# Research Objects: Preserving Scientific Workflows and Provenance

Khalid Belhajjame
Université Paris Dauphine

# Storyline

- Why Research Objects?
- Overview of the Research Objects
- Portfolio of Research Object Management Tools
- Provenance Distillation Through Workflow Summarization

# Electronic papers are not enough

**Notes on writing a scientific paper**

by Paul Krieg (7/21/97, revised 10/3/97, 2/03, 5/05, 5/06)

This is a guide toward successful writing scientific papers in the Krieg Genetics Lab. Scientific writing is not the same as creative writing. This is good as it means that anyone can do it. It is paint-by-the numbers, rather than Monet!

I have come up with this system over two decades and it seems to work well, especially for beginning writers. Point 3 below is the key to the system. START WITH THE RESULTS. We do not write scientific papers in the same order that they appear in the journal (e.g. Introduction, methods, results and discussion). In fact, I advocate writing the introduction last (see point 4).

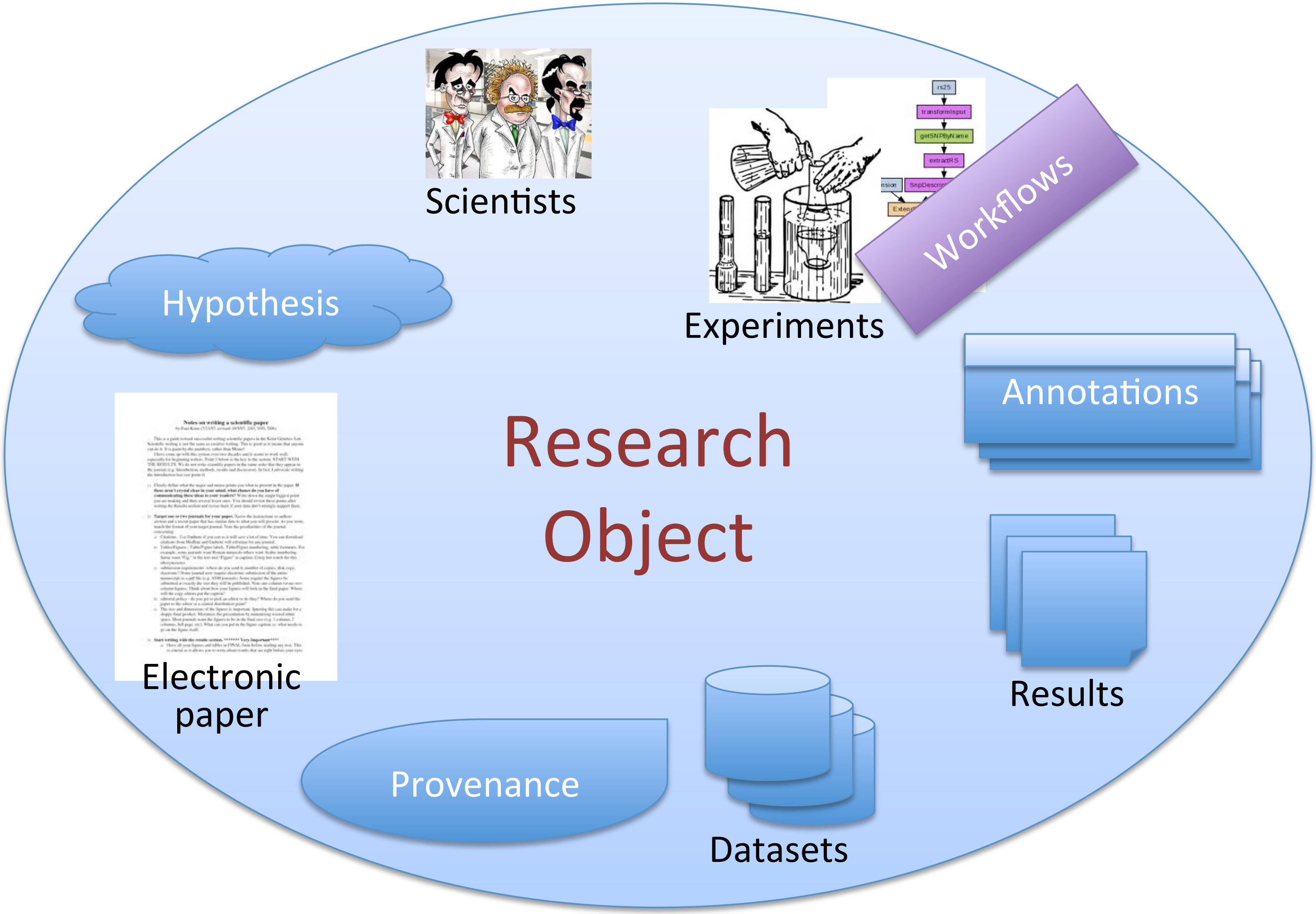
1) Clearly define what the major and minor points you what to present in the paper. **If these aren't crystal clear in your mind, what chance do you have of communicating these ideas to your readers?** Write down the single biggest point you are making and then several lesser ones. You should revisit these points after writing the Results section and revise them if your data don't strongly support them.

2) **Target one or two journals for your paper.** Xerox the instructions to authors section and a recent paper that has similar data to what you will present. As you write, match the format of your target journal. Note the peculiarities of the journal concerning:
   a) Citations.. Use Endnote if you can as it will save a lot of time. You can download citations from Medline and Endnote will reformat for any journal.
   b) Tables/Figures.. Table/Figure labels, Table/Figure numbering, table footnotes. For example, some journals want Roman numerals others want Arabic numbering. Some want "Fig." in the text and "Figure" in caption. Crazy but watch for this idiosyncrasies.
   c) submission requirements -where do you send it, number of copies, disk copy, electronic? Some journal now require electronic submission of the entire manuscript as a pdf file (e.g. ASM journals). Some require the figures be submitted at exactly the size they will be published. Note one column versus two column figures. Think about how your figures will look in the final paper. Where will the copy editors put the caption?
   d) editorial policy - do you get to pick an editor or do they? Where do you send the paper to the editor or a central distribution point?
   e) The size and dimensions of the figures is important. Ignoring this can make for a sloppy final product. Maximize the presentation by minimizing wasted white space. Most journals want the figures to be in the final size (e.g. 1 column, 2 column, full page, etc). What can you put in the figure caption vs. what needs to go on the figure itself.

3) **Start writing with the results section. ******* Very Important******
   a) Have all your figures and tables in FINAL form before starting any text. This is crucial as it allows you to write about results that are right before your eyes.

