From Scientific Workflows to Research Objects: Publication and Abstraction of Scientific Experiments

Daniel Garijo Verdejo

Supervisors: Oscar Corcho, Yolanda Gil

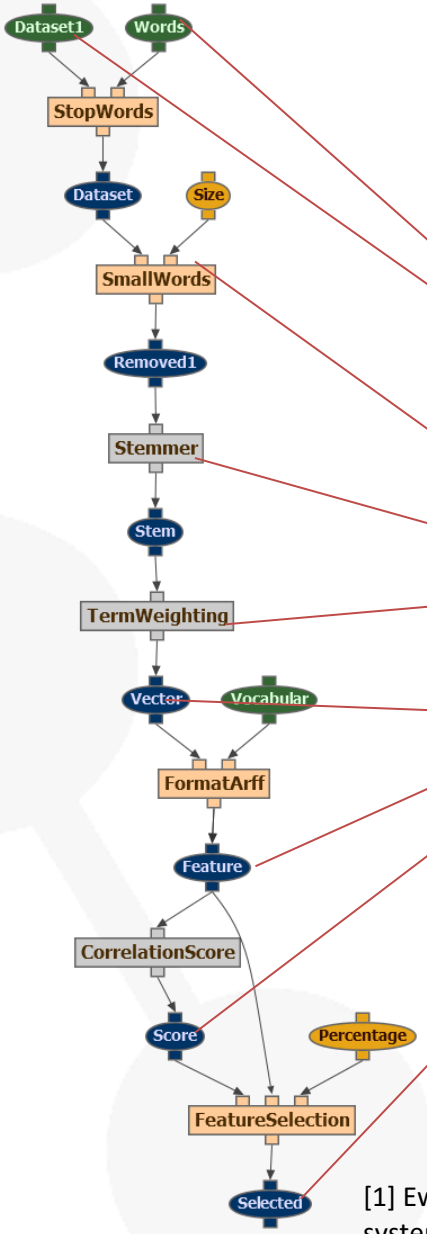
Ontology Engineering Group
Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid

Date: 27/02/2014

Index

1. Background
2. What do I do?
3. Motivation
4. Overview
5. Representing and publishing scientific workflows in the Web
   - Linked Data
   - Templates and provenance traces
   - Standards
6. Common motifs among scientific workflows
   - Workflow motif catalog
7. Detecting common fragments among scientific workflows
8. Workflows as part of an experiment: Research Objects

What are Scientific Workflows?



- "Template defining **the set of tasks needed to carry out a computational experiment**" [1]
- Inputs
- Steps
- Intermediate results
- Outputs
- **Data driven**, usually represented as **Directed Acyclic Graphs** (DAGs)

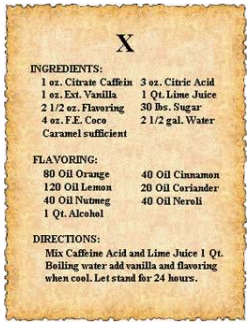
[1] Ewa Deelman, Dennis Gannon, Matthew Shields, Ian Taylor, Workflows and e-science: an overview of workflow system features and capabilities, Future Generation Computer Systems 25 (5) (2009) 528–540.

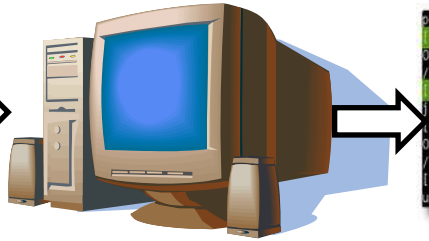
How are scientific workflows created?

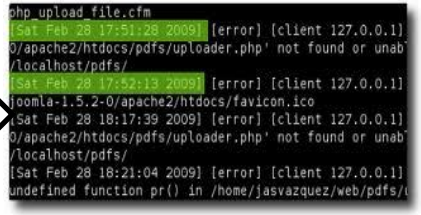
Experiment

Lab book

Laboratory Protocol (recipe)

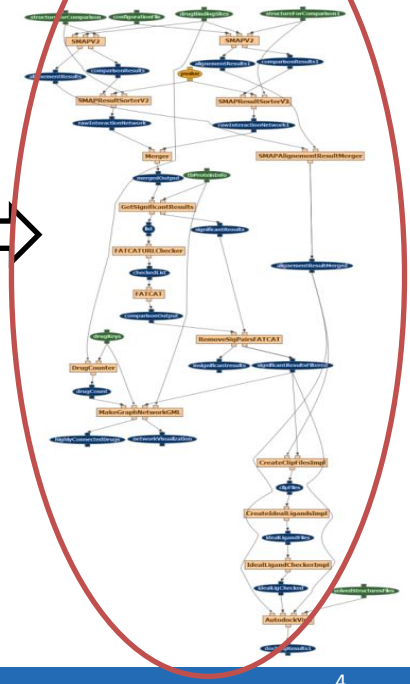






Digital Log

Workflow



- Similar to Laboratory Protocols

What am I working on?

- Workflow representation
  - Plan/template representation
  - Provenance trace representation
  - Link between templates and traces

CH1: Can we export an abstract template of the method being represented?
CH2: How do we interoperate with other workflow results?
CH3: How do we access the workflow results?
CH4: How do we link an abstract method with several implementations?

