myExperiment 2.0 – Preserving digital Research Objects using the Wf4Ever architecture

Stian Soiland-Reyes
myGrid, University of Manchester

EGI/SHIWA Workshops on e-Science Workflows

Budapest, 2012-02-10

Background

Taverna

http://www.taverna.org.uk/

Taverna - Scientific Workflow Management System

~85000 downloads

~EU projects: SCAPE, BioVeL, HELIO,

e-Lico, VPH-SHARE, EGI-INSPiRE….

my experiment

http://www.myexperiment.org/

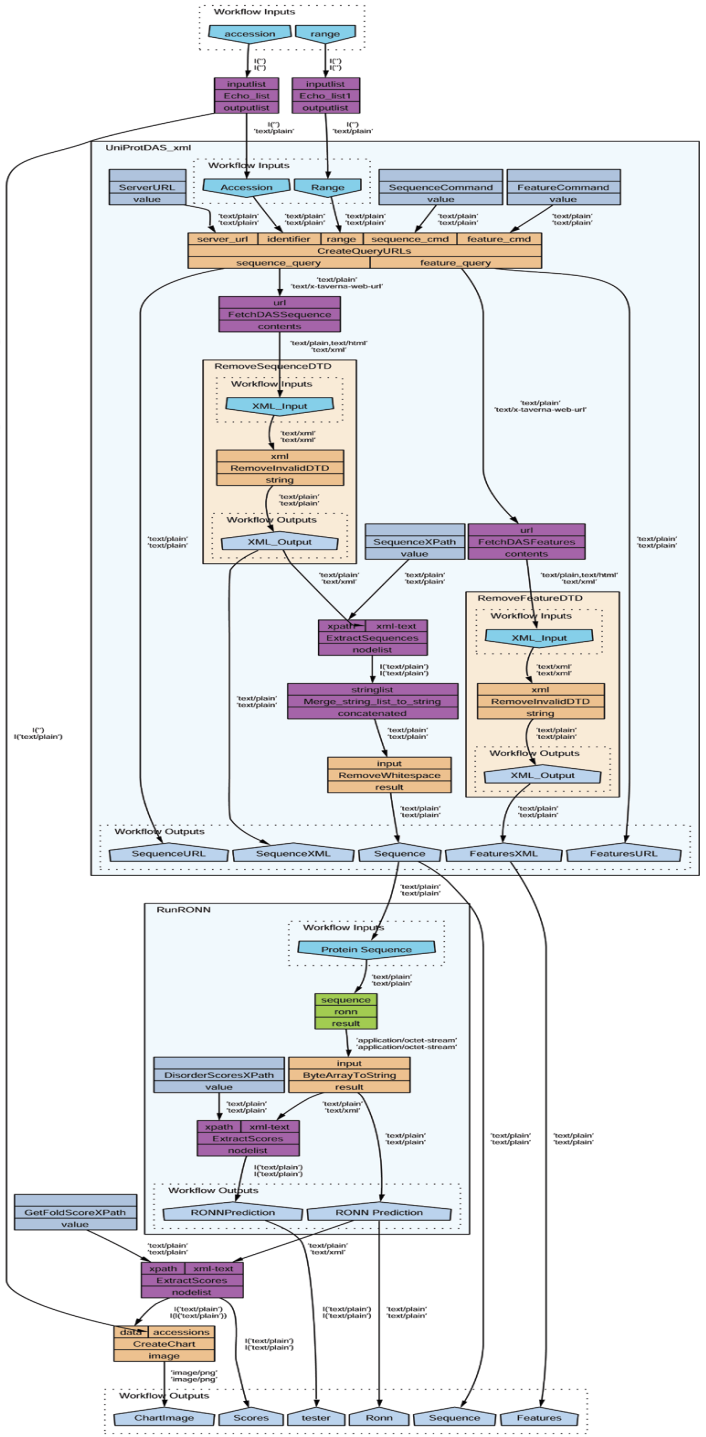
myExperiment - Web 3.0 virtual environment, library and social network for workflows

~5000 registered users

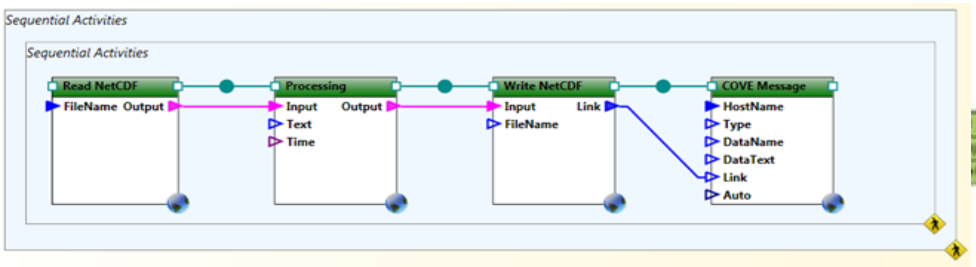
~2200 workflows

~21 different systems

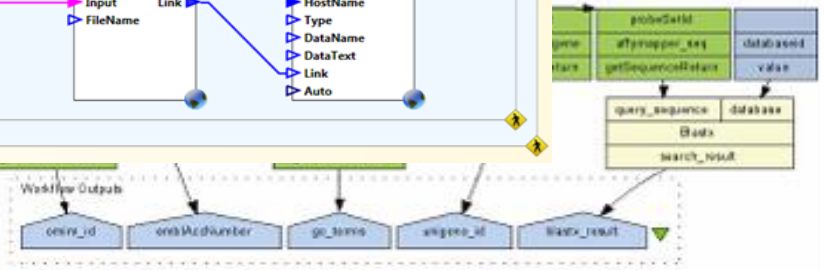
What is a Scientific Workflow?

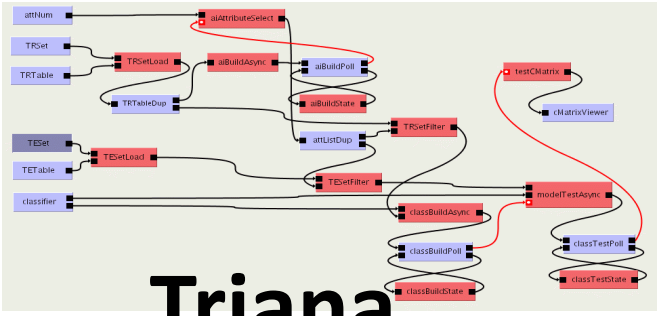


- Workflows coordinate the execution of **services** and **link** together resources.
- **Data-driven** rather than process-driven: «*Send output from A to B and C*»
- Semi-automated computational execution in scientific problem-solving → *repeatable*, *reproducable*, *reusable*
- The implementation of a **scientific method**

