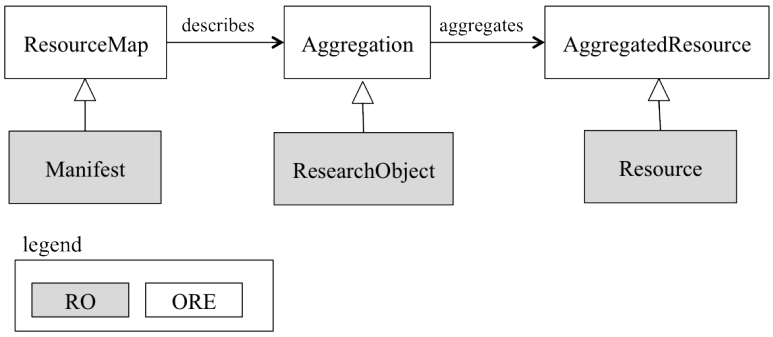

Figure 1: RO as an ORE aggregation.

- `ro:ResearchObject`[1], represents an aggregation of resources. It is a sub-class of `ore:Aggregation` and acts as an entry point to the research object.
- `ro:Resource`, represents a resource that can be aggregated within a research object and is a sub-class of `ore:AggregatedResource`. A resource can be a Dataset, Paper, Software or Annotation. Typically, a `ro:ResearchObject` aggregates multiple `ro:Resource`, and this relationship is specified using the property `ore:aggregates`.
- `ro:Manifest`, a sub-class of `ore:ResourceMap`, represents a resource that is used to describe a `ro:ResearchObject`. It plays a similar role to the manifest in a JAR or a ZIP file, and is primarily used to list the resources that are aggregated within the research object.

The second core requirement that, the Research Object Core Ontology caters for, is the descriptions of the research object and its elements. We chose the Annotation Ontology (AO) release 2.0b2 [**?**].To annotate research objects, we make use of the following three Annotation Ontology terms `ao:Annotation`[2], which represents the annotation itself; `ao:Target`, which is used to specify the `ro:Resource`(s) or `ro:ResearchObject`(s) subject to annotation; and `ao:Body`, which comprises a description of the target.

[1]The namespace of the Research Object Core Ontology `ro` is `http://purl.org/net/wf4ever/ro#`
[2]The namespace of `ao` is `http://purl.org/ao/`

In the case of research objects, we use annotations as a mean for decorating a resource (or a set of resources) with metadata information. The body is specified in the form of a set of RDF statements, which can be used to, e.g., specify the date of creation of the target or its relationship with other resources or research objects. Also, annotations can be provided for human consumption (e.g. a description of a hypothesis that is tested by a workflow-based experiment), or for machine consumption (e.g. a structured description of the provenance of results generated by a workflow run). Both kinds of annotations are accommodated using Annotation Ontology structures.

2.2 RO Extension Ontologies

We present in this section two extensions to the core Research Object ontology. The first specializes the kinds of resources that the research object can aggregate. In particular, we present extensions to specify method and experiments and the traces of their executions. The second kind of extension shows how specific metadata information, specifying the evolution of the research object over time, can be specified by specializing the Research Object core ontology.

Specifying Workflows To describe workflow research objects the workflow description vocabulary *wfdesc*[3] defines several specific resources that are involved in a workflow specification. The choice of these resources was performed by examining the commonalities between major data driven workflows, namely Taverna[4], Wings[5] and Galaxy[6], to cite a few.

