Using a suite of ontologies for preserving workflow-centric research objects

Belhajjame, K.; Zhao, J.; Garijo, D.; Gamble, M.; Hettne, K.; Palma, R.; ... ; Goble, C.

Web Semantics: Science, Services and Agents on the World Wide Web 32 (2015) 16–42

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Using a suite of ontologies for preserving workflow-centric research objects

Khalid Belhajjame [a,*], Jun Zhao [b], Daniel Garijo [c], Matthew Gamble [d], Kristina Hettne [e], Raul Palma [f], Eleni Mina [e], Oscar Corcho [c], José Manuel Gómez-Pérez [g], Sean Bechhofer [d], Graham Klyne [h], Carole Goble [d]

[a] PSL, Université Paris Dauphine, LAMSADE, France
[b] School of Computing and Communications, Lancaster University, UK
[c] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[d] School of Computer Science, University of Manchester, UK
[e] Leiden University Medical Center, Leiden, The Netherlands
[f] Poznan Supercomputing and Networking Center, Poznan, Poland
[g] iSOCO, Madrid, Spain
[h] Department of Zoology, University of Oxford, UK

ARTICLE INFO

Article history:
Received 28 August 2013
Received in revised form
11 December 2014
Accepted 26 January 2015
Available online 11 February 2015

Keywords:
Research object
Scientific workflow
Preservation
Annotation
Ontologies
Provenance

ABSTRACT

Scientific workflows are a popular mechanism for specifying and automating data-driven *in silico* experiments. A significant aspect of their value lies in their potential to be reused. Once shared, workflows become useful building blocks that can be combined or modified for developing new experiments. However, previous studies have shown that storing workflow specifications alone is not sufficient to ensure that they can be successfully reused, without being able to understand what the workflows aim to achieve or to re-enact them. To gain an understanding of the workflow, and how it may be used and repurposed for their needs, scientists require access to additional resources such as annotations describing the workflow, datasets used and produced by the workflow, and provenance traces recording workflow executions.

In this article, we present a novel approach to the preservation of scientific workflows through the application of *research objects*—aggregations of data and metadata that enrich the workflow specifications. Our approach is realised as a suite of ontologies that support the creation of workflow-centric research objects. Their design was guided by requirements elicited from previous empirical analyses of workflow decay and repair. The ontologies developed make use of and extend existing well known ontologies, namely the Object Reuse and Exchange (ORE) vocabulary, the Annotation Ontology (AO) and the W3C PROV ontology (PROVO). We illustrate the application of the ontologies for building Workflow Research Objects with a case-study that investigates Huntington's disease, performed in collaboration with a team from the Leiden University Medial Centre (HG-LUMC). Finally we present a number of tools developed for creating and managing workflow-centric research objects.

1. Introduction

As science becomes increasingly data driven, many scientists have adopted *workflows* as a means to specify and automate repetitive experiments that retrieve, integrate, and analyse datasets using distributed resources [1]. Using a workflow, an experiment can be defined as a graph where the nodes represent analysis operations, which can be supplied locally or accessible remotely, and edges specify dependencies between the operations.

* Corresponding author.
E-mail addresses: Khalid.Belhajjame@dauphine.fr (K. Belhajjame), j.zhao5@lancaster.ac.uk (J. Zhao), dgarijo@fi.upm.es (D. Garijo), matthew.gamble@cs.man.ac.uk (M. Gamble), k.m.hettne@lumc.nl (K. Hettne), rpalma@man.poznan.pl (R. Palma), ocorcho@fi.upm.es (O. Corcho), jmgomez@isoco.com (J.M. Gómez-Pérez), sean.bechhofer@cs.man.ac.uk (S. Bechhofer), Graham.Klyne.@zoo.ox.ac.uk (G. Klyne), carole.goble@cs.man.ac.uk (C. Goble).

http://dx.doi.org/10.1016/j.websem.2015.01.003

The value of a workflow definition is not limited to its original author, or indeed to the original study for which it was created. Once specified, a workflow can be re-used or repurposed by other scientists. This reuse can be as a means of understanding an experimental process, replicating a previous experimental result, or even using the workflow as a building-block in the design of new workflow-based experiments. To support this potential for reuse, public repositories such as myExperiment [2] and CrowdLabs [3] can be used by scientists to publish workflow definitions and share them over the web.

However, sharing just the workflow specifications is not always sufficient to guarantee successfully reuse. Previous empirical analysis of 92 workflows from myExperiment [4] has demonstrated that nearly 80% of the workflows suffered from *decay* in the sense that they could not be understood, or executed when downloaded. These failures were shown to be a result of one or more of the following issues:

(i) **Insufficient documentation**. The user was unable to grasp the analysis or experiment implemented by the workflow due to the lack of descriptions of its inputs, intermediate steps, and outputs.
(ii) **Missing example data**. Even in situations where the users were able to understand the overall analysis implemented by the workflow, it was difficult to determine what kind of data values to use as inputs to successfully execute that workflow.
(iii) **Volatile third-party resources**. Many workflows could not be run because the third party resources they rely on were no longer available (e.g., web services implementing their steps). For example, the SOAP web services provided by KEGG[1] to query its databases have been replaced by Rest Web Services. As a result, a large number of the workflows that use the SOAP services in myExperiment could not be run.
(iv) **Execution environment**. In certain cases, the execution of the workflow required some specific software infrastructures to be installed locally, e.g., the R statistical tool.

It is clear that in order to ensure the successful *preservation* of workflows, there is a need to change how we make them. Specifically we understand successful workflow preservation to be *the immediate and continued ability to understand, run, and reuse the experimental process described by a workflow*.

Issues 1, 2, and 4 above are all introduced at the point of the workflow's publication, through the omission of necessary supporting data or metadata. Issue 3 is instead a consequence of using 3rd party services as part of a workflow, and is a relevant issue in *workflow decay* [4]. Whilst the loss of 3rd party services is out of the control of original authors, there are a number of approaches to remedy this type of workflow decay by making use of metadata – such as additional semantic descriptions about the services used [5], or provenance information [6–8] – all of which can be either provided by the author of the workflow or automatically tracked and computed.

In light of this we propose a novel approach to workflow preservation where workflow specifications are not published in isolation, but are instead accompanied by auxiliary resources and additional metadata. Specifically we have chosen to adopt and extend the *Research Object* approach proposed in [9].

The Research Object approach defines an extendable model of data aggregation, and semantic annotation. At its core, the model allows us to describe aggregations of data and enrich that aggregation with supporting metadata. This aggregation can then be published and exchanged as a single artifact. Using this approach we have built a unit of publication that combines the workflow specification along with the supporting data and metadata required to improve preservation and the potential for reproducibility. Our implementation of workflow-centric research objects is realised as a series of ontologies that support both a core model of aggregation and the domain specific workflow preservation requirements.

In this paper we make the following contributions:

- We present a series of requirements for the data and metadata needed to accompany workflow specifications to support workflow preservation.
- We outline four ontologies that we have developed in response to those requirements, that can be used to describe Workflow-Centric Research Objects.
- We present a collection of tools that make use of those ontologies in the support and management of Workflow Research Objects.
- Finally, we present a series of competency queries that demonstrate how Workflow Research Objects support workflow preservation.

The remainder of this paper is organised as follows. We present the main requirements that guided the ontology development in Section 2. We present a case study from a Huntington's disease investigation for illustrating how the ontologies can be used (in Section 3). We present the ontologies in Section 4. We go on to present the tools we developed around them, and competency queries that can be answered using Workflow Research Objects[2] (in Section 5). We present and compare related work with ours in Section 6. Finally, we present our conclusions and future work in Section 7. The resources used in the paper are available online,[3] and the ontologies are documented online [10].

2. Requirements

Our previous work [4] has identified a need to preserve more than just the workflow specifications in order to preserve their understandability, reusability and reproducibility. Related literature on supporting preservation of software [11,12] and best practice recommendations on supporting scientific reproducibility and computing [13–15] has further confirmed the need to preserve software, data and methods in aggregate. We present 5 requirements in detail that serve to establish the type of data and metadata that we need to support workflow preservation.

R_1 *Example data inputs should be provided.* Of the 92 workflows analysed in [4], 15% of them could no longer be run because they were not accompanied with any data examples. Even when inputs were textually described, it was difficult to establish input data values to be used for their execution. Without input data, both experiment reproducibility and the ability to understand the function the workflow is inhibited.

1 http://www.kegg.jp.
2 Note that in this paper we use the terms Workflow Research Object and Research Object interchangeably.
3 http://purl.org/net/jwsRO.

R_2 *Workflows should be preserved together with provenance traces of their data results.* Provenance traces of executions allow users to track how results were produced by the workflow, and repair broken workflows [6]. Past studies have shown the usefulness of provenance information in supporting workflow reproducibility [6,16–18]. The issues described in Section 1 could all benefit from the availability of detailed provenance information: *issue* 1 by replaying how the workflow functions [16] using the complete trace of all the computational tasks taking place in the workflow; *issue* 2, by finding example inputs data used by the workflow; *issue* 3, by retrieving the intermediate results produced in the original runs to resume workflow runs from the failure point; and finally *issue* 4, by retrieving information about the original computational environment, like the OS, library dependencies as well as their versions.

Extensive provenance tracking is the focus of many reproducibility efforts, like VisTrails [16] and CDE [19], etc. This is also in line with the recommendations of several reproducibility best practice guidelines [13,14], which highlight the need of making all computational steps and parameter settings available.

A caveat to provenance is that the complexity of the traces can make it a challenge to quickly identify all the information to address the above questions, or to track as much provenance information as needed [20]. A well described workflow with good documentation can provide a complimentary solution for understanding how the workflows should work, just like documentation for software tools or code.

R_3 *Workflows should be well described and annotatable.* Insufficient documentation impairs the runnability and understandability of workflows [4]. In the software world, imprecise documentation has similarly been identified as a critical barrier [21] to code reproducibility.

A number of works have approached the issue of describing experimental process and investigations driven by different needs. Related approaches include (1) capturing all the experimental steps and entities involved in using a common vocabulary so as to facilitate an interoperable understanding across investigations [22,23], (2) capturing extensive scientific discourse information around investigations (including hypothesis, claims, evidences, and etc.) in order to achieve automated knowledge discovery and hypothesis generation [24,25], (3) modelling scientists, publications, grants, and etc. about investigations in order to enhance the discovery of collaborators across disciplines and organisations [26], and etc.

To support the documentation of workflows we see the need for:

(1) A structured description of the experimental steps carried out in a workflow in a *system-neutral language*, so that workflows from different systems can be annotated, queried and understood without relying on their specific language. Our description of the workflow therefore needs to provide a simplified high-level description, suitable for describing the steps of the workflow, but without the complexity of a fully functional and operational workflow language.

(2) Functional annotations to workflows as a whole. This is similar to the principle of documenting the "design and purpose" of software code by Wilson et al. [15], by for example, describing hypothesis to be tested by the workflow or providing sketches of the tasks to be carried out by the workflow.

(3) High-level functional annotations to the steps of a workflow, using controlled vocabularies, in order to facilitate domain-specific level understanding about what each task aims to achieve.

R_4. *Changes/Evolutions of workflows, data, and environments should be trackable.* According to our empirical study [4], volatility of third party resources accounts for 50% of the cause to workflow decay. Although changes of third party resources is not always under the author's control, mechanisms can be provided to remedy the issue. At the same time, attempting to re-run a workflow or reproduce in settings different from the original ones (like at a different time, on a different machine, using different datasets), is a common practice in scientific research. Therefore, we must provide support for users to deal with and document changes and subsequently trace through changes, so that users can (1) retrieve the original version of input data and environment setting in order to reproduce/verify the original results; (2) retrieve the different parameter configurations used to generate the different versions of outputs; and (3) identify the different changes made to the workflow specification in the process of experimenting with alternative/replacement services, different parameter settings, and etc.

This requires precise provenance tracking of workflow executions and workflow evolution. Provenance information can be very useful when the workflow is being adapted to run in a new environment, with different local libraries, operating systems or access to third party resources [16].

R_0. *Packaging auxiliary data and information with workflows*. Our requirement analysis highlights a need for publishing more than workflow specifications themselves.

Guidelines for scientific reproducibility, and scientific software development [13–15,21] have also identified a need to share more than data, method and code, to achieve scientific transparency and reproducibility.

We note however that beyond simply making these resources available, there is a need for a mechanism that *links* individual resources to the specific version of a workflow-based experiment, and describes its role in that experiment. The same file, database entry, or piece of data with an identifier, may be used in any number of experiments. It is the contextual information about the role it played that is required for understanding.

Without this important contextual information when workflows are shipped from one lab to another we may lose the link between specific versions and configurations in the sea of trials-and-errors. Being able to share all these resources and auxiliary information about them (like provenance or annotations) as a single entity, and keep this relationship information within, is therefore fundamental.

It is this need to not only aggregate content, but richly describe that aggregation that is the driving motivation for us to adopt the Research Object model for building our workflow-specific extensions.

In response to the requirements outlined above, we have developed four ontologies that support the creation of workflow-centric research objects:

- `ro`: http://purl.org/wf4ever/ro#.
- `wfdesc`: http://purl.org/wf4ever/wfdesc#
- `wfprov`: http://purl.org/wf4ever/wfprov#
- `roevo`: http://purl.org/wf4ever/roevo#.

These ontologies provide the mechanism to describe an aggregation of resources, and enrich that aggregation with the metadata required for workflow preservation.

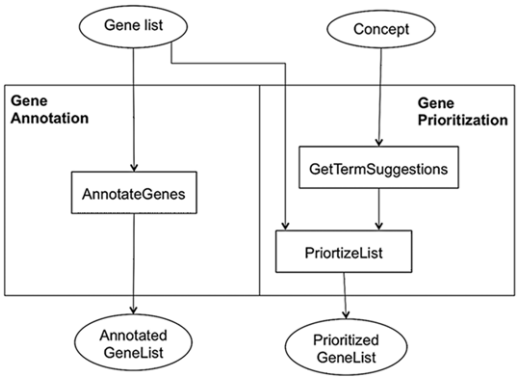

Fig. 1. A sketch of the two main analyses followed for gene interpretation: gene annotation and gene prioritisation.

In brief the ontologies and their responsibilities are as follows:

`ro`: The Research Object Ontology—designed in response to requirement R_0, is domain agnostic and enables the description of an aggregation of resources (described in Section 4.3).

`wfdesc`: The Workflow Description Ontology—in response to R_3, is used to describe the workflow specifications included in a Workflow Research Object (described in Section 4.1 and illustrated in Fig. 5).

`wfprov`: The Workflow Provenance Ontology—in response to R_1 and R_2, is used to describe the provenance traces obtained by executing workflows (described in Section 4.2 and illustrated in Fig. 5).

`roevo`: The Research Object Evolution Ontology—in response to R_4, is used for describing the evolution of Workflow Research Objects and allows to track and describe the changes made to a Workflow at different levels of granularity (described in Section 4.4 and illustrated in Fig. 8).

3. Case study: investigating the epigenetic mechanisms involved in Huntington's disease

In this section we describe our case-study, a workflow based experiment investigating aspects of Huntington's disease. This study was performed with a team of scientists from the Leiden University Medical Centre (HG-LUMC) as part of the EU FP7 Wf4ever project—a project focused on Workflow preservation.

Huntington's disease (HD) is the most commonly inherited neurodegenerative disorder in Europe, that affects 1 out of 10 000 people. Although the genetic mutation that causes HD was identified 20 years ago [27], the downstream molecular mechanisms leading to the HD phenotype are still poorly understood. Transcriptional deregulation is a prominent feature of HD with gene expression changes taking place even before first symptoms arise. Epigenetic alterations can be responsible for such transcriptional abnormalities. Linking changes in gene expression to epigenetic information might shed light on the disease aetiology.