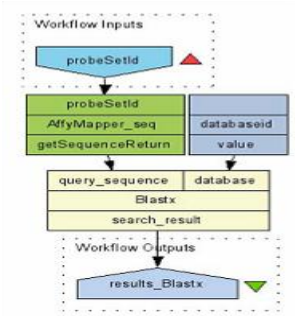


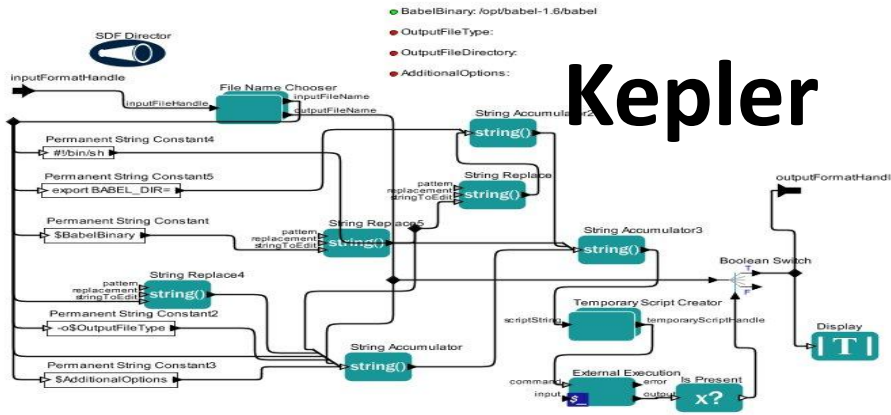
Trident



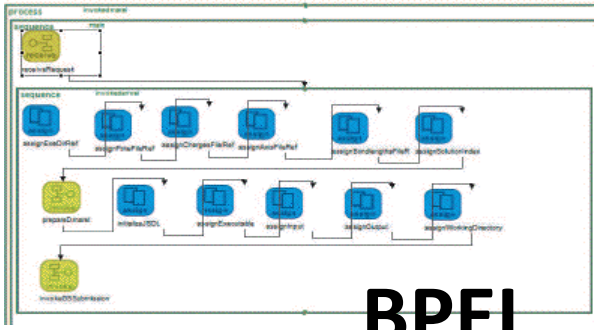


Triana



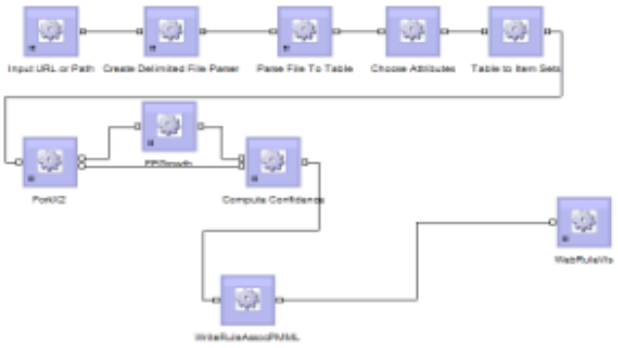


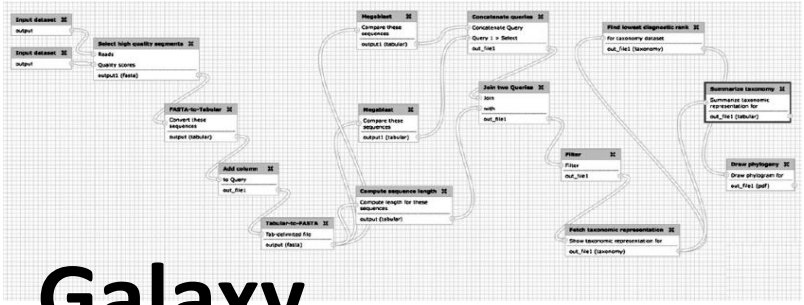
Kepler



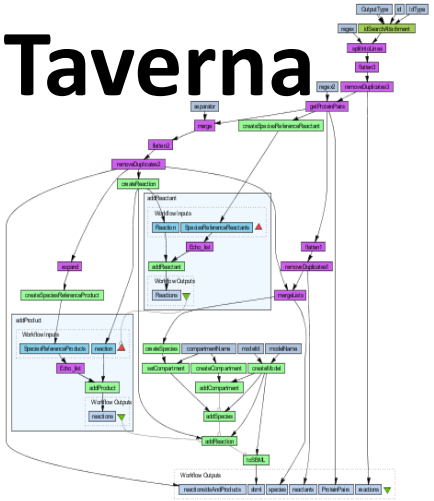
BPEL

Meandre





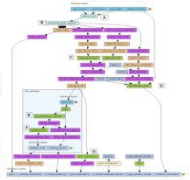
Galaxy



Taverna

Reuse, Recycling, Repurposing

- Paul writes **workflows** for identifying *biological pathways* implicated in *resistance to Trypanosomiasis* in **cattle**
- Paul meets Jo. Jo is investigating *whipworm* in **mouse**.
- Jo reuses one of Paul's workflow **without change**.
- Jo identifies the *biological pathways* involved in sex dependence in the mouse model, believed to be involved in the ability of mice to exp*el the parasite*.
- Previously a manual **two year study** by Jo had failed to do this.



"A biologist would rather share their toothbrush than their gene name"

Mike Ashburner and others
Professor in Dept of Genetics,
University of Cambridge, UK

myexperiment

http://www.myexperiment.org/

- "Facebook for Scientists"
  ...but different to Facebook!
- A repository of research methods
- A social network of people and things
- A Social Virtual Research Environment
- A probe into researcher behaviour
- Open source (BSD) Ruby on Rails app
- REST and SPARQL, Linked Data
- Influenced BioCatalogue, MethodBox and SysMO-SEEK

myExperiment currently has 5378 members, 292 groups, 2273 workflows, 534 files and 217 packs

myGrid MANCHESTER 1824 UNIVERSITY OF Southampton School of Electronics and Computer Science OXFORD e-Research CENTRE e-Research SOUTH JISC EPSRC Microsoft Research

Workflows

Search filter terms

« previous 1 2 3 ... 174 next »

Sort by: Rank

Showing 1739 results. Use the filters on the left and the search box below to refine the results.

Search

Filter by type

- Taverna 2 752
- Taverna 1 562
- RapidMiner 192
- Kepler 43
- Bioclipse Script... 34
- GWorkflowDL 24
- LONI Pipeline 23
- BioExtract Server 16
- Trident (Packa... 10
- LabTrove Tem... 9

Filter by tag

- example 215
- mygrid 103

Taverna 2

Pathways and Gene annotations for QTL region (v7)

View

Download (v7)

Original Uploader

Paul Fisher

Created: 19/11/09 @ 18:18:52 | **Last updated:** 02/09/11 @ 11:44:57

Credits: Paul Fisher

License: Creative Commons Attribution-Share Alike 3.0 Unported License

This workflow searches for genes which reside in a QTL (Quantitative Trait Loci) region in the mouse, Mus musculus. The workflow requires an input of: a chromosome name or number; a QTL start base pair position; QTL end base pair position. Data is then extracted from BioMart to

Workflow Entry: Success-Abandonment-Classification

Created at: 06/02/08 @ 14:35:41 **Last updated:** 02/07/08 @ 17:15:25

| License | Credits (2) | Attributions (0) | Tags (6) | Featured in Packs (1) | Ratings (1) | Attributed By (0) | Favourited By (2) |
| Citations (0) | Version History | Reviews (0) | Comments (0) |

Version 3 (latest) (of 3)

View version: 3 (latest)

Version created on: 06/02/08 @ 14:35:41 **by:** Andrea Wiggins | Revision comments

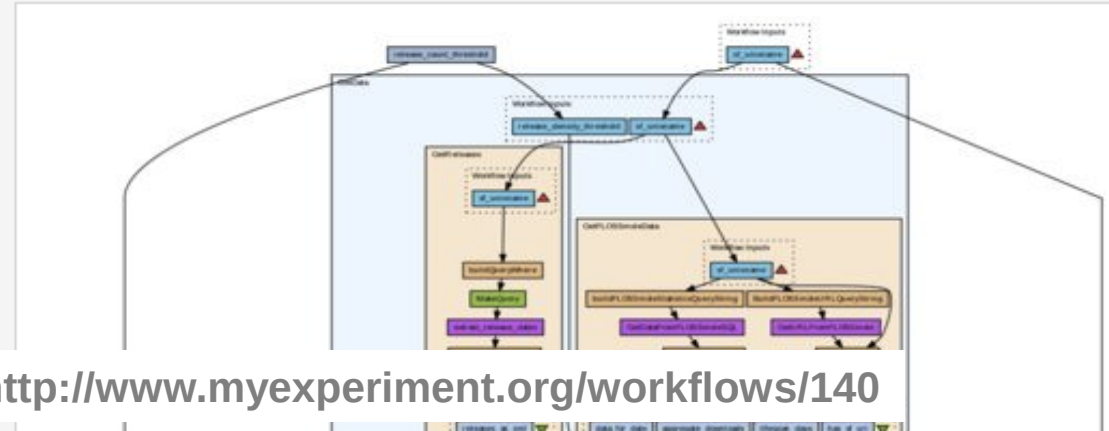
Last edited on: 02/07/08 @ 17:15:25 **by:** Andrea Wiggins