Electronic
paper

# Electronic papers are not enough



Research Object

Scientists

Experiments

Workflows

Hypothesis

Annotations

Electronic paper

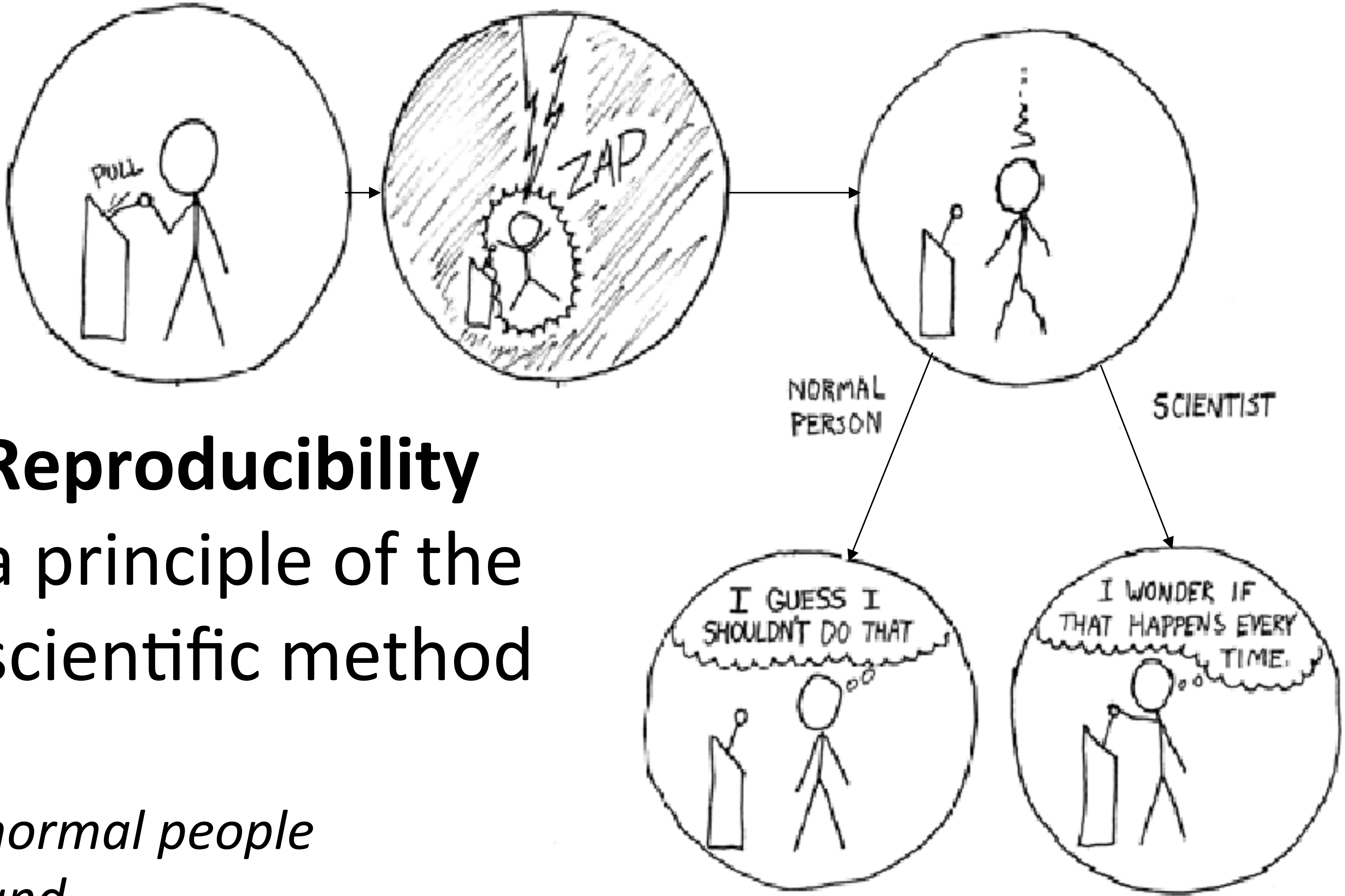
Results

Provenance

Datasets

# Benefits Of Research Objects

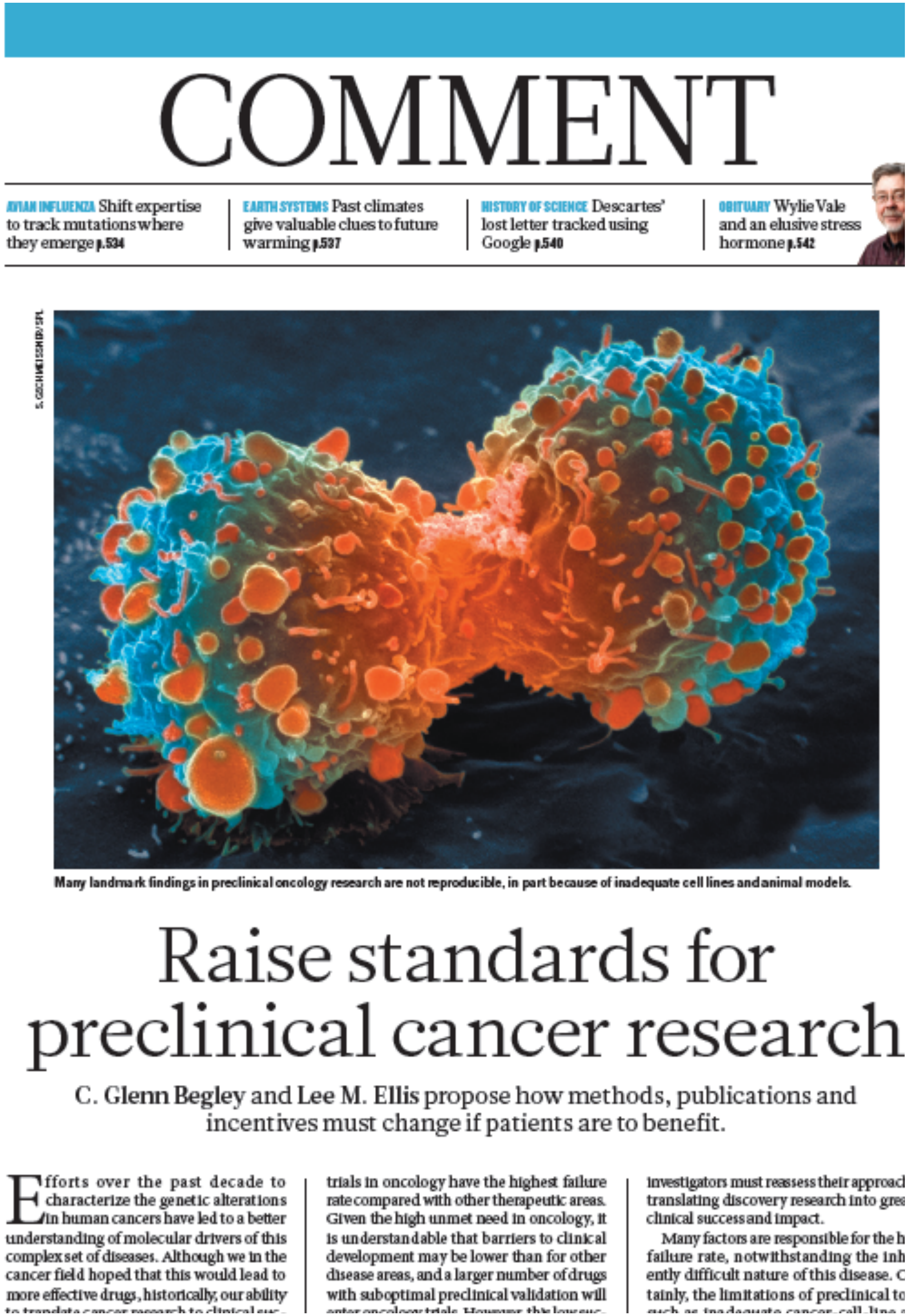
- A research object aggregates all elements that are necessary to understand research investigations.
- Methods (experiments) are viewed as first class citizens
- Promote reuse
- Enable the verification of reproducibility of the results


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

**Reproducibility**
a principle of the scientific method

*normal people*
*and*
*scientist*

http://xkcd.com/242/



# COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534

EARTH SYSTEMS Past climates give valuable clues to future warming p.537

HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540

OBITUARY Wylie Vale and an elusive stress hormone p.542

S. GSCHMEISSNER/SPL

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

## Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical suc- trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low suc- investigators must reassess their approac translating discovery research into grea clinical success and impact.

Many factors are responsible for the h failure rate, notwithstanding the inh ently difficult nature of this disease. C tainly, the limitations of preclinical to such as inadequate cancer-cell-line

# 47 of 53 "landmark" publications could not be replicated

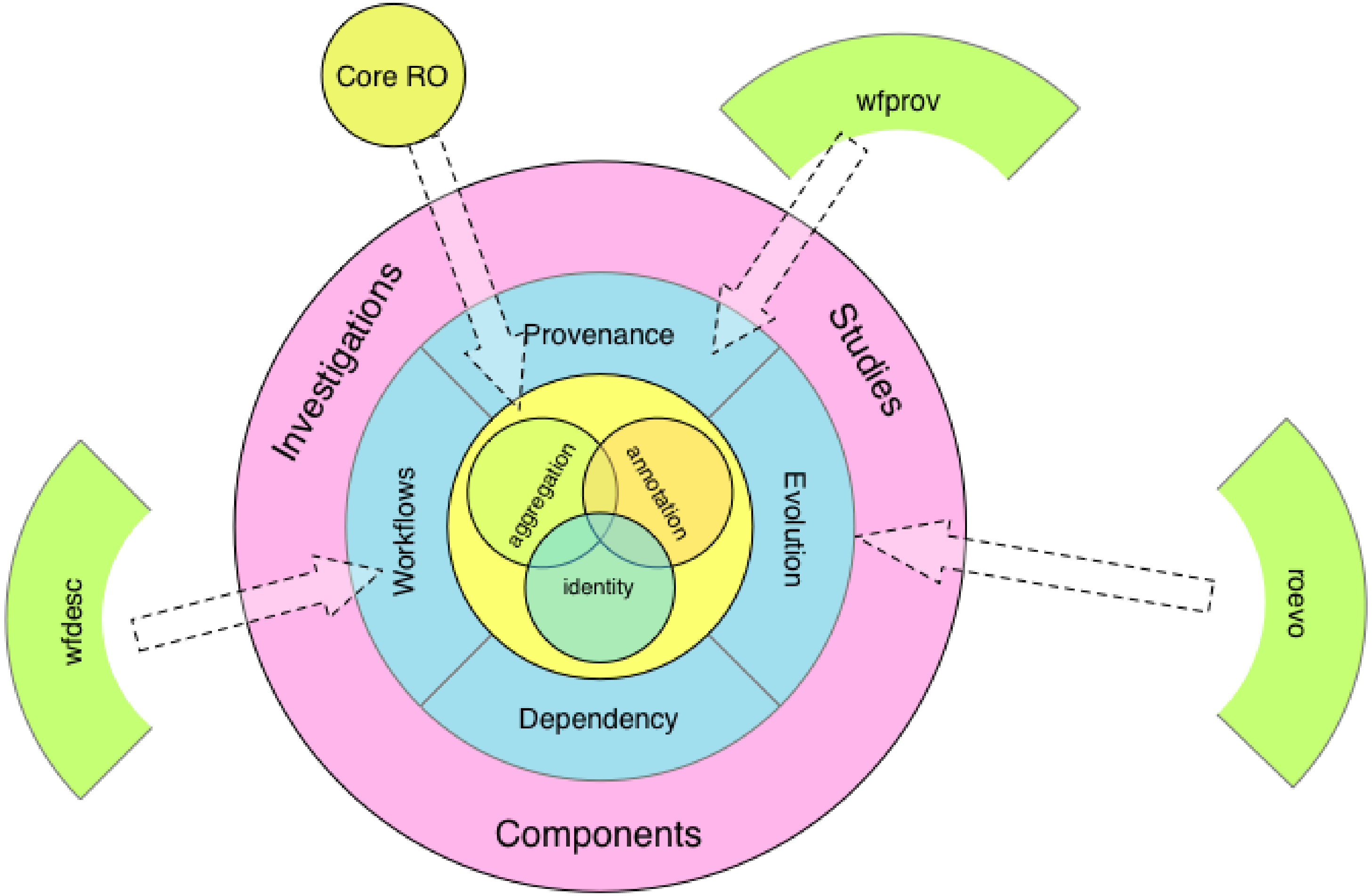
Inadequate cell lines and animal models

Nature, 483, 2012

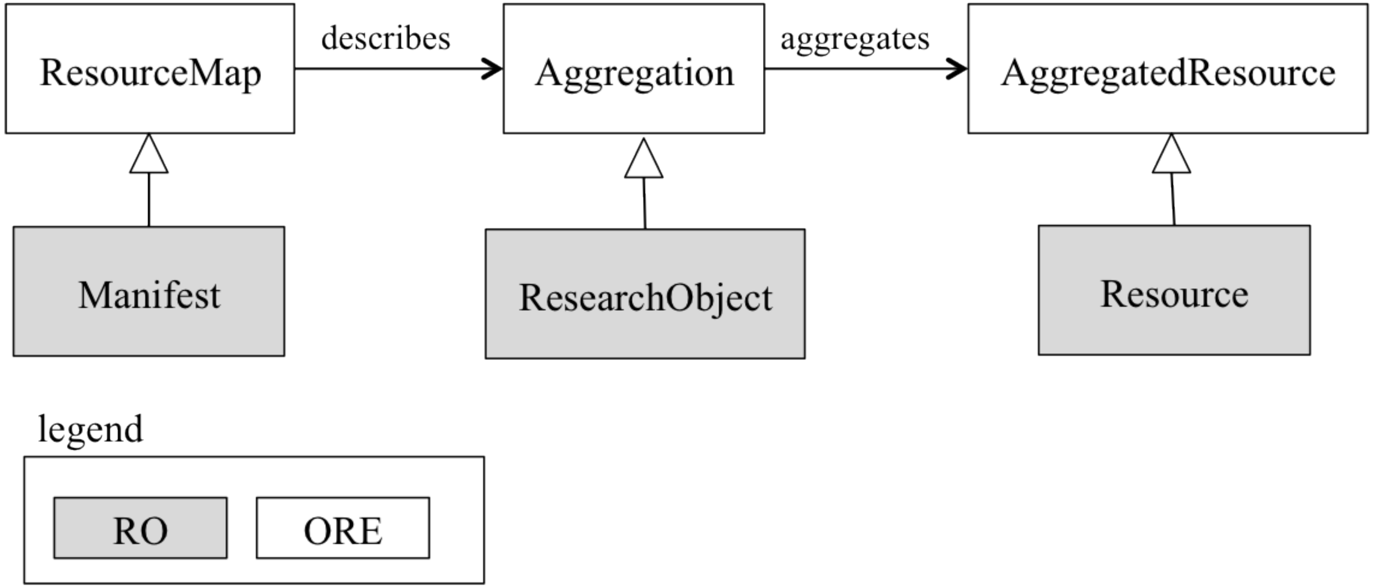
Credit to Carole Goble JCDL 2012 Keynote