- Creation of abstractions/motifs in scientific workflows
  - Abstraction catalog
  - Find how different workflows are related

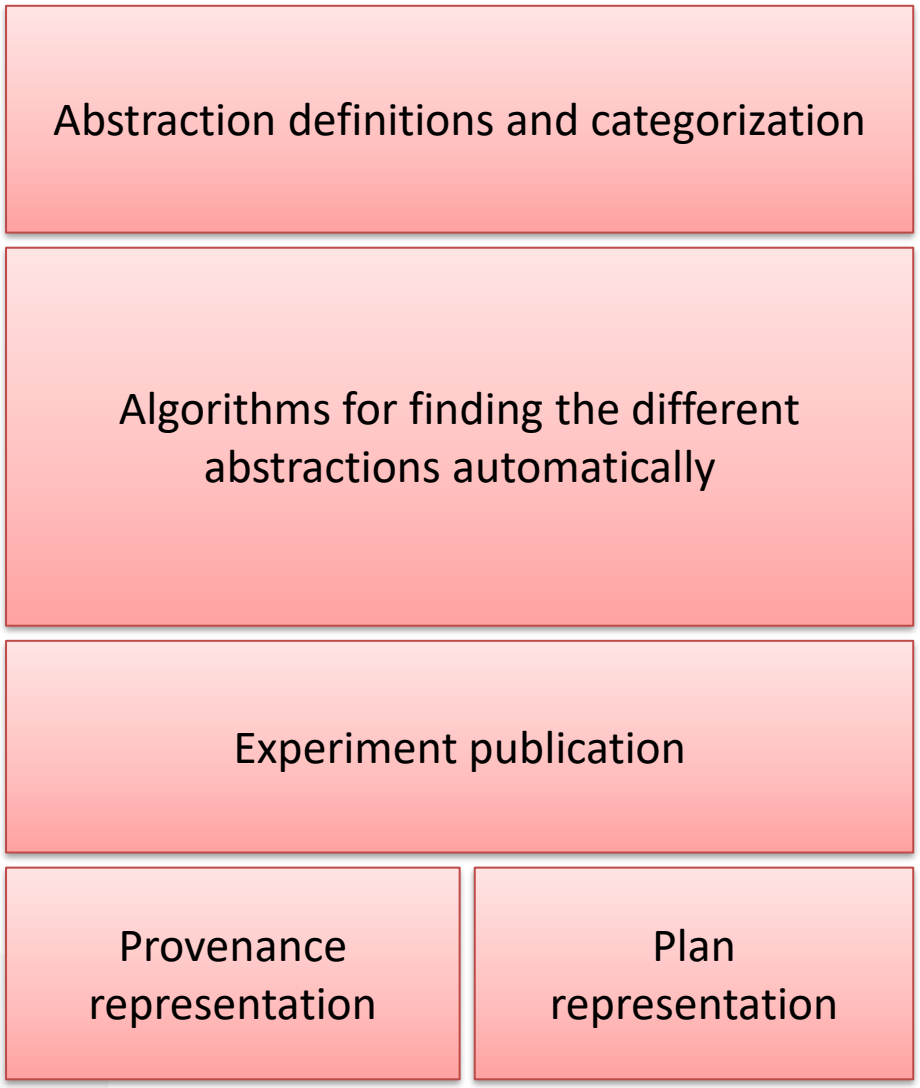
CH5: How can we detect what are the typical operations in scientific workflows?
CH6: How can we detect them automatically?

- Understandability and reuse of scientific workflows
  - Relation between the workflows involved in the same experiment (Research Objects)

CH7: Which workflow parts are related to other workflows?
CH8: How do workflows depend on the other parts of the experiments?

- As a designer: Discovery
  - Workflows with similar functionality fragments/methods
  - Design based in previous templates.
- As user/reuser: Understandability, Exploration
  - Search workflows by functionality
  - Commonalities between execution runs
  - Component categorization

Overview



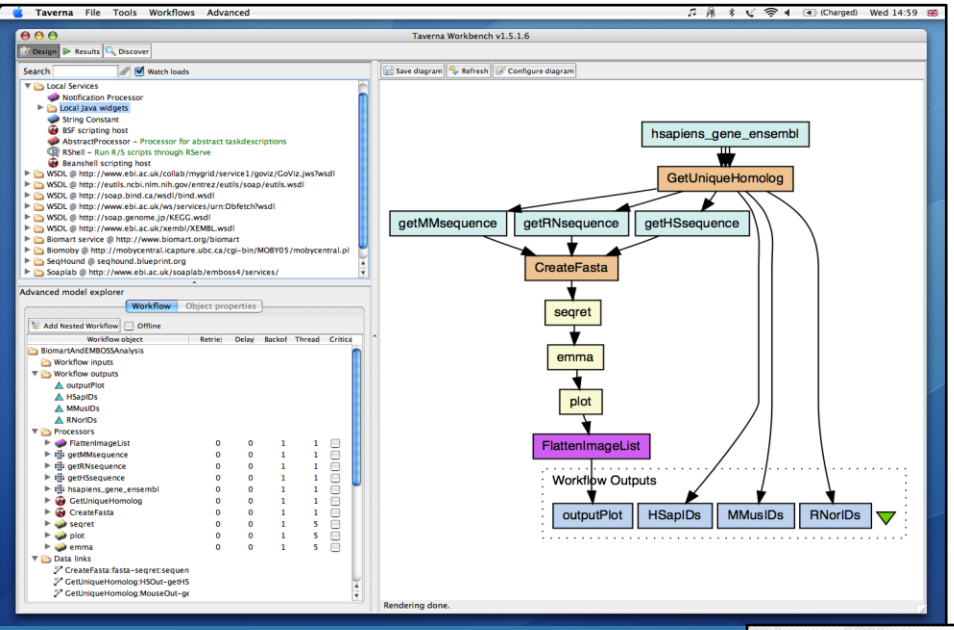
Descriptions/ PSMs/Ontologies

Data mining tools, graph analysis, etc.

