Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

STREP FP7-ICT-2007-6 270192

Objective ICT-2009.4.1 b) – “Advanced preservation scenarios”

D6.3v2: Genome Wide Association Study Workflows v2

Deliverable Co-ordinator: Kristina Hettne

Deliverable Co-ordinating Institution: Leiden University Medical Centre (LUMC)

Other Authors: Kristina Hettne (LUMC), Harish Dharuri (LUMC), Marco Roos (LUMC)

This deliverable provides Genome Wide Association Study Workflows

Document Identifier:	Wf4ever/2010/D6.3v2/v1.0	Date due:	30/09/2012
Class Deliverable:	Wf4ever 270192	Submission date:	30/09/2012
Project start date:	December 1, 2010	Version:	1.0
Project duration:	3 years	State:	Final
		Distribution:	Public

Wf4Ever Consortium

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

Intelligent Software Components S.A. Edificio Testa Avda. del Partenón 16-18, 1º, 7ª Campo de las Naciones, 28042 Madrid Spain Contact person: Dr. Jose Manuel Gómez-Pérez E-mail address: jmgomez@isoco.com	**University of Manchester** Department of Computer Science, University of Manchester, Oxford Road Manchester, M13 9PL United Kingdom Contact person: Professor Carole Goble E-mail address: carole.goble@manchester.ac.uk
Universidad Politécnica de Madrid Departamento de Inteligencia Artificial Facultad de Informática, UPM 28660 Boadilla del Monte, Madrid Spain Contact person: Dr. Oscar Corcho E-mail address: ocorcho@fi.upm.es	**University of Oxford** Department of Zoology University of Oxford South Parks Road, Oxford OX1 3PS United Kingdom Contact person: Dr. Jun Zhao / Professor David De Roure E-mail address: {jun.zhao@zoo.ox.ac.uk, david.deroure@oerc.ox.ac.uk}
Poznań Supercomputing and Networking Center Network Services Department Poznań Supercomputing and Networking Center Z. Noskowskiego 12/14, 61-704 Poznan Poland Contact person: Dr. Raúl Palma de León E-mail address: rpalma@man.poznan.pl	**Instituto de Astrófísica de Andalucía** Dpto. Astronomía Extragaláctica Instituto Astrofísica Andalucía Glorieta de la Astronomía s/n 18008 Granada, Spain Contact person: Dr. Lourdes Verdes-Montenegro E-mail address: lourdes@iaa.es
Leiden University Medical Centre Department of Human Genetics Leiden University Medical Centre Albinusdreef 2, 2333 ZA Leiden The Netherlands Contact person: Dr. Marco Roos E-mail address: M.Roos1@lumc.nl	

Change Log

Version	Date	Amended by	Changes
0.1	05-09-2012	Kristina Hettne	Created document
0.2	10-09-2012	Kristina Hettne	Added Materials and Methods section
0.3	14-09-2012	Kristina Hettne	Added results and discussion sections
0.4	17-09-2012	Kristina Hettne	Added workflow results and figures
0.5	26-09-2012	Kristina Hettne	Made changes according to QA comments from IAA.
1.0	28-09-2012	Kristina Hettne	Made changes according to QA comments from Jose Manuel Gomez Perez

Executive Summary

This document describes the workflows developed during phase II of the project at the Human Genetics Department of the Leiden University Medical Centre (HG-LUMC) for interpreting results from genome-wide association (GWA) studies. The main goal of this deliverable is to produce workflows. At the same time, we applied the tooling and best practices that are emerging from the project to aggregate the workflow and associated material as a preservable 'Research Object' (RO). A detailed description about the state of the current tooling can be found in D1.4v1. Workflows form a crucial part of the data to populate the RO models and software in Wf4Ever, and the HG-LUMC is committed to producing good quality workflows that can be preserved. To promote re-use and combat workflow decay, we developed Best Practices for workflow design. In this document, we describe workflows for interpreting GWA study data, Pest Practices for workflow design and their relation to ROs. Finally, we characterize the workflows according to current state of workflow preservation and archived them according to the project tooling.

Table of contents

List of Figures

1. Introduction

This document describes the genome-wide association study (GWAS) interpretation workflows developed during phase II of the project at the Human Genetics Department of the Leiden University Medical Centre (HG-LUMC). According to Wf4Ever's Description of Work this deliverable will contain the *second version of the selected workflows in the domain and their preservation using Wf4Ever technology*. We have addressed these objectives in the following ways:

- Implemented workflows for interpretation of GWAS data.
- Developed Best Practices for workflow design (described in section 2.1).
- Preserved the workflows, related data and documentation according to the latest version of the Wf4Ever reference implementation [1].

The focus on Best Practices for workflow design can be contrasted with the work reported in WP5, for which the focus is more on achieving Virtual Observatory compliant workflows and ROs where all elements are well identified by using annotations or by using a well-defined folder hierarchy if the current allowed annotations are not enough.

1.1 Background on Metabolic Syndrome case study

The HG-LUMC investigates the genetic background and molecular mechanisms behind a substantial number of rare and common diseases. An important technological focus is on Single Nucleotide Polymorphism (SNP) arrays, which is a type of DNA microarray that is used to detect polymorphisms within a population. The basic principles of the SNP array are the same as the DNA microarray, i.e. the convergence of DNA hybridization, fluorescence microscopy, and solid surface DNA capture. Various other types of data are collected in order to unravel the molecular basis for the progression of diseases and for finding new leads for treatment. Here we will focus on workflows created to analyse whole genome sequencing data from the primary case study dedicated to Wf4Ever which relates to unravelling the genetic background of the Metabolic Syndrome (MetS). In short, MetS is defined by a number of clinical criteria and not by underlying biological phenomena. The biological cause of the development of its associated diseases such as diabetes, is unclear. A GWAS associates genetic variation markers of many individuals with disease or risk factors for disease by statistical tests that have been developed for this purpose. However, in general, these associations explain a relatively small part of the genetic variation and have relatively small effect sizes. In contrast, genetic variants that associate with metabolite levels generally explain a higher percentage of the genetic variation and demonstrate larger effect sizes [2]. The top hits emerging from these studies also show that, frequently, the polymorphisms are present in or near a gene that is biologically relevant to the associated metabolite [2]. To understand the biomolecular basis of the association, scientists typically dwell on identifying genes in the vicinity of the SNP and the possible pathways that the gene participates in. The common objective for users of the GWAS interpretation workflows is to help interpret the results of a GWAS by integrating information from heterogeneous sources. The workflows for this purpose developed within the

HG-LUMC concerns interpreting SNP associations from a GWAS on human metabolite variation, using pathways from metabolic pathway databases and Gene Ontology (GO)[1] biological process associations from next-generation text mining. It has been shown before [3,4] that annotating GO biological processes to a gene by the use of next-generation text mining both reproduces and, more importantly, extends the annotation already provided in the GO database. We hypothesize that SNPs can be functionally annotated using metabolic pathway database queries and complemented by next-generation text mining, and that new hypotheses regarding the role of SNPs in biological processes relevant for human health can be formulated using these functional annotations. The experiment is performed entirely in silico, that is, performed on the computer. We evaluate our hypothesis by comparing our results to a list of SNPs that have been functionally annotated by hand [2].

1.2 **Best Practices for workflow design**

As a part of the preservation aspect of workflow-centric ROs, workflow decay has been studied within the context of the Wf4Ever project [5,6]. The main conclusion drawn from the study was that as much as 80% of all Taverna workflows on myExperiment failed to be either executed or produce the same results. The main reasons for the workflows to break were the following: volatile third-party resources, missing example data, missing execution environment, and insufficient descriptions about the workflows. Volatile third-party resources can apart from breaking a workflow also cause it to produce different results. The different results could be the result of services that are still available by using the same identifiers but their functionality have changed, or services that are updated intentionally or unintentionally (e.g. malware). We believe that all of these issues, with a possible exception of the first one, can be prevented by following workflow design guidelines at the workflow design stage. Unfortunately, to our knowledge no such guidelines exist, which led us to define 10 Best Practices for designing workflows. Our intention was to describe proposed best practices, not observed practices, and we based them on best practices as taught for the design and execution of the scientific method in experimental science as well as best practices for the development of software. Observed practices (e.g. using a more trial and error approach) may produce useful workflows. We however propose a number of steps that we believe create workflows of sufficient quality for scientific discourse, and subsequently tried to follow these steps when creating the GWAS interpretation workflows.