Workflow 4Ever

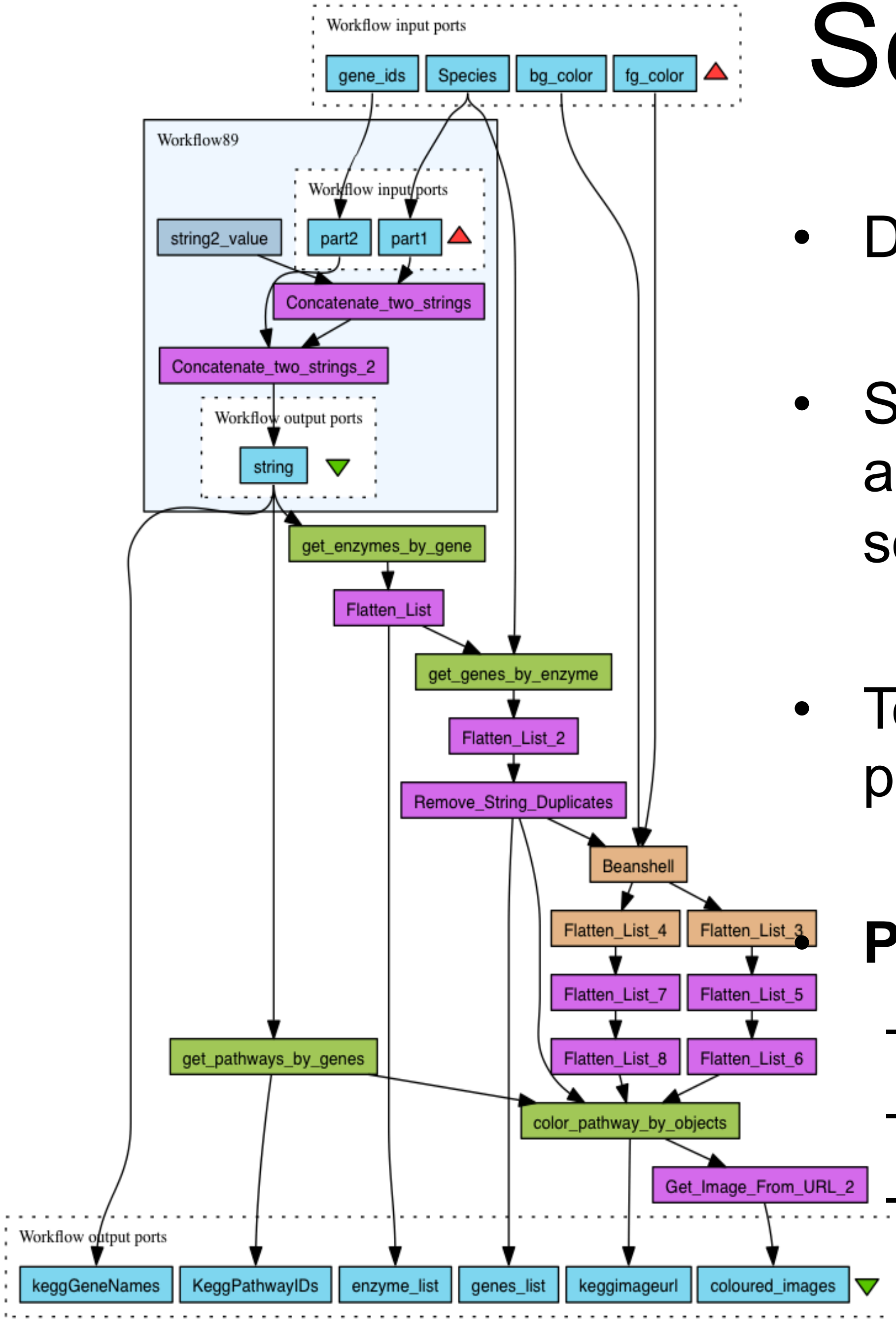
# Overview of the Research Object Model



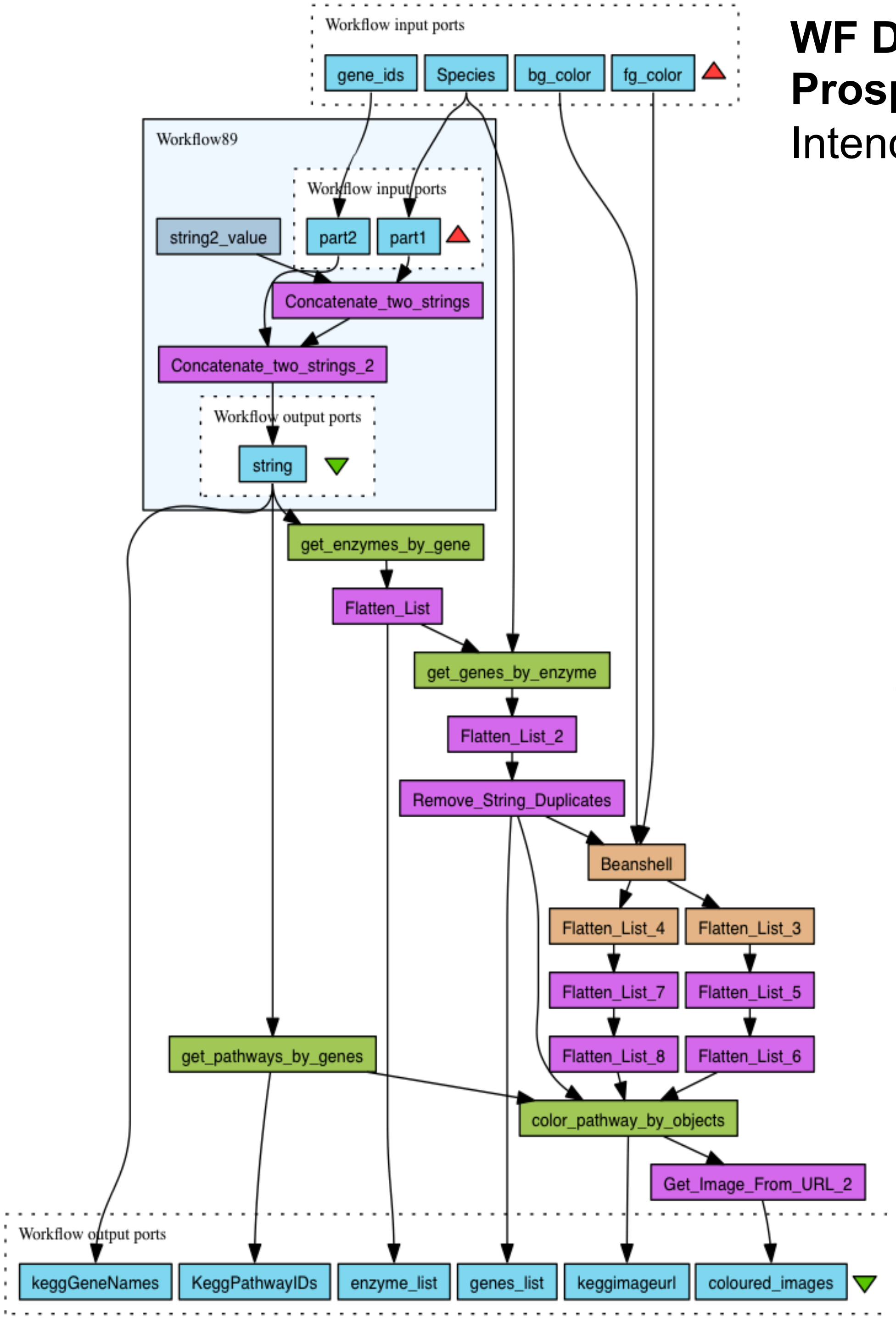
# Research Object as an ORE Aggregation



# Scientific Workflows

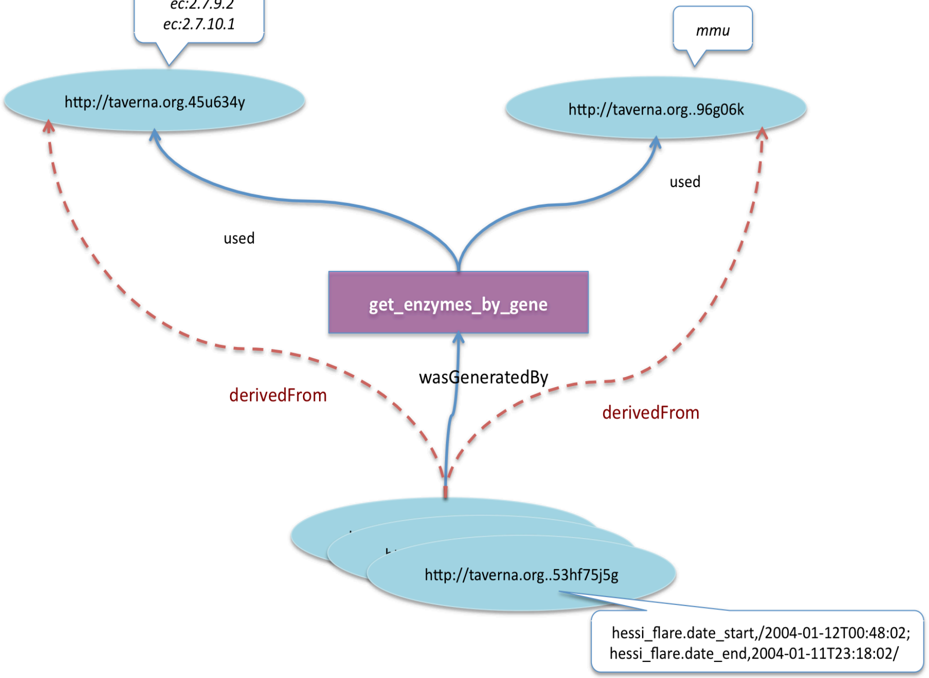


- Data driven analysis pipelines
- Systematic gathering of data and analysis tools into computational solutions for scientific problem-solving
- Tools for automating frequently performed data intensive activities
- **Provenance** for the resulting datasets
  - The method followed
  - The resources used
  - The datasets used



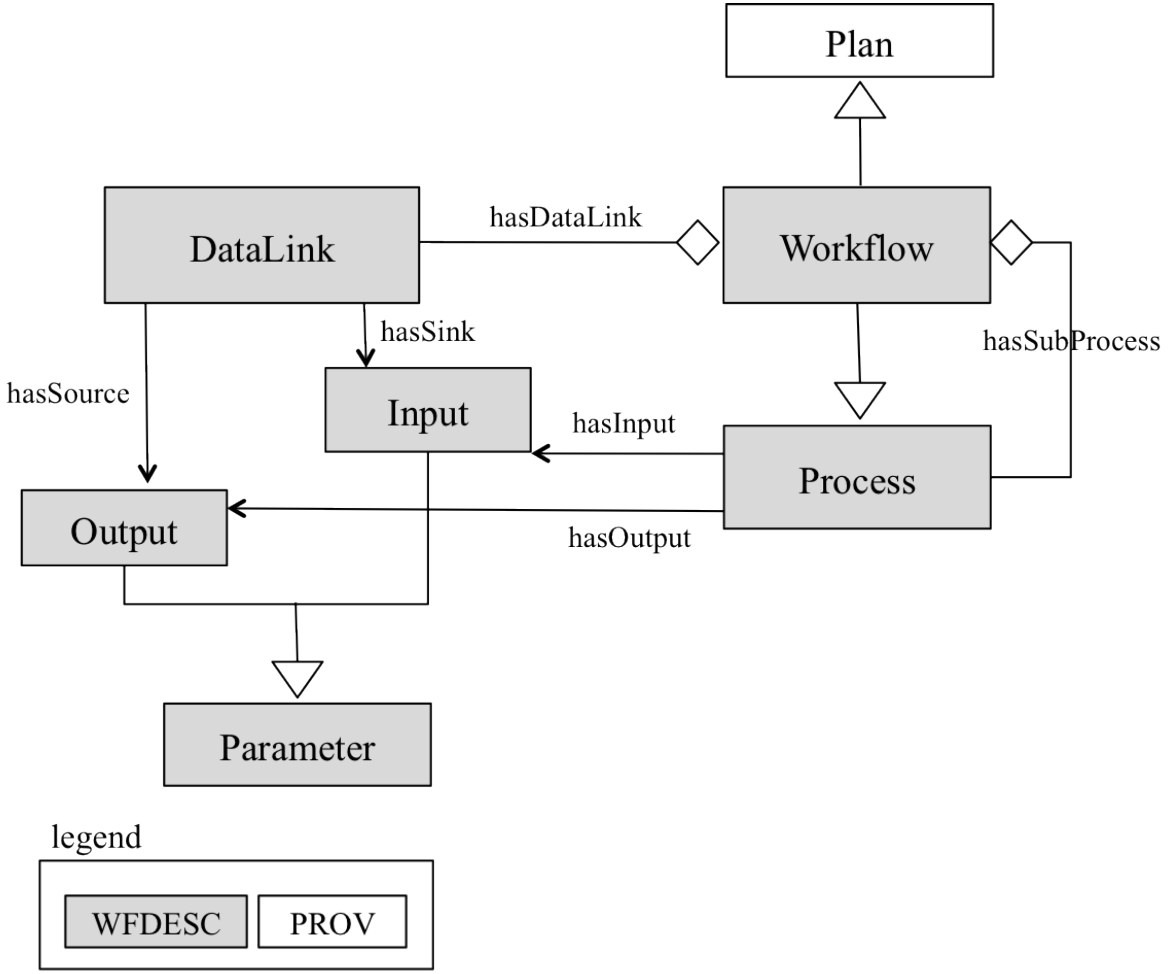
**WF Description**
**Prospective Provenance:**
Intended method for analysis