RDF Stores

Vocabularies

Taverna and Wings

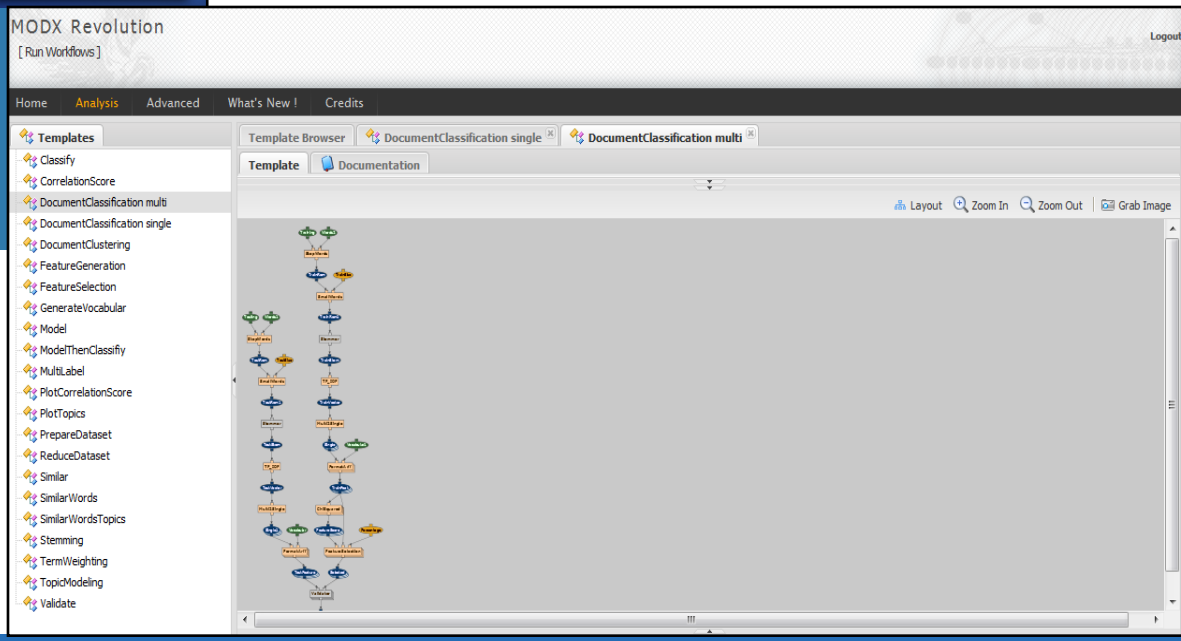


Taverna

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

WINGS

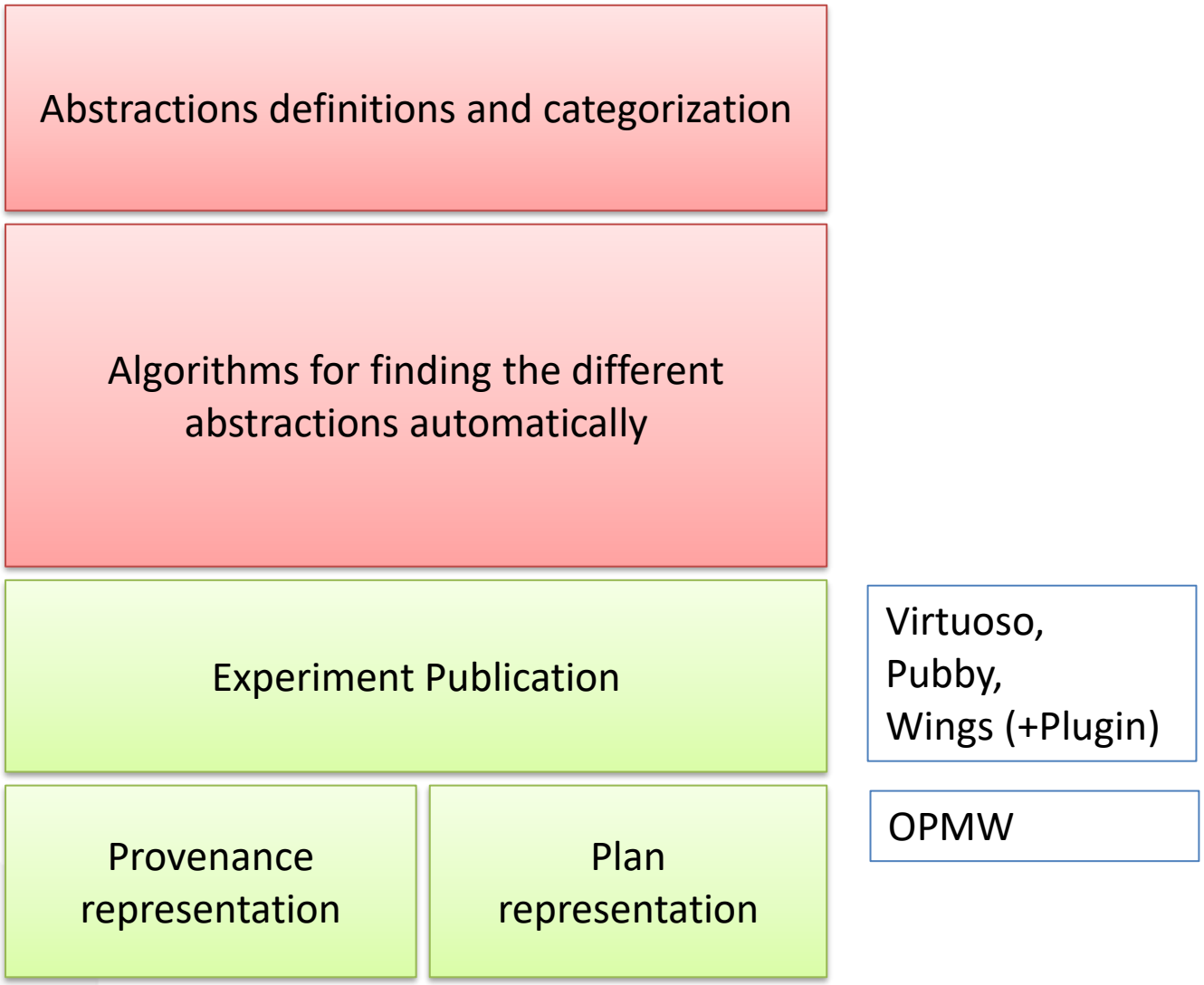
http://www.wings-workflows.org/



Representing and publishing scientific workflows in the Web

A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data. Garijo, D.; and Gil, Y. In *Proceedings of the 6th workshop on Workflows in support of large-scale science*, page 47-56, Seattle, 2011. ACM

