Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

STREP FP7-ICT-2007-6 270192

Objective ICT-2009.4.1 b) – "Advanced preservation scenarios"

D6.3v3: Genome Wide Association Study workflows v3

Deliverable Co-ordinator: Kristina Hettne

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

Other Authors: Marco Roos (LUMC); Eleni Mina (LUMC); Eelke van de Horst (LUMC), Harish Dharuri (LUMC); Mark Thompson (LUMC); Don Cruickshank (University of Oxford); Stian Soiland-Reyes (University of Manchester); Sander van Boom (University of Applied Sciences Leiden)

This deliverable provides Genome Wide Association Study workflows.

Document Identifier:	Wf4ever/2010/D6.3v3/v1.0	Date due:	September 30, 2013
Class Deliverable:	Wf4ever 270192	Submission date:	**September 30, 2013**
Project start date:	December 1, 2010	Version:	v1.0
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@uva.nl	

Change Log

Version	Date	Amended by	Changes
0.1	26-08-2013	Kristina Hettne	Document creation and outline
0.2	30-08-2013	Kristina Hettne, Eleni Mina, Eelke van de Horst, Sander van Boom	Added the introduction and materials and methods section
0.3	31-08-2013	Kristina Hettne	Edited the materials and methods section
0.4	02-09-2013	Kristina Hettne, Harish Dharuri	Edited the materials and methods section
0.5 (QA)	03-09-2013	Kristina Hettne, Eleni Mina, Don Cruickshank, Stian Soiland-Reyes	Changed the materials and methods to Strategy and workflow building, added the RO use scenarios, general discussion, concluding remarks, references, figures, and the executive summary.
0.5 (QA corrected)	23-09-2013	Kristina Hettne, Eelke van de Horst, Mark Thompson	Edited the document following QA comments.
1.0	30-09-2013	Kristina Hettne	Final version.

Executive Summary

This document describes the workflows developed during phase III 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 and for gene expression data related to Huntington's disease. 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 preserved "Research Object" (RO). A detailed description about the state of the current tooling can be found in D1.4v2. 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. 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 workflows developed during phase III of the project at the Human Genetics Department of the Leiden University Medical Centre (HG-LUMC) and their preservation using Wf4Ever technology. These workflows are fitted into three different case studies. Two of these are genomics-oriented; 1) the Metabolic Syndome (MetS) case study on Genome Wide Association Study (GWAS) data, and 2) the Huntington's disease (HD) case study on gene expression data. The third is purely bioinformatics oriented: the concept profile generation and analysis pipeline case study. These three case studies have been the focus for WP6 during the whole course of the project. An introduction to the three case studies can be found below, including a short description of previous efforts reported for these case studies in WP6 deliverables D6.3v1 and D6.3v2, and progress as reported in the current deliverable (D6.3v3).

1.1 Background on the concept profile generation and analysis case study

Previously, the BioSemantics group at the HG-LUMC has been involved in the invention and exploration of the concept profile matching method for biomedical text mining [1]. Concept profile matching is a knowledge discovery method that has proved successful in generating hypotheses about molecular mechanisms explaining the results from genotype-phenotype studies. This technology has previously been implemented in the Anni standalone application[1]. The monolithic tool is difficult to maintain and provides no way for users to save their procedures, results, or related provenance.

At the core of this technology are the concept profiles, which in the past had to be generated using a number of custom scripts and manual operations. We aimed to move towards a more customizable and service oriented architecture of the concept profile generation pipeline by developing a set of components, workflows and services that represent individual steps of the pipeline. To aid interoperability, when possible, we opted for Semantic Web standards to interface with these components. Our efforts related to these aims have been concreted during the third year of the project, and are presented in the current deliverable.

In addition, and as an alternative to the Anni monolithic tool, we have developed Web services to interface with the legacy databases behind Anni for concept profile analysis. The Web services have been designed with the new concept profile generation pipeline components in mind, for an easy transfer to the new technology when it is ready. We first described the development and use of these Web services in D6.3v1, when they were at an early development stage. At that time, the Anni Web tool user operation of performing the concept profile matching had been transformed into a Web service and was described in a prototype workflow. Further developments were described in D6.3v2, where an additional eight Anni Web tool user operations had been transformed to Web services. Also, a first runnable version of a GWAS analysis workflow using the Anni Web services was presented. During phase III of the project, two additional Anni Web tool user operations have been transformed to Web services, and the already existing Web services have been tested and fine-tuned. Web service specifications and the current status of the Anni Web services workflow pack are presented in the current deliverable, with an accompanying RO implementation.

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

1.2 Background on the Metabolic Syndrome (MetS) case study

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]. To understand the biomolecular basis of the association, scientists typically dwell on identifying genes in the vicinity of the genome variant referred to Single Nucleotide Polymorphism (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 concern interpreting SNP associations from a GWAS on human metabolite variation, using pathways from metabolic pathway databases and Gene Ontology (GO)[2] biological process associations from the concept profile analysis Web services. In previous deliverables we have reported on workflows related to this case study. These workflows have had the SNP as a starting point (see D6.3v1 and D6.3v2); in the current deliverable we focus on workflows having the metabolite as a starting point, with an accompanying RO implementation.

1.3 Background on the Huntingtons Disease (HD) case study

The genetic mutation that causes HD was identified 20 years ago but the downstream molecular mechanisms leading to the HD phenotype are still poorly understood. Epigenetic phenomena such as DNA methylation and histone modifications can cause long-term changes in gene expression over generations of dividing cells. Effects at the level of DNA and higher orders of DNA organisation have been shown to play an important role in the HD pathogenesis. It is clear that new hypotheses that take into account epigenetic mechanisms may shed some light on the downstream mechanisms that lead to HD symptoms.

The first workflow developed for this purpose concerned integration of gene expression microarray data with genome locations and was reported in D6.3v1. The developments during phase II and III of the project related to this case study are reported in the current deliverable. For example, the hypothesis has evolved and the previously reported workflow has been decomposed in 3 parts, to fit the new needs of our experiment. Also, the concept profile analysis workflows have been reused in this case study in order to interpret the results in the light of current literature. An accompanying RO implementation is described.

1.4 Outline

A description of the methodology resources used as well as the workflows themselves is provided in section 2. Different RO technology and tools use scenarios originating from the use cases are presented and

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

discussed in section 3. Issues related with workflow development, Wf4Ever tools, semantic annotations, quality and preservation can be found in section 4. Section 5 is dedicated to conclusions.