The team from HG-LUMC analysed HD gene expression data from three different brain regions that they integrated with publicly available epigenetic data to test for overlaps between differentially expressed genes in HD and these epigenetic datasets.

The epigenetic datasets considered in this analysis were CpG islands and chromatin marks. Epigenetic changes can switch genes on and off and control which genes are transcribed. Therefore, they are suspected to be implicated in various diseases. CpG islands and the selected chromatin marks are such areas on the genome where these changes can occur. CpG islands are areas of the genome with a high concentration of the CG dinucleotides that methylation occurs and if they are located near a gene promoter can affect the expression of that particular gene. Methylated areas on the genome are responsible for turning a gene off. Chromatin marks are also playing an important role in gene transcription by making chromatin regions accessible or repressed. The genes that overlapped with each of those epigenetic datasets were interpreted and prioritised using a text mining method called *concept profile matching* [28,29]. For interpreting the gene lists they used annotations of Biological processes to enrich them and export those annotations that describe them the best. In addition to that, the gene list was further prioritised based on its relation with Huntingdon (HTT) (Fig. 1).

Fig. 1 sketches the two main analyses that the scientists followed for gene interpretation, namely *gene annotation* and *gene prioritisation*. The ellipses in the figure represent data artifacts, rectangles represent analyses steps, and the edges specify the dataflow dependencies. Given a list of genes (which are overlapping with an epigenetic feature, CpG islands or one of the four chromatin states), gene annotation is used to gather information about the genes in the list. Gene prioritisation, on the other hand, is a two-step process. Given a concept (gene or biological process) that is provided as input by the scientists, a set of terms describing that concept is retrieved. The list of terms obtained as a result is then used to prioritise the gene list. In the case of this example, the scientists were interested in prioritising the gene list against the concept representing the Huntington (HTT) concept.

3.1. Workflows

The steps illustrated in Fig. 1 were performed using three scientific workflows. Specifically, gene annotation, which consists of one step (*AnnotateGenes* in Fig. 1) was performed using the Taverna workflow *annotate_genes_biological_processes.t2flow* (illustrated in Fig. 2). On the other hand, gene prioritisation was performed using two workflows. The step *getTermSuggestions* (in Fig. 1) was performed

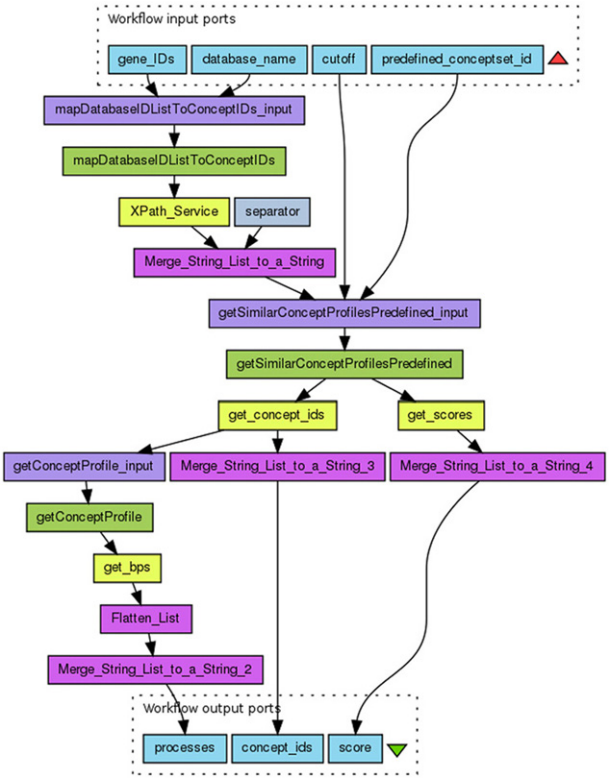

Fig. 2. A Taverna workflow used to annotate genes with biological processes.

using the workflow *getConceptsSuggestionsFromTerm.t2flow*,[4] and the *prioritizeList* step was performed using the workflow *prioritize_gene_list_related_to_a_concept.t2flow*.[5]

Fig. 2 shows the workflow *annotate_genes_biological_processes* that is used to annotate gene list, written for the Taverna workflow system. The workflow uses a local knowledge base that is mined from literature by a text mining tool [30] to enrich their knowledge about a set of input genes. Therefore, this workflow takes as an input a list of comma separated entrez gene identifiers, the database name that will be used to map the gene ids to the local concept profile identifiers (in this case of entrez gene ids, the database name should be "EG"), a cut off parameter for the number of annotations to be obtained, and an identifier for the predefined concept set id that is used in the local database (in this case "5" is used, which stands for Biological processes). The workflow can be found on myExperiment http://www.myexperiment.org/workflows/3921.html.

3.2. Creating a workflow research object

In order to preserve the workflows and their context, a Workflow Research Object is created that aggregates various information resources related to the workflows, including the original hypothesis, example inputs used for running these workflows, the workflow definitions themselves as well as metadata descriptions about them, and finally, execution traces of the workflow runs.

Fig. 3 depicts the process by which the scientists in HG-LUMC created the Research Object to encapsulate the implemented workflows and all resources associated to the *in silico* analysis.

First, a blank "pack" is created in myExperiment through the myExperiment web portal [2]. A pack is a basic aggregation of resources, which can be workflows, files, presentations, papers, or links to external resources. From the viewpoint of myExperiment users, Workflow Research Objects take the form of packs. Indeed, Workflow Research Objects can be viewed as an evolution of myExperiment packs.

As a result of creating a blank pack, a resolvable identifier is allocated by the Research Object Digital Library (RODL) [31] for the new Workflow Research Object. Where myExperiment acts as a front-end for Workflow Research Objects, the RODL acts as a back-end for their storage and retrieval.

The scientists then populate the newly created Workflow Research Object by filling in the title and the description. They also provide a text file specifying the hypothesis that they are investigating using the workflows. The hypothesis for the HD analysis is as follows:

Epigenetic phenomena are implicated in Huntington's disease gene deregulation.

Also included is a sketch that depicts the main steps of the overall investigation, specified using a graphical drawing tool.

Specifications of workflows are provided in their native language, in this case the T2 flow language of Taverna. The workflow specifications are then automatically transformed into the RDF `wfdesc` format, supported by the Workflow Research Object. This format

4 http://www.myexperiment.org/workflows/3722.html.
5 http://www.myexperiment.org/workflows/3891.html.

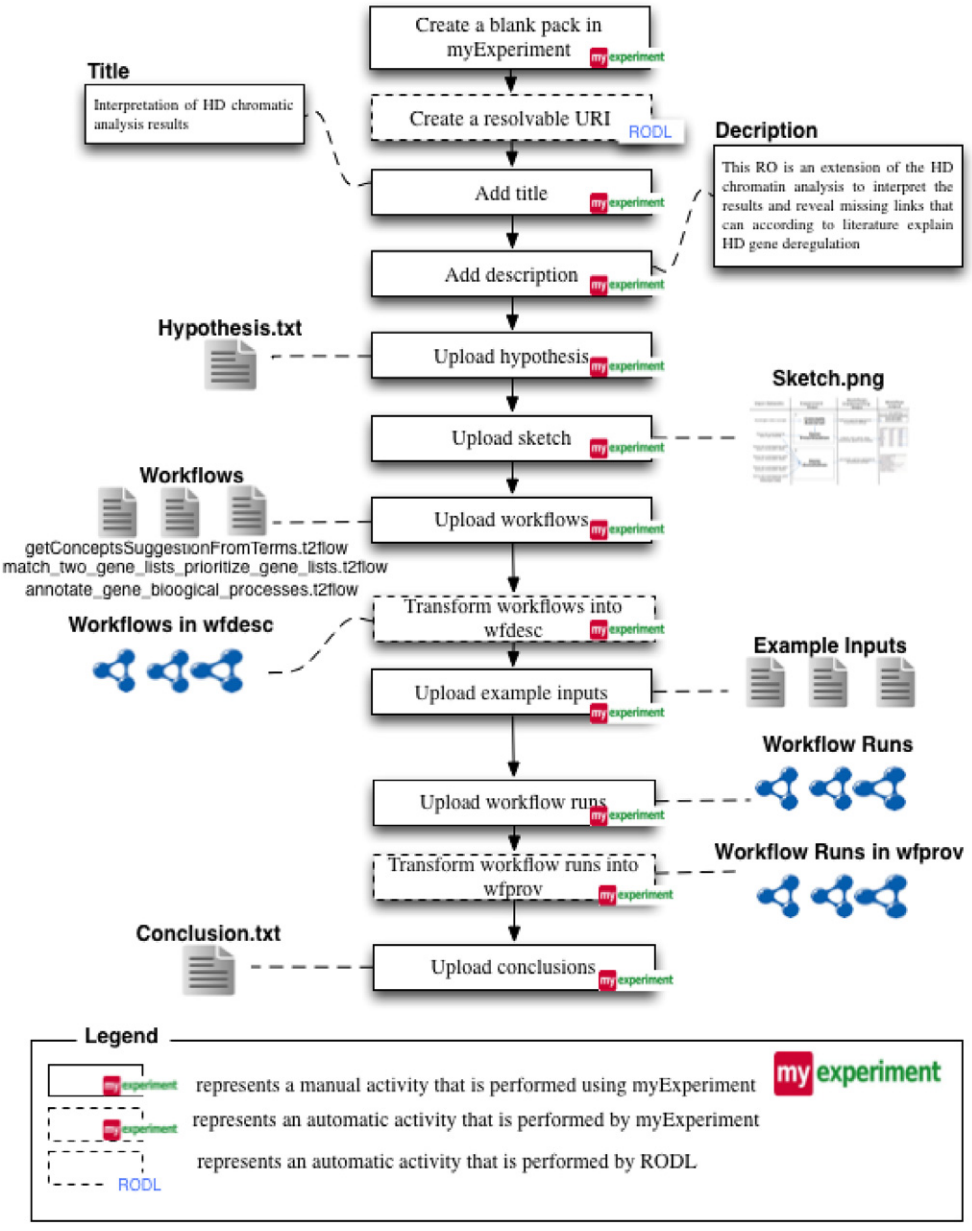

Fig. 3. A depiction of the process that the scientist went through to create the Workflow Research Object for her study.

can be used, for instance, for querying the workflows and retrieving information about their constituent steps using the SPARQL query language [32]. The scientists upload files containing example inputs that can be used to feed the execution of the uploaded workflows, and specify which files can be used as an input for each workflow. These are then followed up with files containing the traces of workflow runs, obtained by executing the uploaded workflows. The traces are again automatically transformed, this time into the `wfprov` format.

Finally, the scientists provide a file summarising the conclusions drawn from the analysis of the workflow results. The contents of the conclusions file in the HD example are as follows:

> *The analysis of the results produced by the workflows we have designed allowed us to identify both known and novel associations with Huntington, and to prioritise mechanisms that are likely to be involved in HD and are associated with epigenetic regulation. Full analysis of the results are presented in Mina et al., 2014* [33].

The Workflow Research Object created for the HD investigation can be accessed online http://purl.org/net/jwsRO446.

4. Workflow Research Object ontologies

An overview of the 4 ontologies is depicted in Fig. 4. This illustrates how our proposed models extend and link existing ontologies. The first two ontologies, `wfdesc` and `wfprov`, specify workflows and their provenance execution traces respectively while extending the W3C `prov-o` ontology [34]. Our third ontology, `ro`, aggregates workflow specifications, their provenance traces and other auxiliary resources, such as data files, images, etc. `ro` extends the `ore` ontology [35] to specify aggregations, and uses the annotation ontology, **ao** [36], to specify

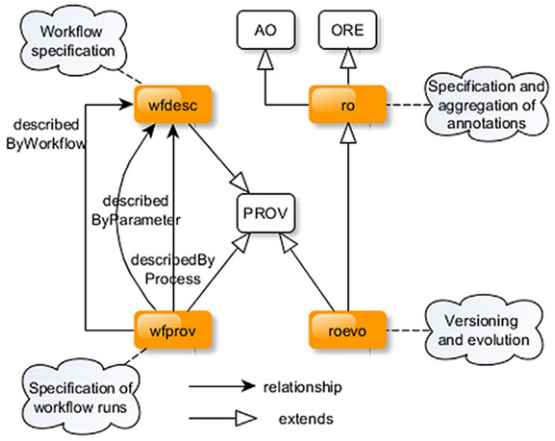

Fig. 4. An overview of the Workflow Research Object suite of ontologies and the ontologies they use and extend.

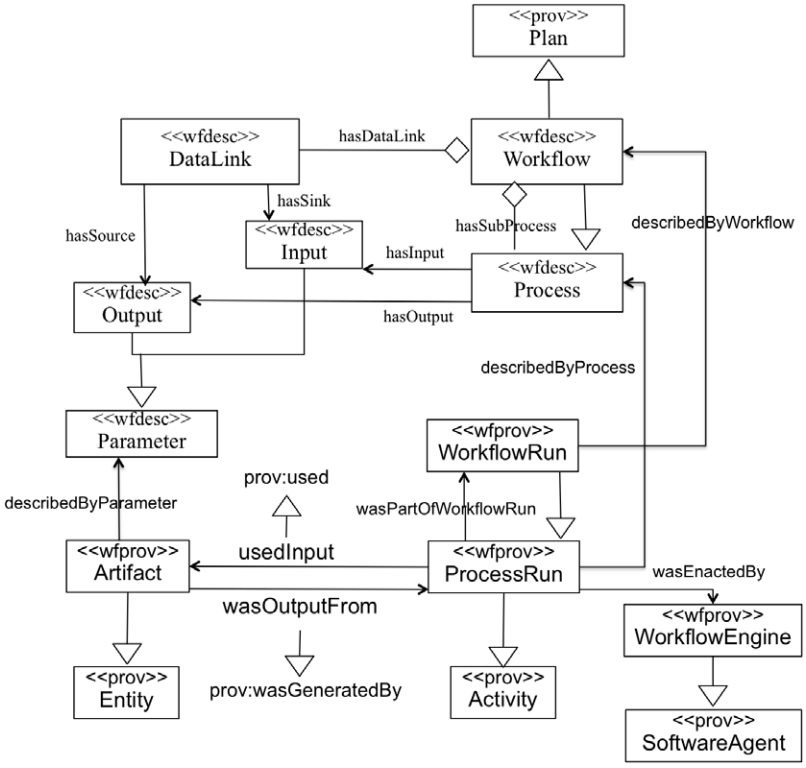

Fig. 5. The *wfdesc* and *wfprov* ontologies and their relations to `prov-o`.

annotations. Finally, the `roevo` ontology is used to specify the evolution of Workflow Research Objects. For this purpose, it extends the `prov-o` and `ro` ontologies. In what follows, we present the four Workflow Research Object ontologies in detail.

4.1. Specifying workflows using wfdesc

The workflow description vocabulary (*wfdesc*)[6] is used to describe the workflow specifications included in a Workflow Research Object. The features of the ontology were established by an examination of the core and overlapping concepts used in 3 major data driven workflow systems, Taverna [37], Wings [38] and Galaxy [39].

The upper part of Fig. 5[7] illustrates the terms that compose the *wfdesc* ontology. Using this ontology, a workflow is described using the following three main terms:

- `wfdesc:Workflow` is used to represent workflows. It is defined as a subclass of the `prov:Plan` [34].
- `wfdesc:Process` is used to represent a step in a workflow.

6 http://purl.org/wf4ever/wfdesc#.