Overview


Abstractions definitions and categorization
Algorithms for finding the different abstractions automatically
Experiment Publication
Provenance representation
Plan representation
Virtuoso,
Pubby,
Wings (+Plugin)
OPMW

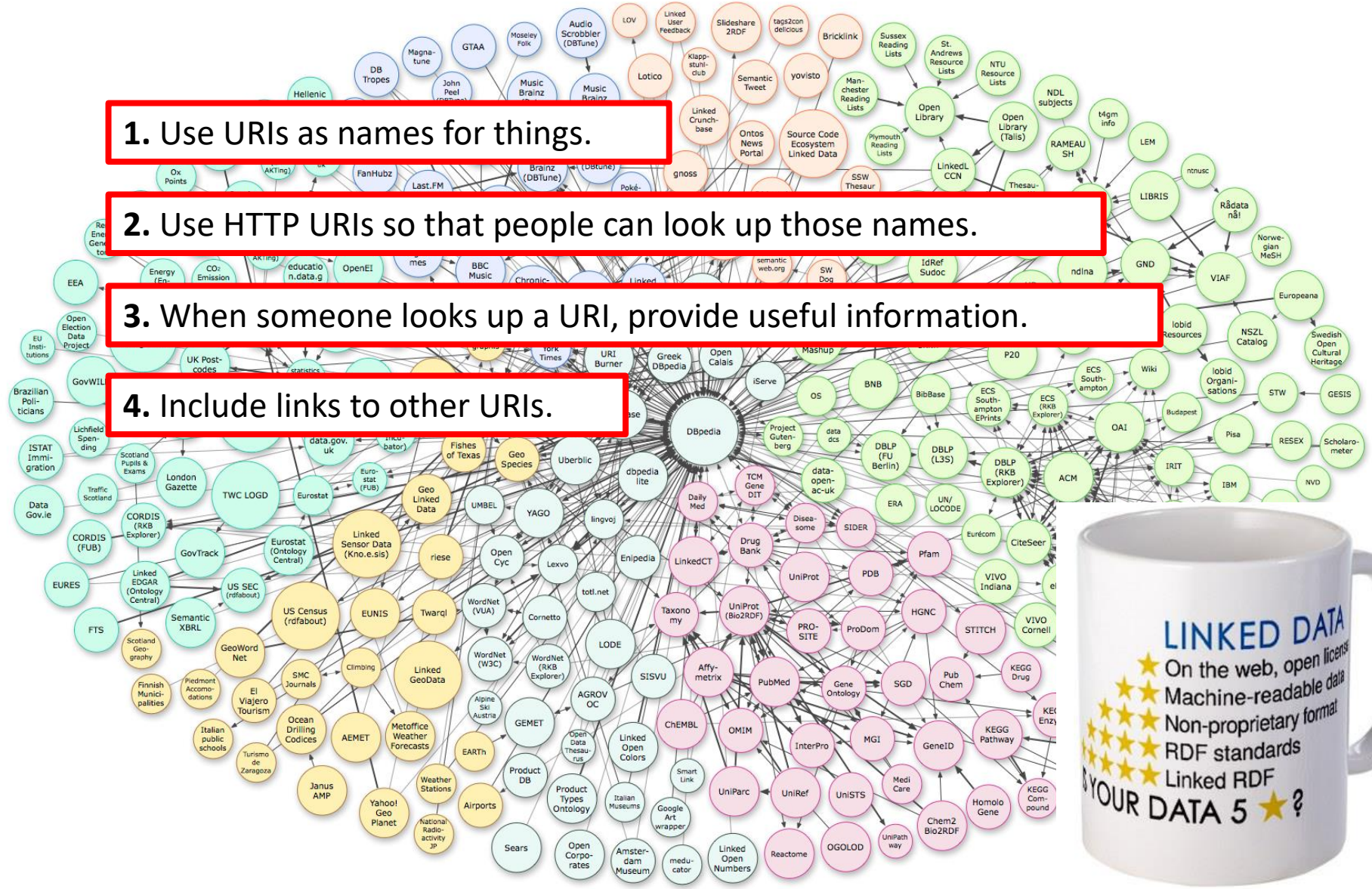
What is Linked Data?