Title: Success-Abandonment-Classification

Type: Taverna 1

Preview

(Click on the image to get the full size)



http://www.myexperiment.org/workflows/140

Workflow Type

Taverna 1

Original Uploader

Andrea Wiggins

License

All versions of this Workflow are licensed under:

Credits (2)

(People/Groups)

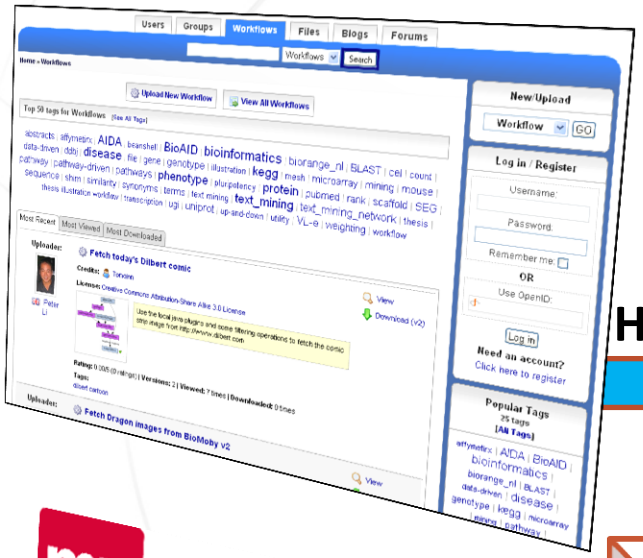
Andrea Wiggins

James Howison

Attributions (0)

David De Roure

APIs for developers



XML

facebook

iGoogle

android

HTML

API config

ORE
FOAF
SIOC

Managed REST API

Search API

Search Engine

SPARQL endpoint

RDF Store

tags

ratings

reviews

profiles

workflows

credits

groups

friendships

files

packs

Eprints
Dropbox
...

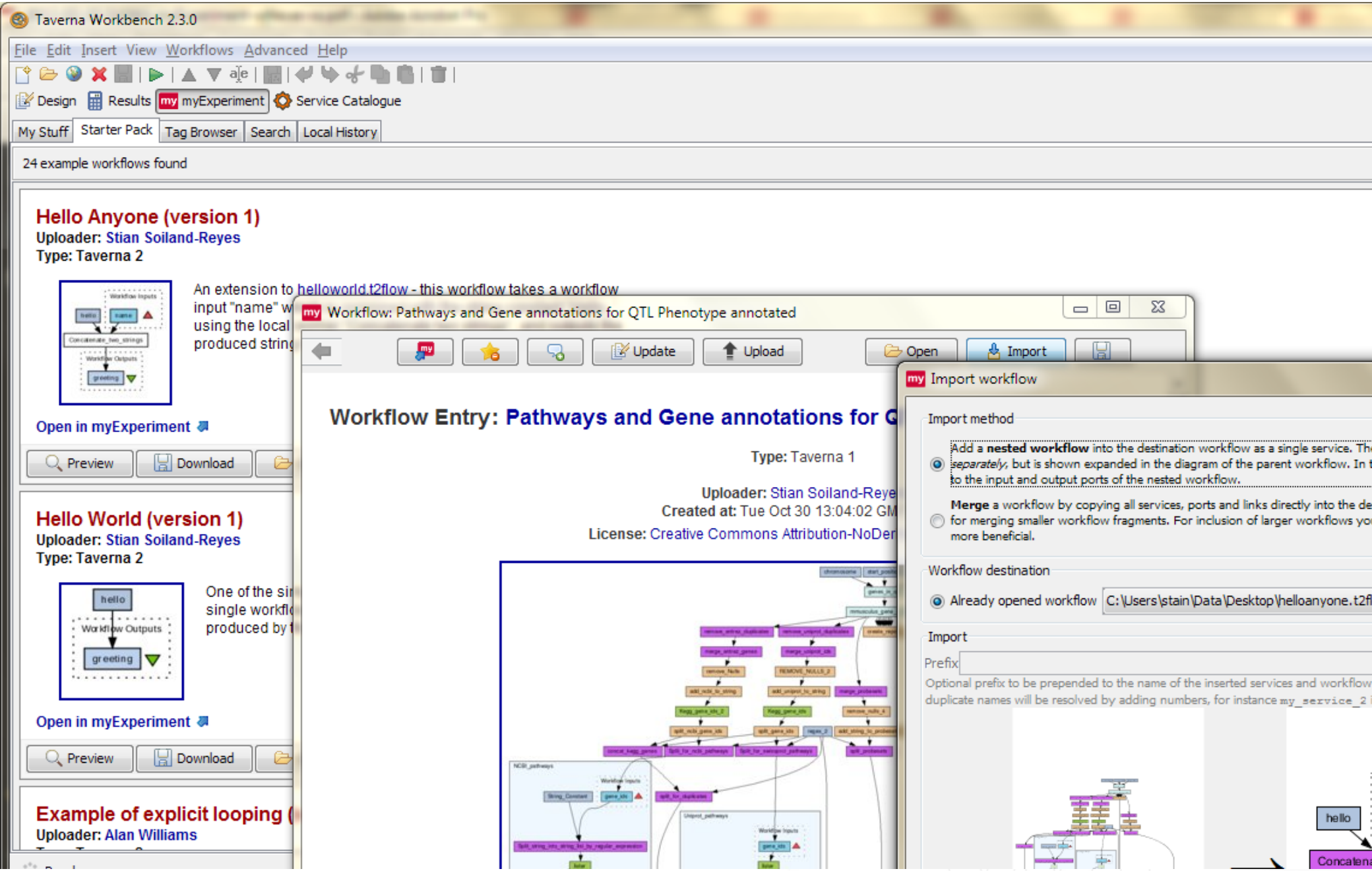
mySQL

Enactor API

Enactor

- myExperiment plugin for Taverna
  - Browse myExperiment workflows
    - My workflows
    - Tags
    - Search
  - Open workflow
    - + Embed in existing workflow
  - Upload workflows
    - Provide metadata

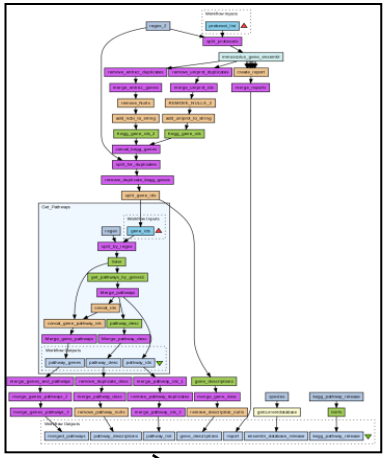
Taverna


Taverna Workbench 2.3.0
File Edit Insert View Workflows Advanced Help
Design Results myExperiment Service Catalogue
My Stuff Starter Pack Tag Browser Search Local History
24 example workflows found
Hello Anyone (version 1)
Uploader: Stian Soiland-Reyes
Type: Taverna 2
An extension to helloworld.t2flow - this workflow takes a workflow
Open in myExperiment
Preview Download
Hello World (version 1)
Uploader: Stian Soiland-Reyes
Type: Taverna 2
hello
Workflow Outputs
greeting
Open in myExperiment
Preview Download
Example of explicit looping (
Uploader: Alan Williams
Workflow: Pathways and Gene annotations for QTL Phenotype annotated
Update Upload Open Import
Workflow Entry: Pathways and Gene annotations for G
Type: Taverna 1
Uploader: Stian Soiland-Reye
Created at: Tue Oct 30 13:04:02 GM
License: Creative Commons Attribution-NoDer
Import workflow
Import method
Add a nested workflow into the destination workflow as a single service. Th
separately, but is shown expanded in the diagram of the parent workflow. In t
to the input and output ports of the nested workflow.
Merge a workflow by copying all services, ports and links directly into the de
for merging smaller workflow fragments. For inclusion of larger workflows yo
more beneficial.
Workflow destination
Already opened workflow C:\Users\stain\Data\Desktop\helloanyone.t2f
Import
Prefix
Optional prefix to be prepended to the name of the inserted services and workflow
duplicate names will be resolved by adding numbers, for instance my_service_2
hello
Concatena