7 We use the UML notation to illustrate the Workflow Research Object ontologies and their instance examples.

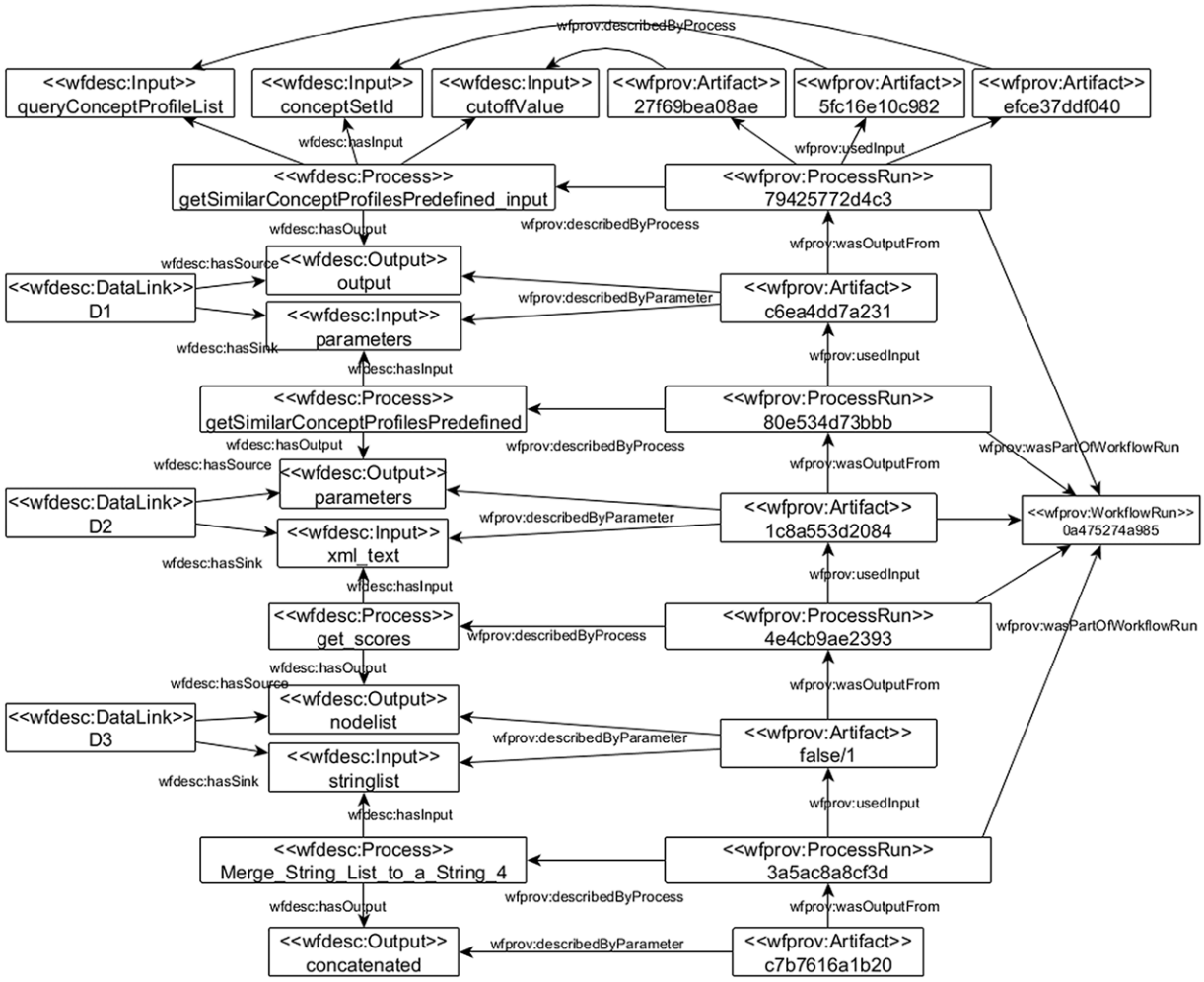

Fig. 6. Fragment of the workflow in Fig. 1-C (the central steps) represented with `wfdesc` (left) and `wfprov` (right).

- `wfdesc:DataLink` is used to specify data dependencies between the processes in the workflow. 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 as input for the latter.

Fig. 6 shows on the left side an example of how a portion of the workflow shown in Fig. 2 can be expressed using `wfdesc`.[8] There are four processes, namely *getSimilarConceptsProfile_input*, *getSimilarConceptsProfilesPredefined*, *get_scores* and *Merge_List_to_a_String_4*. These processes are connected using three data links *D1*, *D2* and *D3*. For example, the data link *D3* connects the output *nodelist* of the process *get_scores* to the input *stringlist* of the process *Merge_String_List_to_a_String_4*. The `wfdesc` RDF representation of the workflow fragment can be found in Appendix A.1, and the complete RDF file is available online at http://purl.org/net/gene_bio_process_wf.

4.2. Describing workflow runs using *wfprov*

The `wfprov` ontology is used to describe the provenance traces obtained by executing workflows.

The lower part of Fig. 5 illustrates the structure of the *wfprov* ontology and its alignments with the `prov-o` ontology.

- `wfprov:WorkflowRun` represents the execution of a workflow.
- `wfprov:ProcessRun` represents the enactment of a process and it is a subclass of `prov:Activity`.
- `wfprov:Artifact` represents an artifact that is used or generated by a given process run and it is a subclass of `prov:Entity`.

Some example *wfprov* provenance information can be found on the right side of Fig. 6, which is obtained by enacting the workflow (*Annotate_gene_list_w*) represented on the left of the figure. It shows the process runs that are part of this workflow run (*0a475274a985*). It also specifies which input files were used (*27f69bea08ae*, *efce37ddf040* and *5fc16e10c982*) and the final result obtained by the workflow fragment (*c7b7616a1b20*), along with intermediate results. Parameter values and process runs are connected to the workflow descriptions using the properties `wfprov:describedByParameter` and `wfprov:describedByProcess` respectively. All the processors are connected to the `wfprov:WorkflowRun` through the property of `wfprov:wasPartOfWorkflowRun`, so we can navigate easily through them.

The `wfprov` RDF representation of the above example can be found in Appendix A.2, and the complete RDF file is available online.[9]

[8] For simplification purposes we illustrate the encoding of a fragment of the workflow.

[9] purl.org/net/jwsWfprov.

Workflow descriptions and provenance can be used to enrich the description of the data produced by the workflow, to indicate the original datasets used by the workflow to produce the results, and the transformations that were applied to the data retrieved from the original data sources. Such information can be used for crediting the authors of the original data sources, for enriching the textual description of the datasets produced as a result of the workflow execution, or even for specifying the creation date and the authors of the dataset, as recommended by the W3C Data Catalog Vocabulary.[10]

4.3. Describing aggregations using the ro ontology

We created the `ro`[11] ontology to create Research Objects that aggregate a workflow, provenance traces, and other auxiliary resources, e.g., hypothesis, conclusions, data files, etc. In our development we used and extended the ORE vocabulary [40], which defines a standard for the description and exchange of aggregations of Web resources. Workflow Research Objects are defined in terms of three main ORE concepts:

- `ore:Aggregation`, which groups together a set of resources so that they can be treated as a single resource.
- `ore:AggregatedResource`, which refers to a resource aggregated in an `ore:Aggregation`. An `ore:AggregatedResource` can be aggregated by one or more `ore:Aggregations` and it does not have to be physically included in an `ore:Aggregation`. An `ore:Aggregation` can aggregate other `ore:Aggregations`.
- `ore:ResourceMap`, which is a resource that provides descriptions of an `ore:Aggregation`.

Using ORE, we defined the following terms for specifying Workflow Research Objects:

- `ro:ResearchObject`, represents a Workflow Research Object. It is a sub-class of `ore:Aggregation`.
- `ro:Resource`, represents a resource that can be aggregated within a Workflow Research Object and is a sub-class of `ore:AggregatedResource`. Typically, a `ro:ResearchObject` aggregates multiple `wro:Resources`, 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 Workflow Research Object.

As well as being able to aggregate resources, we require a general mechanism for annotation. For this purpose, we make use of the Annotation Ontology (AO) [36]. The `ro` ontology reuses three main Annotation Ontology concepts for defining annotations: `ao:Annotation`,[12] used for representing the annotation itself; `ao:target`, used for specifying the `ro:Resource`(s) or `ro:ResearchObject`(s) subject to annotation; and `ao:body`, which comprises a description of the target.

Workflow Research Objects use annotations as a means 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 annotate the date of creation of the target, its relationship with other resources or Workflow Research Objects, etc.

The Workflow Research Object model does not prescribe specific vocabularies to be used for annotations. Users are free to use vocabularies that they deem suitable for encoding their annotations. The intention is to keep the `ro` ontology as domain neutral as possible. Note that for the next release of the `ro` ontology, we intend to use the W3C Open Annotation model,[13] which is a feature compatible successor to the Annotation Ontology.

Fig. 7 illustrates a fragment of the RDF "manifest" file that describes the Workflow Research Object used from our case-study. The node `wfro` represents the Workflow Research Object, which is composed of a number of `ro:Resources`, e.g., a text file specifying the hypothesis, a sketch specifying the overall experiment, and t2flow files specifying Taverna workflows: `match_two_gene_lists_prioritize_gene_list.t2flow` and `explainScoresStringInput2.t2flow`. The Workflow Research Object is described using a manifest file, which extends the ORE term `ore:ResourceMap`. The figure also illustrates an annotation that is labelled by `a1`. It is used to describe the workflow `explainScoresStringInput2.t2flow` using a named graph that is encoded within the `file1.rdf` file.

4.4. Tracking research object evolution using the roevo ontology

The *roevo* ontology[14] is used for describing the evolution of Workflow Research Objects. Specifically, it allows to track and describe the changes made to a Workflow Research Object at different levels of granularity: the changes made as a whole (its creation and current status) and the changes made to the individual resources aggregated (additions, modifications and removals). The *roevo* ontology extends the `prov-o` ontology, which provides the foundational information elements to describe evolution of Research Objects.

Fig. 8 illustrates the concepts of the *roevo* ontology and how it extends `prov-o`:

- Three sub-classes of `ro:ResearchObject` have been created to capture different states of a Workflow Research Object during its life time. A `roevo:LiveRO` represents a Workflow Research Object that is being created and populated. A `roevo:ArchivedRO` can be seen as a production Workflow Research Object to be preserved and archived. Finally, a `roevo:SnapshotRO` represents a live Workflow Research Object at a particular point in time.
- `roevo:VersionableResource` represents a resource that is subject to evolution, which can be a `roevo:SnapshotRO`, a `roevo:ArchivedRO`, a `ro:Resource`, or `ro:AggregatedAnnotation`. Since we want to track the provenance of a `roevo:VersionableResource`, we consider this class to be a sub-class of `prov:Entity`.

10 http://www.w3.org/TR/vocab-dcat.
11 http://purl.org/wf4ever/ro#.
12 http://purl.org/ao/.
13 http://www.w3.org/community/openannotation.
14 http://purl.org/wf4ever/roevo.

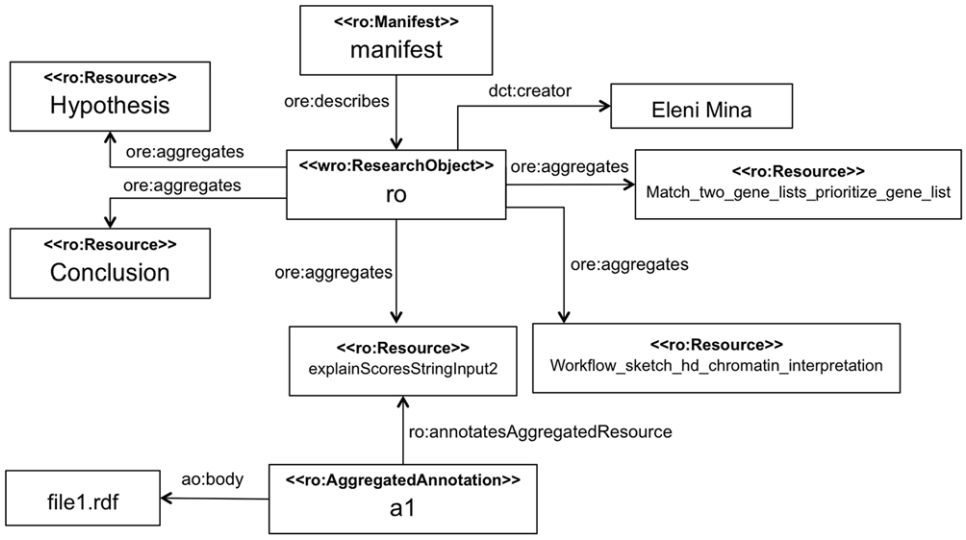

Fig. 7. An example illustrating the use of the `ro` ontology for specifying a Workflow Research Object.

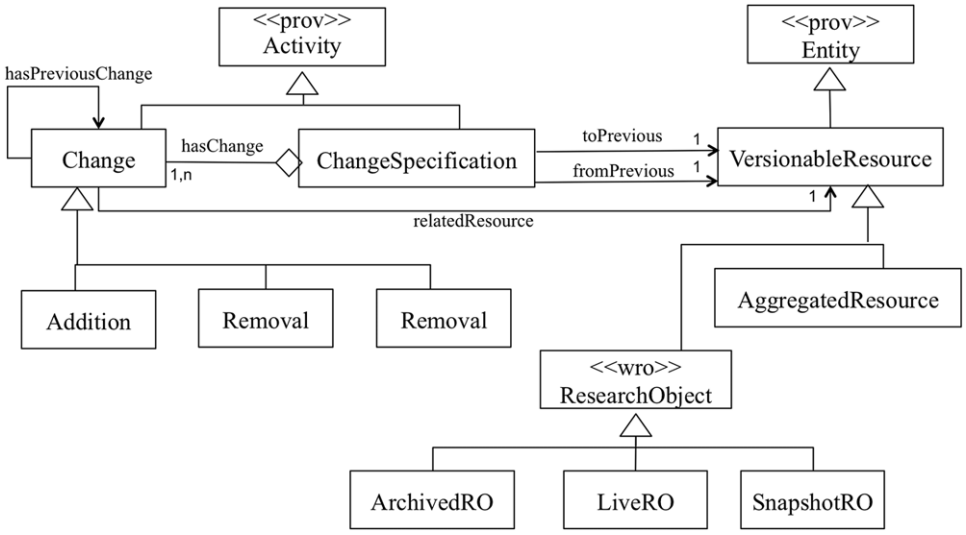

Fig. 8. `roevo` ontology for tracking changes and evolution of worflow research objects.

- `roevo:ChangeSpecification` designates a set of (unit) changes (addition, removal or update) that given a `roevo:VersionableResource` yields a new `roevo:VersionableResource` (see the object properties `roevo:from-version` and `roevo:to-version` in Fig. 8).
- `roevo:change` designates a (unit) change, which can be adding, removing or modifying a resource or a Workflow Research Object. Changes are chronologically ordered using the `roevo:hasPreviousChange` property.

To illustrate how the `roevo` ontology can be used, consider a Workflow Research Object that contains the workflow illustrated in Fig. 9. Such a workflow could not be run after a while from its creation because the web service that implements the process `explainScoresStringInput` was no longer available. To repair such a workflow, the scientist created a new Workflow Research Object that contains a new workflow obtained by replacing the process `ExplainScoresStringInput` with a process associated with an available web service that performs the same task as the unavailable one. The `roevo` ontology allows capturing the evolution of the Workflow Research Object at different granularities, as we describe below.