1.3 Outline

A description of the methodology, resources and Wf4Ever tools used is provided in section 2. Resulting workflows, scientific results as well as the results of following the Best Practices for workflow design are presented and discussed in section 3. Issues risen in the RO-ification process

related with tools, semantic annotations, quality and preservation can be found in section 4. Section 5 is dedicated to conclusions.

[1] http://www.geneontology.org/

2. Materials and Methods

2.1 Strategy and workflow building

All workflows were developed in the open source Taverna Workflow Management System version 2.4 [10]. The first workflow will be referred to as “workflow number 3124” in the rest of this document. It uses the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [9] of manually curated metabolic pathways that are available via KEGG Web Services[2] to annotate the SNPs with metabolic pathways. However, since the KEGG Web Services only understand gene identifiers, we used Biomart Web Services[3] to first map the SNP identifier to its corresponding gene identifier. To facilitate reuse, the step of mapping the SNP identifier to its corresponding gene identifier was also implemented as a separate workflow. Workflow number 3124 was designed and implemented by Harish Dharuri, an end user of Wf4Ever tooling at the HG-LUMC. The workflow was implemented before the 10 Best Practices for workflow design had been released. After release of the guidelines on the Wf4Ever wiki, Harish was asked to take a look at the guidelines and try to adapt his workflow accordingly.

The second set of workflows will be referred to as “pack number 282” in the rest of this document. The workflows in pack 282 use the Anni Web services[4] developed at the HG-LUMC by Reinout van Schouwen in collaboration with Kristina Hettne and Marco Roos. The web services intend to capture the most common analysis steps performed with the life-science focused interactive next-generation text mining tool 'Anni'[56]. Anni was developed to help the researcher in three major information-intensive tasks; 1) finding information on a specific subject, such as “give me all the genes that are directly associated with "prostatic neoplasms".”; 2) exploring the associations between a set of concepts, such as a list of genes that were found to be differentially expressed in a DNA microarray experiment.; and 3) literature-based knowledge discovery, such as prediction of drug-target interactions. The technology behind Anni (concept profile matching based on the vector-space model) has proven popular, but the monolithic tool is difficult to maintain. Therefore, we adopted an e-Science Approach based on (i) Web Services for the common steps available in Anni, (ii) workflows for the common procedures enacted by Anni, (iii) a workflow-to-web tool to leverage the advanced technology for (less technical) biologists. We wish to preserve the most common Anni methods by means of workflows and additional preservation tools from wf4Ever. It is implied in all genomics use cases, but for this study we wanted to tailor the workflow specifically for the GWAS use case. The workflows in pack 282 are developed Kristina Hettne, Marco Roos and Reinout van Schouwen, and designed according to the following 10 Best Practices (see also the Wf4Ever wiki[7] for a detailed description of the Best Practices):

[2] http://www.biocatalogue.org/services/11

[3] http://www.biomart.org/biomart/martservice

[4] http://www.biocatalogue.org/services/3330#overview

[5] http://biosemantics.org/anni

[6] Anni was developed by the BioSemantics group. The Biosemantics Group is a collaboration between the Medical Informatics department of the ErasmusMC University Medical Center of Rotterdam and the HG-LUMC.

[7] http://www.wf4ever-project.org/wiki/display/docs/Workflow+Design+Best+Practices

1. **Make an abstract workflow:** A workflow sketch provides a reference point of the main task(s) of the workflow through the implementation process. We anticipate it to promote sharing between computer and workflow systems due to its non-explicit nature, to help in designing the experiment, and to help in the communication with for example supervisors and colleagues.
2. **Use modules:** We anticipate that implementing all the executable components of the main nested workflows as separate, runnable subworkflows facilitates independent testing and validation of the execution of each of the individual components as well as the main nested workflow.
3. **Think about the output:** Who is the output intended for? Is it supposed to be used as input to another workflow, stored in a database or be presented to the end user? Should it be a graph, a table or text? The reason for thinking about the output of the workflow already at the design stage is that it is easier than trying to adjust a ready workflow. Also, it will help in structuring potentially large output data.
4. **Provide example inputs and outputs:** According to [6], example inputs and outputs are crucial for the understanding of the workflow, for validation, and for maintenance purposes.
5. **Annotate:** We propose to annotate a workflow as much as possible. Annotating a workflow carefully can be seen as doing good science, and helps to record what is needed for a publication later on. It also facilitates use and re-use of workflows [6].
6. **Make it executable from outside the local environment:** This Best Practice is expected to increase the reproducibility of the workflow.
7. **Choose services carefully:** One of the major reasons that cause workflows to break are volatile third-party services [6].
8. **Reuse existing workflows:** Reuse is important for many reasons. It fights redundancy, and may lead to better workflows if the developer repairs it upon feedback from the person that wants to reuse the workflow for his own purposes. It will also help the workflow developer get ideas on methods and workflow patterns. For example, another user that is familiar with one of your workflows is more likely to understand another workflow that you designed. It is also beneficial when repairing workflows: repairing a given workflow may entails repairing the workflows in which it is used as a subworkflow.
9. **Advertise:** It is a duty of science to share your results. It also helps progress by letting others build on your work without reinventing it.
10. **Maintain:** Workflow maintenance is expected to increase the longevity of the workflows.

The results of using these best practices will be reported and discussed in section 3.3.

2.2 RO management

The workflows with their related data and documentation were described and preserved according to the last version of the Wf4Ever reference implementation [1]. In particular, we used the RO manager[8] to create and manage the RO, and the Portal to the Research Object Digital Library (RO Portal)[9] to store and visualize the RO. The RO manager and the RO Portal implement the RO models “wf4ever” and “wfdesc” [5] for workflow lifecycle management developed within WP2. Provenance is captured by the “wfprov” model [7]. RO evolution, sharing and collaboration aspects as reported in [8] were not explored due to lack of functional user interfaces. Similarly, RO completeness and quality as described in [7] are not explored in this deliverable due to lack of functional user interfaces. These aspects will be dealt with in the final WP6 deliverable that is due at end of the project (D6.3v3). We conducted the following steps when creating the RO:

1. All files related to the experiment were sorted into folders in the local file system (see section 4.2).
2. Github[10] was used to create a working copy of the RO[11] that could be used by the developers in the project for test purposes, and to manage all content before creating the actual RO (that is, an RO implemented using the Wf4Ever technology).
3. The RO manager was used to manage the RO and map it to the wfdesc, wf4ever, and wfprov models. All RO manager commands were collected in an executable shell script file, which creates the research object from scratch based on the raw files (see Appendix B). For example, all the elements in the "Workflows" folder are wfdesc:Workflows, all the elements in the "Datasets" folder are wf4ever:Datasets, and all the elements in the "WorkflowRuns" folder are wfprov:WorkflowRuns. A workflow run is linked to the .t2flow file by the relationship "wfprov:describedByWorkflow". The "wfprov:wasEncatedBy" relationship was used to assert that a workflow run was executed in Taverna.
4. The Wf4Ever tool scufl2-wfdesc[12] was used to extract a wfdesc workflow description from a Taverna workflow (.t2flow).
5. Provenance from all workflow runs were exported by using the PROV plugin in Taverna. That version does not expose the intermediate values as files, but their trace (without the actual values) is there in the PROV.
6. We retrieved the PROV to wfprov mapping by first installing cwm[13] and then running the prov-to-wfprov[14] program.
7. The RO manager was used to upload the RO to the RODL[15] for preservation.