2. Materials and methods

All workflows from the three different case studies were created using the Taverna workbench 2.4[3], following the Best Practices for workflow design[4] developed during the course of the project (lead: WP6). In addition, all workflows were stored on myExperiment. However, since some of the workflows are included in manuscripts that are currently either submitted for review or in preparation, they are not public on myExperiment yet. Access will be granted to members of the Wf4Ever project and members of the project review committee upon request. A description of the specific materials and methods related to the different case studies follows.

2.1 Concept profile generation and analysis case study

We start with describing the materials and methods related to concept profile generation and continue with describing the materials and methods related to concept profile analysis. To facilitate collaboration with related project developed by the Netherlands Bioinformatics Center (NBIC)[5], the project is being maintained in the NBIC Development Project Environment under the BioSemantics Beta development project[6].

Concept profile generation

The concept profile generation pipeline (Figure 1) produces concept profiles. It uses a set of concepts that are identified in a document corpus and determines the co-occurrence of concepts in text by means of an indexer and a thesaurus or ontology containing instances. Currently, a prototype pipeline is being developed in which some parts are represented by mock-ups/placeholders. This way, the individual components and the pipeline can be developed in parallel, and the requirements for each part can be analysed at an early stage. We aim to implement the pipeline and its constituting components in a generic and flexible manner in order to promote its extension and repurposing. The Resource Description Framework (RDF) was adopted as the data interchange format between the components. RDF is an open standard endorsed by the W3C for making statements about resources, in particular web resources (such as documents), Input and intermediate results of the pipeline are stored in a RDF triple store. A triple store stores statements about resources as subject-predicate-object triples. Optionally, these triples can be aggregated into graphs. An example would be a triple that states that two documents (subject, object) are related (predicate). Each component is a view on the triple store. The input and output of these components are indicated by graphs. This facilitates the provenance tracking of (intermediate) results. The core entities of the pipeline, and thus in the store, are the concepts, the text resources, and links between them. Concepts and text resources are represented by Uniform Resource Identifiers, and relations between these items as properties of these resources. The choice to build the pipeline on top of a triple store was stirred by its native support for

[3] http://www.taverna.org.uk

[4] http://www.wf4ever-project.org/bestpractices

[5] http://www.nbic.nl

[6] https://trac.nbic.nl/biosemantics_bet_dev/roadmap

Semantic Web technologies and its flexibility. For instance, a database forces one to commit to a fixed schema in advance. As the prototype pipeline matures, a switch from a triple store to a graph database might be considered. An RDF triple store is primarily meant for storing and querying RDF, while a graph database stores any type of graph structure and is optimized for graph operations such as graph traversal. Since most graph databases also support the RDF data model, these databases might be better candidates for analysing and manipulating RDF graphs compared to triple stores, however, graph databases are typically not trivial to set up and use. To aid interoperability, Semantic Web standards such as Simple Knowledge Organization System (SKOS)[7] have been adopted to interface with the components in the pipeline. SKOS is a W3C Semantic Web standard for terminological resources such as controlled vocabularies, thesauri, subject headings, and taxonomies. An advantage of the RDF-based SKOS over other flat-file formats is that vocabularies can be published as part of the emerging web of linked data, easily integrated with other RDF datasets, and processed by Semantic Web applications. We describe work done during year three of the project related to the different components of the pipeline below.

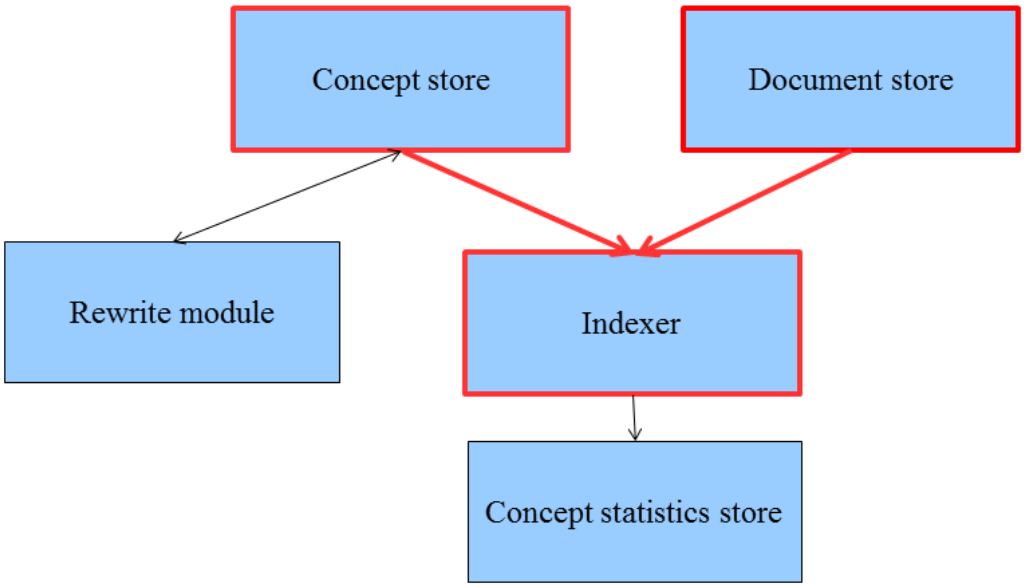

Figure 1. Components of the concept profile generation pipeline.

Indexer

The central component of the pipeline is the indexing engine Peregrine[8]. Peregrine recognizes concepts in human readable text, based on a database (thesaurus) of known terms. Multi-word terms are correctly recognized. If terms can represent multiple concepts, Peregrine will attempt to disambiguate them. Development of Peregrine currently takes place at NBIC.

Concept Store

[7] http://www.w3.org/TR/skos-reference/

[8] https://trac.nbic.nl/data-mining

The concept store is an abstraction of the source of concepts that Peregrine uses. Currently, concepts are imported directly into the underlying triple store. However, with aforementioned SKOS importer, any SKOS vocabulary can be used as a source of concepts for Peregrine. In particular, community curated concept sources, such as the ConceptWiki[9] and BioPortal[10] will be used. Collaboration with ConceptWiki developers has been established to align the SKOS representation of the concepts in the ConceptWiki, their properties and relations, with the format expected by the Peregrine SKOS importer. Refinement of the ConceptWiki SKOS representation is still ongoing and changes to the original model are anticipated. In addition to ConceptWiki, the OpenPHACTS[11] framework has been explored to use as a concept store. Two meetings, one community, one internal, have been attended to evaluate the usefulness for our purpose. Use of BioPortal as a concept source has not yet been investigated. To use existing thesauri that are not in SKOS format, such as Jochem[12], in the new concept profile generation pipeline a tool was developed that translates the legacy ErasmusMC ontology file (EMC) format files to SKOS. This tool will not only help to expand the application domain of thesauri in EMC format, but will also help to create reference tests to compare old concept profile generation pipeline with the one being developed.