Fig. 10 illustrates how the evolution at the level of the Workflow Research Object can be captured using `roevo`. It specifies that the Research Object, named `data_interpretation-2-snapshot`, was revised to give rise to a new Workflow Research Object named `data_interpretation-2-snapshot-1`. Notice that we make use of the `prov-o` object property prov:wasRevisionOf. `data_interpretation-2-snapshot-1` was obtained using a change specification consisting of two (unit) changes that are ordered using the property `roevo:hasPreviousChange`. The first change consists of removing a resource, representing a file containing the specification of the Taverna workflow `annotate_genes_biological_processes.t2flow`. The second change consists in adding a file containing the specification of a new Taverna workflow `annotate_genes_biological_processes_xpath_cpids.t2flow`.

Fig. 11 illustrates how the evolution at a finer grain, i.e., at a workflow level instead of the Research Object, can be captured using `roevo`. It specifies that the workflow `annotate_genes_biological_processes_xpath_cpids.t2flow` was a revision of the workflow

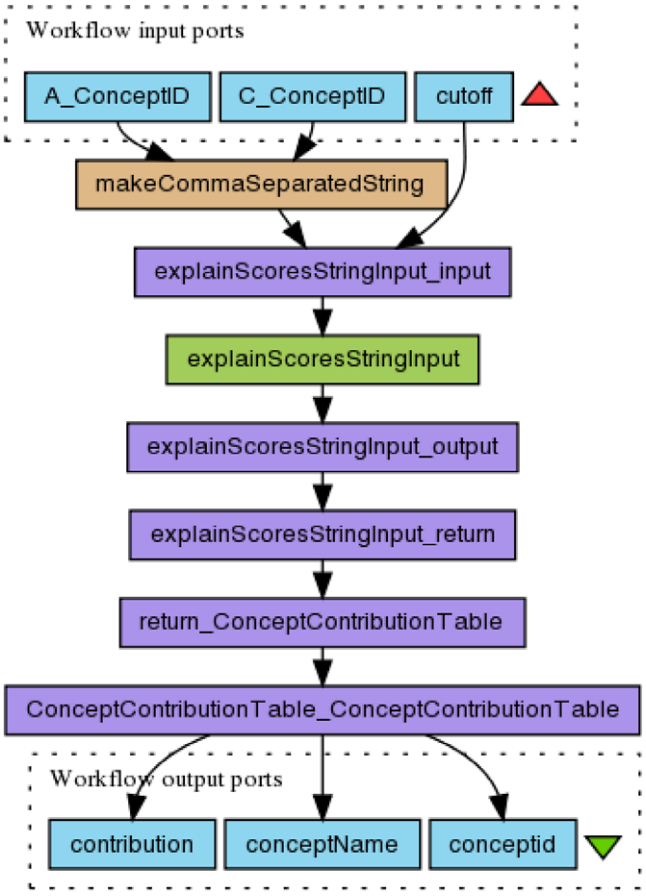

Fig. 9. An example of a Taverna workflow suffering from decay. The figure is automatically generated by the Taverna workflow system. The steps in the workflow are coloured to distinguish those steps that are implemented by web services from those that are locally implemented as Java programs or beanshell scripts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

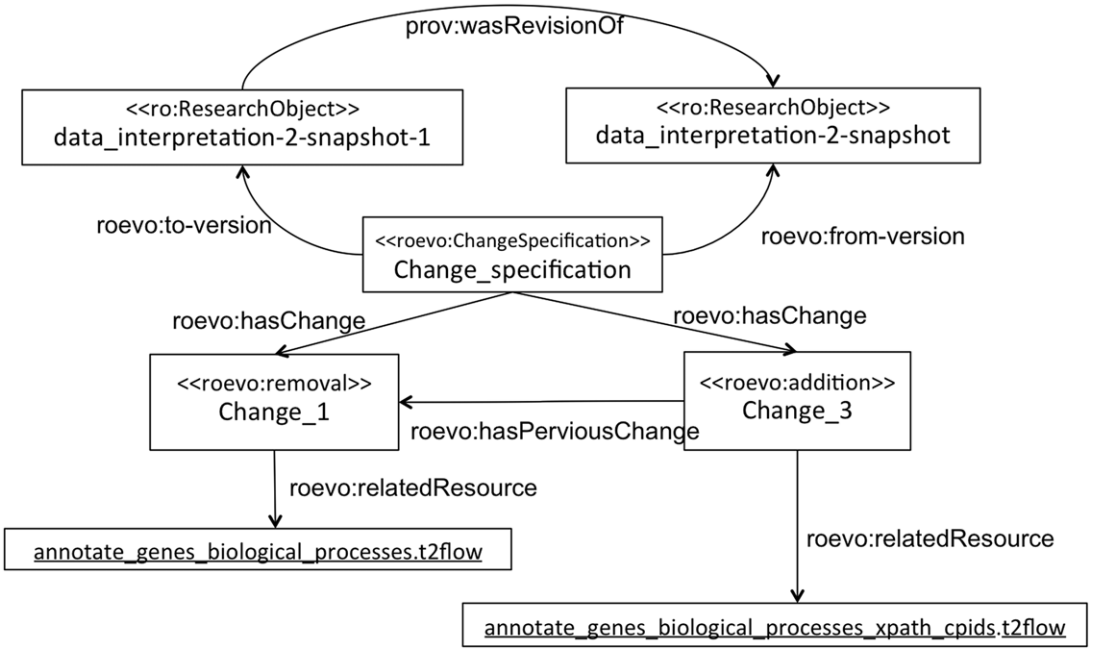

Fig. 10. Specifying Workflow Research Object evolution using *roevo*: Example 1.

`annotate_genes_biological_processes.t2flow`. Such a revision took place using a change specification that consists of 6 (unit) changes that are ordered using the `roevo:hasPreviousChange` property. The 6 changes are as follows:

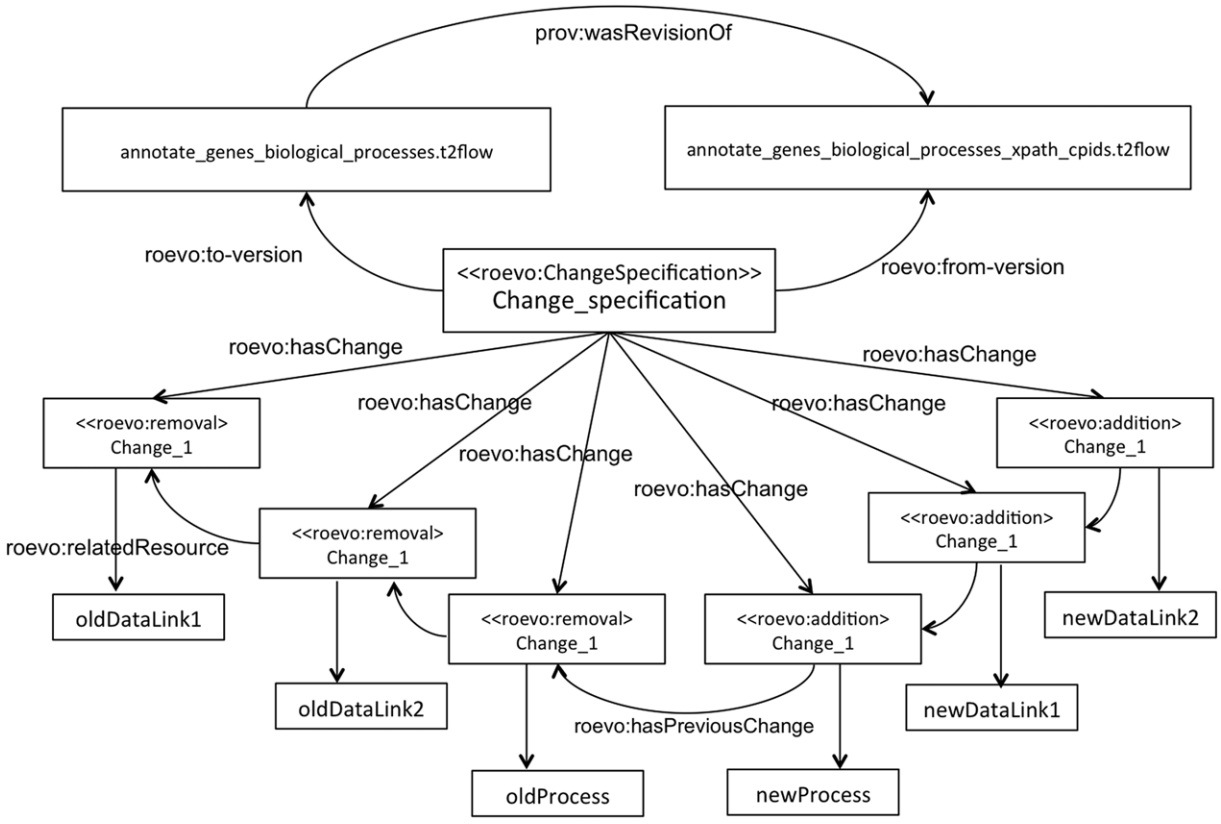

Fig. 11. Specifying Workflow Research Object evolution using *roevo*: Example 2.

(i) `change1` consists of removing the datalink, `oldDataLink1` connecting the process `explainScoresStringInput_input` and the process `explainScoresStringInput` in the workflow (see Fig. 9).
(ii) `change2` consists of removing the datalink, `oldDataLink2` connecting the process `explainScoresStringInput` and the process `explainScoresStringInput_output` in the workflow.
(iii) `change3` consists of removing `oldProcess` representing the process `explainScoresStringInput` in the workflow.
(iv) `change4` consists of adding `newProcess` to the workflow, which represents the new process that is associated with an available web service.
(v) `change5` consists of adding a datalink, `newDataLink1` connecting the process `explainScoresStringInput_input` to the newly added process.
(vi) `change6` consists of adding a datalink, `newDataLink2` connecting the new process to the process `explainScoresStringInput_output` in the workflow.

The RDF turtle listing of the above example can be found in Appendix A.3.

5. The Workflow Research Object family of tools

We have developed a suite of tools to support scientists in creating, annotating, publishing and managing Workflow Research Objects. The Research Object Manager (described in Section 5.1) is a command line tool for creating, displaying and manipulating Workflow Research Objects. The Research Object Manager incorporates the essential functionalities for Workflow Research Object management, especially by developers and a technically skilled audience used to working in a command-line environment. The Research Object Digital Library (RODL, described in Section 5.2) acts as a full-fledged back-end. RODL incorporates capabilities to deal with collaboration, versioning, evolution and quality management of Workflow Research Objects. Finally, we have also extended the popular virtual research environment myExperiment [2] to allow end-users to create, share, publish and curate Research Objects (Section 5.3). The developed tools are interoperable. For example, a user can utilise the Research Object Manager to create Research Objects, and upload them to the RODL portal or the development version of myExperiment, where it can undergo further changes.

5.1. The Research Object Manager

The Research Object Manager is a command line tool for creating, displaying and manipulating Workflow Research Objects. It is primarily designed to support a user working with Workflow Research Objects in the user's local file system. RODL and Research Object Manager can exchange Workflow Research Objects between them, using the Workflow Research Object vocabularies. The Research Object Manager also includes a checklist evaluation functionality, which is used to evaluate if a given Workflow Research Object satisfies

pre-specified properties (e.g., the input data is declared, the hypothesis of the experiment is present, the Workflow Research Object has some examples to play with, etc.).

The Research Object Manager is documented in a user guide that is available online.[15] The source code is maintained in the Wf4ever Github repository.[16]

5.2. Research object digital library (RODL)

RODL is a back-end service that does not directly provide a user interface, but rather interfaces through which client software can interact with RODL and provides different user interfaces for managing Workflow Research Objects.

The main system level interface of RODL is a set of REST APIs, including the Research Object API[17] and the Research Object Evolution API.[18]

The Research Object API, also called the Research Object Storage and Retrieval API, defines the formats and links used to create and maintain Workflow Research Objects in the digital library. Given that the semantic metadata is an important component of a Workflow Research Object, RODL supports content negotiation for the metadata resources, including formats such as RDF/XML, Turtle and TriG.

The Research Object Evolution API defines the formats and links used to change the lifecycle stage of a Workflow Research Object, to create an immutable snapshot or archive from a mutable live Workflow Research Object, and to retrieve the evolution provenance of a Workflow Research Object. The API follows the *roevo* ontology (see Section 4.4), visible in the evolution metadata generated for each state transition.

Additionally, RODL provides a SPARQL endpoint that allows queries over HTTP to the metadata of all stored Workflow Research Objects.[19]

A running instance of RODL is available for testing.[20] At the moment of writing, it holds more than 1100 Workflow Research Objects with 20 225 resources and 7393 annotations. The majority of the Workflow Research Objects have been created from existing workflow entries in myExperiment, through a bulk migration process that used services developed to transform workflows into Workflow Research Objects. The remaining Workflow Research Objects come from new experiments by scientists, partners of the Wf4ever project, from the domains of bioinformatics and astronomy.

The reference client of RODL is **the Research Object Portal (RO Portal)**,[21] developed alongside RODL to test new features and expose all available functionalities. Its main features are Research Object exploration and visualisation. The Portal uses all APIs of RODL. The development version of **myExperiment** (see Section 5.3) also uses RODL as a backend for storing packs.

5.3. Workflow Research Object-enabled myExperiment

myExperiment [2] is a virtual research environment targeted towards collaborations for sharing and publishing workflows (and experiments). While initially targeted towards workflows, the creators of myExperiment were aware that scientists needed to share more than just workflows and experiments. Because of this, myExperiment was extended to support the sharing Packs. At the time of writing, myExperiment had 337 packs. Just like a workflow, a pack can be annotated and shared. The notion of a Research Object, presented in this paper, can be viewed as an extension of the myExperiment pack. A myExperiment pack is like a folder in which the constituent resources can be virtual (not necessarily files). It allows aggregating resources, versioning and specifying the kinds of the constituent resources as well as the relationships between those resources.

In order to support complex forms of sharing, reuse and preservation, we have incorporated the notion of Workflow Research Objects into the development version of myExperiment.[22] In addition to the basic aggregation supported by packs, alpha myExperiment provides the mechanisms for specifying metadata that describes the relationships between the resources within the aggregation. For example, a user is able to specify that a given file represents a hypothesis, a workflow run obtained by enacting a given workflow, or conclusions drawn by the scientists after analysing the workflow run.

5.4. Example competency queries

To illustrate the potential of Workflow Research Objects for preservation, and the value of their structured representation, we have developed a series of competency queries. These queries are designed to evaluate our approach by demonstrating the ability to answer questions about a workflow's data and metadata, and have been drawn from the requirements outlined in Section 2.

The queries are capable of:

(i) Retrieving metadata associated with a workflow description—addressing requirement R_3

(ii) Retrieving information about the relationship between workflow descriptions and workflow runs—addressing requirement R_2

(iii) Retrieving lineage information associating the results of a workflow run with its inputs—addressing requirement R_2

(iv) Detecting differences between two versions of a Workflow Research Object—addressing requirement R_4

(v) Retrieving information about the relationship between a Workflow Research Object and the data artifacts it encompasses—addressing requirement R_1.

15 http://wf4ever.github.io/ro-manager/.

16 https://github.com/wf4ever/ro-manager.

17 http://wf4ever-project.org/wiki/display/docs/RO+API+6.

18 http://wf4ever-project.org/wiki/display/docs/RO+evolution+API.

19 http://sandbox.wf4ever-project.org/portal/sparql.

20 http://sandbox.wf4ever-project.org/rodl/.

21 http://sandbox.wf4ever-project.org/portal.

22 http://alpha.myexperiment.org/packs/.

All queries can all be seen to address requirement R_0 by being predicated on the availability of additional data or metadata.

In this section we evaluate the queries against the structured data and metadata captured in our HD case study Research Object. For each we present a description of each query, their translation into a SPARQL, and the results obtained by evaluating them.