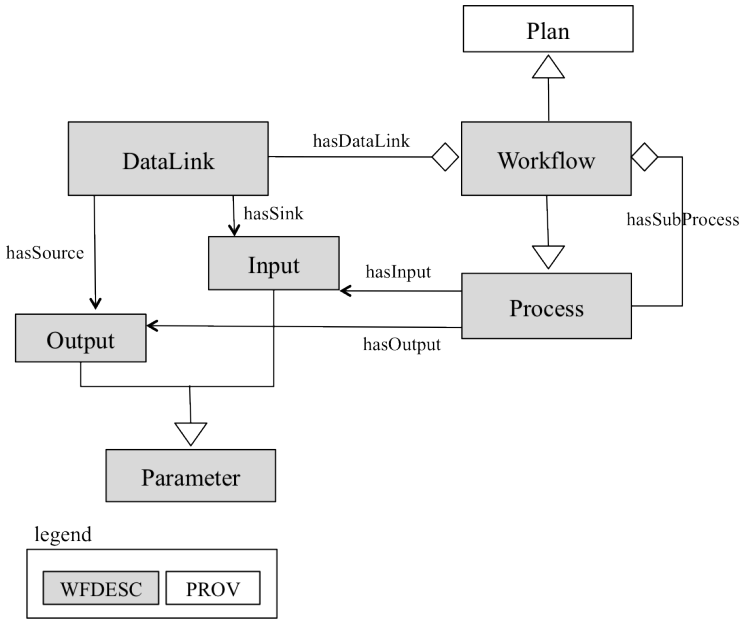

Figure 2: The *wfdesc* ontology.

Figure 2 illustrates the terms that compose the *wfdesc* ontology. Using such ontology, a workflow is described using the following three main terms:

- `wfdesc:Workflow` refers to a network in which the nodes are processes and the edges represent data links. It is defined as a subclass of the *Plan* concept from the PROV-O ontology, which represents a set of actions or steps intended by one or more agents to achieve some goals [**?**].

[3]The name space of *wfdesc* is `http://purl.org/wf4ever/wfdesc#`.
[4]http://www.taverna.org.uk
[5]http://http://wings-workflows.org
[6]http://galaxyproject.org

- `wfdesc:Process` is used to describe a class of actions that when enacted give rise to process runs. Processes specify the software component (e.g., web service) responsible for undertaking those actions.

- `wfdesc:DataLink` is used to encode the data dependencies between the processes that constitute a workflow. Specifically, a data link connects the output of a given process to the input of another process, specifying that the artifacts produced by the former are used to feed the latter.

Describing Experimental Provenance using the *wfprov* Vocabulary The *wfprov* ontology is used to describe the provenance traces obtained by enacting workflows. It is defined as an extension to the ongoing W3C PROV standard ontology - PROV-O[7].

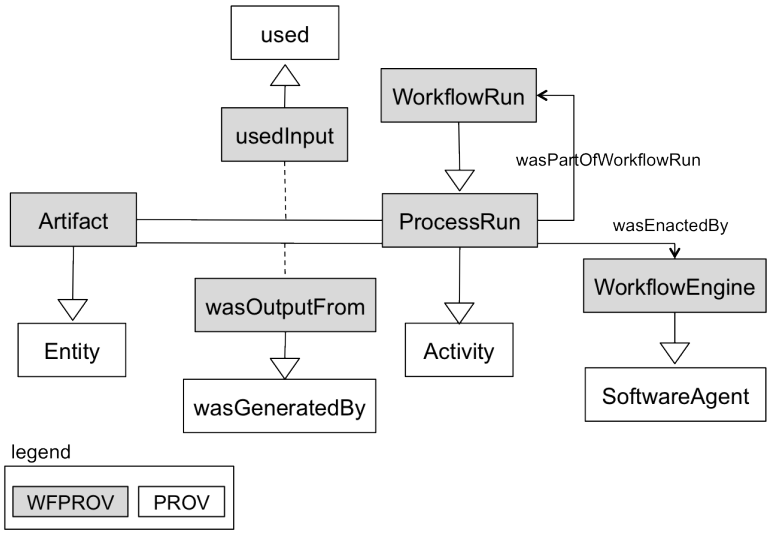

Figure 3: The *wfprov* ontology.

Figure 3 illustrates the structure of the *wfprov* ontology and its alignments with the W3C PROV-O ontology. A a workflow run (`wfprov:WorkflowRun`) represents the enactment of a given workflow. It is composed of a set of process runs (`wfprov:ProcessRun`), each representing the enactment of a process. A process run may use some artifacts (`wfprov:Artifact`) as input and generate others as output. A process run is enacted by a workflow engine (`wfprov:WorkflowEngine`), which can be seen as a PROV software agent.

By chaining the usage and generation of artifact together, the *wfprov* ontology allows scientists to trace the lineage of workflow results. For example the user can identify the input artifacts that were used to feed the wokflow run (as a whole) to obtain a given output that was generated by the workflow run.