[8] https://github.com/wf4ever/ro-manager
[9] http://sandbox.wf4ever-project.org/portal/ro
[10] https://github.com/
[11] https://github.com/wf4ever/ro-catalogue/tree/master/v0.1/concept-profile-matching-golden-exemplar
[12] https://github.com/wf4ever/scufl2-wfdesc
[13] http://www.w3.org/2000/10/swap/doc/cwm.html
[14] https://github.com/wf4ever/ro/blob/master/mapping/prov-o/prov-to-wfprov.n3

8. The RO was visualized in the RO portal (see Appendix C).

[15] http://sandbox.wf4ever-project.org/rodl/ROs/concept-profile-matching-golden-exemplar

3. Workflow results and discussion

We developed workflows to interpret the top hits from a GWAS. Workflow 3124[16] (see Appendix A, Figure 1) uses manually curated metabolic pathway information to annotate SNPs with metabolic pathways, and pack 282[17] (See Appendix A, Figure 2-4) uses information from the scientific literature to predict biological processes associated with the SNPs.

3.1 Workflow 3124: Mining the Kegg pathway database with the top hits from a GWAS

Workflow 3124 takes the name of the SNP as input and gives a report stating the name of the SNP, the Entrez gene[18] identifiers of the associated genes and the KEGG[19] identifiers and names of the associated metabolic pathways as output (see Appendix A, Figure 1). As a result of the Best Practices for workflow design, Harish added example inputs and example outputs to the workflow, and annotated all workflow input and output ports and processes.

3.2 Pack 282: Functional annotation of SNPs using next-generation text mining

The pack implements the concept profile matching pipeline in Anni, and adapts it to interpret the top SNPs from a GWAS (the SNP to gene step is currently not possible in Anni). The underlying literature and concept profile databases are the same as for the current version of Anni (2.1). Pack 282 consists of three main workflows (2999, 2973, 2972) (see Appendix A, Figure 2-4).

Workflow 2999 returns a list of available concept sets to match the SNPs against. For the experiment described in this study, the concept set GO biological processes was used. The ID for the concept set that is chosen by the user can be used as input for workflow 2973.

Workflow 2973 consist of nine nested workflows (eight different components in the pack, for which one is used twice within the main workflow for different purposes). It reuses the SNP to Entrez gene identifier part of workflow 3124, here implemented as a separate workflow. Workflow 2973 takes a SNP ID and a set of GO biological processes as input together with a set of parameter settings. It then goes through a number of steps corresponding to actions taken by the user for a similar task in Anni:

a) Map the input SNP identifiers to Entrez gene identifiers (in Anni, this step would have to be performed by the user as a pre-processing step since Anni currently does not accept SNP identifiers). Workflow: SNP_ID2EntrezGene_ID[20].

b) Map the input Entrez gene identifiers to concept identifiers. Workflow: DatabaseID_to_ConceptID[21].

[16] http://www.myexperiment.org/workflows/3124
[17] http://www.myexperiment.org/packs/282
[18] http://www.ncbi.nlm.nih.gov/gene
[19] http://www.genome.jp/kegg/
[20] http://www.myexperiment.org/workflows/2971
[21] http://www.myexperiment.org/workflows/2969

c) Filter for concepts that have a concept profile in the database. Workflow: Filter_concepts_with_profiles[22].

d) Retrieve concept set to match with. Workflow: Get_Concept_IDs[23].

e) Match the concept profiles of the genes with concept profiles for GO Biological Processes and return a matching score. Workflow: Match_concept_profiles[24].

f) Inspect documents that includes both the query and the match concepts. Workflow: Find_co_occurring_documents[25].

g) Explain the predicted associations between the query concept and matched concepts that have no co-occurring documents by retrieving the intermediate concepts contributing to the matching score. Workflow: Explain_concept_scores[26]. (As an extra step, the literature supporting the association can be retrieved using the separate workflow Find_supporting_documents[27].)

The workflow "Get_concept_information"[28] is used twice in workflow 2973. The workflow implements a common operation in Anni which is not performed by the user but used by Anni when performing other tasked asked by the user. The outputs from workflow 2973 give the following information: 1) information about the GO biological process with the highest matching score against the SNPs, 2) if there are any co-occurring documents supporting this finding, and 3) the concepts contributing most to the match (intermediate concepts). The output parameters can be saved in Taverna as an Excel sheet together with the input parameters. Documents supporting the relation with the intermediate concepts contributing most to the match can be retrieved with workflow 2972.

3.3 Following the Best Practices for Workflow Design

The workflows in pack 282 were implemented according to the Best Practices for workflow design as described in section 2.1, with the following results:

1. **Make an abstract workflow:** A workflow sketch was made in power point, since there is no accepted standard for creating such a sketch (see Appendix A, Figure 5). The sketch tries to capture the whole experiment at a higher level and includes workflow 3124 as well as pack 282. The workflow sketch is aggregated in the RO.

2. **Use modules:** All executable components were implemented as separate, runnable workflows. This indeed facilitated independent testing and validation of the execution of each of the individual components as well as the main nested workflow. We did however notice that Best Practices on how

[22] http://www.myexperiment.org/workflows/3178
[23] http://www.myexperiment.org/workflows/2997
[24] http://www.myexperiment.org/workflows/2620
[25] http://www.myexperiment.org/workflows/3006
[26] http://www.myexperiment.org/workflows/2725
[27] http://www.myexperiment.org/workflows/2972
[28] http://www.myexperiment.org/workflows/2998

to test a workflow and its nested components are missing. We suggest that something similar to unit testing would be useful for testing workflows as well.

3. **Think about the output:** The outputs of the workflows were implemented as output boxes in Taverna. They can be saved to disc from Taverna, for example using the save to Excel option. Even though the output can be saved from Taverna in this way, the limited export options give a scattered impression and it can be difficult to relate the different outputs to each other. Better export options are desired, and a requirement for future versions of Taverna.

4. **Provide example inputs and outputs:** All workflows have example inputs and outputs. The example outputs match the example inputs. However, there is no standard for how to do this if the example is a large data file that does not fit in the example window of Taverna. One way to solve this would be to provide the example files in the RO and use the wfdesc:Input and wfdesc:Output properties to annotate the example files. We did this when needed.

5. **Annotate:** All inputs and outputs, processes (for example Web services) and subworkflows (nested components) of the workflows are annotated. Inputs and outputs are annotated with a brief description and an example value. Processes are annotated based on their purpose in the workflow. Technical details are left out, since they are considered to belong to the original process description (for example, a Web service technical details should be provided in the service catalogue). Workflows are annotated with a purpose, and the names of the workflow creator and workflow contributors. While annotating the workflows, we noticed a lack of standards for how to do this. We used Taverna description fields and example fields, but noticed that myExperiment does not use all this information. For example, Web service descriptions were not propagated. We propose to choose meaningful names for the workflow title, inputs, outputs, and for the processes that constitute the workflow, and to focus on 'how' a component is used in this workflow and 'why' it is in there. We also recommend to include a reference to information about what the component does in general (e.g. by referencing a service on BioCatalogue), and to describe the resource keeping it mind that it may disappear or change at some time in the future. We noticed that when importing services from BioCatalogue or workflows from myExperiment, no reference to the origin seem to be preserved in Taverna. The link to its origins are lost. Another issue is related to subworkflows. Some subworkflows are used multiple times in a workflow. The relationship between the copies is lost, which necessitates annotating the same component more than once. Also, annotations of underlying components (e.g. subworkflows) are not visible at higher levels of a nested workflow. The structure of annotations could be more refined. Authors could have contributors. Descriptions could e.g. be divided into purpose in workflow, results, references, origin. We also missed naming conventions for different parts of a workflow (input node, output node, nested workflow etc). In general, showing annotations from linked resources in a workflow (e.g. links to/from nested or imported components) will make it seem less intimidating to add/edit annotations, and it will be helpful for a user. We propose that myExperiment and Taverna draw from the same linked sources (the ROs), showing appropriate fields at appropriate times. Links that are references may be added manually, but then

you would like to be helped by the interface (e.g. fields for 'see also' references, such as to myExperiment).