Document Store

The document store is an abstraction of the source of documents that are indexed by Peregrine. This component consists of a document manager, Solr 4.4[13], and a set of workflows that retrieve documents and metadata from public and local sources (see section 3.1).

Rewrite module and concept statistics store

The rewrite module is an essential part of the pipeline since it prepares a thesaurus to be used for text mining purposes. Previous work has shown that simple rewrite and suppress rules can have a dramatic improvement on concept recognition [3]. This module, together with the concept statistics store, is the next in priority to be implemented for the pipeline.

Concept profile analysis

Concept profile analysis is the step after the concept profile generation. Concept profile analysis consists of a number of standard user operations applied when using concept profiles for genomics data interpretation. These user operations form a pipeline consisting of multiple workflows. To for example perform pathway analysis for a gene expression experiment, a user would first provide a list of gene names which would be mapped to database identifiers. These database identifiers would in turned be used in the pipeline to query a database for the corresponding concept profiles, which then would be matched with a predefined set of

[9] http://ops.conceptwiki.org

[10] http://bioportal.bioontology.org

[11] http://www.openphacts.org/

[12] http://www.biosemantics.org/index.php?page=jochem

[13] http://lucene.apache.org/solr/4_4_0/

concept profiles of the category “biological process”. The concept profile matching score between the concept profiles for the genes in the uploaded gene list and the concept profiles for the concepts in the concept set “biological process” would be calculated, resulting in a ranked list of biological processes for the gene list. Finally, literature evidence in the form of documents containing co-mentions of the gene and biological processes and/or documents providing enough statistical evidence to support the gene-biological process associations without actually mentioning the gene and the biological process together would be retrieved. Our goals were to create Web services that translate to standard user operations in the Anni standalone application and to make these services available through workflows. Therefore, we adopted an e-Science approach based on (i) Web services for the common operations available in Anni, (ii) workflows for the common procedures enacted by Anni, (iii) a workflow-to-web tool to leverage the functionality of advanced workflows via a simple web interface. The approach stimulates collaboration of specialists: software engineers and computer scientists, bioinformaticians, and biologists. The Web services were designed according to the outcome of an Anni user requirement analysis, where the common user operations were identified. The Web services were implemented using Java, MySQL, Spring 3, and Apache Tomcat following the Java API for XML Web Services (JAX-WS) specifications, and made available through the Life Science Web service registry BioCatalogue[14]. Example user procedures implemented as Taverna workflows can be run on the Web through the t2web tool (see for example http://workflow.biosemantics.org/t2web/workflow/3397).

2.2 MetS case study

One of the workflows[15] reported in D6.3v2 for analysing GWAS data was based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) Web Service Definition Language (WSDL) Simple Object Access Protocol (SOAP) Web services. These were deprecated as of December 31, 2012 and replaced with Representational State Transfer (REST) services[16]. We changed the workflow to be compliant with this new Web service format, following the best practice number 10 from the 10 Best Practices for workflow design, namely “Advertise and Maintain”[17]. Complementing the previously reported GWAS interpretation workflows that approached the problem of explaining SNP-metabolite associations from a GWAS starting with the SNP, map it to one or more genes and retrieve pathways annotated to these genes (see D6.3v1 and D6.3v2), we implemented two workflows that start with the metabolite (see section 3.1

[14] http://www.biocatalogue.org/services/3330

[15] http://www.myexperiment.org/workflows/3124

[16] http://www.kegg.jp/kegg/rest/

[17] http://www.wf4ever-project.org/bestpractices#maintain

2.3 HD case study

The workflows for this case study are available as a pack on myExperiment[18] (access upon request) and described in section 3.1. The previous workflow reported in D6.3v119 (access upon request) has been decomposed in 3 parts, to fit the new needs of the experiment as developed during the course of the project (see section 3.1). The experiment associates gene deregulation in HD with specific genomic regions, namely CpG islands [4] and four chromatin states [5], that we hypothesize to have an epigenetic role in HD gene deregulation. In addition, four workflows from the "Concept profile analysis using Anni Web services" pack described in section 2.1 under "Concept profile analysis" have been reused in this use case (see section 3.1).

[18] http://www.myexperiment.org/packs/485

[19] http://www.myexperiment.org/workflows/2623

3. Results

3.1 Workflows

Concept profile generation and analysis case study

To make Peregrine suitable for use in the concept profile generation workflows, we extended Peregrine to import of SKOS vocabularies in addition to its legacy ErasmusMC ontology file format. This allows us to use any SKOS vocabulary as a source of concepts for indexing. The SKOS import functionality is now integrated in the Peregrine codebase. A prototype workflow entitled "Concept_profile_generation_pipeline_(prototype)"[20] (access upon request) has been implemented that uses the Peregrine Web service without SKOS import. This service will be replaced with Peregrine with SKOS import. The intended input for the above workflow is the articles stored in the document store. To store documents in the document store, we developed the following workflows. The workflow "PubMed_Search_and_Solr_storage"[21] represents a case of workflow reuse, where an existing workflow on myExperiment to retrieve abstracts from biomedical journals using eUtils[22] from MedLine was extended by removing outputs and text extraction and adding an automatic Solr storage process. Because of the limits imposed on document retrieval, a cache of MedLine articles will be used to bootstrap the document store. The workflow "wf4ever_PDF2TXT2Solr_Database"[23] copes with full text articles that are only available in PDF. This workflow uses optical character recognition to extract text from these documents.