**WF Execution Trace**
**Retrospective Provenance:**
Actual data used, actual invocations, timestamps and data derivation trace

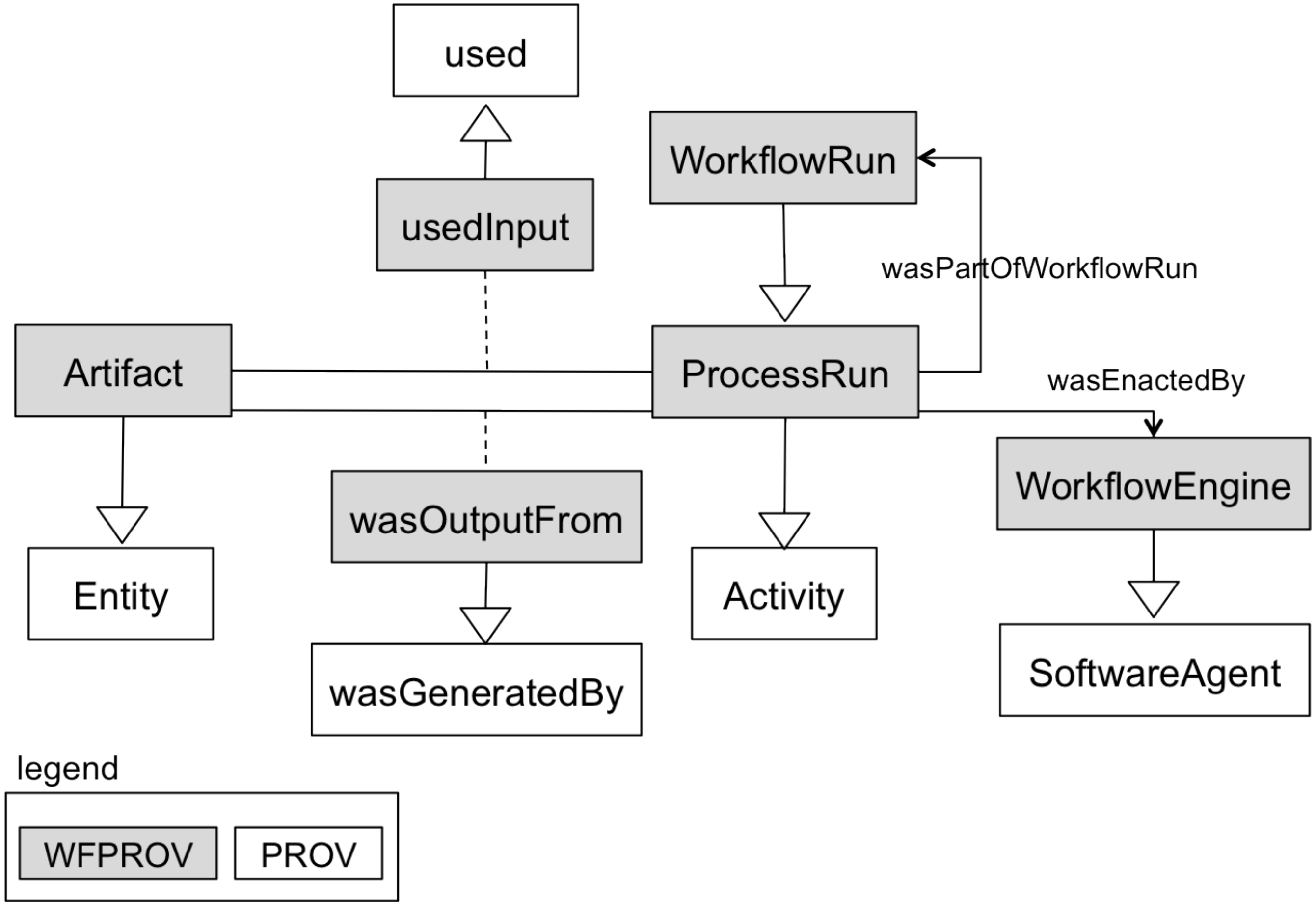

ec:2.7.9.2
ec:2.7.10.1
mmu
http://taverna.org.45u634y
http://taverna.org..96g06k
used
used
get_enzymes_by_gene
wasGeneratedBy
derivedFrom
derivedFrom
http://taverna.org..53hf75j5g
hessi_flare.date_start,/2004-01-12T00:48:02;
hessi_flare.date_end,2004-01-11T23:18:02/

*PROV Primer, Gil et al*

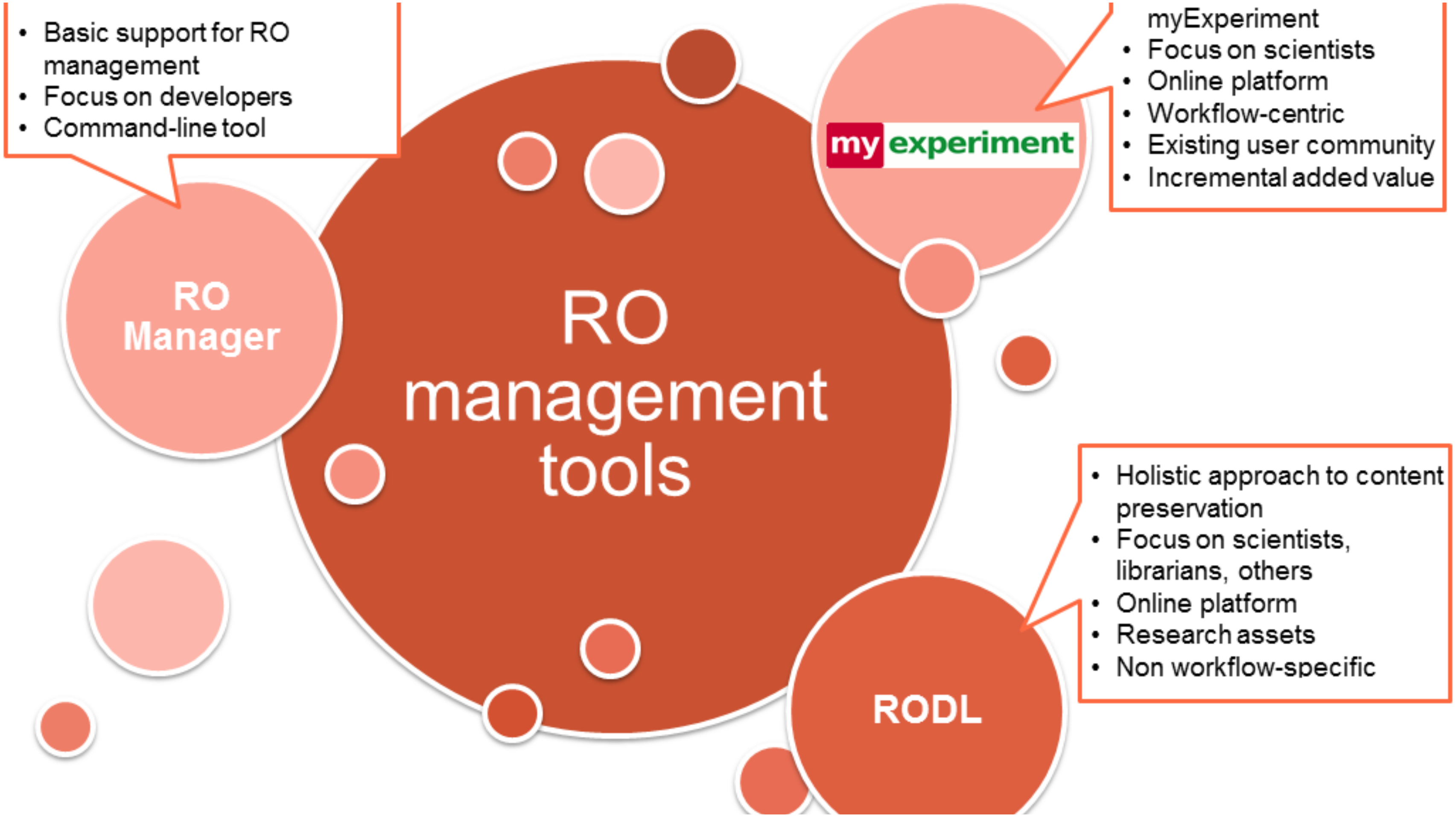
# Specifying Workflows using WfDESC



# Specifying Workflow Provenance using WfPROV



# Portfolio of Research Object Tools



- Basic support for RO management
- Focus on developers
- Command-line tool

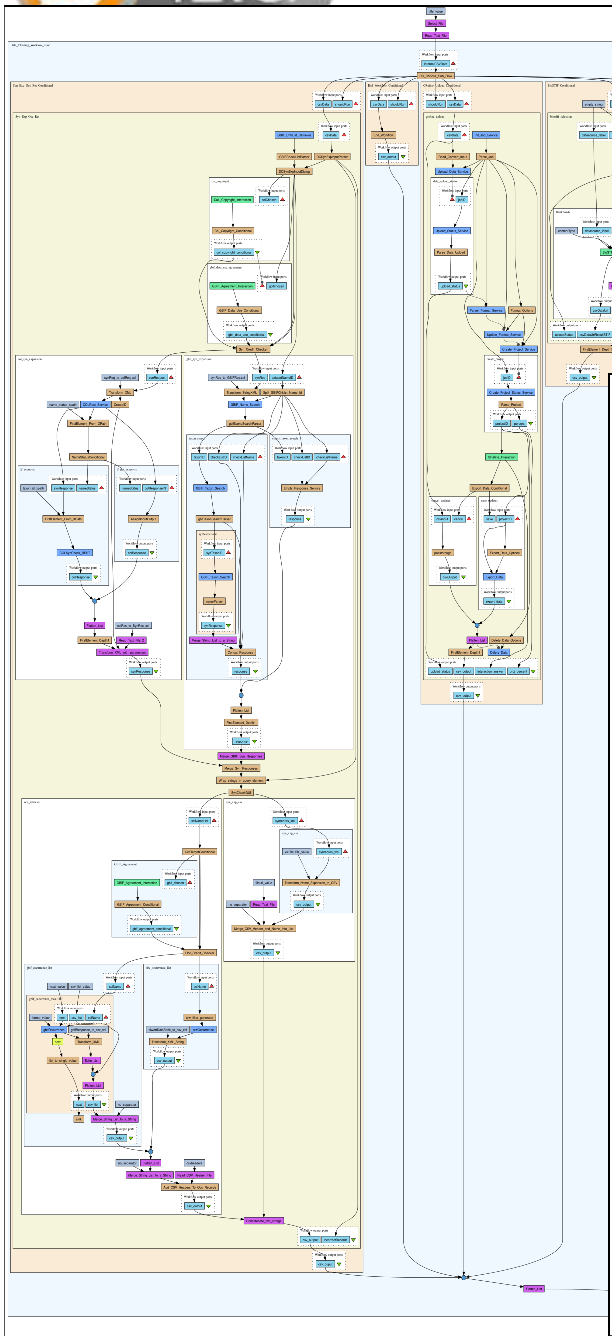
myExperiment
- Focus on scientists
- Online platform
- Workflow-centric
- Existing user community
- Incremental added value