6. **Make it executable from outside the local environment:** The workflows can be run outside the LUMC. They use public Web services and have been tested outside the LUMC.
7. **Choose services carefully:** The workflow uses public Web services that have a green light on BioCatalogue. This green light, together with very limited history on BioCatalogue, is the only available trust-metric for a Web service[29]. More effort to develop and implement reliability statistics for Web services is needed.
8. **Reuse existing workflows:** We made our workflows modular and noticed that the modularity made possible to use one subworkflow twice in a nested workflow.
9. **Advertise:** The workflows were put on myExperiment, both as separate workflows and as a pack. Other ways to advertise workflows could be to provide links to them from scientific publications.
10. **Maintain:** We respond to e-mails from users regarding Web services being down and causing the workflow to break. We plan the maintenance to monthly tests of the workflows by running them with their example values. A schedule for this in myExperiment with build-in reminders (for example automatically generated e-mails alerting the developer that the workflow needs to be run) would help the maintenance.

3.4 Comparison of workflow results

The results from the two different approaches to annotate SNPs, and how well they compared to manual annotations are presented in tables 1-3. The workflow that annotates the SNPs with pathways from the KEGG database (workflow 3124) was able to reproduce 10 out of the 15 manually curated SNP functions, and gave no results for five SNPs (denoted as -1 in table 2). In comparison, the workflows that annotate SNPs with associations from the literature (pack 282) were able to reproduce 10 of the 15 manually curated SNP functions, and gave suggested annotations for the remaining five. For the SNPs that mapped to genes with unknown function, the workflows in pack 282 suggests a gene function that needs to be further investigated and confirmed by laboratory procedures.

[29] http://www.biocatalogue.org/wiki/doku.php?id=public:help:biocatalogue:how_we_monitor_web_services

SNP ID	Gene name	Illig et al annotation
rs174547	FADS1	very long unsaturated fatty acid biosynthesis (desaturation)
rs2014355	ACADS	mitochondrial fatty acid beta-oxidation
rs211718	ACADM	mitochondrial fatty acid beta-oxidation
rs2286963	ACADL	mitochondrial fatty acid beta-oxidation
rs9393903	ELOVL2	very long unsaturated fatty acid biosynthesis (elongation)
rs2216405	CPS1	Glutamate metabolism
rs7156144	PLEKHH1	gene function unknown; suggested SNP function: metabolism of plasmalogen
rs11158519	SYNE2	gene function unknown; suggested SNP function: sphingosine-1-phosphate phosphatase 1 activity.
rs168622	SPTLC3	sphingolipid biosynthesis
rs8396	ETFDH	fatty acid beta-oxidation, suggested SNP function: acylcarnitine hydroxylation and carboxylation
rs7094971	SLC16A9	solute carrier, orphan receptor (study suggests carnitine to be the substrate)
rs2046813	ACSL1	The gene is not discussed further in the paper
rs603424	SCD	mono-unsaturated long chain fatty acid biosynthesis (desaturation)
rs272889	SLC22A4	active transport of carnitine and acetyl-carnitine
rs541503	PHGD	serine biosynthesis

Table 1: Manual annotation of the SNPs.

SNP ID	Gene name	KEGG
rs174547	FADS1	path:hsa01040 Biosynthesis of unsaturated fatty acids - Homo sapiens (human)
rs2014355	ACADS	path:hsa00071 Fatty acid metabolism - Homo sapiens (human)
rs211718	ACADM	path:hsa00071 Fatty acid metabolism - Homo sapiens (human)
rs2286963	ACADL	path:hsa00071 Fatty acid metabolism - Homo sapiens (human)
rs9393903	ELOVL2	path:hsa00062 Fatty acid elongation - Homo sapiens (human)
rs2216405	CPS1	path:hsa00250 Alanine aspartate and glutamate metabolism - Homo sapiens (human)
rs7156144	PLEKHH1	-1
rs11158519	SYNE2	-1
rs168622	SPTLC3	path:hsa00600 Sphingolipid metabolism - Homo sapiens (human)
rs8396	ETFDH	-1
rs7094971	SLC16A9	-1
rs2046813	ACSL1	path:hsa00071 Fatty acid metabolism - Homo sapiens (human)
rs603424	SCD	path:hsa01040 Biosynthesis of unsaturated fatty acids - Homo sapiens (human)
rs272889	SLC22A4	-1
rs541503	PHGD	path:hsa00260 Glycine serine and threonine metabolism - Homo sapiens (human)

Table 2: KEGG annotation of the SNPs.

SNP ID	Gene name	Concept profile matching
rs174547	FADS1	unsaturated fatty acid biosynthesis
rs2014355	ACADS	fatty acid beta-oxidation
rs211718	ACADM	fatty acid beta-oxidation
rs2286963	ACADL	fatty acid beta-oxidation
rs9393903	ELOVL2	unsaturated fatty acid biosynthesis
rs2216405	CPS1	urea cycle
rs7156144	PLEKHH1	stimulation of tumor necrosis factor production
rs11158519	SYNE2	centrosome positioning
rs168622	SPTLC3	sphingolipid biosynthesis
rs8396	ETFDH	choline metabolism
rs7094971	SLC16A9	glucokinase regulator
rs2046813	ACSL1	cutin biosynthesis
rs603424	SCD	unsaturated fatty acid biosynthesis
rs272889	SLC22A4	organic cation transport
rs541503	PHGD	l-serine biosynthesis

Table 3: Annotation of the SNPs by the use of concept profile matching (next-gen text mining).

4. General discussion

4.1 Workflow development

This deliverable describes the workflows developed during phase II of the project. The workflow for annotating SNPs with KEGG pathways developed during phase I of the project (see D6.3v1) has been completely redesigned with computational performance and the Best Practices for workflow design in mind. It has changed workflow number on myExperiment from 2622 to 3124, reflecting the fundamental changes to the workflow (minor changes would only cause the workflow to be updated with a new version number on myExperiment). The workflow for annotating SNPs with associations from the literature developed during phase I of the project (see D6.3v1) was only at the design stage at the time of the D6.3v1 deliverable. At that time, it contained two subworkflows (workflow 2622 and workflow 2620), and could not be executed. In comparison, the suite now contains 11 executable workflows designed with the Best Practices for workflow design in mind. The Best Practices for workflow design helped us in the workflow design and implementation process and to formulate requirements for future versions of Taverna and MyExperiment, as described in section 3.3. We intend to continue using these Best Practices when developing the workflows that will be reported in the next WP6 deliverable, and to report on the results of their use.

4.2 Impact

SNP annotation is often performed manually by searching curated databases and/or reading the scientific literature. The steps taken during this process are often not documented or preserved. We have tried to automate much of this process by using workflows that draw information from the curated database KEGG and the literature mining tool Anni. The workflows were implemented using Best Practices for workflow design, with the aim to combat workflow decay already at the design stage. The Best Practices for workflow design have been presented internally at the LUMC, and at the 2nd BioVeL Workshop on taxonomic and phylogenetic workflows[30] in Sweden, May 10-11, 2012.

4.3 Quality and completeness

The quality and completeness of the RO is expected to be expressed using the metrics described in [7], when functional user interfaces have been developed. In the meantime, we try to minimize workflow decay by following the Best Practices for workflow design and to work towards RO completeness by using a high-level tree-folder structure (see section 4.4).

4.4 Annotations and RO Building

Development of Best Practices for RO users are currently proposed as a showcase in the showcase backlog[31] (the showcase backlog is part of the Wf4Ever project work methodology[32]). Until these are ready, we believe that workflow-centric ROs for genomics are best described following a high-level tree-folder structure, since earlier attempts with trying to create too restrictive standards within the genomics domain has failed. We expect that Wf4Ever tooling will make the structure for us (the genomics users) and allow us to be flexible, instead of the other way around. In line with this reasoning, we propose the use of at least four different folders:

1) Datasets – for optional workflow input and output files that are too big to give as *examples* in Taverna, and for large *actual* workflow input and output files. Annotations should make their purpose and workflow and/or workflow run relations clear.