The workflows implementing the Anni Web services are available on myExperiment as a pack, with the title "Concept profile analysis using Anni Web services"[24]. The pack consists of 12 workflows. 11 of these can be seen as building blocks, or components, that can be used to create user analysis pipelines mimicking the way a user would interact with the Anni standalone application. An example is the "DatabaseID_to_ConceptID"[25] workflow, which matches a specific database identifier for a gene to the type of id that is used by the databases behind the Anni standalone application and the Anni Web services. The Web service itself is called "mapDatabaseIDtoConceptIDs"[26]. The pack also contains one example of a user analysis pipeline, entitled "GWAS_to_biomedical_concept"[27]. This workflow takes one SNP and a concept set (such as "GO Biological Processes", determined by the user) as input, maps the SNP to a gene,

[20] http://www.myexperiment.org/workflows/3724

[21] http://www.myexperiment.org/workflows/3659

[22] http://www.ncbi.nlm.nih.gov/books/NBK25500

[23] http://www.myexperiment.org/workflows/3656

[24] http://www.myexperiment.org/packs/368

[25] http://www.myexperiment.org/workflows/2969

[26] https://www.biocatalogue.org/soap_operations/28108

[27] http://www.myexperiment.org/workflows/3522

calculates the concept profile matching score between the concept profile for the gene and the concept profiles for the concepts in the concept set, returns the ranked list of concepts in the concept profile set based on concept profile matching score against the gene profile, finds co-occurring documents between the query concept and the match concept with the highest rank (cut-off determined by the user), finds the concept that contributes the most to the match, and the documents that support this finding.

MetS case study

We implemented two workflows to interpretation SNP-metabolite associations resulting from a GWAS, starting from the metabolite: "Kegg:_Pathway_Scheme"[28], and "Kegg:_Reactions_Scheme"[29]. The overall idea behind the design of these workflows was to generate a set of genes that potentially influence the levels of a metabolite due to the common pathways that they share.

The "Kegg:_Pathway_Scheme" uses the KEGG Rest services to determine all the genes operating in the metabolic pathways that the input metabolite participates in. The input for the workflow is a KEGG compound id, and it produces a summary text file of the results that is stored in a local directory.

The "Kegg:_Reactions_Scheme" determines all the enzymes/genes that participate in a radius of two reaction steps around a given metabolite. Similar to the previous workflow, the input for the workflow is a KEGG compound id, and it produces a summary text file of the results that is stored in a local directory. Broadly, the scheme involves the following steps:

1. Determine all the reactions that the given metabolite participates in
2. Determine all the compounds that participate in these reactions
3. Filter certain compounds like H2O, ATP etc. to avoid non-specific connections
4. Determine all the reactions that the compounds passing through step 3 participate in
5. Determine the enzymes that drive the reactions from step 4
6. Determine genes corresponding to the enzymes in step 5
7. Store the Entrez database gene ids as a text file

HD case study

Two workflows ([30],[31] (access upon request)) in the pack were implemented to get differentially expressed genes for two different brain regions. Required inputs are mRNA expression profiles from human brain data of 44 Huntington's Disease-gene-positive cases and 36 age- and sex-matched controls for three brain areas

[28] http://www.myexperiment.org/workflows/3086

[29] http://www.myexperiment.org/workflows/3124

[30] http://www.myexperiment.org/workflows/3716

[31] http://www.myexperiment.org/workflows/3717

(caudate nucleus, frontal lobe, cerebellum) and meta-data describing samples in the experiment (phenotype file). Differential expression was computed using the bioconductor package limma. The workflow tests for differential expression in each brain region separately. The input parameter "cf" is responsible for exporting the file with the differentially expressed probes in each region (1: caudate nucleus, 2: frontal cortex, 3: cerebellum). The probes are mapped to Entrez (*Global Query Cross-Database Search System*) gene ids using the Affymetrix Human Genome U133 Set annotation data, (packages hgu133a for array A and hgu133b for array B). In the case where multiple probes correspond to the same gene id, the values of the probe with the most significant changes are used. Final outcome of this workflow is a report where each row contains a gene id, a fold change and its corresponding *p*-value indicating the significance of every change in gene expression and adjusted *p*-values, generated by Benjamini and Hochberg's method for multiple-testing correction.

The workflow "map_genes_to_chr_location"[32] (access upon request), uses the biomart library in Bioconductor to map each gene to its corresponding genomic location. We query the biomart database to export information related to each gene's transcription start and transcription end site, official gene symbol, the strand that the transcription initiates and the chromosome name.

Next, the workflow "get_promoter_region_compute_overlaps"[33] (access upon request), computes a promoter region for each gene according to the workflow input parameters, "upstream" and "downstream", that the user has to define. Subsequently, overlapping genes with each dataset are computed. This workflow was reused multiple times to compute overlapping genes with each of the datasets and in order to test different sets of parameters (promoter region values and overlap) and decide the best combination for each dataset.

In the pack there is also the workflow "Download_data_from_array_express_+_create_expressionset_object"[34] (access upon request) that can be used to download data from the array express and create files to be read by the other workflows in the pack.

The workflow pack "HD data interpretation"[35] (access upon request), was created in order to further analyse and interpret the results from the HD chromatin analysis.

The workflow "Annotate_gene_list_with_top_ranking_concepts"[36] (access upon request) uses three of the component workflows from the Anni Web services pack. The workflow annotates a comma separated gene list with a predefined concept set, as for example Biological processes or Disease/syndrome.

---

[32] http://www.myexperiment.org/workflows/3712

[33] http://www.myexperiment.org/workflows/3718

[34] http://www.myexperiment.org/workflows/3719

[35] http://www.myexperiment.org/packs/486

[36] http://www.myexperiment.org/workflows/3721

The workflow “Annotate_gene_list_with_top_ranking_concepts_+_explain_concept_associations”[37] (access upon request) uses four of the component workflows from the Anni Web services pack. The workflow annotates a comma separated gene list with a predefined concept set, as for example Biological processes or Disease/syndrome, and also returns the overlapping concepts from the match that contribute most to the association.

The workflow “Get_concept_suggestions_from_term”[38] (access upon request) uses one of the Anni Web services to suggest concept ids that match the query term. The user can run this workflow with any term of interest as for example "human", "htt", "Transcription" etc, and will get suggestions for concept ids together with descriptions. Then she/he can choose the concept id that matches the best to her/his needs and use it to the rest of the concept profile analysis workflows.

The workflow “Prioritize_gene_list_related_to_a_concept”[39] (access upon request) prioritizes genes that are related to a specific concept, e.g. HTT. In order to obtain the concept id of the term that is going to be matched against the gene list, the workflow “Get_concept_suggestions_from_term” (see above), needs to be run first.