Query 1 *Find the creator of the Workflow Research Object*. This query is useful, e.g., for the (re)-user of the workflow to identify the person to credit.

The SPARQL [32] query that can be used for answering this query can be formulated as follows:

```
PREFIX
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX wro: <http://purl.org/wf4ever/ro#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
select distinct ?name
where {
<http://sandbox.wf4ever-project.org/rodl/ROs/
data_interpretation-2/>
   dct:creator ?creator ;
           rdf:type wro:ResearchObject .
 ?creator foaf:name ?name . }
```

The results obtained by evaluating the query can be found in Appendix A.4. Specifically, the results point out that the Workflow Research Object was created by `Eleni Mina`.

Query 2 *Find the workflow used to generate the gene annotation result reported.* This workflow can be used to identify the experiment (workflow) that generated a given data (result).

The SPARQL query that can be used for answering the above query can be formulated as follows:

```
PREFIX
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX wfprov: <http://purl.org/wf4ever/wfprov#>
PREFIX wfdesc: <http://purl.org/wf4ever/wfdesc#>
select distinct ?workflow ?def
where {?output wfprov:describedByParameter ?parameter ;
prov:wasGeneratedBy ?processrun .
filter regex(str(?parameter), "concept_ids") .
?processrun rdf:type ?type; wfprov:wasPartOfWorkflowRun
    [wfprov:describedByWorkflow ?workflow ].
?workflow wfdesc:hasWorkflowDefinition
    [<http://purl.org/pav/importedFrom> ?def] }
```

Evaluating the above SPARQL query returns a list of workflows, which can be found in the SPARQL results listed in Appendix A.4.

Query 3 Find the inputs used to feed the execution of the workflow that generated a given result. This is an example of lineage query that is used to identify the input data values that contributed to a given result that was obtained as a result of a workflow execution.

The SPARQL query that can be used for answering the above query can be formulated as follows:

```
PREFIX
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX wfprov: <http://purl.org/wf4ever/wfprov#>
PREFIX wfdesc: <http://purl.org/wf4ever/wfdesc#>
PREFIX ta: <http://ns.taverna.org.uk/2012/tavernaprov/>
select distinct ?inputValue
where {?output wfprov:describedByParameter ?parameter ;
prov:wasGeneratedBy ?processrun .
filter regex(str(?parameter), "concept_ids") .
?processrun rdf:type ?type;
   wfprov:wasPartOfWorkflowRun ?workflowrun.
?workflowrun prov:used ?inputs ;
   prov:startedAtTime ?start.
?inputs ta:content ?inputValue }
```

Evaluating the above SPARQL query returns the URIs of the input data values in question (the SPARQL results can be found in Appendix A.4).

Query 4 *Find all the workflows that have been modified between the two versions of a Workflow Research Object.*

The SPARQL query that can be used for answering the above query can be formulated as illustrated below. Notice that the query retrieves the workflows that are associated with a change that is part of a change specification that is associated with the Workflow Research Object identified by the URI <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/>.

```
PREFIX
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX wfprov: <http://purl.org/wf4ever/wfprov#>
PREFIX wfdesc: <http://purl.org/wf4ever/wfdesc#>
PREFIX ta: <http://ns.taverna.org.uk/2012/tavernaprov/>
PREFIX roevo: <http://purl.org/wf4ever/roevo#>
select distinct ?workflow
where{?snapshot roevo:isSnapshotOf
<http://sandbox.wf4ever-project.org/rodl/ROs/
data_interpretation-2/> ; roevo:wasChangedBy ?spec .
?spec rdf:type roevo:ChangeSpecification ;
roevo:hasChange
[roevo:relatedResource ?workflow] .
?workflow rdf:type wfdesc:Workflow .
}
```

Evaluating the above SPARQL query returns the URIs of the following workflows: annotate_genes_biological_processes_xpath_cpids, explainScoresStringInput2, explainScoresStringInput and annotate_genes_biological_processes. The SPARQL results of the above query can be found in the Appendix in Appendix A.4.

Query 5 *Find the Workflow Research Objects that use a given gene association file as input.*

The SPARQL query presented below is used to retrieve Workflow Research Objects that use a given gene ontology association file, identified by < *http://example.com/gaf_1* >, as input. More specifically, the query retrieves the Workflow Research Objects that contain a workflow run, such that the gene ontology association file in question is used by a process run that belongs to that workflow run.

```
PREFIX wfprov: <http://purl.org/wf4ever/wfprov#>
PREFIX ore: <http://www.openarchives.org/ore/terms/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX wro: <http://purl.org/wf4ever/ro#>
select distinct ?wro ?title
where {?wro a wro:ResearchObject .
?wro dct:title ?title .
<http://example.com/gaf_1> wfprov:wasUsedBy ?processrun .
?processrun wfprov:wasPartOfWorkflowRun ?wfrun .
?wro ore:aggregates ?wfrun
}
```

6. Related work

Three parts of related work are presented in this section, including existing approaches for preserving workflows, for representing additional information about workflows driven by different motivation requirements and for representing bundle structure.

6.1. Scientific workflow preservation

Preservation of digital objects is a long studied topic in the digital preservation community. There is a growing recognition that preserving scientific or business workflows requires new features to be introduced, particularly given the dynamic nature of these objects. A few recent proposals from this community have also taken a similar approach to ours [41,42], by preserving more than process objects themselves and capturing additional contextual information about the processes, data, as well as the human actors involved in the processes. So far these works have more emphasis on preservation of software or business processes, which is a nice complement to our focus on scientific workflows.

Another aspect of workflow preservation is to provide the infrastructure to support the enactment and execution of workflows in the long term. Virtual Machines can be used to package up all the original settings and dependency libraries that are need to re-enact a workflow. Similarly packaging tools such as Docker [21] or ReproZip [17], can also help users to create relatively lightweight packages that include all the dependencies required to reproduce a workflow or a computational experiment. We too have a zip-based serialisation of our Workflow Research Objects described in the RO Bundle Specification [43]. These approaches differ with respect to ours in that they lack a structured description of the aggregation. As a result they lack a convenient mechanism to attach arbitrary annotations. They are also limited to aggregating resources that can be directly serialised, and lack the ability to describe an aggregation that includes remote resources, such as large third party databases.

A number of existing scientific workflow systems take a similar approach to ours in making use of provenance tracking for supporting reproducibility and enabling workflow preservation (as in e.g. VisTrails [44], Wings [38], etc.). Provenance information is particularly helpful when a workflow can no longer be executed [6,45], due to changes of third party resources used by the workflow or the execution environment (like the OS or dependency libraries).

6.2. Workflow/experiment descriptions

Analogous to software documentation, descriptions about workflows, like its main function and how it is divided in smaller steps, are also critical for understanding and preserving workflows. This is particularly useful when information like provenance is unavailable, incomplete or incomprehensible. Existing workflow systems use different languages to specify their workflows, which presents a challenge for interpreting a workflow description or querying workflow execution traces. Several attempts have been made to represent workflows from different systems in a unified language, but driven by different requirements than ours. For example, the IWIR model [46] was designed as an interchange language to make interoperable workflow templates among workflow systems. In our work, we focus on the descriptions of workflows, their steps and resources for their proper preservation, leaving out of scope whether the template can be imported by another workflow system or not.

Other related efforts are D-PROV [47] and OPMW [48], developed in parallel to our vocabularies. Their scope is similar to ours, but the complexity of the workflow patterns that is covered by each model is different. This is partially due to the different types of workflow systems that were used to drive their design requirements. D-PROV aims at representing complex scientific workflows which may include loops and optional branches. OPMW takes a simple approach by modelling just pure data flow workflows. Driven by our requirements, `wfdesc` does not cover the patterns of loops or branches, which are uncommon in the majority of scientific workflow systems. But it does provide descriptions for sub-workflows, i.e. nested workflows, included as part of a given workflow, which is a pattern that is not covered by OPMW.

`wfdesc` is essentially aimed for capturing the core structure of scientific workflows. If one needs to capture the bigger context, for example, about the scientific experiments or investigations, some existing vocabularies can be used for this purpose. For example, OBI (Ontology for Biomedical Investigations) and the ISA (Investigation, Study, Assay) model are two widely used community models from the life science domain for describing experiments and investigations. OBI provides common terms, like investigations or experiments to describe investigations in the biomedical domain [22]. It also allows the use of domain-specific vocabularies or ontologies to characterise experiment factors involved in the investigation. ISA structures the descriptions about an investigation into three levels: *Investigation*, for describing the overall goals and means used in the experiment, *Study* for documenting information about the subject under study and treatments that it may have undergone, and *Assay* for representing the measurements performed on the subjects. We have shown how the ISA framework can be used together with Research Object to capture the bigger context about a scientific investigation and boost the reproducibility of its results [49].

6.3. Scientific investigation preservation and packaging

The Knowledge Engineering from Experimental Design (KEfED) model aims to capture more than the process of a scientific investigation. The model provides a formalism of the process of observational reasoning and interpretational reasoning [50]. It is driven by the need for enabling reasoning over scientific observations by curating the observations and the process leading to the experimental results. While KEfED allow designing workflow-like processes, it is not build on a standard vocabulary. Moreover, it does not capture the evolution of workflow descriptions.

The Core Scientific Metadata Model (CSMM) [51] is a model to organise data by studies. It is aimed to capture high level information about scientific studies and the data that they produce. Currently it is deployed and used in data management infrastructure developed for the large scale scientific facilities, such as the ISIS Neutron Source [52] and the Diamond Light Source [53]. The model provides a hierarchical way to manage scientific investigations, by its research programme, projects and studies, and a way to categorise datasets into collections and files and associate them with individual investigations. Compared with Workflow Research Objects, CSMM does not provide constructs for specifying workflows or capturing their provenance traces.

6.4. Representation of packaging structure

There are several efforts that have been proposed to allow scientists packaging resources that are relevant to a given investigation. For example, scientific Publishing Packages (SPP) [54] are compound digital objects that encapsulate a collection of digital scientific objects, including raw data, derived products, algorithms, software and textual publications, in order to provide a context for the raw data. Its initial goal was to enable digital libraries to consume all these diverse information objects related to the scientific discovery process as one compound digital object. Its model has a strong notion of data lineage, enabling the expression of provenance of derived data results. However, there is no large adoption of this work and no active development exists to our knowledge. Unlike Workflow Research Objects, SPP does not cater for the description of workflows, their provenance traces, or their evolution.

ReproZip [55] is a tool that record workflows of command-line executions and associated resources, including files, dependency libraries, and variables. It then create package that can be used to rerun and verify the reproducibility of such workflows. Compared with Workflow Research Objects, Reprozip is confined to capturing command-line executions that invoke local programs. In Workflow Research Objects, we are targeting workflows that make use of distributed services that are not necessarily accessible locally. Moreover, Reprozip does not capture information about the evolution of workflows over time. They adopt proprietary language for workflow specifications.

Provenance-To-Use (PTU) [18] is similar to ReproZip. PTU relies on a user-space tracking mechanism for better portability instead of using a kernel-based provenance tracing mechanism. Similar to ReproZip, PTU adopts proprietary language for specification, and do not capture evolution of the workflow specification.

Science Object Linking and Embedding (SOLE) is a system that allows linking articles with science objects [56]. A science object can be a source code of a software, a dataset or a workflow. SOLE allows the reader (curator) to specify human-readable tags that links the paper with science objects. It transforms each tag into a URI and points to a representation of the corresponding science object. While their objective is similar to ours, the authors of SOLE take the view that the scientific article is the main object that contains links to other (science) objects. In our case, we focus on scientific workflows and link them to other resources, e.g., their provenance traces.

Our goal of workflow preservation is also related to facilitating reproducibility. It is aimed to complement many other existing efforts, which look at reproducibility, through policy and infrastructure (such as runmycode.org [57]), preservation of computation environment (such as SHARE [58]), the creation of executable papers (e.g., Utopia [59], Sweave [60], or iPython [61]), organisation of actual reproducibility case studies or assessment (e.g., the Reproducibility Initiative[23] or Mozilla Science Code Review [62]).

To enable reproducibility, the above proposals tend to opt for having locally all the data sources and computational tools necessary for executing the steps of the computation. While this approach is desirable, it is not always possible. In many cases, scientists want to use datasets and tools that are remote and cannot be locally deployed, either because the providers of those datasets and tools do not wish to provide access to their resources or because they are large or computationally expensive and cannot be deployed locally. Our approach allows scientists to describe the use of remote resources to perform their analysis, and the means to gather information about those resources and their relationships. While this approach does not guarantee full reproducibility, we think that it is more realistic, and is a step towards enabling reproducibility.

7. Conclusions

We have presented in this paper a novel approach to scientific workflow preservation that makes use of a suite of ontologies for specifying Workflow Research Objects. These Research Objects contain workflow specifications, provenance traces obtained by executing the workflows, information about the evolution of the workflow Research Object and its components elements, and annotations describing the aggregation as a whole using existing ontologies. We have also reported on available tools that can be used to create and preserve Workflow Research Objects through repositories like myExperiment.

While the notion of Workflow Research Object was initially developed as part of the Wf4ever project, its ethos, models and tools are being adopted and exploited by other communities, such as digital preservation (e.g. the EU Scape[24] project or Timbus[25] project) or workflow-based scientific research (e.g., the EU BioVel[26] project). In our ongoing work, we seek to collaborate with these communities, as well as others, such as Open Access publishers (e.g., GigaScience[27]) and Digital Libraries (e.g. FigShare[28] or Dataverse [63]) to improve the Workflow Research Object concept and vocabularies. We also intend to align our ontologies with existing similar standards and initiatives, such as ISA, OPMW and D-PROV.

We believe that the work presented in this paper has the potential to:

- Facilitate the process by which they package and annotate the resources necessary for preserving their scientific workflows.
- Encourage them to reuse existing workflows. For example, users will have elements that allow them to understand the workflow that they will be reusing, e.g., example inputs and provenance traces.
- Emphasise the importance of associating datasets, with computations (workflows), their provenance, and the people involved. As such, we think that the work presented in this paper has the potential of promoting data citation and its associated advantages, such as encouraging data sharing, tracking data usage, encouraging enriching publications, assuring long-term availability of data and increasing trust in research findings.

Acknowledgements

We warmly thank Stian Soiland-Reyes from the University of Manchester who was instrumental in the development of the ontologies, and Pinar Alper from the same university who gave us useful feedback on the wfdesc and the wfprov ontologies. We would also like to thank the members of the Wf4ever and myGrid projects, and the anonymous reviewers for their comments. The research reported in this paper was supported by the Wf4ever project (EU FP7 STREP 270192) and the myGrid platform for e-Biology (award EP/G026238/1).

Appendix

In this section, we present RDF listings of the examples presented in this paper, as well as the results obtained for competency queries that we ran. Note that some URIs were long, and we had to show them on more than one line in the listings. The URIs in question are identifiers generated by the Taverna workflow system. Rather than inventing our own identifiers, we tried as much as possible to reuse the existing identifiers. Note also that such URIs are not resolvable.

23 https://www.scienceexchange.com/reproducibility.
24 http://www.scape-project.eu.
25 http://timbusproject.net.
26 http://www.biovel.eu.
27 http://www.gigasciencejournal.com.
28 http://figshare.com.

A.1. wfdesc RDF example

The following code represents the wfdesc RDF representation of the workflow fragment shown in Fig. 6

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix wfdesc: <http://purl.org/wf4ever/wfdesc#> .
@prefix wf4ever: <http://purl.org/wf4ever/wf4ever#> .
@base <http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/
Annotate_gene_list_w/>.