2) Documents – at least a HOWTO file describing how to use the RO, and a README files describing how to interpret the RO content. At the moment this folder also contains the shell script that was used to make the RO using the RO manager. This could also be a place to put a reference bibliography file.

3) Workflows – executable workflow files. How the workflows are related to each other is described in the HOWTO and README files, and by looking at annotations for the individual workflows.

4) Workflow-runs – Provenance export from workflow runs, in separate subfolders for each workflow.

With regards to the RO-ification process, much has happened within the project since the D6.3v1 deliverable. The main changes are visible in the RDF manifest: while the manifest for the RO created for the D6.3v1 deliverable only contained Dublin Core metadata terms, the manifest for the RO created for the current deliverable refers to the RO models developed within the project. The large effort being put into the development of these models has not yet left room for implementing a friendly user-interface for RO management. The RO manager and the RO portal were the only tools that were available to us to manage

[30] http://www.biovel.eu/images/events/MS6WorkshopPix/presentations/hettne.ppt

[31] http://www.wf4ever-project.org/wiki/display/docs/Show+case+backlogs#Showcasebacklogs-35.ROGuidelines%2FROPrimerforUsers

[32] http://www.wf4ever-project.org/wiki/display/docs/Work+methodology

our ROs. In order to realise the RO-ification process using these tools, we devoted a showcase in the Wf4Ever backlog to the process[33]. We provided structured feedback[34] on technical issues, which resulted in a new version of the RO manager coupled with extensive documentation (including Frequently Asked Questions (FAQ[35])). The annotations were realised in close collaboration with the technical staff from the different WPs dealing with each of the Wf4Ever models. Except for the RO itself, the results from these efforts can be found in the Annotation Mapping table in the Wf4Ever wiki[36].

4.5 Preservation and versioning

The workflows and their attached data and meta-data are preserved in the RODL. The workflows map identifiers for SNPs with different identifiers for genes, KEGG pathways and GO biological processes. These identifiers are presumed to be stable, but given the growing landscape of naming standards in biology this is an unlikely scenario. For preservation, the provenance (i.e. naming schemes, Web service versions and database versions) related to the use of these identifiers by the relevant Web services at the time of the workflow run needs to be recorded. One could discuss where this should be annotated. We propose this information to be recorded in the Web service catalogue, and automatically propagated to a preserved version of the RO in the RODL. RO evolution, sharing and collaboration aspects as reported in [8] were not explored due to lack of functional user interfaces. Similarly, RO completeness and quality as described in [7] is not explored in this deliverable due to lack of functional user interfaces.

[33] http://www.wf4ever-project.org/wiki/display/docs/RO-ification+of+showcase+62a

[34] http://www.wf4ever-project.org/wiki/display/docs/RO-ification+of+showcase+62a#RO-ificationofshowcase62a-Trialanderror

[35] http://www.wf4ever-project.org/wiki/display/docs/RO+Manager+FAQ

[36] http://www.wf4ever-project.org/wiki/display/docs/Annotation+mapping

5. Concluding remarks

We have created and evaluated workflows for interpretation of GWAS data. The workflows together with their related data, documentation, and metadata are preserved in the RODL. We have explored the RO models “wf4ever”, “wfdesc” and “wfprov” using the RO manager. We noticed a clear need for a mapping of the wf4Ever models to Taverna and MyExperiment, which would enable structured annotations and RO creation and management in a user-friendly environment. Due to lack of functional user interfaces we could not explore aspects of RO evolution, sharing, collaboration, completeness and quality evaluation satisfactory. During the next and final phase of the project we will implement more workflows in the genomics domain and transform them to ROs. As the Wf4Ever tooling develops, we will continue to explore all aspects of the RO using both our already developed ROs and the ones still to come.

References

[1] R. Palma, P. Holubowicz, G. Klyne, A. Garrido, “Reference Wf4Ever Implementation – Phase 1. Deliverable D1.4v1, Wf4Ever Project 2012”, 2012

[2] T. Illig et. al., A genome-wide perspective of genetic variation in human metabolism. Nature Genetics 2010, 42(2):137-41.

[3] R. Jelier, P.A.C. 't Hoen, E. Sterrenburg, J.T. den Dunnen, G.J. van Ommen, J.A. Kors, B. Mons. “Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease”. BMC Bioinformatics, vol. 9, 2008 p. 291.

[4] R. Jelier, J.J. Goeman, K.M. Hettne, M.J. Schuemie, J.T. den Dunnen, and P.A.C. ’t Hoen, “Literature-aided interpretation of gene expression data with the weighted global test,” *Briefings in Bioinformatics*, vol. 12, 2011, p. 518-29.

[5] S. Bechhofer, K. Belhajjame, E. Garcia, “Design, implementation and deployment of workflow lifecycle management components – Phase I. Deliverable D2.2v1, Wf4Ever Project 2012”, 2012

[6] J. Zhao, J. M. Gomez-Perez, K. Belhajjame, G. Klyne, E. Garcia-Cuesta, A. Garrido, K. Hettne, M. Roos, D. De Roure and C. Goble. “Why workflows break - Understanding and combating decay in Taverna workflows”, 8th IEEE International Conference on eScience 2012, *accepted*

[7] E. Garcia-Cuesta, J. Zhao, G. Klyne, A. Garrido, J. M. Gomez-Perez, “Design, implementation and deployment of Workflow Integrity and Authenticity Maintenance components – Phase I. Deliverable D4.2, Wf4Ever Project 2012”, 2012

[8] R. Gonzalez-Cabero, E. Garcia-Cuesta, “Design, implementation and deployment of Workflow Evolution, Sharing and Collaboration components – Phase I. Deliverable D3.2v1, Wf4Ever Project 2012”, 2012

[9] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38(Database issue):D355-60.

[10] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, MR. Pocock , A. Wipat, P. Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004, 20(17):3045-54