Paul's Research Object

Workflow 16

QTL



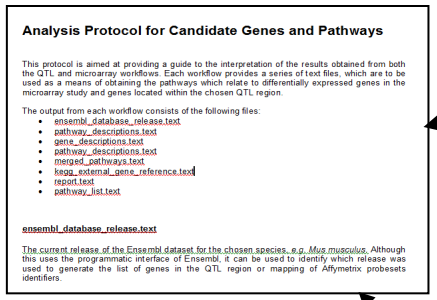
Results

```
path:mmu04060 Cytokine-cytokine receptor interaction - Mus musculus (mouse)
path:mmu00970 Aminoacyl-tRNA biosynthesis - Mus musculus (mouse)
path:mmu00240 Pyrimidine metabolism - Mus musculus (mouse)
path:mmu03010 Ribosome - Mus musculus (mouse)
path:mmu04080 Neuroactive ligand-receptor interaction - Mus musculus (mouse)
path:mmu04210 Apoptosis - Mus musculus (mouse)
path:mmu05220 Chronic myeloid leukemia - Mus musculus (mouse)
path:mmu04612 Antigen processing and presentation - Mus musculus (mouse)
path:mmu00271 Methionine metabolism - Mus musculus (mouse)
path:mmu04912 GnRH signaling pathway - Mus musculus (mouse)
path:mmu04330 Notch signaling pathway - Mus musculus (mouse)
path:mmu04640 Hematopoietic cell lineage - Mus musculus (mouse)
path:mmu00561 Glycerolipid metabolism - Mus musculus (mouse)
path:mmu04110 Cell cycle - Mus musculus (mouse)
path:mmu04530 Tight junction - Mus musculus (mouse)
path:mmu02010 ABC transporters - General - Mus musculus (mouse)
```

produces

Included in

Published in



Analysis Protocol for Candidate Genes and Pathways

This protocol is aimed at providing a guide to the interpretation of the results obtained from both the QTL and microarray workflows. Each workflow provides a series of text files, which are to be used as a means of obtaining the pathways which relate to differentially expressed genes in the microarray study and genes located within the chosen QTL region.

The output from each workflow consists of the following files:

- ensembl_database_release.text
- pathway_descriptions.text
- gene_descriptions.text
- pathway_descriptions.text
- merged_pathways.text
- kegg_external_gene_reference.text
- report.text
- pathway_list.text

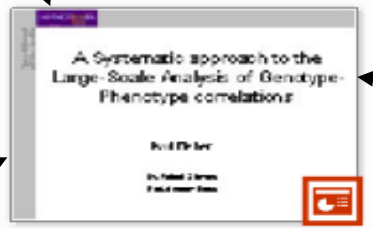
ensembl_database_release.text

The current release of the Ensembl dataset for the chosen species, *e.g. Mus musculus*. Although this uses the programmatic interface of Ensembl, it can be used to identify which release was used to generate the list of genes in the QTL region or mapping of Affymetrix probesets identifiers.

Included in

Feeds into

Included in



A Systematic approach to the Large-Scale Analysis of Genotype-Phenotype correlations

complete.ppt

Included in

Logs

produces

Metadata

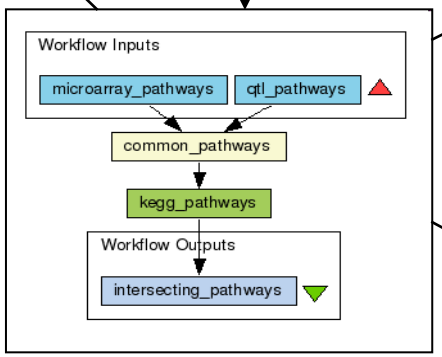
Tags (19)

Creator tags

affymetrix | african trypanosomiasis | cattle | data-driven | disease | entrez | genotype | Kegg Pathways | KeggID | link-integration | microarray | mouse | pathway | pathway-driven | phenotype | sleeping sickness | swissprot | uniprot | web services

[edit]

Add Tags



Workflow Inputs

microarray_pathways

qtl_pathways

common_pathways

kegg_pathways

Workflow Outputs

intersecting_pathways

Slides

Paper

produces

Published in

Primer Name	Left Flank Nucleotide sequence	Right Flank Nucleotide sequence
DAXX_1274_1812	CAGGAGGAATGGCGAGTG	AGCTTAGTCCTTCCCAAGCC
DAXX_140754_456_1070	CTTGTAGGATTGGGACTGGG	TCTCCTCCTCTTCCTCCTCC
DAXX_2270_2720	TGGGCAGGAGAGATGGTTC	ATGGTTCAAGGGAAGGGAAA
DAXX_2644_3187	TGTGTGATTGGCTGGTTGTT	GCAAATACGAGGAGTCTGGG
DAXX_exon5	TCCTCCTCCTACCAATCAAA	AGCAGAACTAACACCACAAGG
Daxx_Upst_479_1104	CAGGCTTCCTCATCAACACC	TGTCTCTATGGCTGTGCAGG

Common pathways

Workflow 13

Results

mygrid

The R.* dimensions

Reusable. The key tenet of Research Objects is to support the sharing and reuse of data, methods and processes.

Repurposeable. Reuse may also involve the reuse of constituent parts of the Research Object.

Repeatable. There should be sufficient information in a Research Object to be able to repeat the study, perhaps years later.

Reproducible. A third party can start with the same inputs and methods and see if a prior result can be confirmed.

Replayable. Studies might involve single investigations that happen in milliseconds or protracted processes that take years.

Referenceable. If research objects are to augment or replace traditional publication methods, then they must be referenceable or citeable.

Revealable. Third parties must be able to audit the steps performed in the research in order to be convinced of the validity of results.

Respectful. Explicit representations of the provenance, lineage and flow of intellectual property.

Replacing the Paper: The Twelve Rs of the e-Research Record" on http://blogs.nature.com/eresearch/

Pack analysis

- ***Workflow*** – pack contains a number of workflows
- ***Presentation*** - encapsulation of a single presentation
- ***Collection*** - a number of things: workflows, presentations, papers
- ***Heterogeneous*** - where the workflows do not appear to have a clear common purpose
- ***Homogeneous*** - workflows appear to be designed to work together

- ***Paper*** - source for a paper
- ***Tutorial*** - tutorial material
- ***Data*** - collection of data files
- ***Derived data*** - results of workflow
- ***Benchmark*** - benchmarking data
- ***Supplementary*** - stuff associated with a paper
- ***Noise*** - tests, tryouts, rubbish
- ***Oddity*** - none of the above

Analysis by Sean Bechhofer

- Workflow Preservation
- Research Objects
- Provenance
- Recommendation
- Astronomy and Genomics

Project ID card

Funded under: 7th FWP (Seventh Framework Programme)

Area: Digital Libraries and Digital Preservation. (ICT-2009.4.1)