3.2 RO use scenarios

We explored the use of the RO technology in different ways for our different case studies, with the aim to mimic how a researcher and/or group would use the technology at different stages in their research. Each scenario will be described below in the beginning of every use case. We aim to explore RO usage during the different states of the lifecycle of a scientific experiment (Figure 2), as defined by the project[40]. We tested three different end-user implementations of the RO technology: the myExperiment alpha 1[41] and myExperiment alpha 2[42], and the RO Digital Library Portal[43]. RO evolution (Live, Snapshot or Archive, see deliverable D3.2v2) was explored in the RO Digital Library Portal

[37] http://www.myexperiment.org/workflows/3720

[38] http://www.myexperiment.org/workflows/3722

[39] http://www.myexperiment.org/workflows/3723

[40] http://www.wf4ever-project.org/a-day-in-the-life-of-a-scientist

[41] http://alpha.myexperiment.org/

[42] http://alpha2.myexperiment.org

[43] http://sandbox.wf4ever-project.org/portal/home

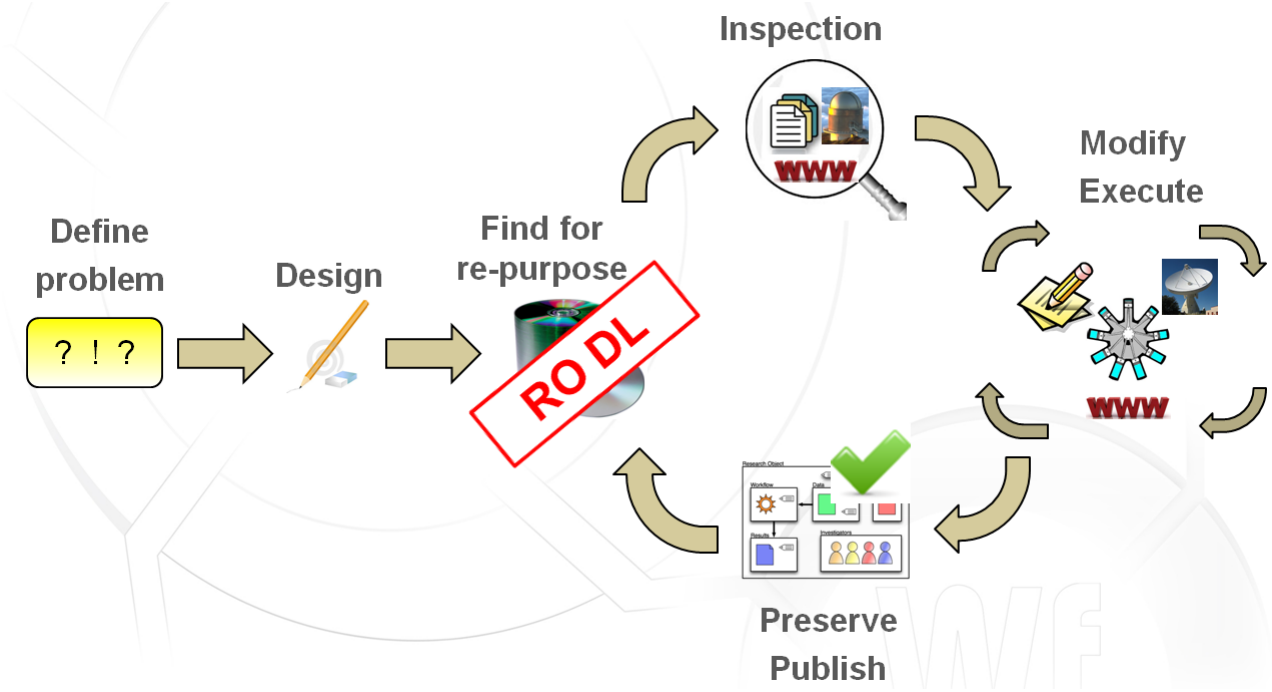

Figure 2. The lifecycle of a scientific experiment.

Concept profile generation and analysis case study

- Scenario: scientific method release by a research group
- State in the scientific experiment lifecycle (Fig. 2): Modify
- Motivation: to provide a set of services and workflows to a scientific community with similar research interest
- Used tools: myExperiment and myExperiment alpha 2

On myExperiment, groups can be created for researchers with similar interests. We created such a group named "Concept profile generation pipeline"[44]. As the work on the concept profile generation pipeline proceeds, workflows from the group can be released as ROs. The group currently has three members, with access to four workflows and one pack. The four workflows that are not in a pack are under development and not yet ready for release. New workflow versions will be uploaded as the development proceeds. MyExperiment stores and tracks these workflow versions. The pack "Concept profile analysis using Anni Web services" is almost ready for release as a workflow-centric research object. The Web services used by the workflows are still in Beta stage, but they run and can be used by the group and others for testing and preliminary analysis. Therefore, the first RO was created. In a future version of myExperiment, the action of creating a pack would automatically create the RO. Currently; myExperiment alpha 2 holds the RO implementation. We explored the RO-enabled features in myExperiment alpha 2 in the following way:

[44] http://www.myexperiment.org/groups/1129

- Created a new pack[45], providing a title and a description. The action of creating an RO consists of generating the container for the items that will be aggregated, and getting a resolvable identifier for it. In myExperiment the action of creating an RO is similar to creating a pack.
- Uploaded the main workflow.
- Added the main workflow to the pack using the functionality “Quick add: (from your stuff)” as type “Workflow”, under folder “Workflow/main”.
- Uploaded the 10 component workflows.
- Added the component workflows to the pack using the functionality “Quick add: (from your stuff)” as type “Workflow”, under folder “Workflow/components”.
- Added the doi url for the Anni Web tool publication to the pack using the functionality “Quick add: (a link)” as type “Paper”, under folder “biblio/used”.
- Edited the previous entry, adding a descriptive title and description.
- Uploaded a file: a poster describing the use of the concept profile analysis Web services.
- Added the file to the pack using the functionality “Quick add: (from your stuff)” as type “Resource”, under folder “biblio/produced”.
- Uploaded a workflow run bundle as type “Workflow run”.