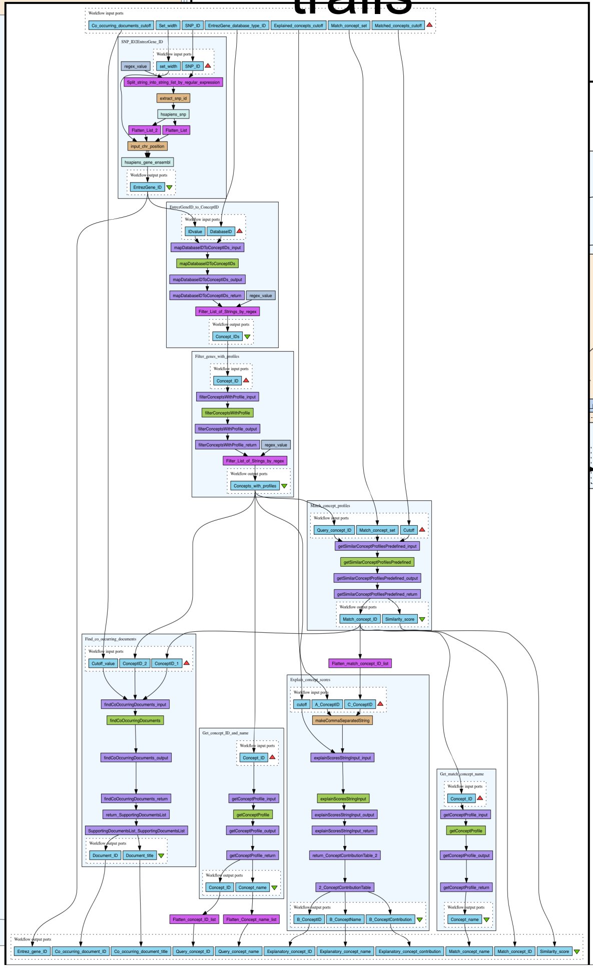
- Holistic approach to content preservation
- Focus on scientists, librarians, others
- Online platform
- Research assets
- Non workflow-specific

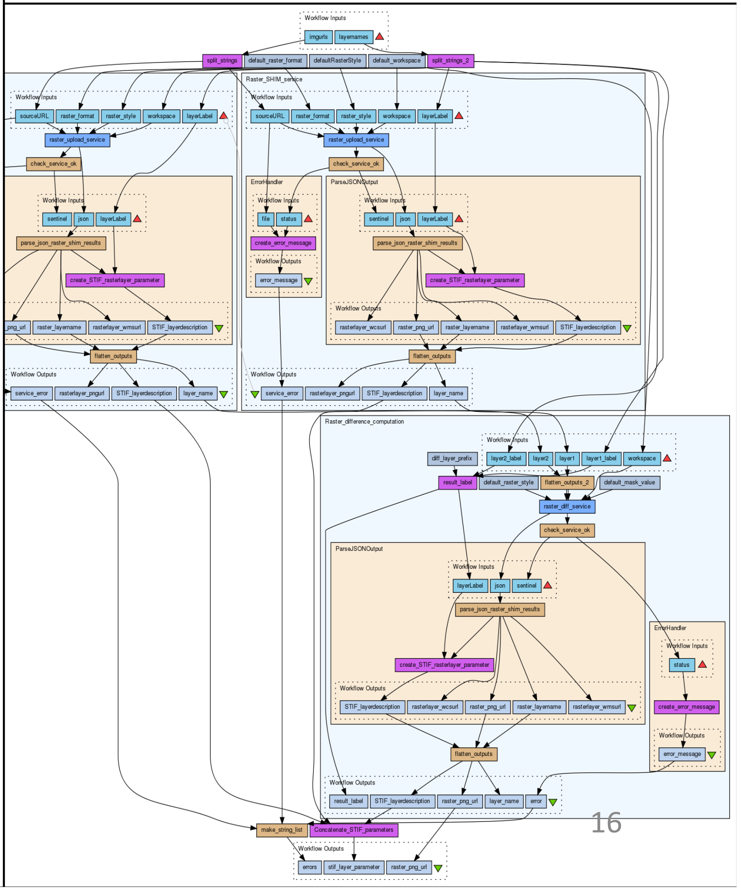
# DEMO

# Workflows can get complex!

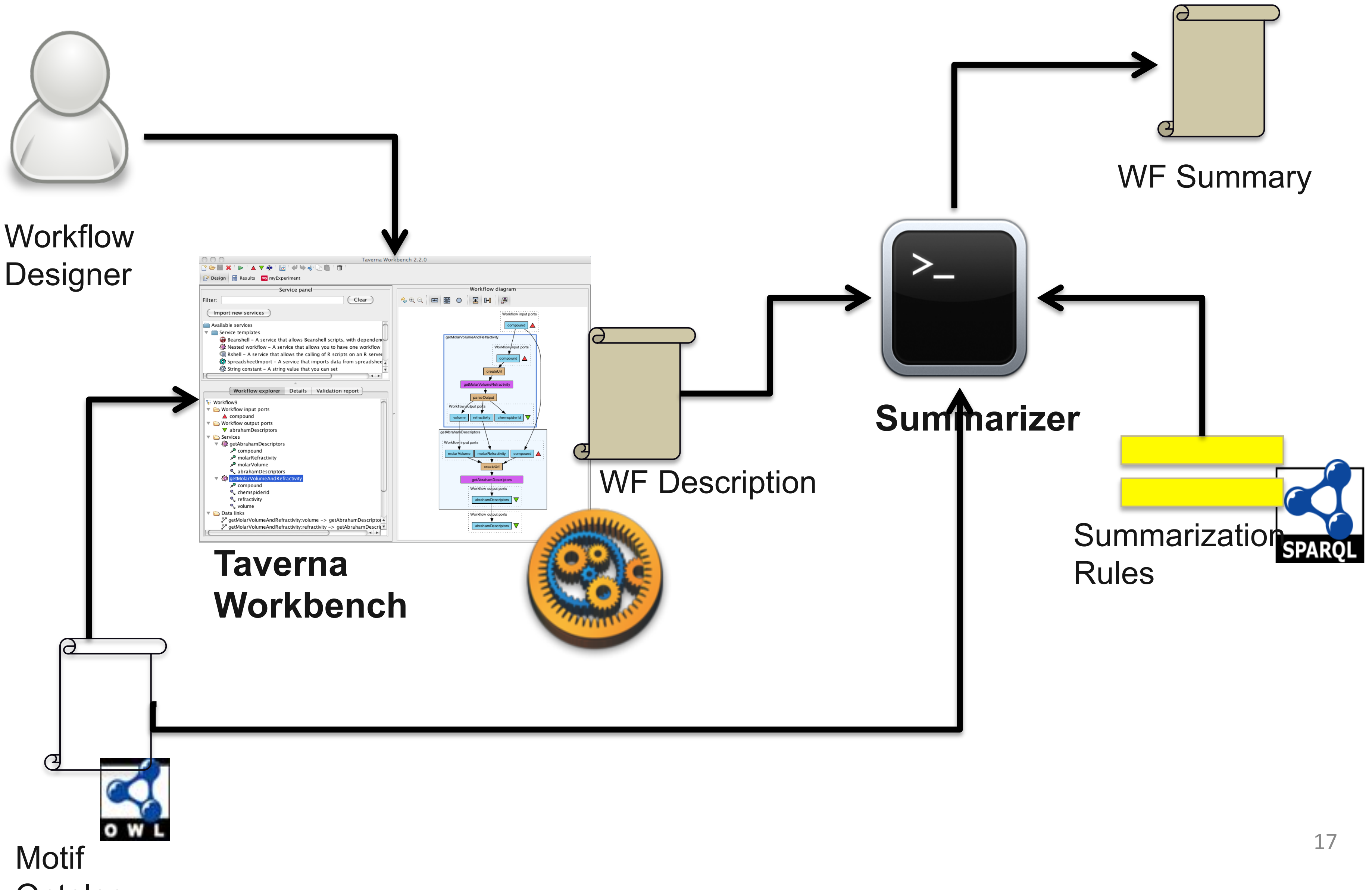
- Overwhelming for users who are not the developers
- Abstractions required for reporting
- Lineage queries result in very long trails