Project reference: 270192

Total cost: 3.86 million euro

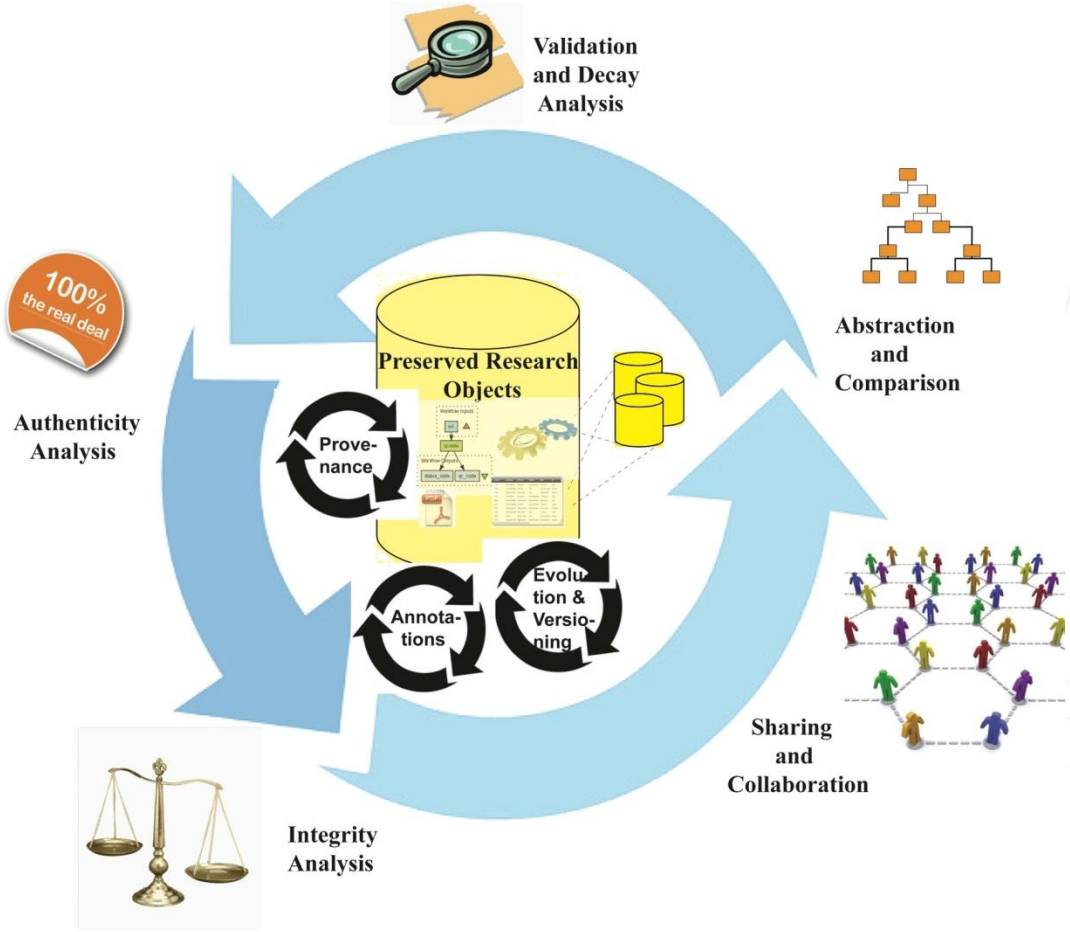
EU contribution: 2.94 million euro

Execution: From 2010-12-01 to 2013-11-30

Duration: 36 months

http://www.wf4ever-project.org/

Preservation of scientific workflows in data-intensive science



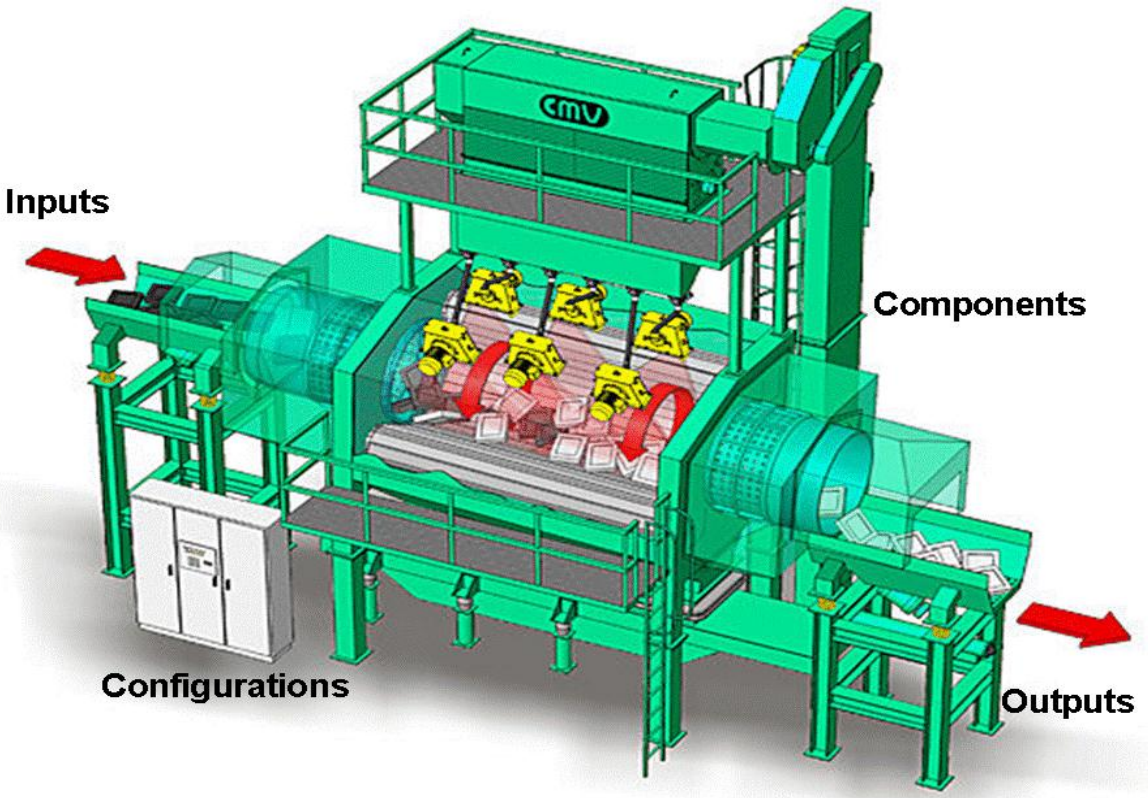
- Scientific workflows aim at the heart of experimental science
  - Enable **automation** of scientific methods
  - Encourage **best practices**
- Need to be **preserved**
  - **Reuse** is fundamental for incremental scientific development
  - **Method reproducibility** is key for credit and publication
- …but workflow preservation is complex
  - Heterogeneous types of information need to be **aggregated**, including workflows and related resources into **research objects**
  - Research objects need to be **trusted** and **understandable** *n* years from now
  - **Social aspects** need to be addressed in order to support reuse in scientific communities

Workflow Decay

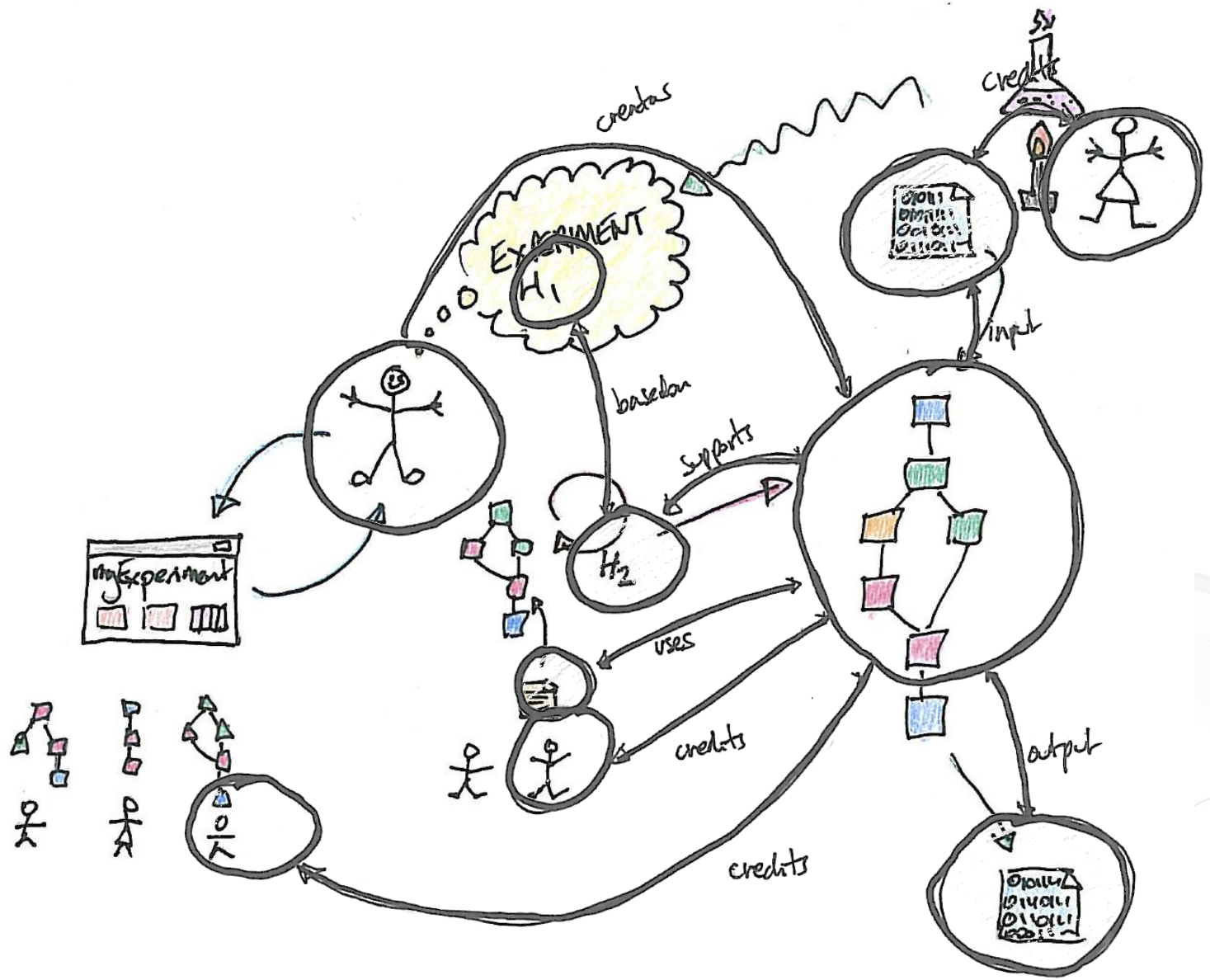
- Component level
- flux/decay/unavailability
- Data level
  - formats/ids/standards
- Infrastructure level
  - platform/resources