This RO can now be shared with collaborators for workflow and Web service testing purposes. The most significant improvements in the myExperiment alpha 2 (which implements the RO model) over the regular myExperiment would be the annotations that you can store on pack items, which gives the ability to create a tailored pack page for a particular purpose and also for other tools to store and retrieve their own annotations even if myExperiment does not understand what they mean. From the user point, the organization of pack resources into folders[46], following the folder structure suggested by WP6 and WP5[47], is a major improvement. The possibility to upload workflow runs from Taverna gives a unique opportunity to store the provenance of a workflow run. To be recognized as a workflow run, the exported provenance should have the structure of a workflow bundle[48]. The workflow bundle is a ZIP-based media format that formalizes how to create a single file that bundles both the RO descriptions and annotations according to the RO models. Currently, workflow bundles can be exported from Taverna by following these steps:

- To install the plugin in Taverna, add the plugin site[49]

---

[45] http://alpha2.myexperiment.org/packs/8

[46] http://alpha2.myexperiment.org/packs/8/items

[47] http://www.wf4ever-project.org/wiki/display/docs/RO+tree+folder+structure

[48] https://w3id.org/bundle

[49] http://build.mygrid.org.uk/taverna/updates/2.4.0/plugins/experimental/

- Then install Taverna-PROV 2.01.1-SNAPSHOT.
- After a restart of Taverna, save the provenance of a run by clicking on "Save all" in the Result panel, click "Provenance PROV". Enter a file name, for instance experiment5.robundle.
- The zip file can then be uploaded as a "Workflow Run" on alpha2.myexperiment.org or explored by unzipping locally

MetS case study

Interpretation of SNP-metabolite associations resulting from a GWAS, starting from the SNP

- Scenario: hypothesis-driven scientific research in the field of genomics where workflows play a central role as methodology
- State in the scientific experiment lifecycle (Fig. 2): Publish
- Motivation: use RO technology to enhance the understanding of the method leading to, and the reproducibility of, the scientific results
- Used tools: explored myExperiment alpha 1

MyExperiment alpha 1, preceding alpha 2, was released in May 2013, and was the state-of-the-art RO end-user technology at the time. We had a real need to use the technology since we were submitting our first RO-supported publication. This publication includes the workflows Alpha 1 also communicated with the RO Digital Library, which is the currently not yet the case for myExperiment alpha 2. We explored the RO-enabled features in myExperiment alpha 1 in the following way:

- Created a new pack[50]. We filled in a title and description of the RO at the point of creation and got a confirmation that the RO had been created and had been assigned a resolvable identifier in the RO Digital Library.
- Added the experiment sketch. Using a popular office presentation tool, we made an experiment sketch and saved it as a PNG image. We then uploaded the image to the pack, selecting the type “Sketch”. As a result, the image gets stored in the Digital Library and aggregated in the RO. In addition, an annotation was added to the RO to specify that the image is of type “Sketch”. A miniature version of the sketch is shown within the myExperiment pack.
- Added the hypothesis. To specify the hypothesis, we created a text file that describes the hypothesis, and then upload it to the pack as type “Hypothesis”. The file gets stored in the Digital Library and aggregated in the RO, this time annotated to be of type “Hypothesis”.
- Added workflows. We uploaded them to the pack as type “Workflow”. MyExperiment then automatically performed a workflow-to-RDF transformation in order to extract the workflow structure according to the RO model, this includes user descriptions and metadata created within the Taverna

[50] http://alpha.myexperiment.org/packs/405

workbench. The descriptions and the extracted structure gets stored in the RO Digital Library and associated with the workflow files as annotations.

- Checklist evaluation: At this point we checked how far we were from satisfying the Minim model (see D4.2v2). The tool informed us that we needed to add the workflow inputs and the experiment conclusions in order to fully satisfy the checklist.
- Added the workflow input file. The data values were stored in files that were then uploaded into the pack as “Example inputs”. Such files gets stored in the RO Digital Library and aggregated in the RO, and as “Example inputs”.
- Added the workflow provenance. Using the Taverna-Prov extension to Taverna, we exported the workflow run provenance to a file that we uploaded to the pack as type “Workflow run”. Similar to other resources, the provenance file gets stored in the digital library with the type “Workflow run”.
- Added the results. We summarized the different workflow outputs to a result file in table format, uploaded to the pack as type “Results”. The file gets stored in the digital library and aggregated in the RO, annotated to be of the type “Results”.
- Added the conclusions. To specify the hypothesis, we created a text file that describes the hypothesis, and then upload it to the pack as type “Hypothesis”. The file gets stored in the digital library and aggregated in the RO, annotated to be of type “Conclusions”.
- Checklist evaluation. At this point we checked how far we were from satisfying the Minim model, and were informed by the tool that the RO now fully satisfies the checklist.
- Annotated and linked the resources. We linked the example input file to the workflows that used the file by the property “Input_selected”. In this particular case, both workflows have the same inputs but they need to be configured in different ways. This is described in the workflow description field in Taverna.

This RO does now fully satisfy the minimal checklist for workflow-centric ROs, and can be published alongside the research article describing the experiment.

Interpretation of SNP-metabolite associations resulting from a GWAS, starting from the metabolite

- Scenario: hypothesis-driven scientific research in the field of genomics where workflows play a central role as methodology
- State in the scientific experiment lifecycle (fig. 2): Preserve
- Motivation: use RO technology to archive the scientific method after manuscript submission stage
- Used tools: RO Digital Library Portal, version 4.8.1

We had a real need to use the technology since we had submitted a workflow-centric publication, and wanted to preserve the workflows with example values as they were at the stage of submission. We explored the RO features in the RO Digital Library Portal in the following way:

- Created a new RO[51] by uploading the workflows as a zip file. The RO was automatically created from the zip and has the status LIVE.
- Edited the title and the description of the RO.
- Navigated to the "Quality" tab and selected the checklist "RO basic requirements".
- Since the RO minimally satisfies the checklist for ready-to-release, we archived the RO by selecting "Evolution: release" from the main RO page. A new RO[52] was created, with the status ARCHIVE.
- Navigated to the "History" tab to inspect the RO evolution.

This RO does now fully satisfy the minimal checklist for basic ROs. If changes are asked by the reviewers during the manuscript review process, the LIVE RO can be updated and snapshots created along the way until the final ARCHIVE RO at the time of publication, keeping track of changes through the RO evolution implementation.

HD case study

- Scenario: hypothesis-driven scientific research in the field of genomics where workflows play a central role as methodology
- State in the scientific experiment lifecycle (fig. 2): Preserve
- Motivation: use RO technology to preserve the scientific method before manuscript submission stage
- Used tools: RO Digital Library Portal, version 4.8.1

We had a real need to use the technology since we are preparing a workflow-centric manuscript, and wanted to preserve the workflows with example values during this pre-submission state. As a follow up experiment we wanted to interpret the results from the above RO using the concept profile analysis workflows (see section 2.3). The RO for this experiment was created similarly to the RO above, and the details for this experiment are included in the RO[53]. The input file of this RO is actually the results of the RO that was used to analyse our gene expression data (**Error! Reference source not found.**), illustrating reuse existing resources and their provenance that have been described previously in a machine readable format from another experiment.