Appendix A – Workflow graphical representations

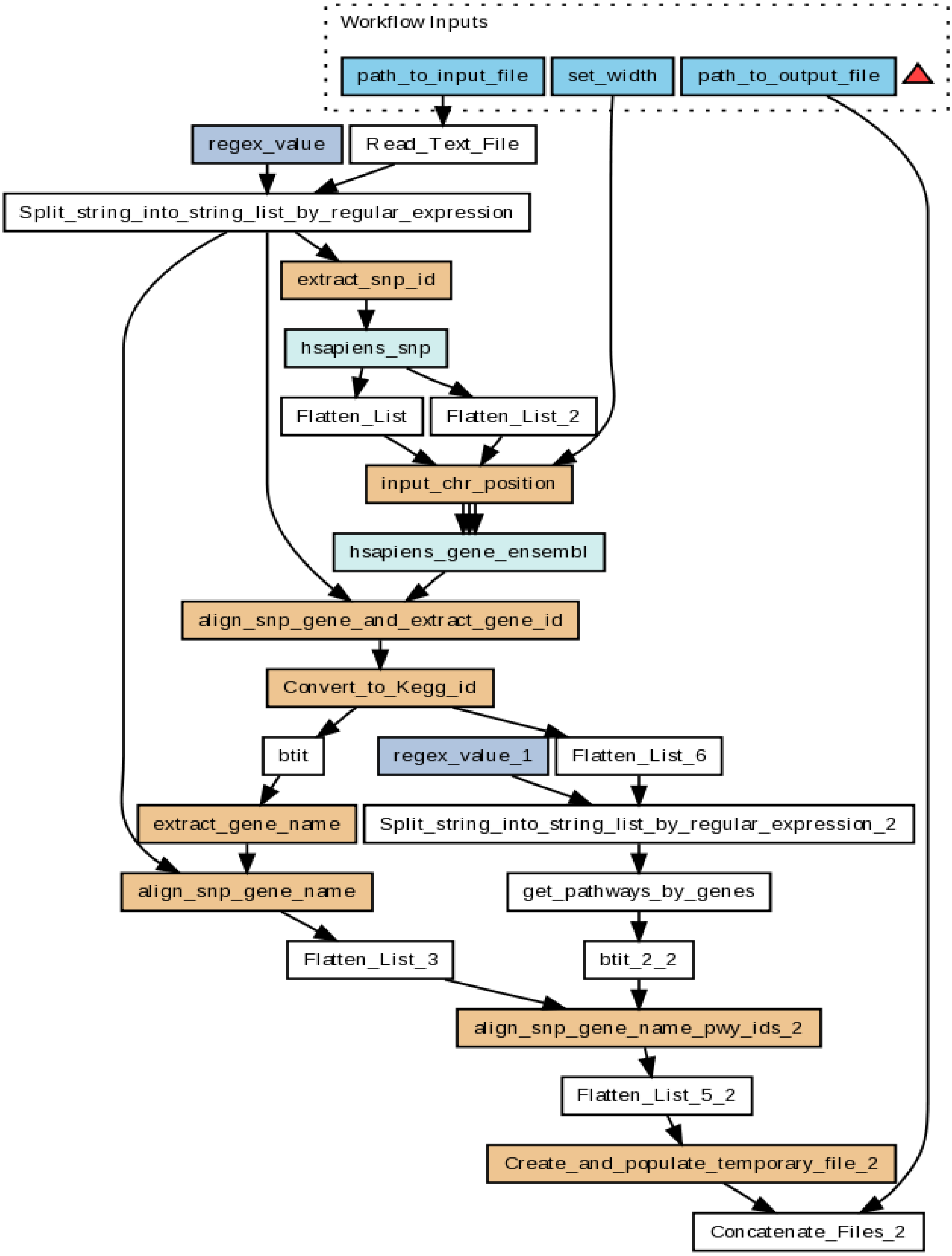


Figure 1: Workflow for annotating SNPs with pathways from KEGG.

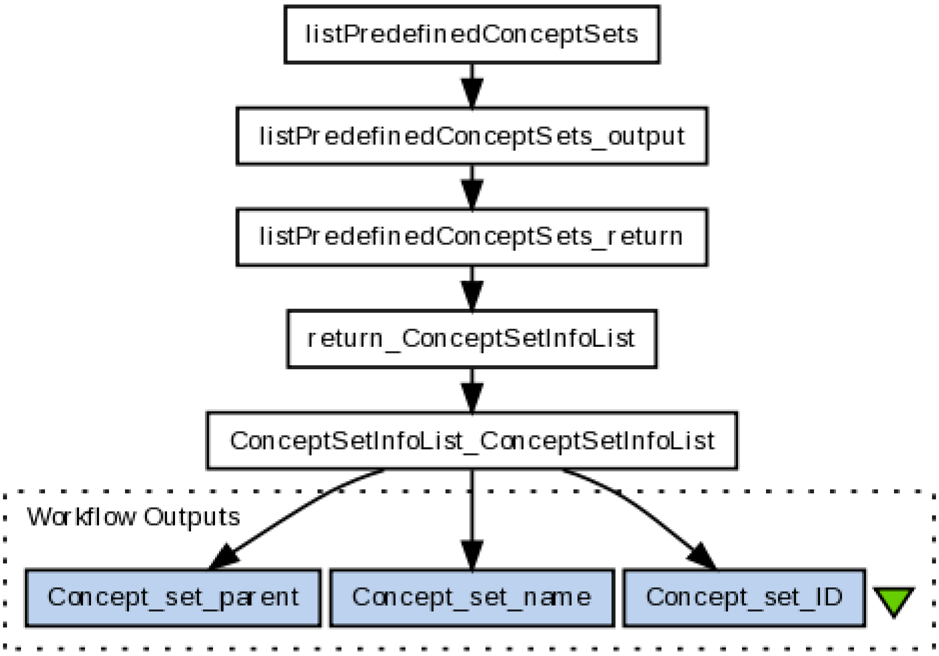

listPredefinedConceptSets
listPredefinedConceptSets_output
listPredefinedConceptSets_return
return_ConceptSetInfoList
ConceptSetInfoList_ConceptSetInfoList
Workflow Outputs
Concept_set_parent
Concept_set_name
Concept_set_ID

Figure 2: Workflow for listing predefined concept sets from the database.

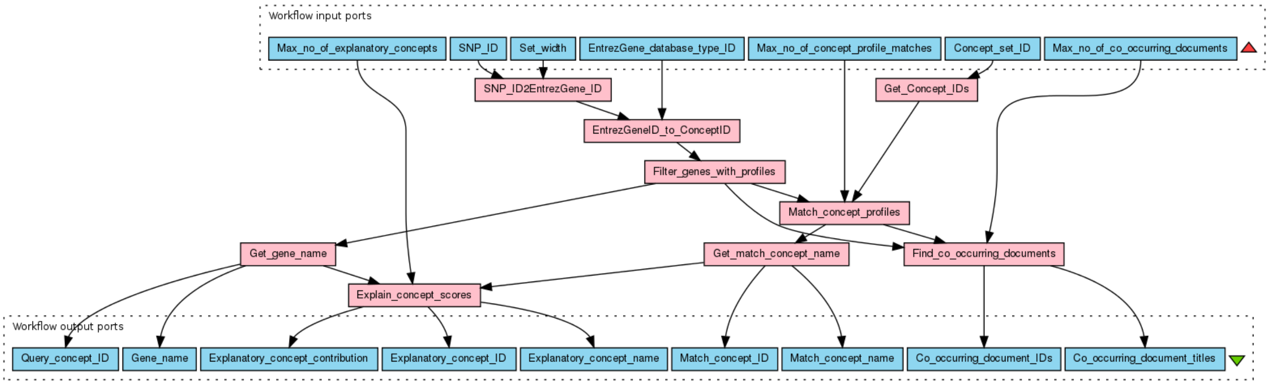


Figure 3: Workflow for interpreting SNPs from a GWAS using next-generation text mining.

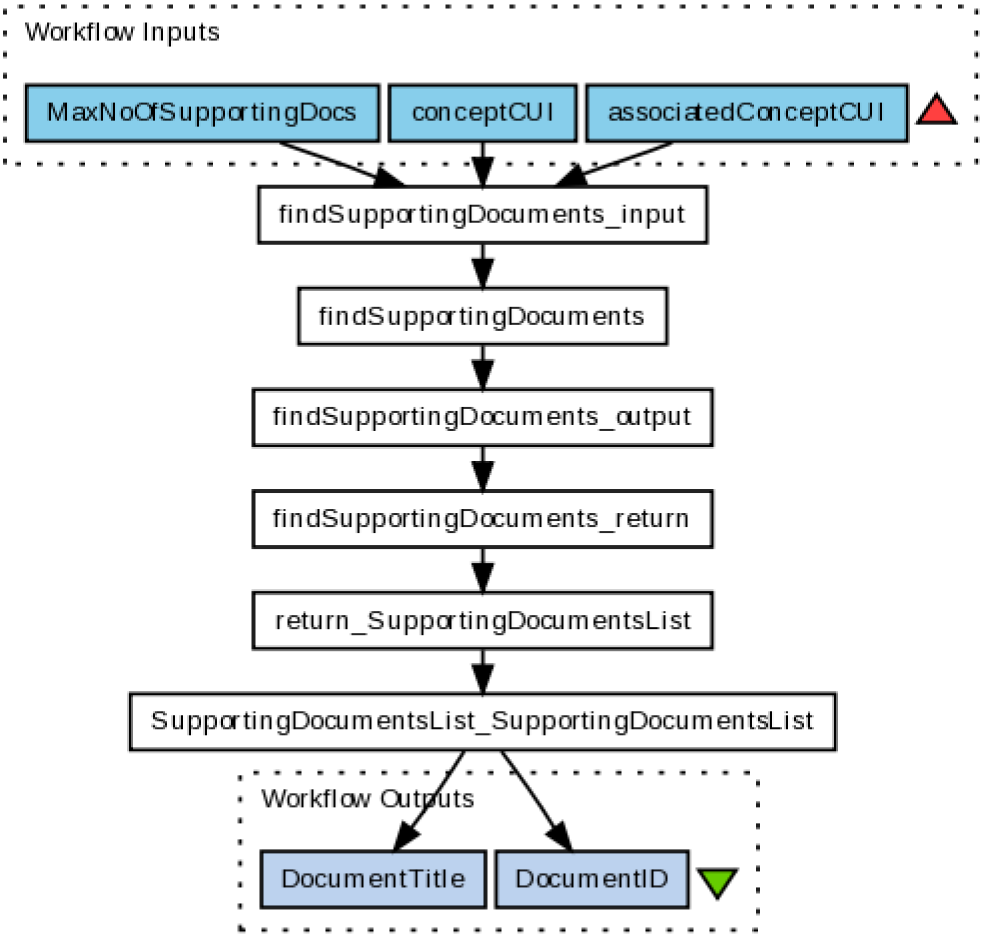

Workflow Inputs
MaxNoOfSupportingDocs
conceptCUI
associatedConceptCUI
findSupportingDocuments_input
findSupportingDocuments
findSupportingDocuments_output
findSupportingDocuments_return
return_SupportingDocumentsList
SupportingDocumentsList_SupportingDocumentsList
Workflow Outputs
DocumentTitle
DocumentID

Figure 4: Workflow for displaying supporting documents for a concept association.