#Inputs/Outputs
<processor/getSimilarConceptProfilesPredefined_input/in/queryConceptProfileList>
        a       wfdesc:Input ;
        rdfs:label "queryConceptProfileList" .
<processor/getSimilarConceptProfilesPredefined/out/parameters>
        a       wfdesc:Output  ;
        rdfs:label "parameters" .
<processor/get_scores/out/nodelist>
        a       wfdesc:Output  ;
        rdfs:label "nodelist" .
<processor/get_scores/in/xml_text>
        a       wfdesc:Input  ;
        rdfs:label "xml_text" .
<processor/Merge_String_List_to_a_String_4/in/stringlist>
        a       wfdesc:Input  ;
        rdfs:label "stringlist" .
<processor/Merge_String_List_to_a_String_4/out/concatenated>
        a       wfdesc:Output  ;
        rdfs:label "concatenated" .
## Processors
<processor/getSimilarConceptProfilesPredefined_input/>
        a                   wfdesc:Process  ;
        rdfs:label "getSimilarConceptProfilesPredefined_input" ;
        wfdesc:hasInput   <processor/getSimilarConceptProfilesPredefined_input/in/queryConceptProfileList> ,
                          <processor/getSimilarConceptProfilesPredefined_input/in/conceptSetId> ,
                           <processor/getSimilarConceptProfilesPredefined_input/in/cutoffValue> ;
        wfdesc:hasOutput  <processor/getSimilarConceptProfilesPredefined_input/out/output> .
<processor/getSimilarConceptProfilesPredefined/>
        a                    wfdesc:Process , wf4ever:SOAPService ;
        rdfs:label                  "getSimilarConceptProfilesPredefined" ;
        wf4ever:rootURI             "http://sequenomics.biosemantics.org/"^^<http://www.w3.org/2001/XMLSchema#anyURI> ;
        wf4ever:wsdlOperationName  "getSimilarConceptProfilesPredefined" ;
        wf4ever:wsdlURI             "http://sequenomics.biosemantics.org/ConceptProfileMiningService?wsdl"
                                     ^^<http://www.w3.org/2001/XMLSchema#anyURI> ;
        wfdesc:hasInput             <processor/getSimilarConceptProfilesPredefined/in/parameters> ;
        wfdesc:hasOutput            <processor/getSimilarConceptProfilesPredefined/out/parameters> .
<processor/get_scores/>
        a                   wfdesc:Process  ;
        rdfs:label        "get_scores" ;
        wfdesc:hasInput   <processor/get_scores/in/xml_text> ;
        wfdesc:hasOutput  <processor/get_scores/out/nodelist> .
<processor/Merge_String_List_to_a_String_4/>
        a                   wf4ever:BeanshellScript , wfdesc:Process ;
        rdfs:label "Merge_String_List_to_a_String_4" ;
        wf4ever:script    "String seperatorString = \"\\n\";\nif (seperator != void) {\n\tseperatorString =
                           seperator;\n}\nStringBuffer sb = new StringBuffer();\nfor (Iterator i =
                           stringlist.iterator(); i.hasNext();) {\n\tString item = (String) i.next();\n
                           \tsb.append(item);\n\tif (i.hasNext()) {\n\t\tsb.append(seperatorString);\n\t}\n
                           }\nconcatenated = sb.toString();\n" ;
        wfdesc:hasInput   <processor/Merge_String_List_to_a_String_4/in/stringlist> ;
        wfdesc:hasOutput  <processor/Merge_String_List_to_a_String_4/out/concatenated> .
##Datalinks
<datalink?
  from=processor/getSimilarConceptProfilesPredefined_input/out/output&
  to=processor/getSimilarConceptProfilesPredefined/in/parameters>
        a                   wfdesc:DataLink ;
        wfdesc:hasSink    <processor/getSimilarConceptProfilesPredefined/in/parameters> ;
        wfdesc:hasSource  <processor/getSimilarConceptProfilesPredefined_input/out/output> .
<datalink?from=in/predefined_conceptset_id&to=processor/getSimilarConceptProfilesPredefined_input/in/
conceptSetId>
        a                   wfdesc:DataLink ;
        wfdesc:hasSink    <processor/getSimilarConceptProfilesPredefined_input/in/conceptSetId> ;
        wfdesc:hasSource  <in/predefined_conceptset_id> .
<datalink?from=processor/getSimilarConceptProfilesPredefined/out/parameters&to=processor/get_scores/in/xml_text>
        a                   wfdesc:DataLink ;
        wfdesc:hasSink    <processor/get_scores/in/xml_text> ;
        wfdesc:hasSource  <processor/getSimilarConceptProfilesPredefined/out/parameters> .
<datalink?from=processor/get_scores/out/nodelist&to=processor/Merge_String_List_to_a_String_4/in/stringlist>
        a                   wfdesc:DataLink ;
```

```
        wfdesc:hasSink    <processor/Merge_String_List_to_a_String_4/in/stringlist> ;
        wfdesc:hasSource  <processor/get_scores/out/nodelist> .
<datalink?from=processor/Merge_String_List_to_a_String_4/out/concatenated&to=out/score>
        a                 wfdesc:DataLink ;
        wfdesc:hasSink    <out/score> ;
        wfdesc:hasSource  <processor/Merge_String_List_to_a_String_4/out/concatenated> .
```

A.2. wfprov RDF example

The following code represents the wfprov RDF representation of the fragment of the workflow shown in Fig. 6.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dct:   <http://purl.org/dc/terms/> .
@prefix wfdesc: <http://purl.org/wf4ever/wfdesc#> .
@prefix wfprov: <http://purl.org/wf4ever/wfprov#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix : <http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/
Annotate_gene_list_w/> .
@base <http://ns.taverna.org.uk/2011/run/666bd1c0-9c70-48a0-bb82-0a475274a985/>.

<process/16509e5c-b0c6-425f-90af-3a5ac8a8cf3d/>
        prov:startedAtTime           "2014-02-24T13:29:39.511+01:00"^^xsd:dateTime ;
        prov:qualifiedAssociation    _:b171 ;
        prov:qualifiedEnd            _:b172 ;
        wfprov:usedInput             <list/c24b8892-afae-4a8d-864d-e3ca3af02897/false/1> ;
        prov:used                    <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/list/
        c24b8892-afae-4a8d-864d-e3ca3af02897/false/1> ;
        rdfs:label                   "Processor execution Merge_String_List_to_a_String_4"@en ;
        prov:wasAssociatedWith       <#taverna-engine> ;
        prov:endedAtTime             "2014-02-24T13:29:39.802+01:00"^^xsd:dateTime ;
        wfprov:describedByProcess    :processor/Merge_String_List_to_a_String_4 ;
        rdf:type                     wfprov:ProcessRun ;
        wfprov:wasPartOfWorkflowRun  <> ;
        prov:qualifiedUsage          _:b119 ;
        wfprov:wasEnactedBy          <#taverna-engine> ;
        prov:qualifiedStart          _:b4 .

<http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/list/
c24b8892-afae-4a8d-864d-e3ca3af02897/false/1>
        wfprov:wasOutputFrom         <process/3f5678c5-b70b-45b9-bfaf-4e4cb9ae2393/> ;
        prov:hadDictionaryMember     _:b48 ;
        wfprov:describedByParameter  :processor/get_scores/out/nodelist ;
        prov:hadMember               <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bfe6a34c-ab66-4e81-9659-a2f7ded388ca> ;
        prov:hadDictionaryMember     _:b139 ;
        prov:hadMember               <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bef18979-4c44-47d4-a535-42a46e1eae80> ;
        prov:qualifiedGeneration     _:b217 ;
        rdf:type                     prov:Dictionary ;
        rdf:type                     wfprov:Artifact ;
        prov:hadDictionaryMember     _:b242 ;
        wfprov:describedByParameter  :processor/Merge_String_List_to_a_String_4/in/stringlist ;
        rdf:type                     prov:Entity ;
        prov:hadDictionaryMember     _:b243 ;
        prov:wasGeneratedBy          <process/3f5678c5-b70b-45b9-bfaf-4e4cb9ae2393/> ;
        prov:hadDictionaryMember     _:b230 ;
        rdf:type                     prov:Collection ;
        prov:hadMember               <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        97099d75-f9f4-434f-8a2a-67393e1fdee6> ;
        prov:hadMember               <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        caafda42-a850-465d-a3ed-0c3367fded7a> ;
        prov:hadMember               <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        9520996f-a4cb-478b-b8ee-0a51189c46fe> .

<process/3f5678c5-b70b-45b9-bfaf-4e4cb9ae2393/>
        prov:qualifiedUsage          _:b200 ;
        prov:qualifiedEnd            _:b22 ;
        prov:used                    <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        ecce5803-c8e4-4ffb-b4ee-1c8a553d2084> ;
        prov:qualifiedStart          _:b201 ;
        wfprov:wasEnactedBy          <#taverna-engine> ;
        wfprov:wasPartOfWorkflowRun  <> ;
        prov:wasAssociatedWith       <#taverna-engine> ;
        prov:endedAtTime             "2014-02-24T13:29:39.243+01:00"^^xsd:dateTime ;
        wfprov:describedByProcess    :processor/get_scores/ ;
        wfprov:usedInput             <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        ecce5803-c8e4-4ffb-b4ee-1c8a553d2084> ;
        rdfs:label                   "Processor execution get_scores"@en ;
        rdf:type                     wfprov:ProcessRun ;
```

```
        prov:qualifiedAssociation    _:b202 ;
        prov:startedAtTime           "2014-02-24T13:29:37.772+01:00"^^xsd:dateTime .

<http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/ecce5803-c8e4-4ffb-b4ee-1c8a553d2084>
        tavernaprov:content          <intermediates/ec/ecce5803-c8e4-4ffb-b4ee-1c8a553d2084.txt> ;
        wfprov:describedByParameter  :processor/get_concept_ids/in/xml_text ;
        wfprov:describedByParameter  :processor/get_scores/in/xml_text ;
        wfprov:describedByParameter  .processor/getSimilarConceptProfilesPredefined/out/parameters ;
        wfprov:wasOutputFrom         <process/494d6358-7422-4b6a-af37-80e534d73bbb/> ;
        prov:qualifiedGeneration     _:b60 ;
        prov:wasGeneratedBy          <process/494d6358-7422-4b6a-af37-80e534d73bbb/> ;
        rdf:type                     wfprov:Artifact ;
        rdf:type                     prov:Entity .

<process/494d6358-7422-4b6a-af37-80e534d73bbb/>
        prov:used                    <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        49dd6b6d-49ee-43be-b854-c6ea4dd7a231> ;
        prov:qualifiedUsage          _:b276 ;
        prov:startedAtTime           "2014-02-24T13:29:11.462+01:00"^^xsd:dateTime ;
        prov:wasAssociatedWith       <#taverna-engine> ;
        prov:qualifiedAssociation    _:b190 ;
        wfprov:describedByProcess    :processor/getSimilarConceptProfilesPredefined ;
        rdfs:label                   "Processor execution getSimilarConceptProfilesPredefined"@en ;
        wfprov:wasPartOfWorkflowRun  <> ;
        prov:qualifiedStart          _:b254 ;
        prov:qualifiedEnd            _:b277 ;
        prov:endedAtTime             "2014-02-24T13:29:37.632+01:00"^^xsd:dateTime ;
        wfprov:wasEnactedBy          <#taverna-engine> ;
        wfprov:usedInput             <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        49dd6b6d-49ee-43be-b854-c6ea4dd7a231> ;
        rdf:type                     wfprov:ProcessRun .

<http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/49dd6b6d-49ee-43be-b854-c6ea4dd7a231>
        tavernaprov:content          <intermediates/49/49dd6b6d-49ee-43be-b854-c6ea4dd7a231.txt> ;
        wfprov:describedByParameter  :processor/getSimilarConceptProfilesPredefined/in/parameters ;
        wfprov:describedByParameter  :processor/getSimilarConceptProfilesPredefined_input/out/output ;
        wfprov:wasOutputFrom         <process/fab9fbf2-9b7b-49d0-8c34-79425772d4c3/> ;
        prov:qualifiedGeneration     _:b63 ;
        prov:wasGeneratedBy          <process/fab9fbf2-9b7b-49d0-8c34-79425772d4c3/> ;
        rdf:type                     wfprov:Artifact ;
        rdf:type                     prov:Entity .

<process/fab9fbf2-9b7b-49d0-8c34-79425772d4c3/>
        wfprov:usedInput             <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bf4c7b43-6318-429c-bc69-efce37ddf040> ;
        prov:qualifiedEnd            _:b329 ;
        prov:qualifiedAssociation    _:b183 ;
        prov:qualifiedUsage          _:b316 ;
        rdf:type                     wfprov:ProcessRun ;
        prov:qualifiedUsage          _:b92 ;
        prov:wasAssociatedWith       <#taverna-engine> ;
        wfprov:wasPartOfWorkflowRun  <> ;
        prov:startedAtTime           "2014-02-24T13:29:11.119+01:00"^^xsd:dateTime ;
        prov:qualifiedStart          _:b245 ;
        wfprov:wasEnactedBy          <#taverna-engine> ;
        prov:used                    <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bf40c31c-b6e6-4b9b-abb9-5fc16e10c982> ;
        prov:used                    <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bf4c7b43-6318-429c-bc69-efce37ddf040> ;
        wfprov:usedInput             <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        748b8d4c-d4a1-4d12-b85f-27f69bea08ae> ;
        wfprov:describedByProcess    :processor/getSimilarConceptProfilesPredefined_input ;
        prov:qualifiedUsage          _:b328 ;
        prov:endedAtTime             "2014-02-24T13:29:11.381+01:00"^^xsd:dateTime ;
        rdfs:label                   "Processor execution getSimilarConceptProfilesPredefined_input"@en ;
        prov:used                    <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        748b8d4c-d4a1-4d12-b85f-27f69bea08ae> ;
        wfprov:usedInput             <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bf40c31c-b6e6-4b9b-abb9-5fc16e10c982> .

<http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/bf40c31c-b6e6-4b9b-abb9-5fc16e10c982>
        tavernaprov:content          <inputs/predefined_conceptset_id.txt> ;
        wfprov:describedByParameter  :processor/getSimilarConceptProfilesPredefined_input/in/conceptSetId ;
        wfprov:describedByParameter  :in/predefined_conceptset_id ;
        rdf:type                     wfprov:Artifact ;
        rdf:type                     prov:Entity .

<http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/bf4c7b43-6318-429c-bc69-efce37ddf040>
        tavernaprov:content          <intermediates/bf/bf4c7b43-6318-429c-bc69-efce37ddf040.txt> ;
        wfprov:describedByParameter  :processor/getSimilarConceptProfilesPredefined_input/in/queryConceptProfileList ;
```

```
        wfprov:describedByParameter  :processor/Merge_String_List_to_a_String/out/concatenated ;
        wfprov:wasOutputFrom         <process/16a65585-8763-4d17-966f-c0a636839a86/> ;
        prov:qualifiedGeneration     _:b64 ;
        prov:wasGeneratedBy          <process/16a65585-8763-4d17-966f-c0a636839a86/> ;
        rdf:type                     wfprov:Artifact ;
        rdf:type                     prov:Entity .

<http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/748b8d4c-d4a1-4d12-b85f-27f69bea08ae>
        tavernaprov:content          <inputs/cutoff.txt> ;
        wfprov:describedByParameter  :processor/getSimilarConceptProfilesPredefined_input/in/cutoffValue ;
        wfprov:describedByParameter  :in/cutoff ;
        rdf:type                     wfprov:Artifact ;
        rdf:type                     prov:Entity .

<http://ns.taverna.org.uk/2011/run/666bd1c0-9c70-48a0-bb82-0a475274a985/>
        wfprov:describedByWorkflow  <http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-
        a0b0f4e85935/workflow/Annotate_gene_list_w/> ;
        dct:hasPart                 <process/56feb844-3f84-4dae-99d1-7e8469207414/> ;
        dct:hasPart                 <process/ad944e1b-902a-4e07-abd0-51347c3a744d/> ;
        dct:hasPart                 <process/a7eaa124-7d50-4126-bf33-3c411a9f975f/> ;
        prov:used                   <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bf40c31c-b6e6-4b9b-abb9-5fc16e10c982> ;
        prov:qualifiedUsage         _:b154 ;
        dct:hasPart                 <process/16509e5c-b0c6-425f-90af-3a5ac8a8cf3d/> ;
        wfprov:usedInput            <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        bf40c31c-b6e6-4b9b-abb9-5fc16e10c982> ;
        dct:hasPart                 <process/637fc768-cee5-40fd-bf77-3f59c7dacbde/> ;
        dct:hasPart                 <process/9305bf57-48fd-4770-a923-7694f9c2e414/> ;
        dct:hasPart                 <process/67493fdb-4ff5-4b39-9168-0ab8669b6ad4/> ;
        prov:qualifiedAssociation   _:b155 ;
        wfprov:wasEnactedBy         <#taverna-engine> ;
        dct:hasPart                 <process/9187b610-8f97-43da-9822-352e4af1c816/> ;
        dct:hasPart                 <process/8dc61a2a-9663-40e2-b813-f41b4f111bf3/> ;
        dct:hasPart                 <process/16a65585-8763-4d17-966f-c0a636839a86/> ;
        dct:hasPart                 <process/494d6358-7422-4b6a-af37-80e534d73bbb/> ;
        dct:hasPart                 <process/df50cf7c-ea72-442b-b93b-ef88cd02667b/> ;
        prov:qualifiedUsage         _:b156 ;
        prov:qualifiedUsage         _:b66 ;
        prov:qualifiedEnd           _:b157 ;
        dct:hasPart                 <process/1b7d5932-a2f8-4708-adac-5c450aa64116/> ;
        dct:hasPart                 <process/0f4f157c-03c2-4ab4-a2cd-6c4b991098bb/> ;
        prov:used                   <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        748b8d4c-d4a1-4d12-b85f-27f69bea08ae> ;
        dct:hasPart                 <process/a21c8dd3-a51b-4a31-9fbd-e4be545127e0/> ;
        prov:endedAtTime            "2014-02-24T13:29:48.004+01:00"^^xsd:dateTime ;
        dct:hasPart                 <process/f4804d71-4cb6-4773-b883-a56f40b23a1d/> ;
        rdfs:label                  "Workflow run of Annotate_gene_list_w"@en ;
        prov:qualifiedUsage         _:b158 ;
        dct:hasPart                 <process/fab9fbf2-9b7b-49d0-8c34-79425772d4c3/> ;
        dct:hasPart                 <process/f2a30c58-0a36-43ce-aab1-a309c326c68e/> ;
        dct:hasPart                 <process/f8090fa4-758e-4f50-9a9a-68a5b2fa48e7/> ;
        wfprov:usedInput            <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        748b8d4c-d4a1-4d12-b85f-27f69bea08ae> ;
        prov:qualifiedStart         _:b159 ;
        prov:used                   <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        a30b37d0-9b8b-40f8-aaf7-729a3545e798> ;
        dct:hasPart                 <process/62abd0da-19b8-4825-ad57-245bc5b7b4ca/> ;
        dct:hasPart                 <process/08fa081a-5cb6-4842-845d-d567aa12db01/> ;
        wfprov:usedInput            <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        a30b37d0-9b8b-40f8-aaf7-729a3545e798> ;
        prov:wasAssociatedWith      <#taverna-engine> ;
        dct:hasPart                 <process/bb1639e2-7c83-4b32-9381-298bc5f5702d/> ;
        dct:hasPart                 <process/78e6094f-330e-47e2-96d5-f959c680d34b/> ;
        prov:startedAtTime          "2014-02-24T13:28:58.912+01:00"^^xsd:dateTime ;
        dct:hasPart                 <process/dda43746-653a-43b1-ba22-5f10d23cb543/> ;
        dct:hasPart                 <process/46e8ac2e-7e32-4146-98ea-01eafc00a079/> ;
        dct:hasPart                 <process/3f5678c5-b70b-45b9-bfaf-4e4cb9ae2393/> ;
        dct:hasPart                 <process/120c553b-57f6-4192-ae3b-b59955367579/> ;
        wfprov:usedInput            <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        8a72c8ab-d0b5-4ec5-a77d-a29729f9a317> ;
        rdf:type                    wfprov:WorkflowRun ;
        prov:used                   <http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/
        8a72c8ab-d0b5-4ec5-a77d-a29729f9a317> ;
        dct:hasPart                 <process/3d681747-072a-4b99-9d5f-cb024dad2631/> .

<http://ns.taverna.org.uk/2011/data/666bd1c0-9c70-48a0-bb82-0a475274a985/ref/9ff884a3-3747-4400-bec0-c7b7616a1b20>
        tavernaprov:content          <outputs/score.txt> ;
        wfprov:describedByParameter  :processor/Merge_String_List_to_a_String_4/out/concatenated ;
        wfprov:wasOutputFrom         <process/16509e5c-b0c6-425f-90af-3a5ac8a8cf3d/> ;
        prov:qualifiedGeneration     _:b285 ;
```

```
        wfprov:wasOutputFrom          <> ;
        prov:wasGeneratedBy           <process/16509e5c-b0c6-425f-90af-3a5ac8a8cf3d/> ;
        wfprov:describedByParameter   :out/score ;
        rdf:type                      wfprov:Artifact ;
        rdf:type                      prov:Entity .
```