[51] http://sandbox.wf4ever-project.org/rodl/ROs/mining_kegg_workflows/

[52] http://sandbox.wf4ever-project.org/rodl/ROs/mining_kegg_workflows-release/

[53] http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation/

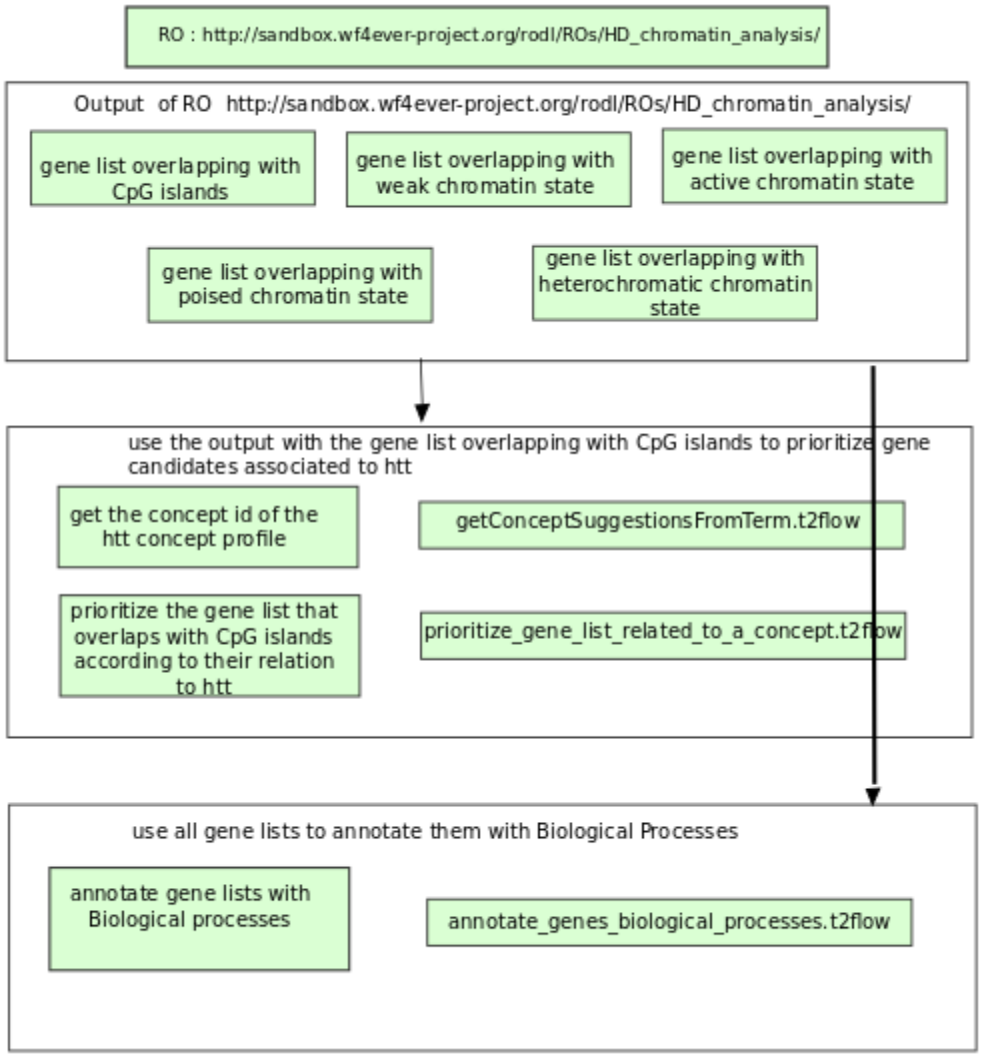

Figure 3. Sketch explaining the connection between the two HD ROs.

We explored the RO features in the RO Digital Library Portal in the following way:

- Created a new RO[54] by uploading the workflows as a zip file. The RO was automatically created from the zip and has the status LIVE.
- Edited the title and the description of the RO.
- Checked the “Research Object quality bar”, which is backed by the Minim model for workflow-centric ROs. We noticed that in order to satisfy the model, we needed to upload the hypothesis or research question, design sketch, and conclusions.

[54] http://sandbox.wf4ever-project.org/rodl/ROs/HD_chromatin_analysis/

- Added the hypothesis. To specify the hypothesis, we created a text file that describes the hypothesis, and then uploaded the file as type “Hypothesis”. The file gets stored in the Digital Library and aggregated in the RO, this time annotated to be of type “Hypothesis”.
- Added the design sketch: using a popular office presentation tool, we made an experiment sketch and saved it as a PNG image. We then uploaded the image, selecting the type “Sketch”. As a result, the image gets stored in the Digital Library and aggregated in the RO.
- Added the conclusions. To specify the conclusions, we created a text file that describes the hypothesis, and then uploaded the file as type “Conclusions”. The file gets stored in the Digital Library and aggregated in the RO, this time annotated to be of type “Conclusions”.
- Checked the “Research Object quality bar”. We noticed that the RO now minimally satisfies the checklist for a workflow-centric RO.
- Added additional resources to the RO, with the aim to increase the understanding of the experiment (datasets as type “Dataset”, background reading papers as type “Paper”, summary results as type “Results”).
- Preserved the RO by selecting “Evolution: snapshot” from the main RO page. A new RO was created, with the status SNAPSHOT[55].
- Navigated to the “History” tab to inspect the RO evolution.