Experiment Decay

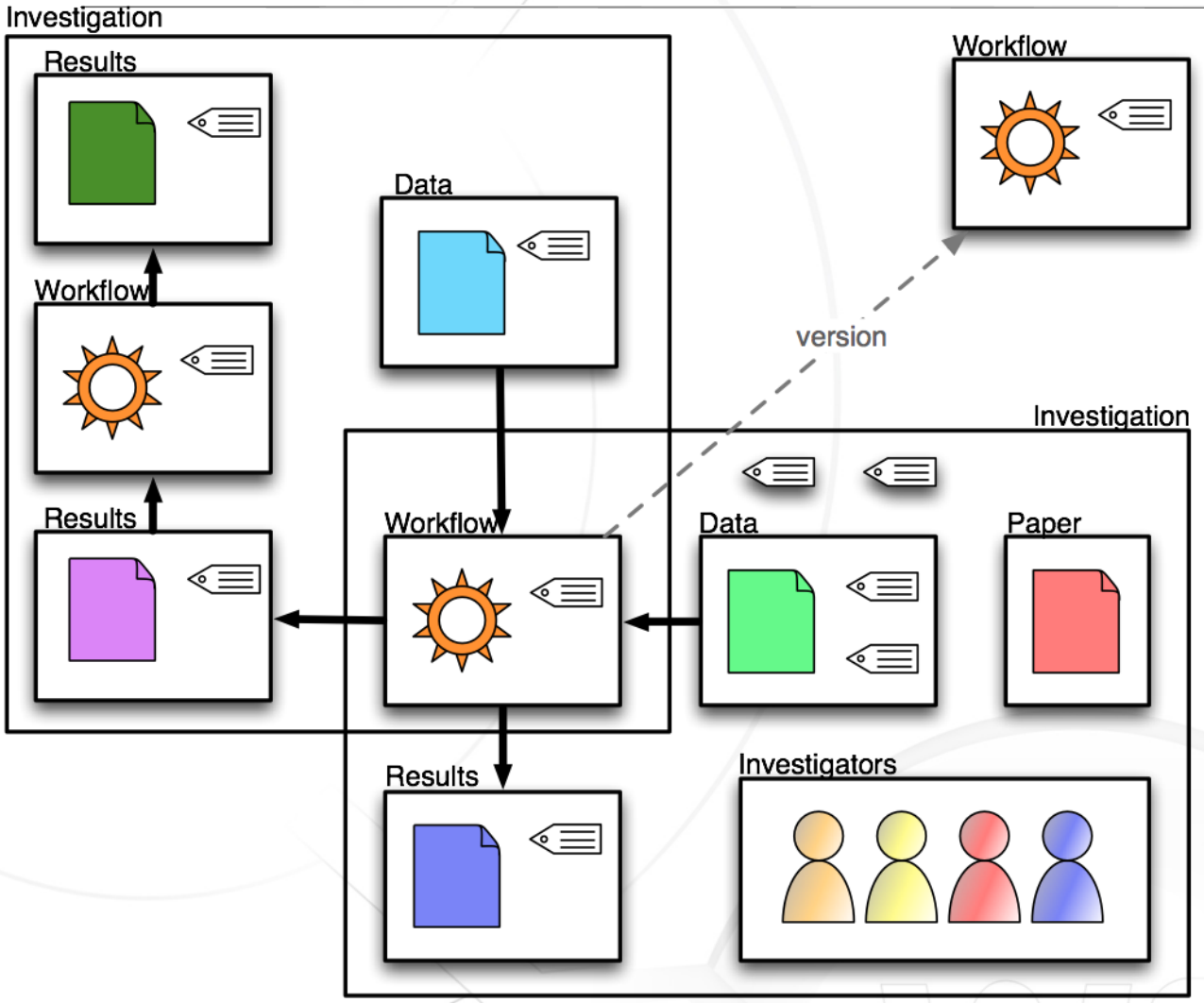
- Methodological changes
- New technologies
- New resources/components
- New data