1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

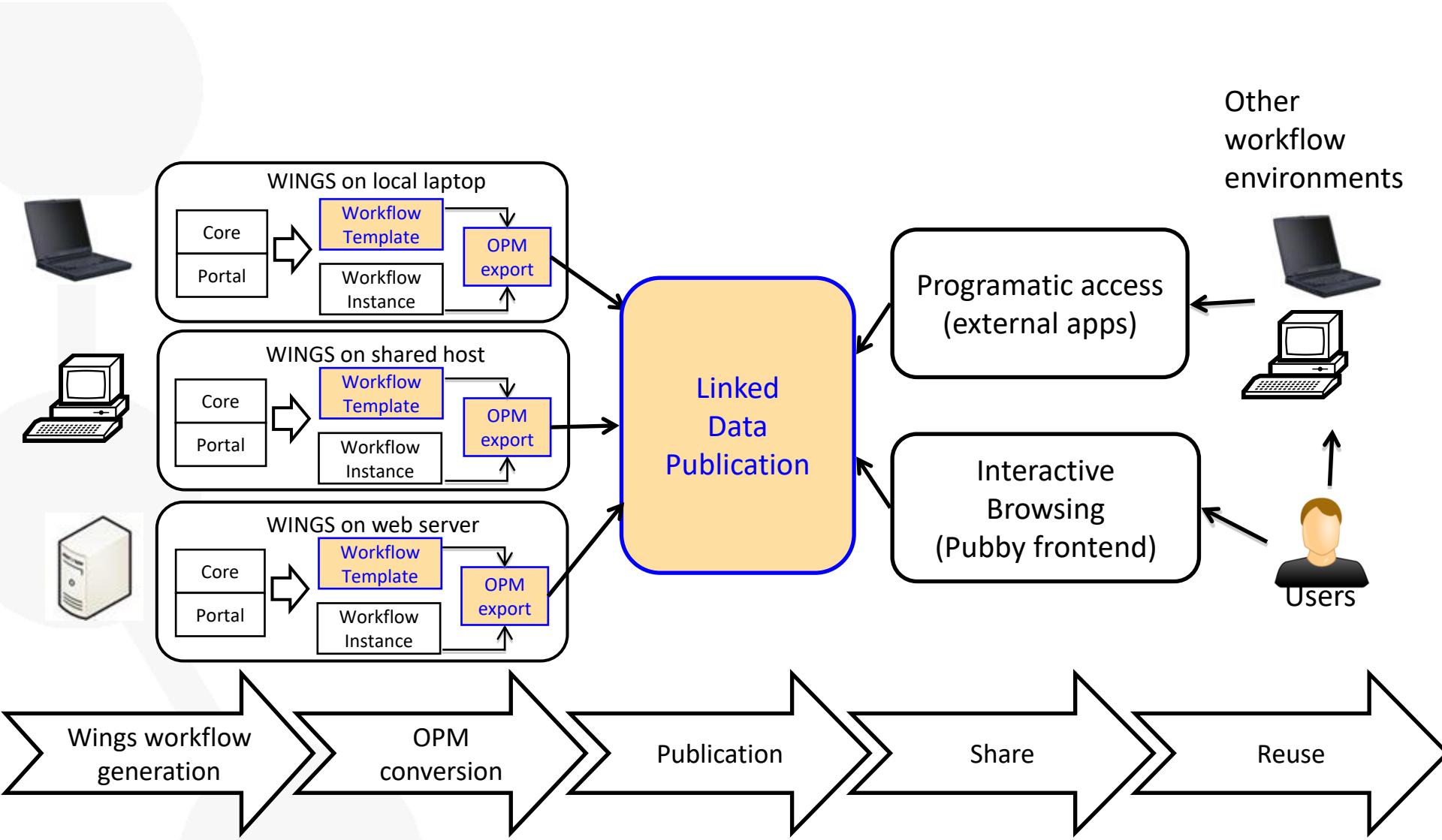
3. When someone looks up a URI, provide useful information.

4. Include links to other URIs.

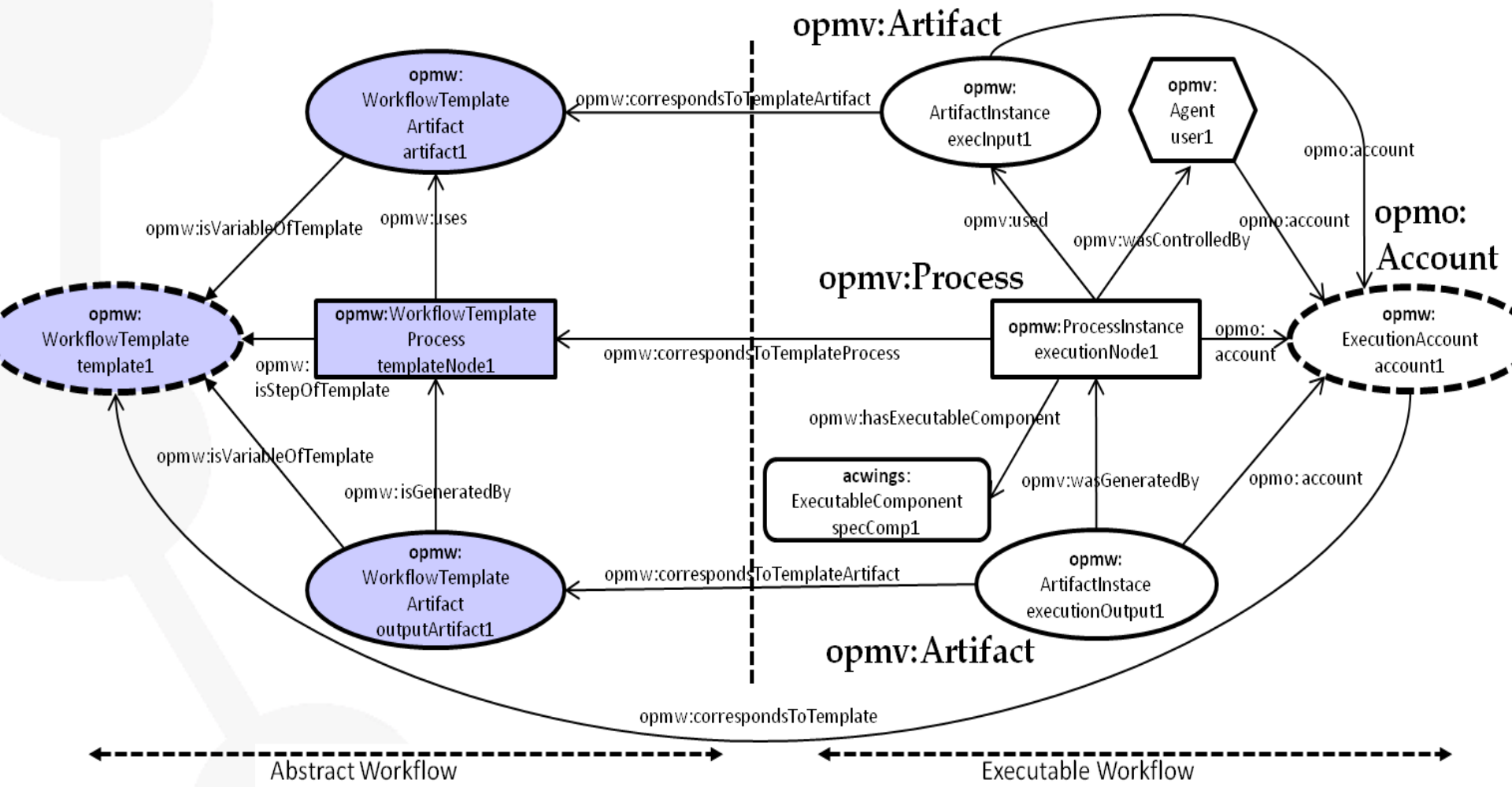


"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Publishing workflows: high level architecture