[55] http://sandbox.wf4ever-project.org/rodl/ROs/HD_chromatin_analysis-snapshot/

4. General discussion

4.1 Workflow development

This deliverable describes the workflows developed during phase III of the project. During this phase a total of 20 new workflows were created by five different users at the HG-LUMC, spread across the different case studies. We aimed to follow the Best Practices for workflow design[56] developed during phase II of the project when creating all these workflows. Three of the users had been involved in the development of the Best Practices, while two had not. Reported experiences include that point six, nine and ten of the Best Practices can be especially challenging to realize. Point six, to make the workflow executable outside the local environment, is a challenge when locally installed software is being used. For example, setting up the R statistical analysis environment so that it runs on a Taverna server is less than trivial because of the following reasons:

- Authentication mechanism in the Taverna credential manager is problematic when used in a non-interactive environment (unlike the Taverna Workbench)
- Taverna does not support components to communicate R datatypes
- There is no support for saving the state of an R workspace to share between components communitcating R datastructures
- There is no way to ensure all required R packages are available on a given system (e.g. biomaRt)

Point nine, to test and validate the workflows, is difficult since there are no specific guidelines for how to test workflows. Something similar to Unit testing in computer programming where different part of an application (from individual units of source code to operating procedures) are tested to determine if they are fit for use could be beneficial. Point ten, to advertise and maintain the workflows, is mostly a matter of allocating time and resources, which is a problem in a competitive research environment where new discoveries are the most important result. The other points were experienced by the users as time-consuming, but rewarding in the end. A citation from one of the users might illustrate this: “Although it is sometimes hard to follow all these guidelines, it is possible and drastically increases the value of your workflow. Creating annotated, modular and maintained workflows has a lot of value for other scientist who then can successfully use your analysis to get results from their data.”

For the Metabolic Syndrome case study, focus lies on workflow maintenance after publication. For the Huntington’s Disease case study, the workflows might still change since the manuscript is still in preparation. Focus will therefore lie on point five of the Best Practices, to annotate the workflows, and on point nine and ten.

[56] http://www.wf4ever-project.org/bestpractices

4.2 Impact

During phase III of the project we have created a large number of workflows (20) incorporated in four different ROs, which have been presented at different national and international meetings[57], conferences[58], and workshops[59]. We have submitted two journal publications related to these workflows and ROs, where we either describe the RO [6], or the workflows only [7], and are preparing the third one [8]. We have also undergone efforts to bring the workflow and RO paradigms to genomics users by initiating projects and organizing workshops. Details of these efforts will be reported in D6.2 "Final report on the creation of a Community of Users in Genomics".

4.3 Quality and completeness

In D6.3v2 we reported that the only way to minimize workflow decay and work towards high quality and completeness was by following the Best Practices for workflow design and by using a high level tree-folder structure. During phase III of the project, functional user interfaces to measure quality and completeness have been developed. We have explored these user interfaces that have been implemented at various stages in the end-user platforms myExperiment (alpha 1 and 2, see section 3.1) and the RO Digital Library Portal (see sections 3.2 and 3.3). The checklist evaluations based on the Minim models for ROs were especially helpful at the time of RO creation as a means to guide what to include in the RO. It gives a certain satisfaction to be able to tick a list and see that you are creating something that is considered valid. However, neither the myExperiment alpha versions nor the RO Digital Library Portal provides any explanation as to why the checklist is being used. It is just there. Light-weight documentation with links to further information, together with tutorials and examples on RO creation is needed. These user requirements have been communicated to the other WPs. In addition to implementing this documentation in the RO tools, the newly launched website reserachobject.org could be a place for this type of documentation. One can also imagine more Minim models for ROs than those already developed. Using the mkminim utility[60], users can actually create such checklists from a checklist description presented in spreadsheet tabular form. The mkminim tool is installed as part of the RO Manager[61].

---

[57] Dutch Huntingtons Disease Meeting in Leiden May 2013, BioAssist meeting in Utrecht February 2013, American Society of Human Genetics Annual Meeting in San Francisco November 2012, The Annual Meeting of the ISMB BioLINK Special Interest Group in Berlin July 2013

[58] ISCB-Asia/SCCG 2012 (Shenzhen Conference on Computational Genomics) in Shenzhen December 2012, Netherlands Bioinformatics Conference in Luntheren April 2013, International Conference on Intelligent Systems for Molecular Biology in Berlin July 2013

[59] Semantic Web Applications and Tools for the Life Sciences Workshop in Paris November 2012

[60] https://github.com/wf4ever/ro-manager/blob/master/src/checklist/mkminim.md

[61] https://pypi.python.org/pypi/ro-manager/

4.4 Annotation and RO building

In D6.3v2 we described RO annotation and building using the RO Manager. During phase III of the project, the two other RO tools myExperiment and RO Digital Library Portal have matured and also they now allow for RO creation and maintenance. Both user interfaces support the final template folder structure[62] suggested by WP5 and WP6. Great effort has also been made to implement the annotation guidelines[63] arising from extensive user requirement analysis[64] during phase III of the project. There is however still room for improvement. Many of the machine-generated annotations that users consider important are hidden from view in robundle or wfbundle files created by services developed in the project for automatic RO creation, according to the RO models, from Taverna workflow or workflow run files.

4.5 Preservation and versioning

We have used the myExperiment alpha versions and the Portal to the RO Digital Library to preserve our workflows and related data and metadata as ROs. The user interfaces are still in alpha or beta stage, but functional for an expert user, or a user with support from an expert user. RO evolution aspects were explored using the RO Digital Library Portal. It was possible to create snapshots and to archive an RO, which was not possible at the time of D6.3v2. Although this is a great progress, these actions need documentation to make sense to a non-expert user that is not aware of the RO lifecycle.

[62] http://www.wf4ever-project.org/wiki/display/docs/RO+tree+folder+structure

[63] http://www.wf4ever-project.org/wiki/display/docs/Annotations+implementation+guidelines+release+1

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

5. Concluding remarks

We have created workflows and ROs for three use cases within the context of genomics and bioinformatics according to the Best Practices for workflow design, and evaluated the impact of these Best Practices on the workflow design process. We have explored aspects of RO evolution, sharing, completeness and quality evaluation using the RO tools myExperiment alpha (1 and 2) and the RO Digital Library Portal, and provided directions for further development and research. Due to lack of a functional and integrated user interface, we could not explore aspects of collaboration satisfactory.

References

[1] Jelier R et al., Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. 2008;9(6):R96.

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

[3] Hettne KM et al., Rewriting and suppressing UMLS terms for improved biomedical term identification. J Biomed Semantics. 2010 Mar 31;1(1):5.

[4] Gardiner-Garden M and Frommer M, CpG Islands in Vertebrate Genomes. Journal of Molecular Biology 196, no. 2 (July 20, 1987): 261–282.

[5] Ernst J et al., Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types. Nature 473, no. 7345 (March 23, 2011): 43–49.

[6] Hettne KM et al., Structuring research methods and data with the Research Object model: genomics workflows as a case study. Submitted to the Journal of Biomedical Semantics

[7] Dharuri H et al., Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles. Submitted to BMC Genomics

[8] Mina E et al., Hypotheses for epigenetic mechanisms in Huntington's Disease exposed by an e-Science approach. In preparation