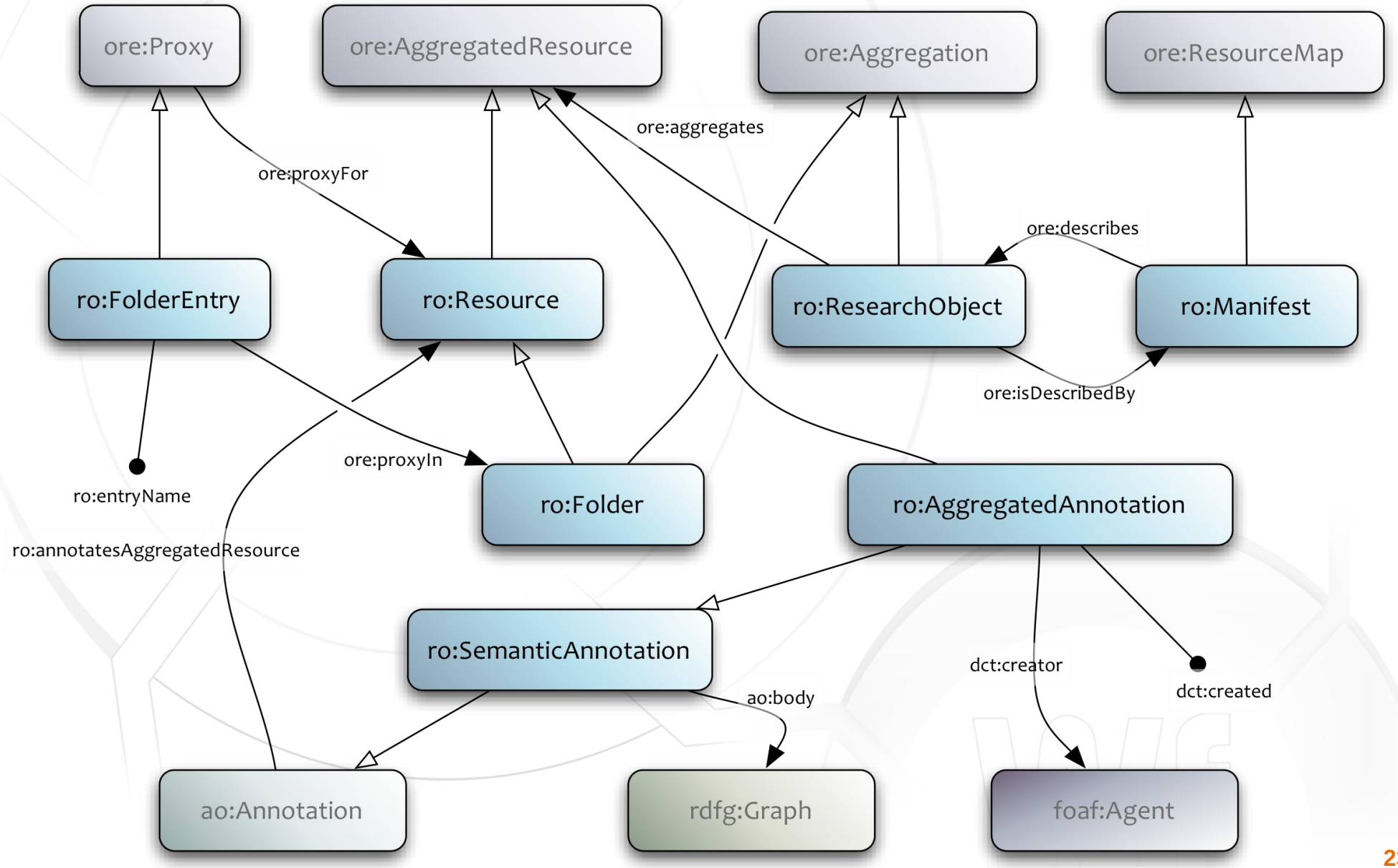
Research Objects as Social Objects



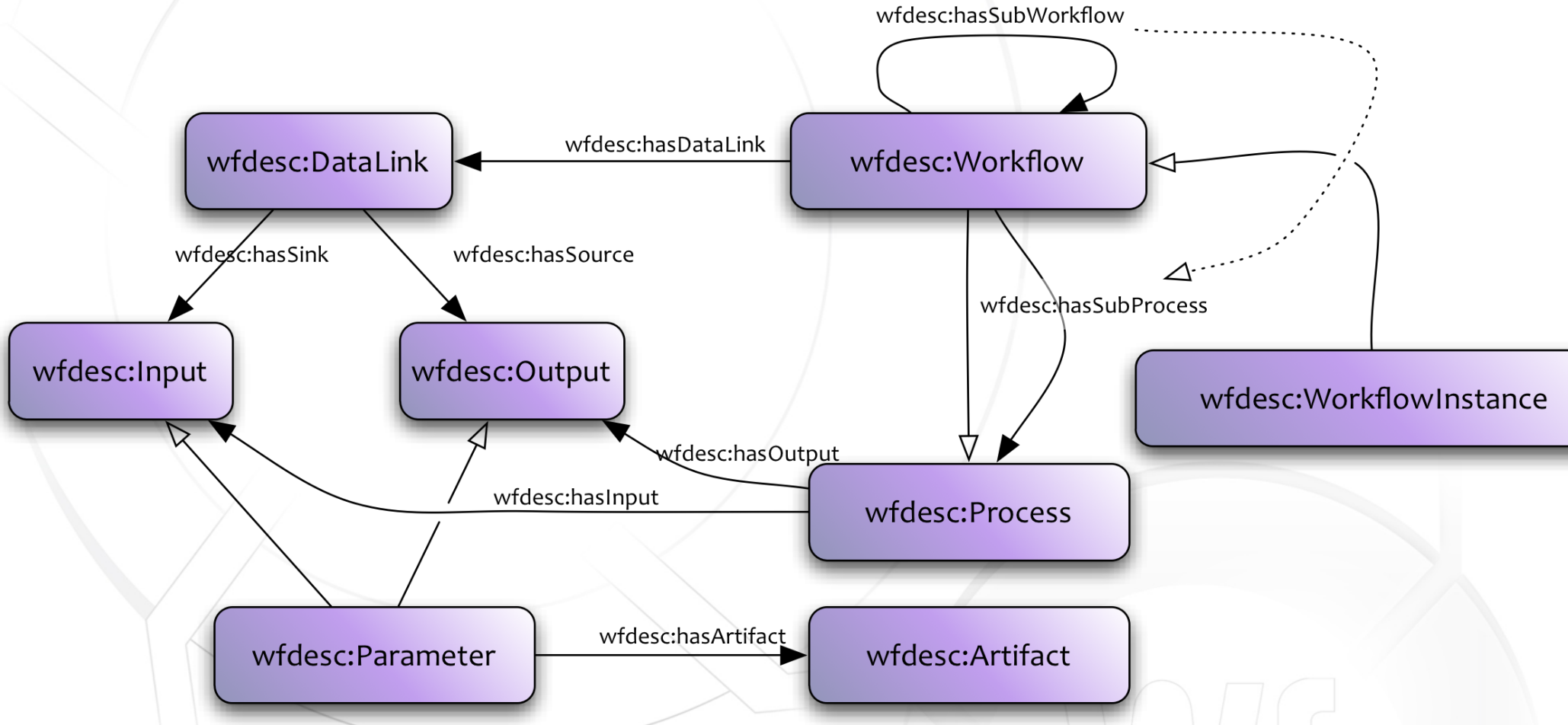
Research objects



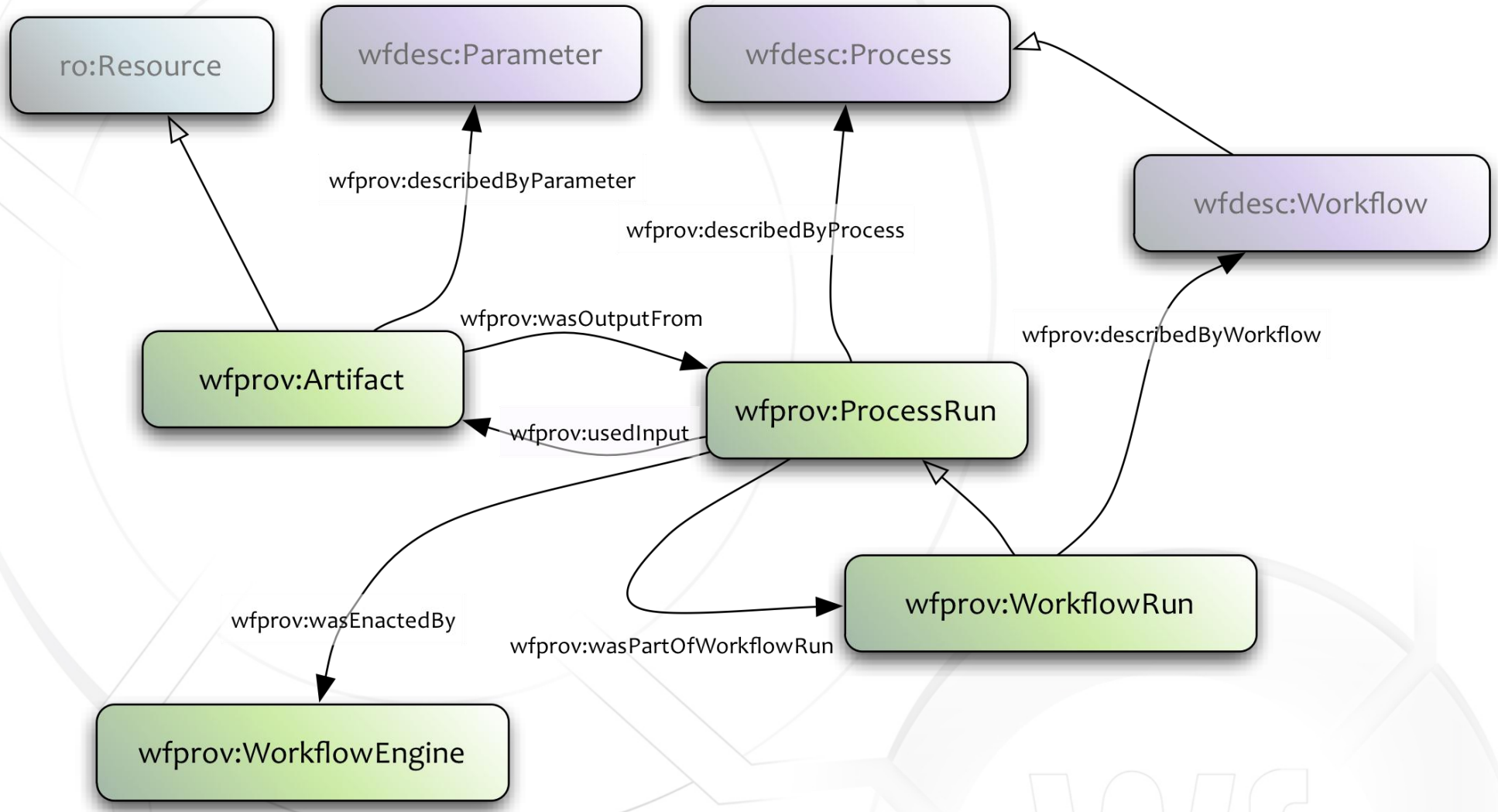
Research Object Core (ro)