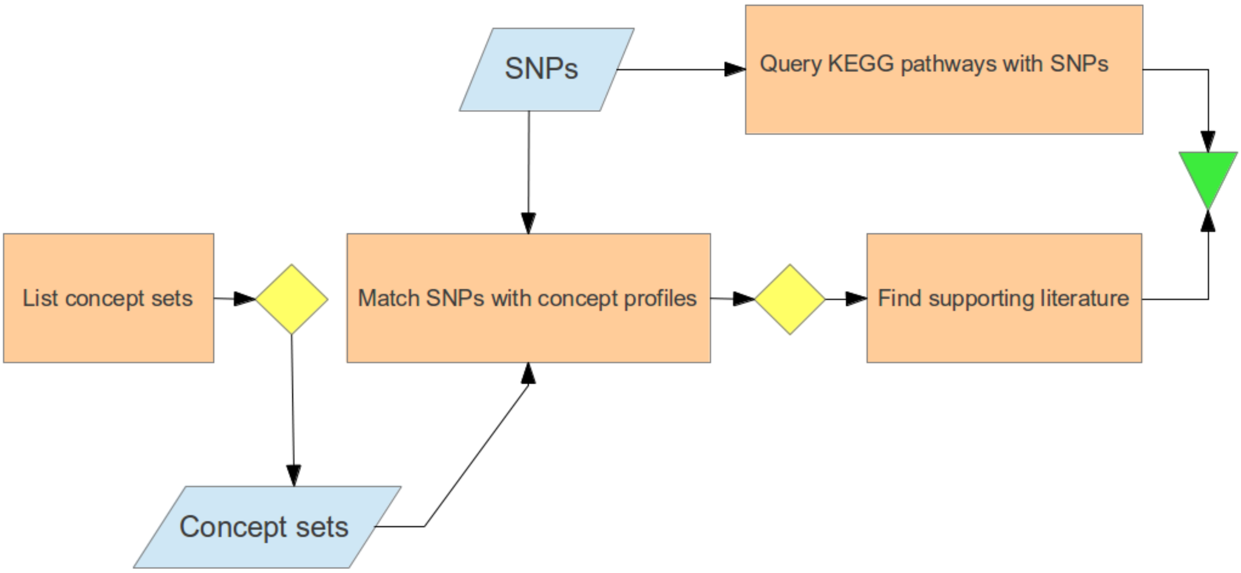

SNPs
Query KEGG pathways with SNPs
List concept sets
Match SNPs with concept profiles
Find supporting literature
Concept sets

Figure 5: Workflow sketch/flowchart.

Appendix B – Make.sh file for creating the RO

```
#!/bin/bash
#
# RO manager script to create trivial RO
#
source ../../ro.sh
# Fail on first error
set -e
TESTRO="."
RONAME="concept-profile-matching-golden-exemplar"
echo "--------"
$RO config -v \
  -b $ROBASE \
  -r http://sandbox.wf4ever-project.org/rosrs5 \
  -n "https://www.google.com/accounts/o8/id?id=AItOawl4I-H_ask7ollkwUzMrjTDrgD2oj_Qu2Q" \
  -t "a5c489aa-b413-47ad-b" \
  -e "k.m.hettne@lumc.nl"
echo "--------"
rm -rf .ro
$RO create -v "Concept Profile Matching Golden Exemplar" -d . -i concept-profile-matching-golden-exemplar
$RO add -v -a -d .
$RO status -v -d .
$RO list -v -d .
echo "--------"
$RO list -v -a -d .
```

```
echo "--------"
# Make wfdesc from t2flow
pushd Workflows
    ../../../lib/scufl2-wfdesc/bin/scufl2-to-wfdesc *t2flow
    # Move away, we don't want to aggregate these annotation bodies
    # as ro:Resources
    mv *wfdesc.ttl ../.ro/
    for t2flow in *t2flow ; do
        $RO annotate -v $t2flow -g ../.ro/$(echo $t2flow|sed s/.t2flow$/.wfdesc.ttl/)
        $RO annotate -v $t2flow rdf:type 'http://purl.org/wf4ever/wfdesc#Workflow'
    done
popd
echo "--------"
pushd Datasets
    for data in * ; do
        $RO annotate -v $data rdf:type 'http://purl.org/wf4ever/wf4ever#Dataset'
    done
popd
echo "--------"
pushd Documents
    for doc in * ; do
        $RO annotate -v $doc rdf:type 'http://purl.org/wf4ever/wf4ever#Document'
    done
popd
echo "--------"
```

```
# Make zip files out of all runs and remove the folders according to the following example commands:

#zip -r Workflow-runs/SNP2KEGG-prov-export.zip Workflow-runs/SNP2KEGG-prov-export

#rm -rf SNP2KEGG-prov-export



$RO annotate -v Workflow-runs/SNP2KEGG-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/List-Concept-Sets-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs8396-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs168622-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs174547-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs211718-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs272889-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs541503-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs603424-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2014355-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2046813-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2216405-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2286963-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs7094971-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs7156144-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'
```

```
$RO annotate -v Workflow-runs/Main-nested-workflow-rs9393903-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs11158519-prov-export.zip rdf:type
'http://purl.org/wf4ever/wfprov#WorkflowRun'



$RO annotate -v Workflow-runs/SNP2KEGG-prov-export.zip wprov:enactorAgent 'Kristina Hettne'

$RO annotate -v Workflow-runs/List-Concept-Sets-prov-export.zip wprov:enactorAgent 'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs8396-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs168622-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs174547-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs211718-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs272889-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs541503-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs603424-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2014355-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2046813-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2216405-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs2286963-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs7094971-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs7156144-prov-export.zip wprov:enactorAgent
'Kristina Hettne'

$RO annotate -v Workflow-runs/Main-nested-workflow-rs9393903-prov-export.zip wprov:enactorAgent
'Kristina Hettne'
```

```
$RO annotate -v Workflow-runs/Main-nested-workflow-rs11158519-prov-export.zip wprov:enactorAgent
'Kristina Hettne'
```

```
$RO link Datasets/output_snp_gene_kegg.txt wfprov:wasOutputFrom Workflows/SNP2KEGG.t2flow

$RO link Datasets/List_Concept_Sets_output.xls wfprov:wasOutputFrom
Workflows/List_Predefined_Concept_Sets.t2flow

$RO link Datasets/rs8396_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs168622_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs174547_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs211718_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs272889_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs541503_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs603424_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs2014355_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs2046813_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs2216405_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs2286963_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs7094971_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs7156144_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs9393903_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow

$RO link Datasets/rs11158519_output.xls wfprov:wasOutputFrom Workflows/main_nested_workflow.t2flow
```

```
$RO link Datasets/top_snps_to_annotate_input.txt wfprov:usedInput
Workflows/main_nested_workflow.t2flow

$RO link Datasets/top_snps_to_annotate_input.txt wfprov:usedInput Workflows/SNP2KEGG.t2flow
```

```
$RO annotate -v Documents/HOWTO.txt rdfs:comment 'Text file describing the protocol for the
experiment, including the order of workflow execution.'
```

```
$RO annotate -v Documents/README.txt rdfs:comment 'Text file describing the background of the
experiment, and the formulated hypothesis.'
```