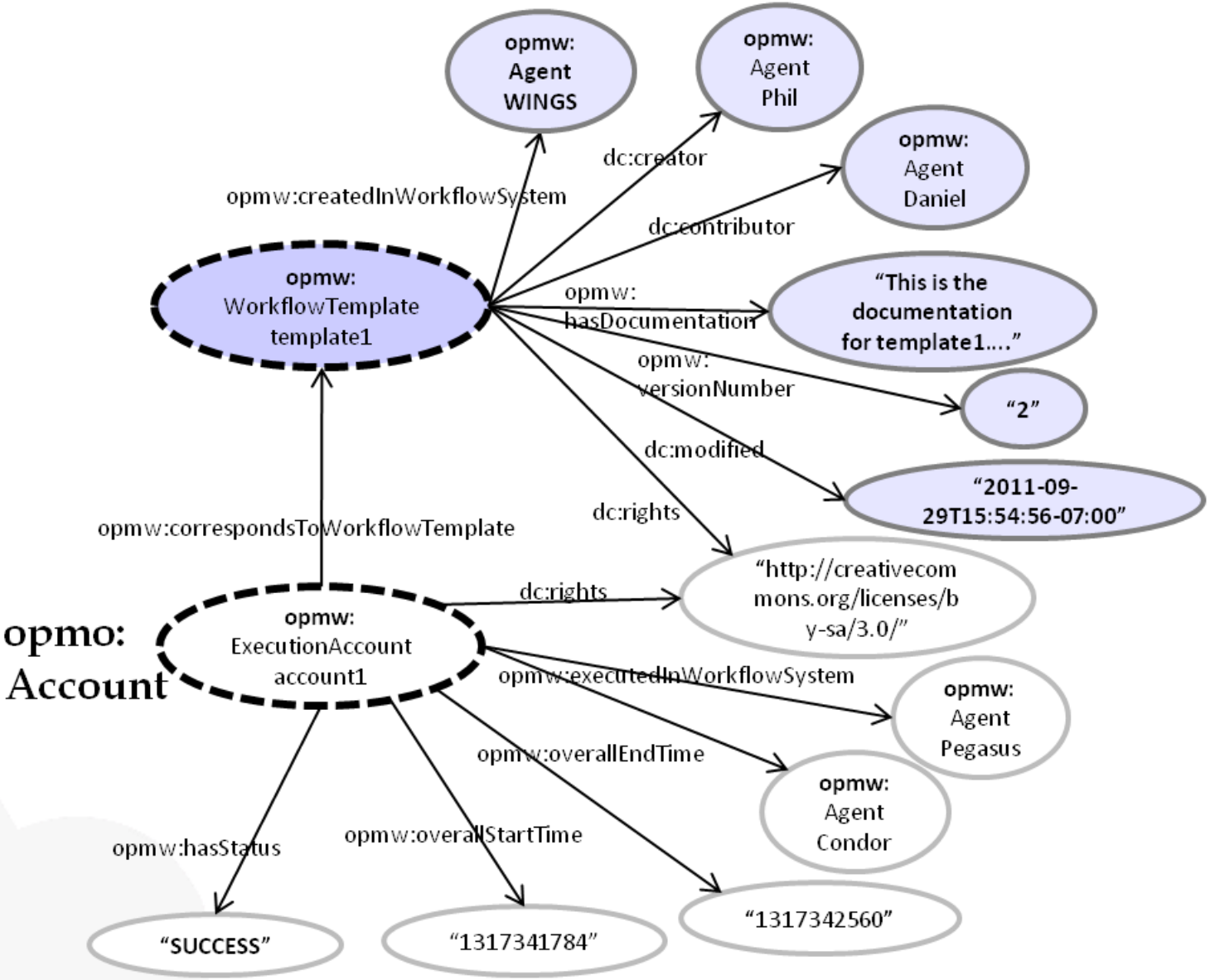


OPMW: Process view



http://www.opmw.org/ontology/

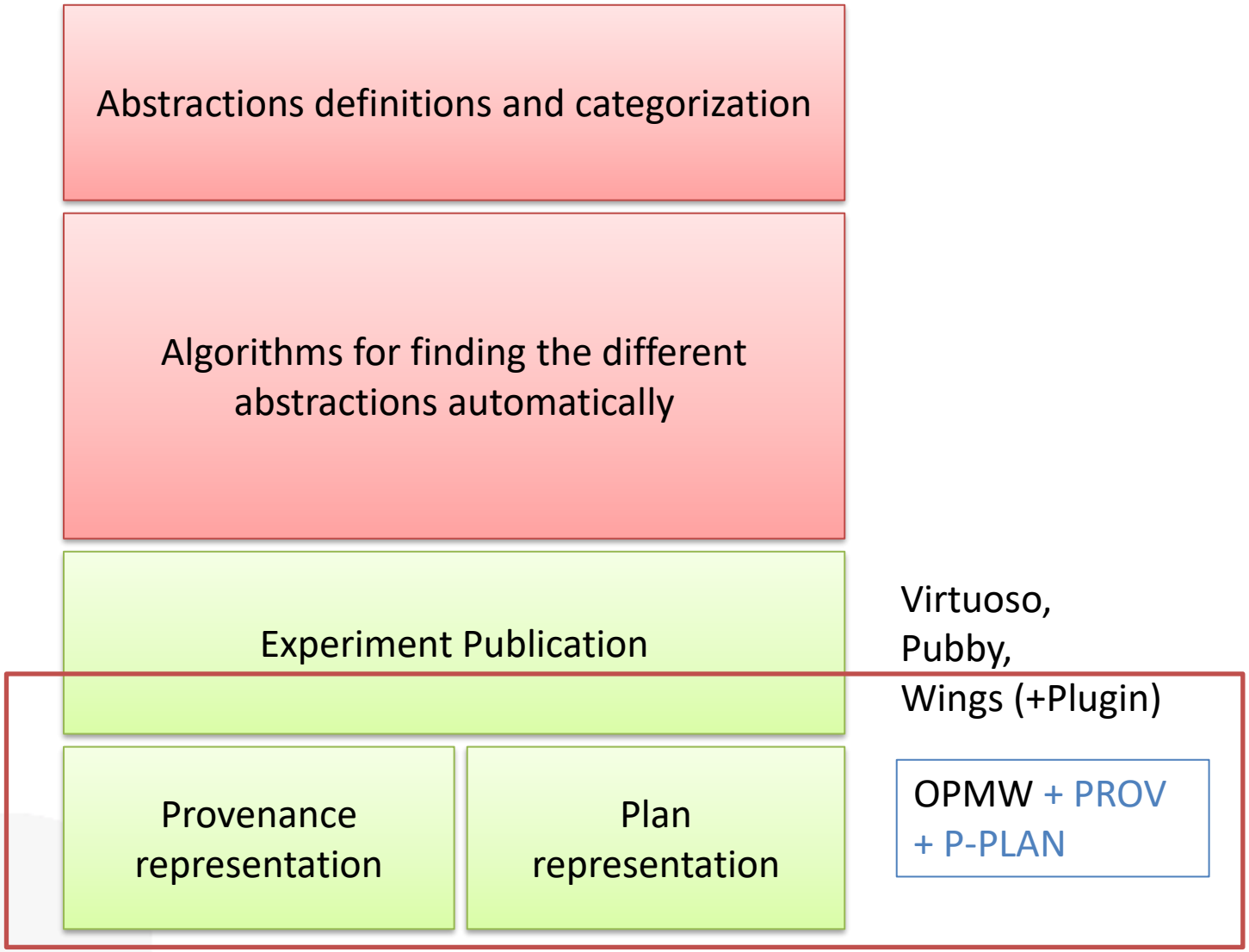
OPMW: Attribution view

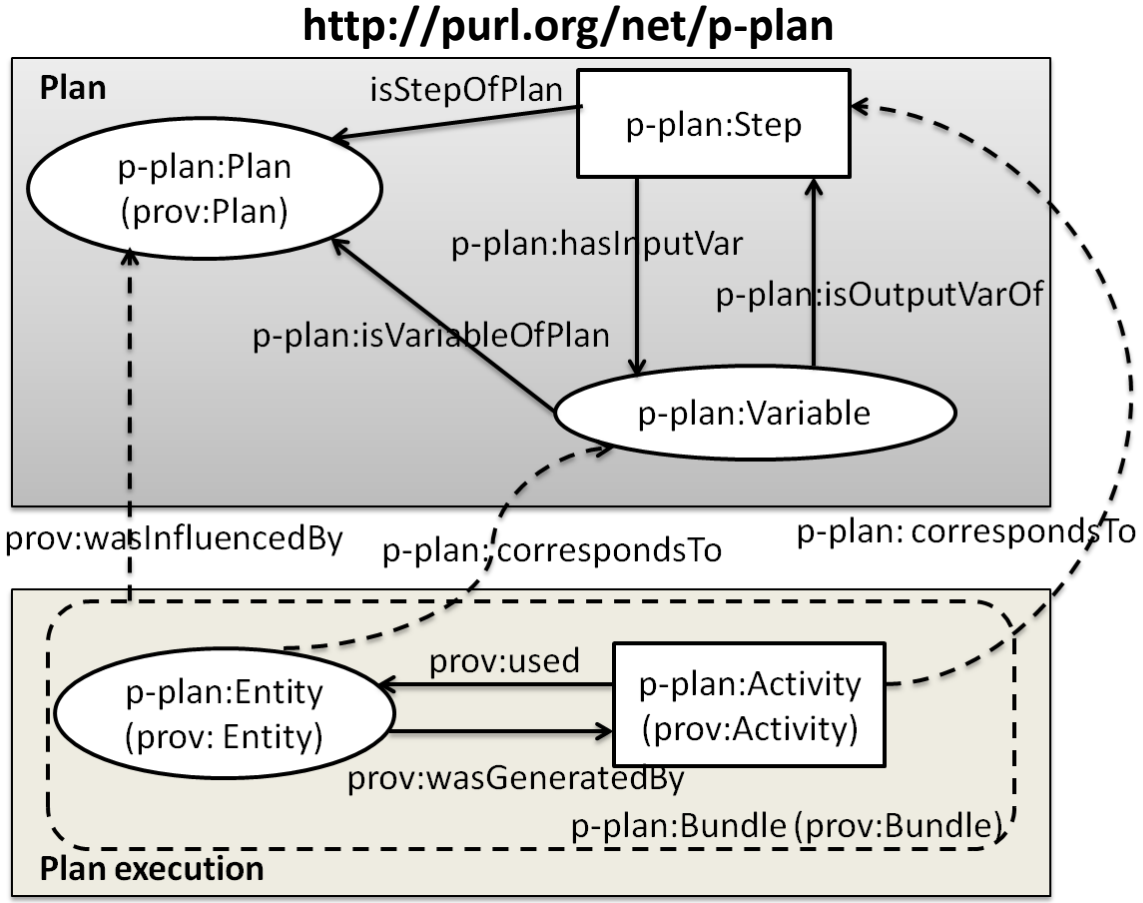


Standards

PROV-O: The PROV Ontology. Lebo, T.; Sahoo, S.; McGuinness, D.; Belhajjame, K.; Corsar, D.; Cheney, J.; Garijo, D.; Soiland-Reyes, S.; Zednik, S.; and Zhao, J. W3C Consortium. 2012.