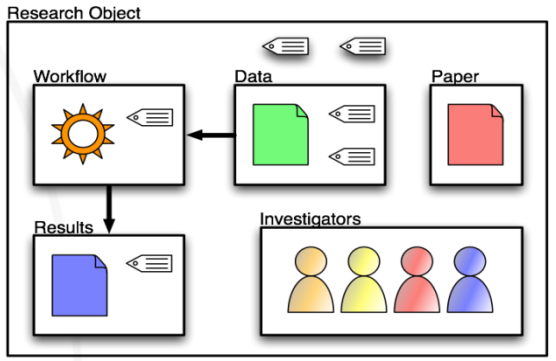
Workflow Description (wfdesc)

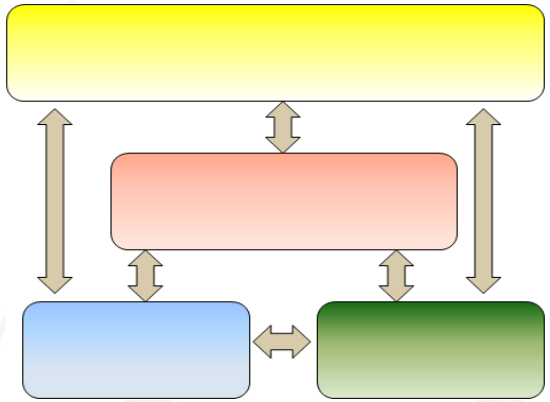


Workflow Provenance (wfprov)

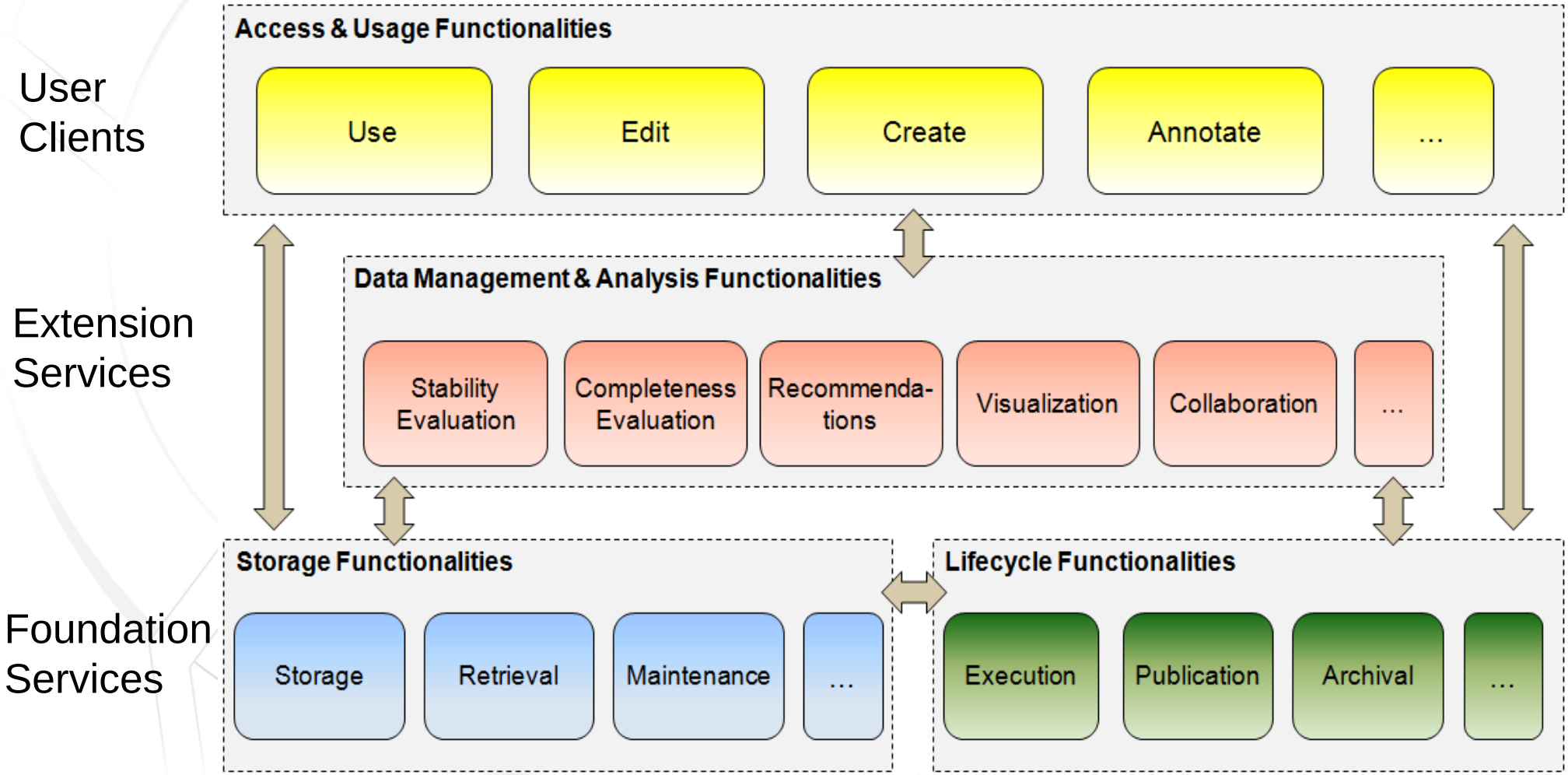


- **Models**
  - Research Object
  - Annotation
  - Provenance
  - Evolution and Versioning
  - *Semantic Web Encoding*
- **Services**
  - Foundational, Extension, User
  - APIs, Architecture
  - *Web protocols/services*
- **Principles**
  - Map into standards
  - Adopt standards
  - Lightweight components
- **Ecosystem**
  - Command line
  - Third party systems





Services


User Clients
Extension Services
Foundation Services
Access & Usage Functionalities
Use
Edit
Create
Annotate
…
Data Management & Analysis Functionalities
Stability Evaluation
Completeness Evaluation
Recommenda-tions
Visualization
Collaboration
…
Storage Functionalities
Storage
Retrieval
Maintenance
…
Lifecycle Functionalities
Execution
Publication
Archival
…

Next steps

- Analyse decay within all myExperiment workflows
  - Estimate: Roughly 50% don't run correctly
  - But why? Services gone? Did they ever work?
  - How have the community evolved those workflows?
- Service and wf substitutions; recommendations
- Provenance analysis and use
  - e.g. verifiability, replayability
- SHIWA approach – can handle "What if the workflow system stops working"

Any Questions?

http://www.myexperiment.org/

http://www.wf4ever-project.org/

http://www.mygrid.org.uk/

http://www.taverna.org.uk/