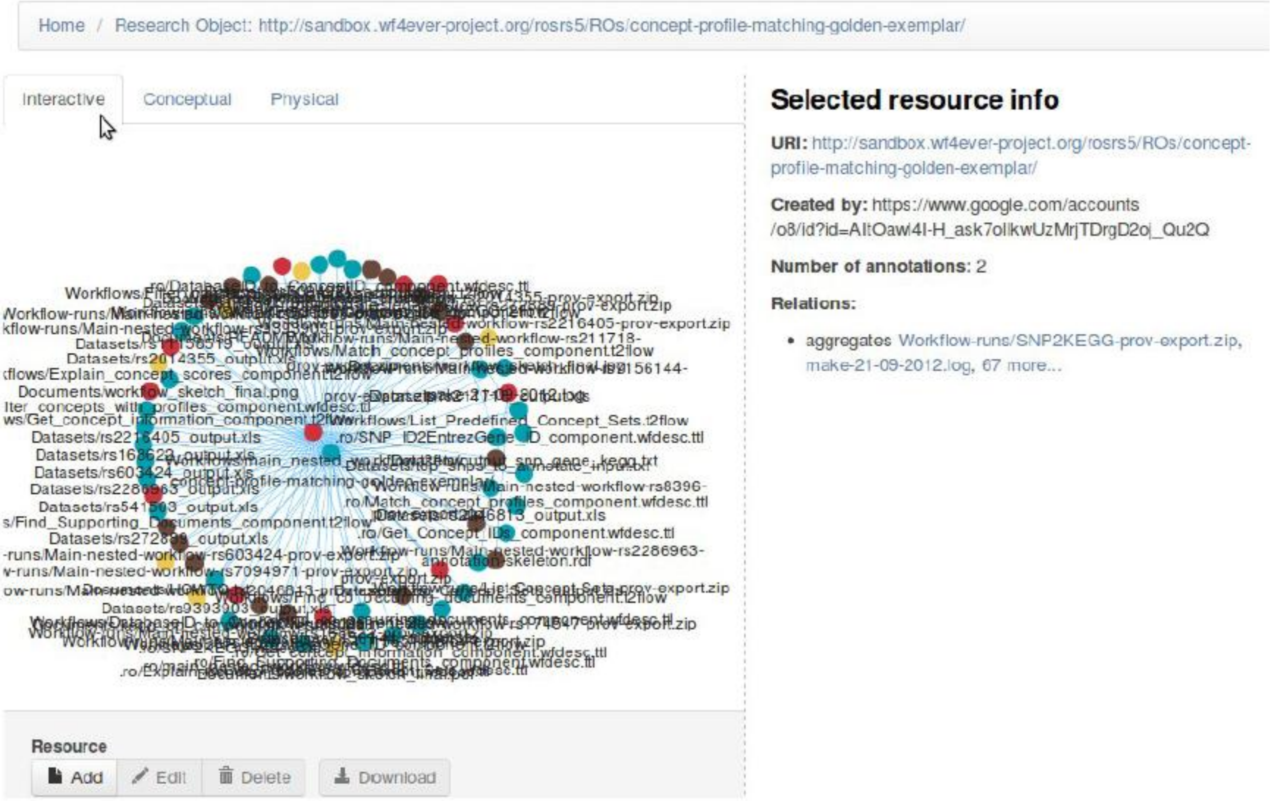
```
$RO annotate -v Documents/kegg_cp_comparison_results.xls rdf:type
'http://purl.org/wf4ever/wfdesc#Results'

$RO annotate -v Documents/kegg_cp_comparison_results.xls rdfs:comment 'Excel file comparing the
results from the workflows.'

$RO annotate -v Documents/workflow_sketch_final.jpg rdfs:comment 'High-level overview of experiment
and order of workflow execution.'

$RO annotate -v Documents/workflow_sketch_final.odp rdfs:comment 'High-level overview of experiment
and order of workflow execution.'

$RO annotate -v Documents/workflow_sketch_final.pdf rdfs:comment 'High-level overview of experiment
and order of workflow execution.'

$RO annotate -v Documents/workflow_sketch_final.png rdfs:comment 'High-level overview of experiment
and order of workflow execution.'
```

```
$RO link Workflow-runs/List-Concept-Sets-prov-export.zip wfprov:describedByWorkflow
Workflows/List_Predefined_Concept_Sets.t2flow

$RO link Workflow-runs/SNP2KEGG-prov-export.zip wfprov:describedByWorkflow Workflows/SNP2KEGG.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs8396-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs168622-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs174547-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs211718-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs272889-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs541503-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs603424-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO link Workflow-runs/Main-nested-workflow-rs2014355-prov-export.zip wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow
```

```
$RO   link   Workflow-runs/Main-nested-workflow-rs2046813-prov-export.zip   wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO   link   Workflow-runs/Main-nested-workflow-rs2216405-prov-export.zip   wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO   link   Workflow-runs/Main-nested-workflow-rs7094971-prov-export.zip   wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO   link   Workflow-runs/Main-nested-workflow-rs7156144-prov-export.zip   wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO   link   Workflow-runs/Main-nested-workflow-rs9393903-prov-export.zip   wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO  link  Workflow-runs/Main-nested-workflow-rs11158519-prov-export.zip  wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow

$RO   link   Workflow-runs/Main-nested-workflow-rs2286963-prov-export.zip   wfprov:describedByWorkflow
Workflows/main_nested_workflow.t2flow



$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/Filter_concepts_with_profiles_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/Find_co_occurring_documents_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/Get_concept_information_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/Match_concept_profiles_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/DatabaseID_to_ConceptID_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/SNP_ID2EntrezGene_ID_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/Get_Concept_IDs_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/Explain_concept_scores_component.t2flow

$RO         link         Workflows/main_nested_workflow.t2flow          wfdesc:hasSubWorkflow
Workflows/Find_Supporting_Documents_component.t2flow



$RO annotate -v . -g ro_annotation.rdf
```

```
echo "--------"

$RO annotations -v

# End.
```

Appendix C. Visualization of the RO in the RO portal



Home / Research Object: http://sandbox.wf4ever-project.org/rosrs5/ROs/concept-profile-matching-golden-exemplar/

Interactive
Conceptual
Physical
Workflows
Documents/workflow_sketch_final.png
Datasets/rs2214355_output.xls
Datasets/rs2216405_output.xls
Datasets/rs603424_output.xls
Datasets/rs541593_output.xls
Datasets/rs272889_output.xls
Datasets/rs9393903_output.xls
Workflows/List_Predefined_Concept_Sets.t2flow
.ro/SNP_ID2EntrezGene_ID_component.wfdesc.ttl
.ro/Match_concept_profiles_component.wfdesc.ttl
.ro/Get_Concept_IDs_component.wfdesc.ttl
Workflows/Match_concept_profiles_component.t2flow
Workflow-runs/Main-nested-workflow-rs2216405-prov-export.zip
Workflow-runs/Main-nested-workflow-rs603424-prov-export.zip
Workflow-runs/Main-nested-workflow-rs7094971-prov-export.zip
s/Find_Supporting_Documents_component.t2flow
annotation_skeleton.rdf
concept-profile-matching-golden-exemplar

Selected resource info

URI: http://sandbox.wf4ever-project.org/rosrs5/ROs/concept-profile-matching-golden-exemplar/

Created by: https://www.google.com/accounts/o8/id?id=AItOawl4I-H_ask7olIkwUzMrjTDrgD2oj_Qu2Q

Number of annotations: 2

Relations:

- aggregates Workflow-runs/SNP2KEGG-prov-export.zip, make-21-09-2012.log, 67 more...

Figure 6: The RO visualized in the RO portal. Colors refer to classes in the Wf4Ever RO models.