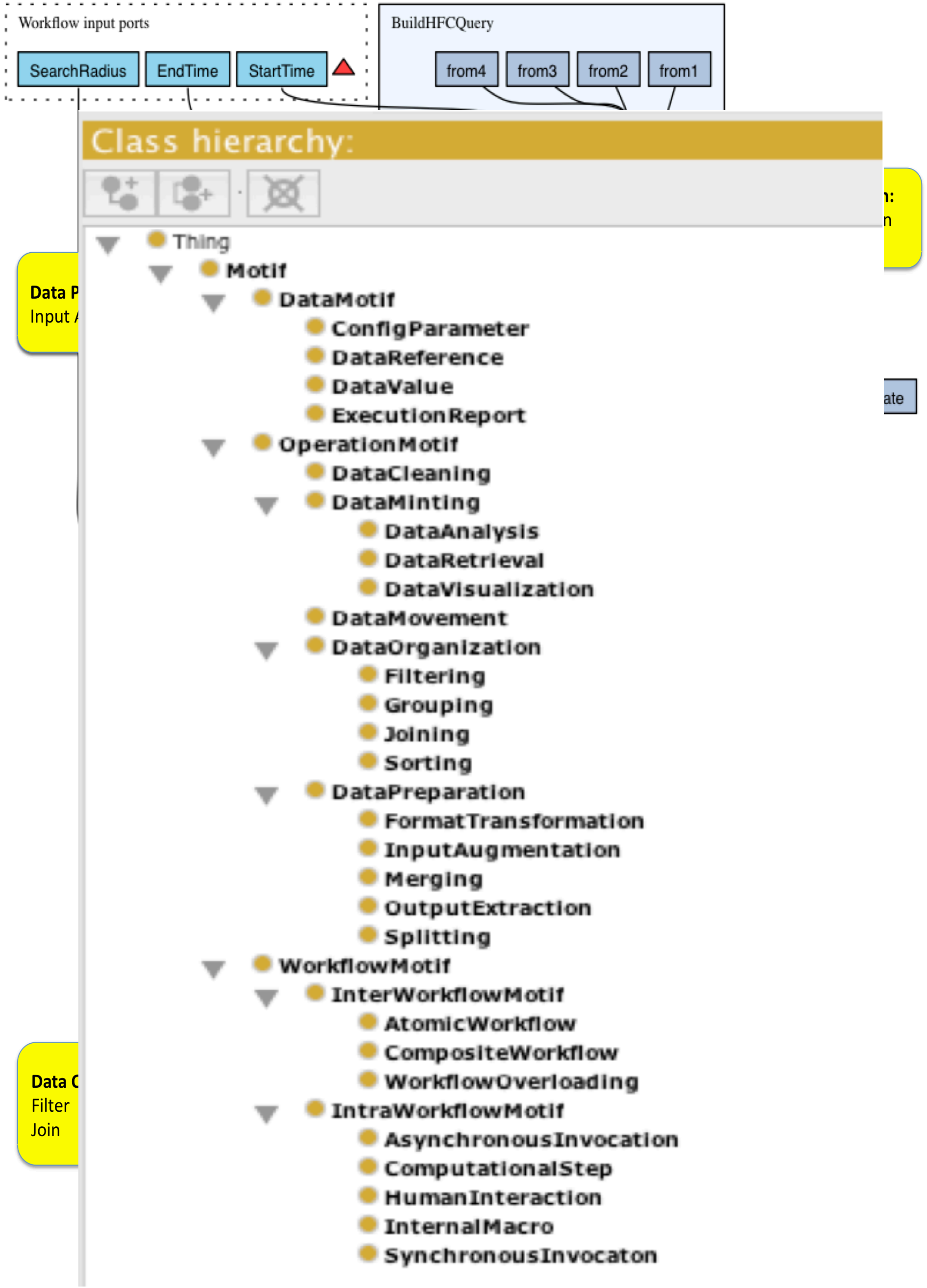


# Overall Approach



# PART-1: Scientific Workflow Motifs

- Domain Independent categorization
  - Data-Oriented Nature
  - Resource/Implementation-Oriented Nature
- Captured In a lightweight OWL Ontology

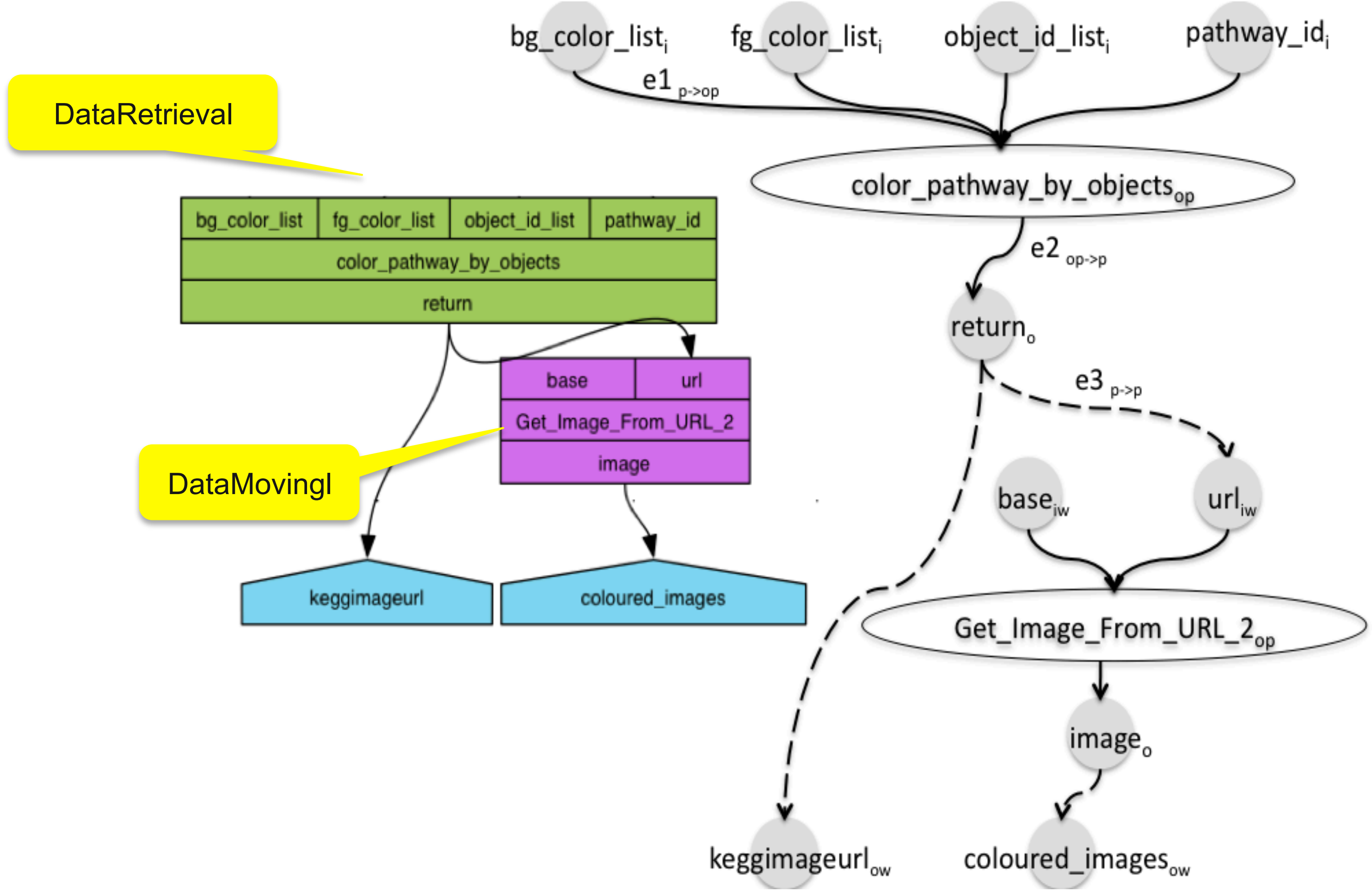


Class hierarchy:

```
Thing
  Motif
    DataMotif
      ConfigParameter
      DataReference
      DataValue
      ExecutionReport
    OperationMotif
      DataCleaning
      DataMinting
        DataAnalysis
        DataRetrieval
        DataVisualization
      DataMovement
      DataOrganization
        Filtering
        Grouping
        Joining
        Sorting
      DataPreparation
        FormatTransformation
        InputAugmentation
        Merging
        OutputExtraction
        Splitting
    WorkflowMotif
      InterWorkflowMotif
        AtomicWorkflow
        CompositeWorkflow
        WorkflowOverloading
      IntraWorkflowMotif
        AsynchronousInvocation
        ComputationalStep
        HumanInteraction
        InternalMacro
        SynchronousInvocaton
```

http://purl.org/net/wf-motifs

# Motif annotations over operations

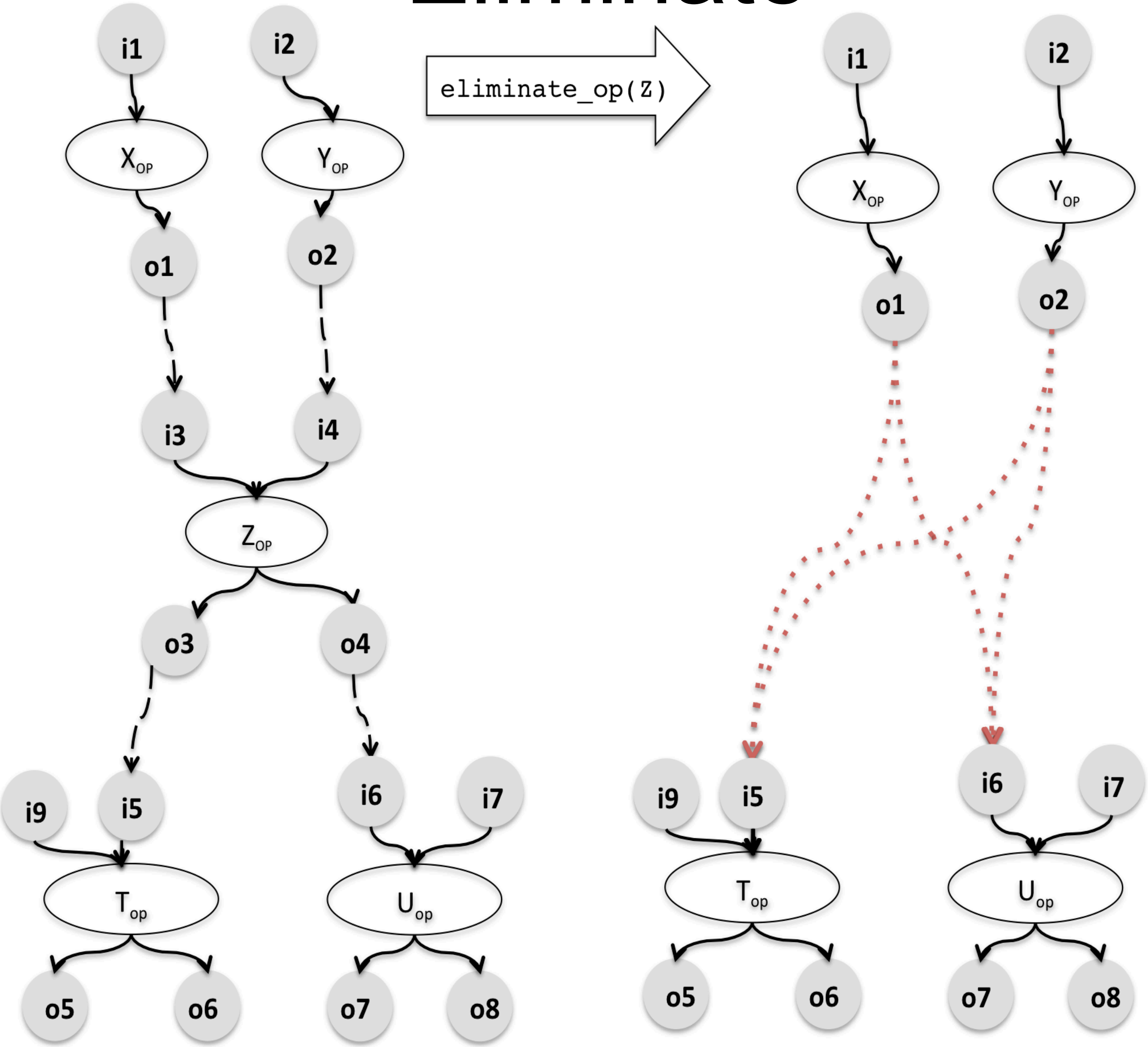


```
motifs(color_pathway_by_objects) = {m1:DataRetrieval}
motifs(Get_Image_From_URL_2) = {m2:DataMoving}
```

# PART-2: Workflow reduction primitives

- Collapse (Up/Down)
- Compose
- Eliminate

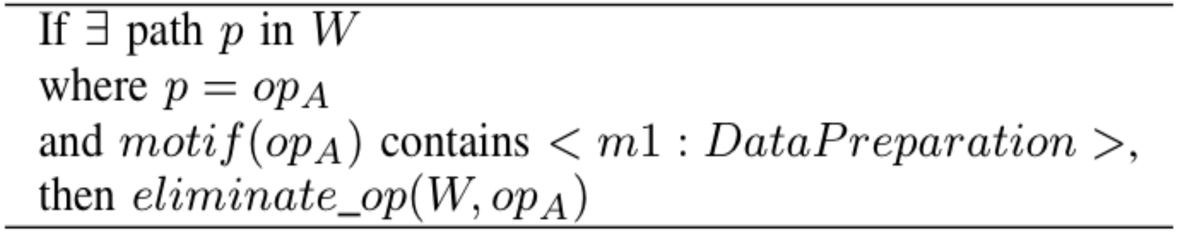
# Eliminate



eliminate_op(Z)