A.3. ROEVO RDF example

The following code represents the ROEVO RDF representation of the example presented in Section 4.4.

```
@prefix wro:      <http://purl.org/wf4ever/ro#> .
@prefix prov:    <http://www.w3.org/ns/prov#> .
@prefix foaf:    <http://xmlns.com/foaf/0.1/> .
@prefix roevo:   <http://purl.org/wf4ever/roevo#> .
@prefix wfdesc:  <http://purl.org/wf4ever/wfdesc#> .
@prefix pav:  <http://purl.org/pav/> .

<https://www.google.com/accounts/o8/id?id=AItOawl_w_9JVAQsyNOviIgiYNzPYAGPNYlvVe4>
a       prov:Agent ;
foaf:name "Eleni Mina" .

## RO level changes
<http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/>
a       wro:ResearchObject , roevo:SnapshotRO ;
roevo:isSnapshotOf <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/> ;
roevo:snapshotedAtTime "2014-02-26T19:18:29.578+01:00" ;
roevo:wasSnapshotedBy <https://www.google.com/accounts/o8/id?id=AItOawl_w_9JVAQsyNOviIgiYNzPYAGPNYlvVe4> ;
roevo:wasChangedBy <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da> ;
prov:wasRevisionOf <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da>
a       roevo:ChangeSpecification ;
roevo:fromVersion <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/> ;
roevo:toVersion <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/> ;
prov:wasAssociatedWith <https://www.google.com/accounts/o8/id?id=AItOawl_w_9JVAQsyNOviIgiYNzPYAGPNYlvVe4> ;
roevo:hasChange <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/
8c6ff8a3-8031-4828-b03a-ff0f5238c6fc> ,
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/7355d3be-6506-4f38-a3bb-e457e0ebe4a8> ,
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-0b2cdd218b20>  .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/8c6ff8a3-8031-4828-b03a-ff0f5238c6fc>
a       roevo:Change , roevo:Removal ;
roevo:relatedResource <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/
annotate_genes_biological_processes.t2flow> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/7355d3be-6506-4f38-a3bb-e457e0ebe4a8>
a       roevo:Change , roevo:Addition ;
roevo:hasPreviousChange <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/
8c6ff8a3-8031-4828-b03a-ff0f5238c6fc> ;
roevo:relatedResource <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
annotate_genes_biological_processes_xpath_cpids.t2flow> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-0b2cdd218b20>
a       roevo:Modification , roevo:Change ;
roevo:hasPreviousChange <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/
7355d3be-6506-4f38-a3bb-e457e0ebe4a8> ;
roevo:relatedResource <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
workflow_sketch_hd_chromatin_interpretation.png> .


## Workflow level changes
<http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
annotate_genes_biological_processes_xpath_cpids.t2flow>
a       wro:Resource , wfdesc:Workflow ;
prov:generatedAtTime "2014-02-24T21:15:29.578+01:00" ;
prov:wasAttributedTo <https://www.google.com/accounts/o8/id?id=AItOawl_w_9JVAQsyNOviIgiYNzPYAGPNYlvVe4> ;
roevo:wasChangedBy <change_specifications/59d8e3d7-db05-4bc3-b689-123456789abc> ;
prov:wasRevisionOf <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/
annotate_genes_biological_processes.t2flow> .

<change_specifications/59d8e3d7-db05-4bc3-b689-123456789abc>
a       roevo:ChangeSpecification ;
roevo:fromVersion <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/
annotate_genes_biological_processes.t2flow> ;
roevo:toVersion <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
annotate_genes_biological_processes_xpath_cpids.t2flow> ;
prov:wasAssociatedWith <https://www.google.com/accounts/o8/id?id=AItOawl_w_9JVAQsyNOviIgiYNzPYAGPNYlvVe4> ;
roevo:hasChange
```

```
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/8c6ff8a3-8031-4828-b03a-abc111def111> ,
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/7355d3be-6506-4f38-a3bb-0987654321ab> ,
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-rstuvpxyz123> ,
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-x34p28muy321> ,
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-rast25987hm2> ,
<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-kdisia984727>  .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/8c6ff8a3-8031-4828-b03a-abc111def111>
a       roevo:Change , roevo:Removal ;
roevo:relatedResource <http://ns.taverna.org.uk/2010/workflowBundle/60a0cc53-cd8d-443f-b303-51c2773b0de8/
workflow/Annotate_gene_list_w/datalink?from=processor/get_scores/out/nodelist&to=processor/
Merge_String_List_to_a_String_4/in/stringlist> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/7355d3be-6506-4f38-a3bb-0987654321ab>
a       roevo:Change , roevo:Removal ;
roevo:hasPreviousChange
  <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/8c6ff8a3-8031-4828-b03a-abc111def111> ;
roevo:relatedResource
  <http://ns.taverna.org.uk/2010/workflowBundle/60a0cc53-cd8d-443f-b303-51c2773b0de8/workflow/
  Annotate_gene_list_w/datalink?from=processor/Merge_String_List_to_a_String_4/out/concatenated&to=out/score> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-rstuvpxyz123>
a       roevo:Change , roevo:Removal ;
roevo:hasPreviousChange <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/
7355d3be-6506-4f38-a3bb-0987654321ab> ;
roevo:relatedResource <http://ns.taverna.org.uk/2010/workflowBundle/60a0cc53-cd8d-443f-b303-51c2773b0de8/
workflow/Annotate_gene_list_w/processor/Merge_String_List_to_a_String_4/> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-x34p28muy321>
a       roevo:Change , roevo:Addition ;
roevo:hasPreviousChange <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/
f9e0164e-9e7a-4bad-b44c-rstuvpxyz123> ;
roevo:relatedResource <http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/
workflow/Annotate_gene_list_w/processor/Merge_String_List_to_a_String_5/> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-rast25987hm2>
a       roevo:Change , roevo:Addition ;
roevo:hasPreviousChange <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/
f9e0164e-9e7a-4bad-b44c-x34p28muy321> ;
roevo:relatedResource <http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/
workflow/Annotate_gene_list_w/datalink?from=processor/Merge_String_List_to_a_String_5/out/
concatenated&to=out/score> .

<change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/f9e0164e-9e7a-4bad-b44c-kdisia984727>
a       roevo:Change , roevo:Addition ;
roevo:hasPreviousChange <change_specifications/59d8e3d7-db05-4bc3-b689-7bca7733f7da/changes/
f9e0164e-9e7a-4bad-b44c-rast25987hm2> ;
roevo:relatedResource <http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935
/workflow/Annotate_gene_list_w/datalink?from=processor/get_scores/out/nodelist&to=processor/
Merge_String_List_to_a_String_5/in/stringlist> .

## Resources
<http://ns.taverna.org.uk/2010/workflowBundle/60a0cc53-cd8d-443f-b303-51c2773b0de8/workflow/Annotate_gene_list_w/
datalink?from=processor/get_scores/out/nodelist&to=processor/Merge_String_List_to_a_String_4/in/stringlist>
a wro:Resource .

<http://ns.taverna.org.uk/2010/workflowBundle/60a0cc53-cd8d-443f-b303-51c2773b0de8/workflow/Annotate_gene_list_w/
datalink?from=processor/Merge_String_List_to_a_String_4/out/concatenated&to=out/score>
a wro:Resource .

<http://ns.taverna.org.uk/2010/workflowBundle/60a0cc53-cd8d-443f-b303-51c2773b0de8/workflow/Annotate_gene_list_w/
processor/Merge_String_List_to_a_String_4/>
a wro:Resource .

<http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/Annotate_gene_list_w/
processor/Merge_String_List_to_a_String_5/>
a wro:Resource .

<http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/Annotate_gene_list_w/
datalink?from=processor/Merge_String_List_to_a_String_5/out/concatenated&to=out/score>
a wro:Resource .

<http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/Annotate_gene_list_w/
datalink?from=processor/get_scores/out/nodelist&to=processor/Merge_String_List_to_a_String_5/in/stringlist>
a wro:Resource .

## Link between t2flow workflow and taverna workflow id
## Old workflow

<http://ns.taverna.org.uk/2010/workflowBundle/60a0cc53-cd8d-443f-b303-51c2773b0de8/workflow/Annotate_gene_list_w/>
```

```
      wfdesc:hasWorkflowDefinition
              <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/Annotate_gene_list_w.wfbundle> .

<http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/Annotate_gene_list_w.wfbundle>
pav:importedFrom
              <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/
              annotate_genes_biological_processes.t2flow> .

## New workflow
<http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/Annotate_gene_list_w/>
      wfdesc:hasWorkflowDefinition
              <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
              Annotate_gene_list_w.wfbundle> .

<http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/Annotate_gene_list_w.wfbundle>
pav:importedFrom
              <http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
              annotate_genes_biological_processes_xpath_cpids.t2flow> .
```

A.4. Results of the example competency queries

This section contains the results of the competency SPARQL queries presented in Section 5.4.