Tracking Research Object Evolution using the *roevo* Vocabulary The *roevo* ontology is another extension to the minimal core ontology for describing an important aspect of research objects, its life cycle. To track the life cycle of a research object, we need to describe its changes at different levels of granularity, about the research object as a whole and about the individual resources. Also, we want to provide sufficient details to track the changes in order to roll back to a particular version or to quality control changes. Therefore, we need to describe when the change took place, who performed the change, and dependency relationships between the changes. Change is closely related to the provenance of a particular version of a research object or a resource. A study of the latest PROV-O ontology shows that it indeed provides all the foundational information elements for us to build the evolution ontology.

Figure 4 illustrates the core concepts of this ontology and how it extends the PROV-O:

[7]Note that the *wfprov* is reported in the W3C PROV Working Group implementation report.

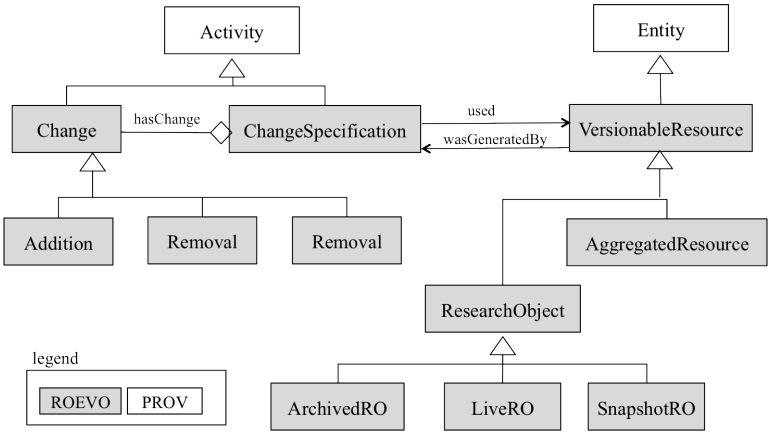

Figure 4: The *roevo* ontology extending PROV-O core terms.

- To capture different status of a research object we create three sub-classes of `ro:ResearchObject`: the `roevo:LiveRO` is a research object to capture research findings during a live investigation and it can changed, and it can either archived or snapshotted. The `roevo:ArchivedRO` can be regarded as a production research object to be preserved and archived, such as one describing findings published in an article, and it can no longer be changed; the `roevo:SnapshotRO` represents a live Research Object at a particular time.

- Both a snapshot of a live Research Object and an archived Research Object can be regarded as a versioned Research Object, i.e. a `roevo:VersionableResource`, Because it is a sub-class of `prov:Entity`, we can reuse PROV-O properties to describe the provenance or changes of this entity, such as pointing to the activity leading to any of its changes, the source research object that it was derived from, and the agent involved in its change.

- A change is a `prov:Activity`, which means that it has a start time, an end time, an input entity and a resulting entity. Also a change leading to a new Research Object can constitute a series of changes. Therefore, we have a composite `roevo:ChangeSpecification` activity, which has a number of unit `roevo:Changes`. A unit change can be adding, removing or modifying a resource or a research object. But these different changes share the same pattern of taking an input entity and producing an output entity, which can all be nicely covered by properties from PROV-O.

3 Research Object Storage and Retrieval

This section presents the components that constitute the RODL, using a UML class diagram, and show how the user can utilize RODL using a UML sequence diagram.

4 Research Object Manager

This section presents the RO manager architecture, and presents the functionalities it provides using a UML sequence diagram, if that is plausible.

5 Research Object-Enabled myExperiment

This section describes the efforts that went into incorporating research objects within myExperiment. In particular, how the notion of myExperiment pack was used as a starting point to incorporate new features/-

functionalaities. We will also discuss the diferent iterations that involved Wf4ever and Biovel users in those developments.

6 Workflow Abstraction using Motifs

This section presents the motif ontology, again using a UML class diagram, and provides an example of a workflow that was annotated using the motifs.

7 Workflow Indexation

This section shows how workflows can be indexed using the trie structure. It presents the approach as well as an example workflow that is indexed.

8 Conclusions

References