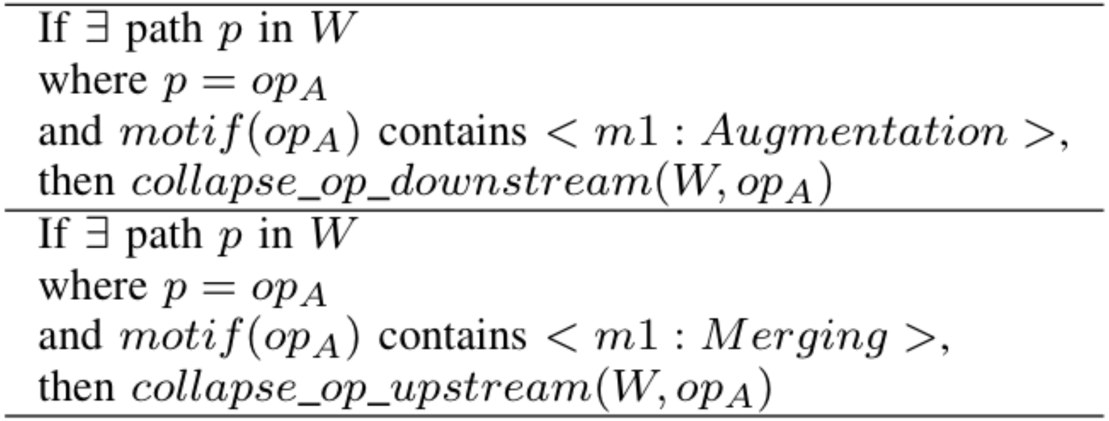
# Two sample strategies

- By-Elimination
  - Minimal annotation effort
  - Single rule



If $\exists$ path $p$ in $W$
where $p = op_A$
and $motif(op_A)$ contains $< m1 : DataPreparation >$,
then $eliminate\_op(W, op_A)$

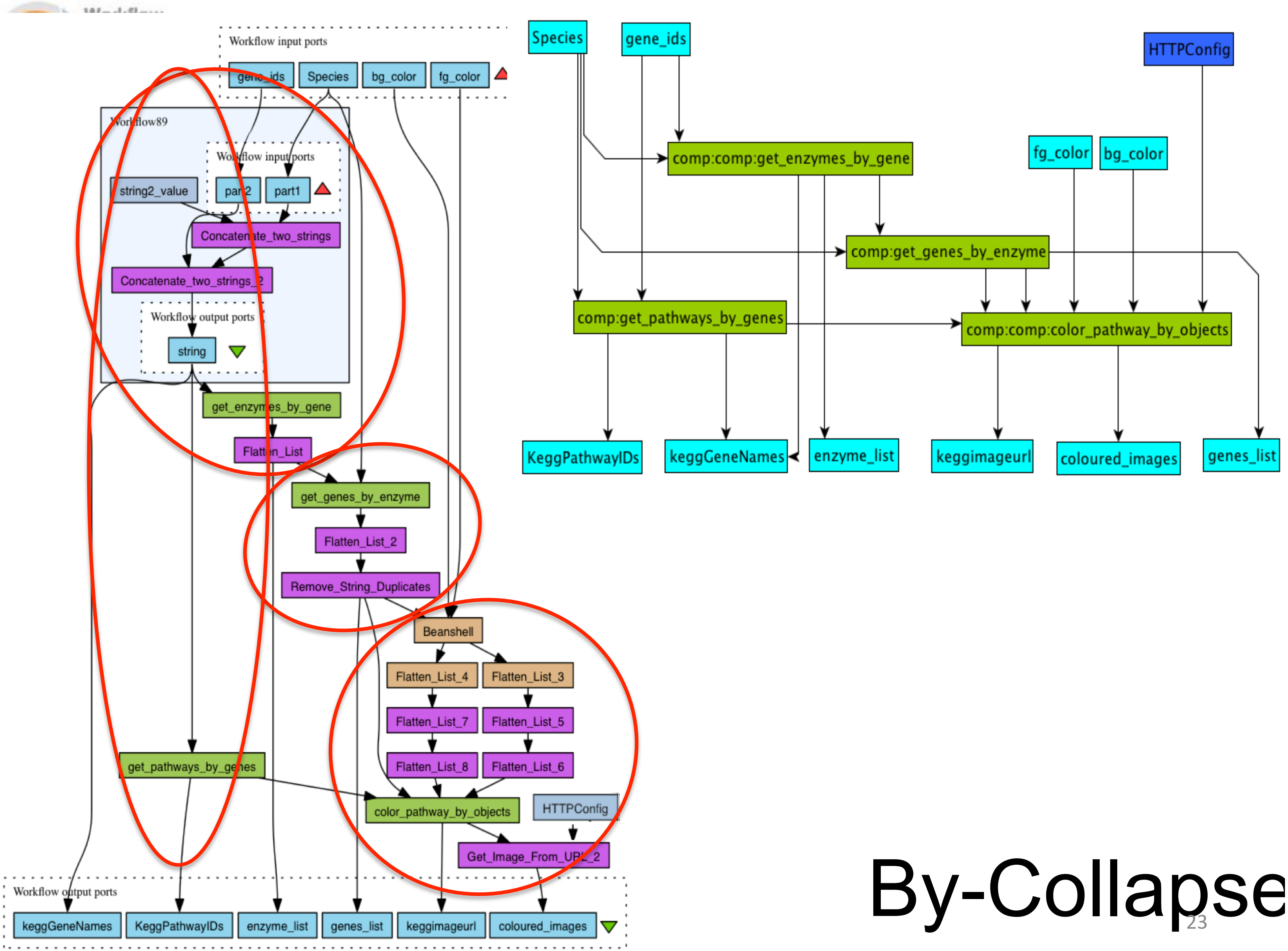
- By Collapse
  - More specific annotation
  - Multiple rules



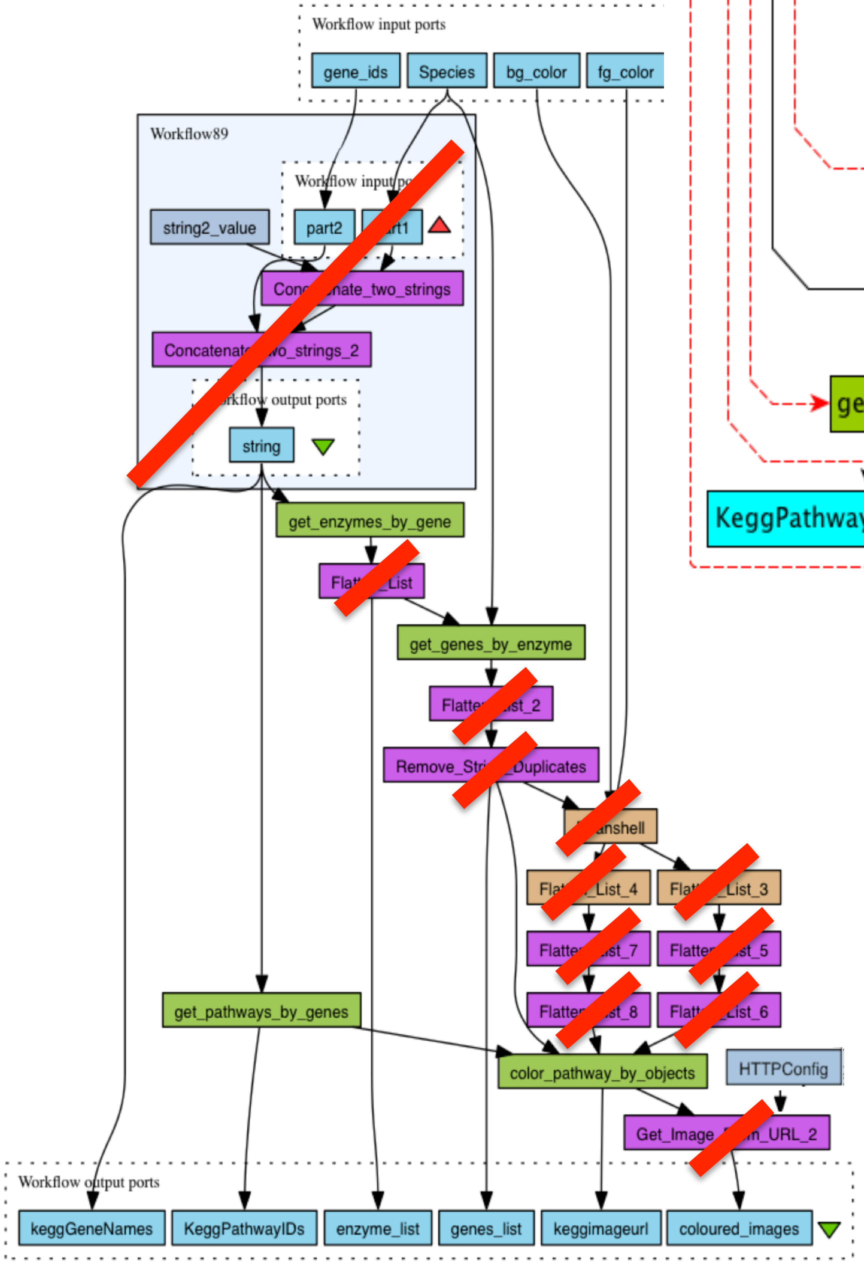
If $\exists$ path $p$ in $W$
where $p = op_A$
and $motif(op_A)$ contains $< m1 : Augmentation >$,
then $collapse\_op\_downstream(W, op_A)$

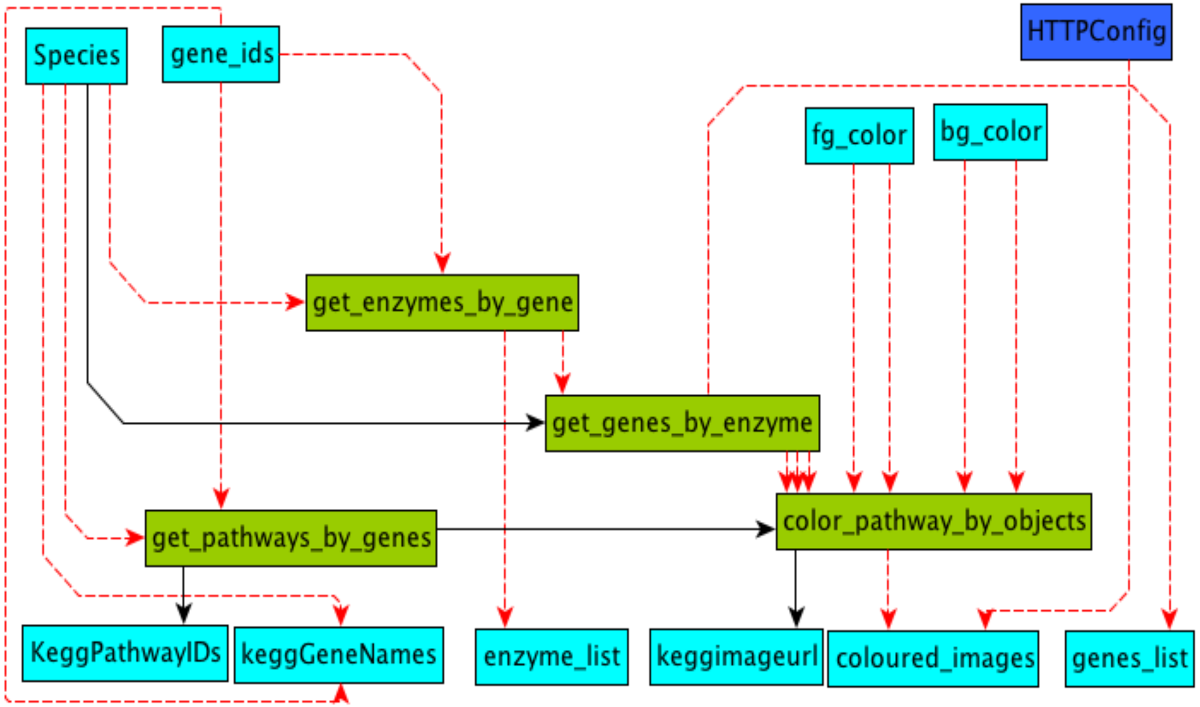
If $\exists$ path $p$ in $W$
where $p = op_A$
and $motif(op_A)$ contains $< m1 : Merging >$,
then $collapse\_op\_upstream(W, op_A)$