Results obtained by evaluating Query 1

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="name"/>
  </head>
  <results>
    <result>
      <binding name="name">
        <literal>Eleni Mina</literal>
      </binding>
    </result>
  </results>
</sparql>
```

Results obtained by evaluating Query 2

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="workflow"/>
    <variable name="def"/>
  </head>
  <results>
    <result>
      <binding name="workflow">
        <uri>http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/
        Annotate_gene_list_w/</uri>
      </binding>
      <binding name="def">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation/
        annotate_genes_biological_processes_xpath_cpids.t2flow</uri>
      </binding>
    </result>
    <result>
      <binding name="workflow">
        <uri>http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/
        Annotate_gene_list_w/</uri>
      </binding>
      <binding name="def">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/
        annotate_genes_biological_processes_xpath_cpids.t2flow</uri>
      </binding>
    </result>
    <result>
      <binding name="workflow">
        <uri>http://ns.taverna.org.uk/2010/workflowBundle/f7b17c46-a5c5-428f-ad45-a0b0f4e85935/workflow/
        Annotate_gene_list_w/</uri>
      </binding>
      <binding name="def">
```

```
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
        annotate_genes_biological_processes_xpath_cpids.t2flow</uri>
      </binding>
    </result>
  </results>
</sparql>
```

Results obtained by evaluating Query 3

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="inputValue"/>
  </head>
  <results>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/mypack-2-annotate_gene_list.bundle/
        inputs/cutoff.txt</uri>
      </binding>
    </result>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/mypack-2-annotate_gene_list.bundle/
        inputs/gene_IDs.txt</uri>
      </binding>
    </result>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/mypack-2-annotate_gene_list.bundle/inputs/
        predefined_conceptset_id.txt</uri>
      </binding>
    </result>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/mypack-2-annotate_gene_list.bundle/inputs/
        database_name.txt</uri>
      </binding>
    </result>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-
        annotate_genes_biological_processes_xpath_cpids_only_cpids-run.bundle/inputs/
        predefined_conceptset_id.txt</uri>
      </binding>
    </result>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-
        annotate_genes_biological_processes_xpath_cpids_only_cpids-run.bundle/inputs/
        database_name.txt</uri>
      </binding>
    </result>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-
        annotate_genes_biological_processes_xpath_cpids_only_cpids-run.bundle/inputs/cutoff.txt</uri>
      </binding>
    </result>
    <result>
      <binding name="inputValue">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-
        annotate_genes_biological_processes_xpath_cpids_only_cpids-run.bundle/inputs/gene_IDs.txt</uri>
      </binding>
    </result>
  </results>
</sparql>
```

Results obtained by evaluating Query 4

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
```

```
    <variable name="workflow"/>
  </head>
  <results>
    <result>
      <binding name="workflow">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
        annotate_genes_biological_processes_xpath_cpids.t2flow</uri>
      </binding>
    </result>
    <result>
      <binding name="workflow">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot-1/
        explainScoresStringInput2.t2flow</uri>
      </binding>
    </result>
    <result>
      <binding name="workflow">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/
        explainScoresStringInput.t2flow</uri>
      </binding>
    </result>
    <result>
      <binding name="workflow">
        <uri>http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2-snapshot/
        annotate_genes_biological_processes.t2flow</uri>
      </binding>
    </result>
  </results>
</sparql>
```

References

[1] Ewa Deelman, Dennis Gannon, Matthew S. Shields, Ian Taylor, Workflows and e-science: An overview of workflow system features and capabilities, Future Gener. Comput. Syst. 25 (5) (2009) 528–540.
[2] David De Roure, Carole A. Goble, Robert Stevens, The design and realisation of the myExperiment virtual research environment for social sharing of workflows, Future Gener. Comput. Syst. 25 (5) (2009) 561–567.
[3] Phillip Mates, Emanuele Santos, Juliana Freire, Cláudio T. Silva, Crowdlabs: Social analysis and visualization for the sciences, in: Judith Bayard Cushing, James C. French, Shawn Bowers (Eds.), SSDBM, in: Lecture Notes in Computer Science, vol. 6809, Springer, 2011, pp. 555–564.
[4] Jun Zhao, José Manuél Gómez-Pérez, Khalid Belhajjame, Graham Klyne, Esteban García-Cuesta, Aleix Garrido, Kristina M. Hettne, Marco Roos, David De Roure, Carole A. Goble, Why workflows break—understanding and combating decay in taverna workflows, in: eScience, IEEE Computer Society, 2012, pp. 1–9.
[5] Khalid Belhajjame, Semantic replaceability of escience web services, in: Third International Conference on e-Science and Grid Computing, e-Science 2007, 10–13 December 2007, Bangalore, India, IEEE, 2007, pp. 449–456.
[6] Khalid Belhajjame, Carole A. Goble, Stian Soiland-Reyes, David De Roure, Fostering scientific workflow preservation through discovery of substitute services, in: eScience, IEEE Computer Society, 2011, pp. 97–104.
[7] Sven Köhler, Sean Riddle, Daniel Zinn, Timothy M. McPhillips, Bertram Ludäscher, Improving workflow fault tolerance through provenance-based recovery, in: Scientific and Statistical Database Management—23rd International Conference, SSDBM 2011, Portland, OR, USA, July 20–22, 2011. Proceedings, Springer, 2011, pp. 207–224.
[8] Daniel Crawl, Ilkay Altintas, A provenance-based fault tolerance mechanism for scientific workflows, in: Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17–18, 2008. Revised Selected Papers, Springer, 2008, pp. 152–159.
[9] Sean Bechhofer, John D. Ainsworth, Jiten Bhagat, Iain E. Buchan, Philip A. Couch, Don Cruickshank, David De Roure, Mark Delderfield, Ian Dunlop, Matthew Gamble, Carole A. Goble, Danius T. Michaelides, Paolo Missier, Stuart Owen, David R. Newman, Shoaib Sufi, Why linked data is not enough for scientists, in: Sixth International Conference on e-Science, e-Science 2010, 7–10 December 2010, Brisbane, QLD, Australia, IEEE, 2010, pp. 300–307.
[10] S. Soiland-Reyes, S. Bechhofer, K. Belhajjame, G. Klyne, D. Garijo, O. Corcho, E. Garcí a Cuesta, R. Palma, Wf4ever research object model 1.0., November 2013. http://dx.doi.org/10.5281/zenodo.12744.
[11] Software Sustainability Institute and Curtis+Cartwrigh. Software preservation benefits framework, Technical report, 2010, Last accessed: December 2014.
[12] Brian Matthews, Brian McIlwrath, David Giaretta, Esther Conway, The significant properties of software: A study. Technical report, JISC report, 2008, March.
[13] Quantifying reproducibility in computational biology: The case of the tuberculosis drugome, PLoS ONE 8 (11) (2013) e80278.
[14] Best practices for computational science: Software infrastructure and environments for reproducible and extensible research, J. Open Res. Softw. 2 (1) (2014) e21.
[15] Greg Wilson, D.A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H.D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, et al., Best practices for scientific computing, PLoS Biology 12 (1) (2014) e1001745.
[16] Juliana Freire, Cláudio T. Silva, Making computations and publications reproducible with vistrails, Comput. Sci. Eng. 14 (4) (2012) 18–25.
[17] Fernando Chirigati, Dennis Shasha, Juliana Freire, Reprozip: Using provenance to support computational reproducibility, in: Proc. of the 6th USENIX Workshop on Theory and Practice of Provenance, 2013.
[18] Ian Foster Quan Pham, Tanu Malik, Using provenance for repeatability, in: Proceedings of the 5th USENIX Workshop on Theory and Practice of Provenance 2013, 2013.
[19] Quan Pham, Tanu Malik, Ian T. Foster, Using provenance for repeatability, in: Proc. of the 6th USENIX Workshop on Theory and Practice of Provenance, 2013.
[20] Sara Magliacane, Paul T. Groth, Towards reconstructing the provenance of clinical guidelines, in: Adrian Paschke, Albert Burger, Paolo Romano, M. Scott Marshall, Andrea Splendiani (Eds.), Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences, in: CEUR Workshop Proceedings, vol. 952, CEUR-WS.org, 2012.
[21] Carl Boettiger, An introduction to docker for reproducible research, with examples from the R environment. CoRR, abs/1410.0846, 2014.
[22] Ryan R. Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M. Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, et al., Modeling biomedical experimental processes with obi, J. Biomed. Semant. 1 (Suppl 1) (2010) S7.
[23] Philippe Rocca-Serra, Marco Brandizi, Eamonn Maguire, et al., Isa software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, Bioinformatics 26 (18) (2010) 2354–2356.
[24] Paolo Ciccarese, Elizabeth Wu, Gwen Wong, Marco Ocana, June Kinoshita, Alan Ruttenberg, Tim Clark, The swan biomedical discourse ontology, J. Biomed. Inform. 41 (5) (2008) 739–751.
[25] Paul Groth, Andrew Gibson, Jan Velterop, The anatomy of a nanopublication, Inform. Serv. Use 30 (1) (2010) 51–56.
[26] Dean B. Krafft, Nicholas A. Cappadona, Brian Caruso, Jon Corson-Rikert, Medha Devare, Brian J. Lowe, et al. Vivo: Enabling national networking of scientists, in: Proceedings of the WebSci10, Raleigh, US, 2010, pp. 1310–1313.
[27] The Huntington's disease collaborative research group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes, Cell 72 (6) (1993) 971–983.
[28] R. Jelier, M.J. Schuemie, A. Veldhoven, L.C.J. Dorssers, G. Jenster, J.A. Kors, Anni 2.0: a multipurpose text-mining tool for the life sciences, Genome Biol. 9 (R96) (2008).

[29] K.M. Hettne, A. Boorsma, D.A. van Dartel, J.J. Goeman, E. de Jong, A.H. Piersma, R.H. Stierum, J.C. Kleinjans, J.A. Kors, Next-generation text-mining mediated generation of chemical response-specific gene sets for interpretation of gene expression data, BMC Med Genom. 6 (2) (2013).
[30] K.M. Hettne, R. van Schouwen, E. Mina, et al., Explain your data by concept profile analysis web services [v1; ref status: approved with reservations 2], F1000Research 3 (173) (2014).
[31] Rudolf Mayer, Stefan Pröll, Andreas Rauber, Raúl Palma, Daniel Garijo, From preserving data to preserving research: Curation of process and context (demo), in: TPDL, Springer, 2013, pp. 490–491.
[32] Eric PrudHommeaux, Andy Seaborne, et al. Sparql query language for rdf, in: W3C Recommendation, vol. 15, 2008.
[33] Eleni Mina, Willeke van Roon-Mom, Peter A.C. 't Hoen, Mark Thompson, Reinout van Schouwen, Rajaram Kaliyaperumal, Kristina Hettne, Erik Schultes, Barend Mons, Marco Roos, Prioritizing hypotheses for epigenetic mechanisms in huntington's disease using an e-science approach, BioData Mining (2014) in press.
[34] Timothy Lebo, Satya Sahoo, Deborah McGuinness, et al., Prov-o: The prov ontology. Technical report, W3C Recommendation, 2013.
[35] Carl Lagoze, Herbert Van de Sompel, ORE specification—abstract data model. http://www.openarchives.org/ore/1.0/datamodel.html (Accessed on February 28, 2014).
[36] Paolo Ciccarese, Marco Ocana, Leyla J. Garcia Castro, Sudeshna Das, Tim Clark, An open annotation ontology for science on web 3.0, J. Biomed. Semant. 2 (Suppl 2) (2011) S4.
[37] K. Wolstencroft, R. Haines, D. Fellows, et al., The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud, Nucl. Acids Res. (2013).
[38] Yolanda Gil, Varun Ratnakar, Jihie Kim, et al., Wings: Intelligent workflow-based design of computational experiments, IEEE Intell. Syst. 26 (1) (2011) 62–72.
[39] Jeremy Goecks, Anton Nekrutenko, James Taylor, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol. 11 (8) (2010) R86.
[40] Carl Lagoze, Herbert Van de Sompel, Ore specification—vocabulary. http://www.openarchives.org/ore/1.0/vocabulary.html (Accessed on February 28, 2014).
[41] Rudolf Mayer, Andreas Rauber, Martin Alexander Neumann, John Thomson, Gonçalo Antunes, Preserving scientific processes from design to publications, in: Theory and Practice of Digital Libraries, Springer, 2012, pp. 113–124.
[42] Angela Dappert, Sébastien Peyrard, Carol C.H. Chou, Janet Delve, Describing and preserving digital object environments, New Rev. Inf. Netw. 18 (2) (2013) 106–173.
[43] Stian Soiland-Reyes, Matthew Gamble, Robert Haines, Research Object Bundle 1.0. November 2014. http://dx.doi.org/10.5281/zenodo.12586.
[44] Steven P. Callahan, Juliana Freire, Emanuele Santos, et al., Vistrails: Visualization meets data management, in: ACM SIGMOD, ACM Press, 2006, pp. 745–747.
[45] Khalid Belhajjame, Annotating the behavior of scientific modules using data examples: A practical approach, in: Proc. of International Conference on Extending Database Technology, 2014, pp. 726–737.
[46] Dagmar Krefting, Tristan Glatard, Vladimir Korkhov, Johan Montagnat, Silvia Olabarriaga, Enabling grid interoperability at workflow level. Grid Workflow Workshop 2011, 2011.
[47] Paolo Missier, Saumen Dey, Khalid Belhajjame, Víctor Cuevas-Vicenttín, Bertram Ludäscher, D-PROV: extending the PROV provenance model with workflow structure, in: Computing Science, Newcastle University, 2013.
[48] Daniel Garijo, Yolanda Gil, A new approach for publishing workflows: Abstractions, standards, and linked data, in: Proceedings of the 6th Workshop on Workflows in Support of Large-scale Science, ACM, 2011, pp. 47–56.
[49] Alejandra Gonzalez-Beltran, Peter Li, Jun Zhao, Maria Susana Avila-Garcia, Marco Roos, Mark Thompson, Eelke van der Horst, Rajaram Kaliyaperumal, Ruibang Luo, Lee Tin-Lap, Lam Tak-wah, Scott C. Edmunds, Susanna-Assunta Sansone, Philippe Rocca-Serra, From peer-reviewed to peer-reproduced: a role for data standards, models and computational workflows in scholarly publishing, in: bioRxiv, 2014.
[50] Thomas Russ, Cartic Ramakrishnan, Eduard Hovy, Mihail Bota, Gully Burns, Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case, BMC Bioinform. 12 (1) (2011) 351.
[51] Brian Matthews, Shoaib Sufi, Damian Flannery, Laurent Lerusse, Tom Griffin, Michael Gleaves, Kerstin Kleese, Using a core scientific metadata model in large-scale facilities, Int. J. Digit. Curation 5 (1) (2010) 106–118.
[52] Science and Technology Facilities Council. Isis. http://www.isis.stfc.ac.uk/index.html. Accessed on the 20th of June 2014.
[53] Science and Technology Facilities Council. Diamond light source. http://www.diamond.ac.uk. Accessed on the 20th of June 2014.
[54] J. Hunter, Scientific publication packages: A selective approach to the communication and archival of scientific output, Int. J. Digit. Curation 1 (1) (2006).
[55] Fernando Chirigati, Dennis Shasha, Juliana Freire, Packing experiments for sharing and publication, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013, pp. 977–980.
[56] Quan Pham, Tanu Malik, Ian Foster, Roberto Di Lauro, Raffaele Montella, Sole: linking research papers with science objects, in: Provenance and Annotation of Data and Processes, Springer, 2012, pp. 203–208.
[57] Victoria Stodden, Christophe Hurlin, Christophe Pérignon, Runmycode.org: a novel dissemination and collaboration platform for executing published computational results, in: Proc. of IEEE 8th International Conference on e-Science, IEEE, 2012, pp. 1–8.
[58] Pieter Van Gorp, Steffen Mazanek, Share: a web portal for creating and sharing executable research papers, Proc. Comput. Sci. 4 (2011) 589–597.
[59] Terri K. Attwood, Douglas B. Kell, Philip McDermott, James Marsh, Stephen Pettifer, David Thorne, Utopia documents: linking scholarly literature with research data, Bioinformatics 26 (18) (2010) 568–574.
[60] Friedrich Leisch, Sweave: Dynamic generation of statistical reports using literate data analysis, in: Compstat, Springer, 2002, pp. 575–580.
[61] Fernando Pérez, Brian E. Granger, IPython: a system for interactive scientific computing, Comput. Sci. Eng. 9 (3) (2007) 21–29.
[62] Marian Petre, Greg Wilson, Plos/mozilla scientific code review pilot: Summary of findings, 2013.
[63] M. Crosas, The dataverse network: An open-source application for sharing, discovering and preserving data, D-Lib. Mag. 17 (1/2) (2011).