# By-Collapse



Workflow input ports
gene_ids
Species
bg_color
fg_color
Workflow89
Workflow input ports
string2_value
part2
part1
Concatenate_two_strings
Concatenate_two_strings_2
Workflow output ports
string
get_enzymes_by_gene
Flatten_List
get_genes_by_enzyme
Flatten_List_2
Remove_String_Duplicates
Beanshell
Flatten_List_4
Flatten_List_3
Flatten_List_7
Flatten_List_5
Flatten_List_8
Flatten_List_6
get_pathways_by_genes
color_pathway_by_objects
HTTPConfig
Get_Image_From_URL_2
Workflow output ports
keggGeneNames
KeggPathwayIDs
enzyme_list
genes_list
keggimageurl
coloured_images
Species
gene_ids
HTTPConfig
comp:comp:get_enzymes_by_gene
fg_color
bg_color
comp:get_genes_by_enzyme
comp:get_pathways_by_genes
comp:comp:color_pathway_by_objects
KeggPathwayIDs
keggGeneNames
enzyme_list
keggimageurl
coloured_images
genes_list



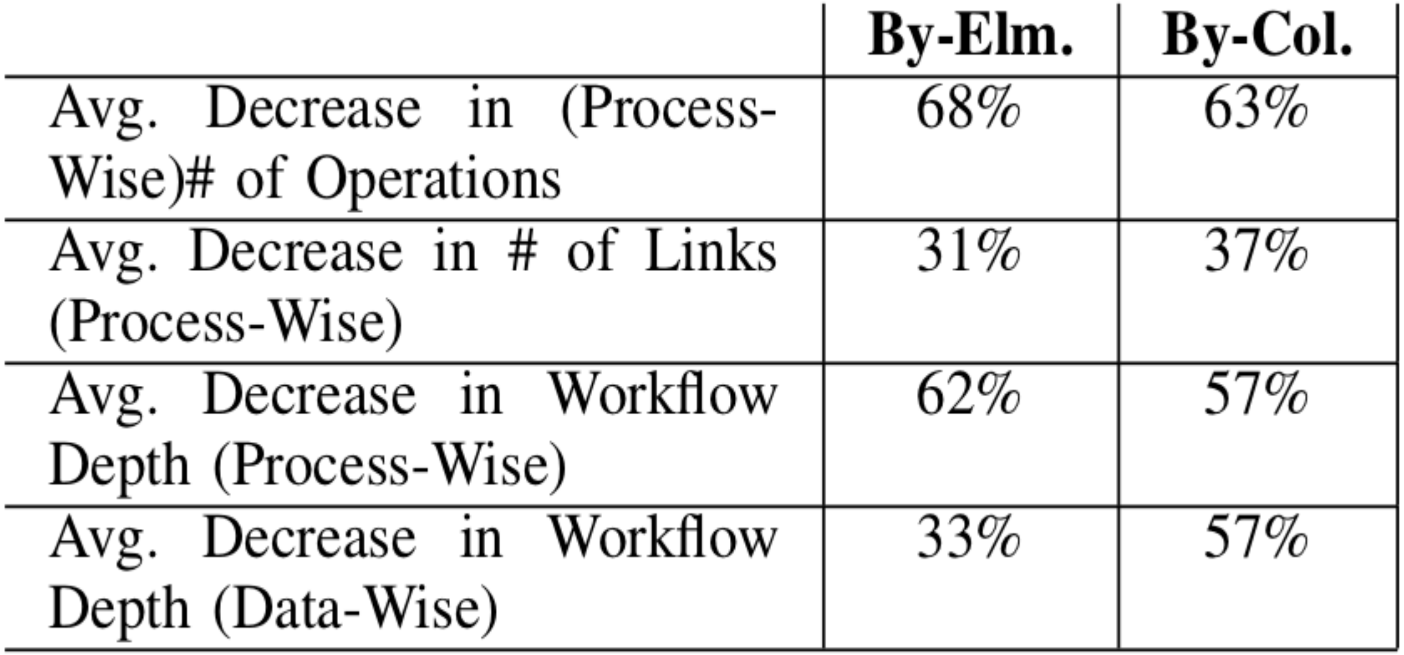


# By-Elimination

# Analysis Data Set

- 30 Workflows from the Taverna system
- Entire dataset & queries accessible from http://www.myexperiment.org/packs/467.html
- Manual Annotation using Motif Vocabulary

# Mechanistic Effect of Summarization



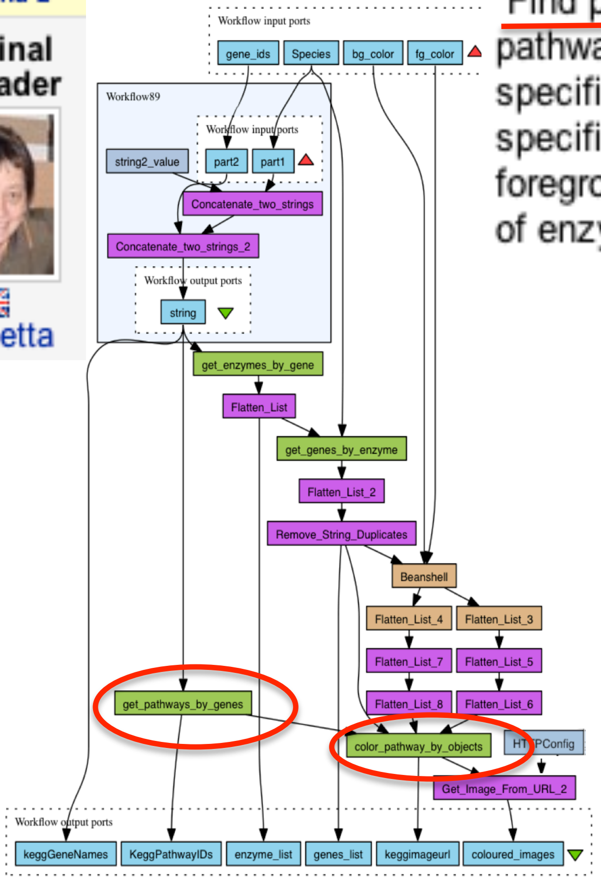
| | By-Elm. | By-Col. |
|---|---|---|
| Avg. Decrease in (Process-Wise)# of Operations | 68% | 63% |
| Avg. Decrease in # of Links (Process-Wise) | 31% | 37% |
| Avg. Decrease in Workflow Depth (Process-Wise) | 62% | 57% |
| Avg. Decrease in Workflow Depth (Data-Wise) | 33% | 57% |

# User Summaries vs. Summary Graphs

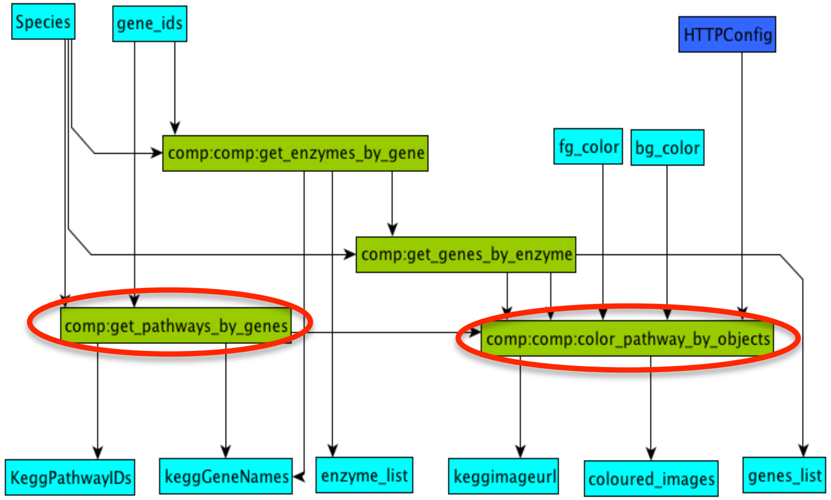
Taverna 2

Original Uploader

Jannetta



"Find pathways in which all the genes in the list are involved. For each pathway draw the pathway diagram. Color all enzyme boxes with colors specified. This workflow still has one problem. The list of colors have to be specified. I would like ideally to only except one background and one foreground color and expand that to a list with length equivalent to the number of enzymes found - just duplicating the specified colors."



| | By-Elm. Precision | By-Elm. Recall | By-Col. Precision | By-Col. Recall |
|---|---|---|---|---|
| Process-Wise | 0.74 | 0.92 | 0.65 | 0.93 |
| Data Wise | 0.14 | 0.55 | 0.33 | 0.43 |

# Highlights

- Research Object model and associated management tools
- Annotations of Workflow Using Motifs
- Methods for Summarizing Workflow and distilling their provenance traces
- Algorithms for Repairing Workflows

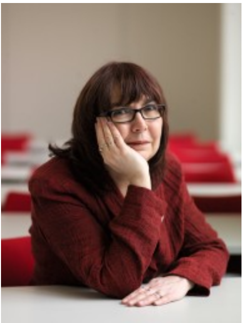
# Ongoing Work

- Validation of the workflow summarization
- Querying of Workflow Execution Provenance using summaries.

# References

- P Alper, K Belhajjame, C Goble, P Karagoz. Small Is Beautiful: Summarizing Scientific Workflows Using Semantic Annotations. IEEE International Congress on Big Data, 2013
- Pinar Alper, Khalid Belhajjame, Carole A. Goble, Pinar Karagoz: Enhancing and abstracting scientific workflow provenance for data publishing. EDBT/ICDT Workshops 2013.
- Belhajjame K, Corcho O, Garijo D, et al. Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse. In proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012), Heraklion, Greece, May 2012
- Belhajjame K, Zhao J, Garijo D, et al. The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web, submitted to the Journal of Web Semanics.
- Daniel Garijo, Pinar Alper, Khalid Belhajjame, Óscar Corcho, Yolanda Gil, Carole A. Goble: Common motifs in scientific workflows: An empirical analysis. eScience 2012.
- Zhao J, Gómez-Pérez JM, Belhajjame K, Klyne G, et al. Why workflows break - Understanding and combating decay in Taverna workflows. IEEE eScience 2012.

# Acknowledgement



# Acknowledgement

EU Wf4Ever project (270129)
funded under EU FP7 (ICT- 2009.4.1).
(http://www.wf4ever-project.org)