Overview


Abstractions definitions and categorization
Algorithms for finding the different abstractions automatically
Experiment Publication
Provenance representation
Plan representation
Virtuoso,
Pubby,
Wings (+Plugin)
OPMW + PROV + P-PLAN



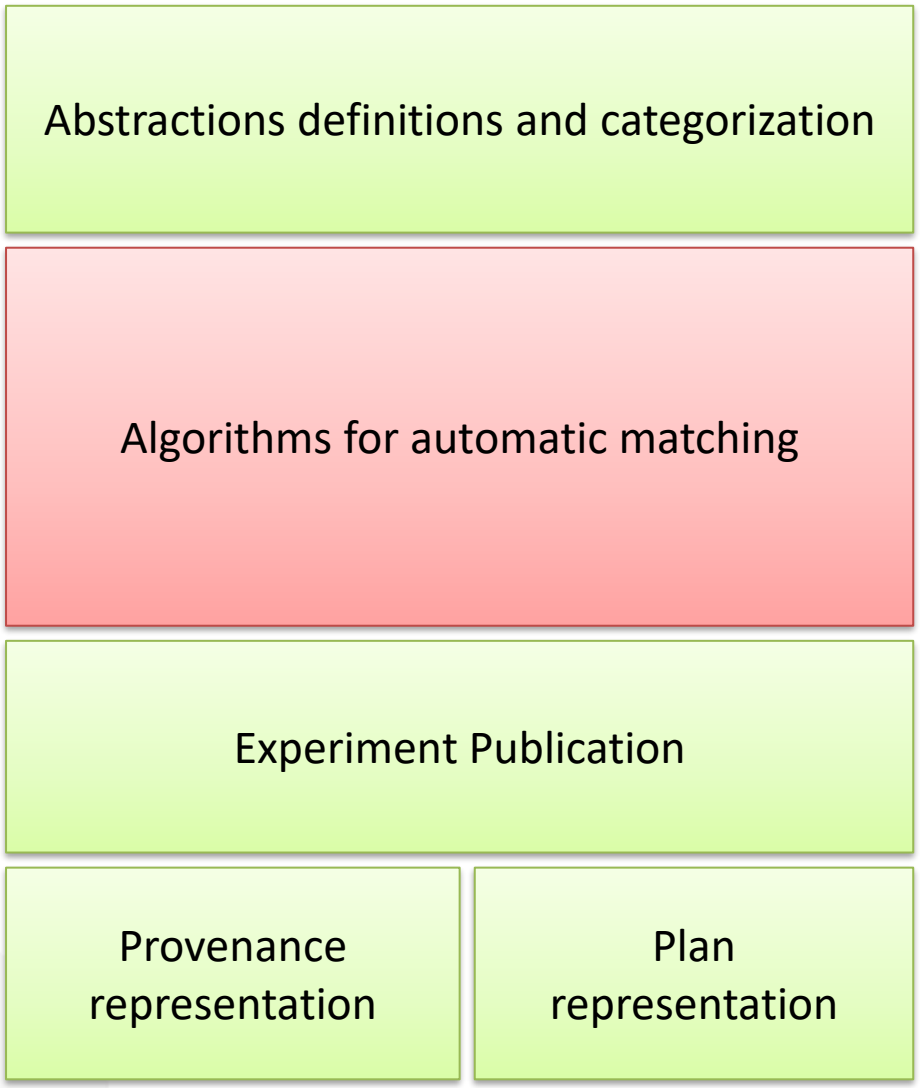
- Plans are not provenance
- P-PLAN: Simple plan model for binding traces to template representations
- Aligned with OPMW and PROV (W3C Provenance Standard)

Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data. Garijo, D.; and Gil, Y. In *Proceedings of the 2nd International Workshop on Linked Science 2012*, Boston, 2012.

Common motifs among scientific workflows

Common motifs in scientific workflows: An empirical analysis. Garijo, D.; Alper, P.; Belhajjame, K.; Corcho, O.; Gil, Y.; and Goble, C. *Future Generation Computer Systems*, . 2013

Overview



Motif Detection

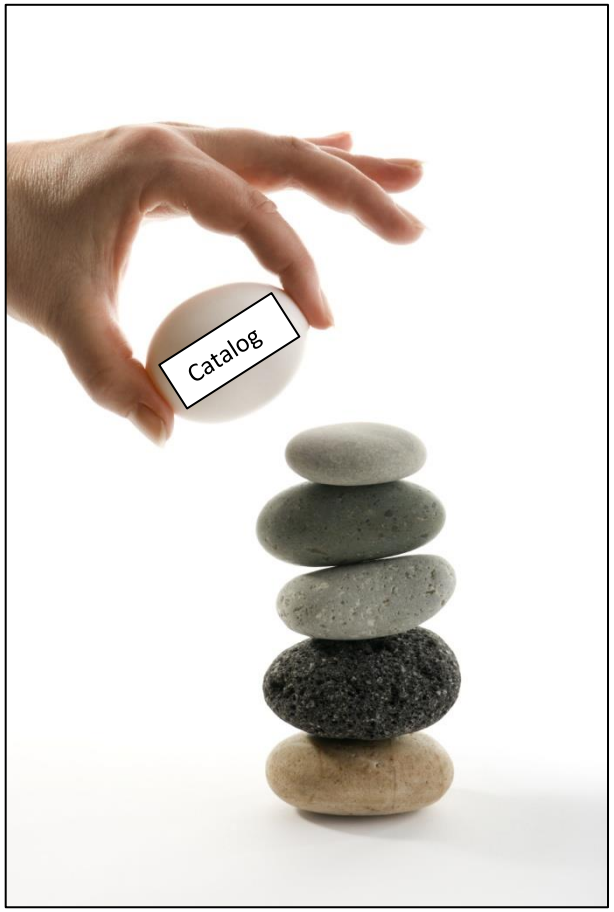
Virtuoso,
Pubby,
Wings (+Plugin)

OPMW

- Empirical analysis on 260 workflow templates from Taverna, Wings, Galaxy and Vistrails

- Catalog of recurring patterns: scientific workflow *motifs*.
  - Data Oriented Motifs
  - Workflow Oriented Motifs

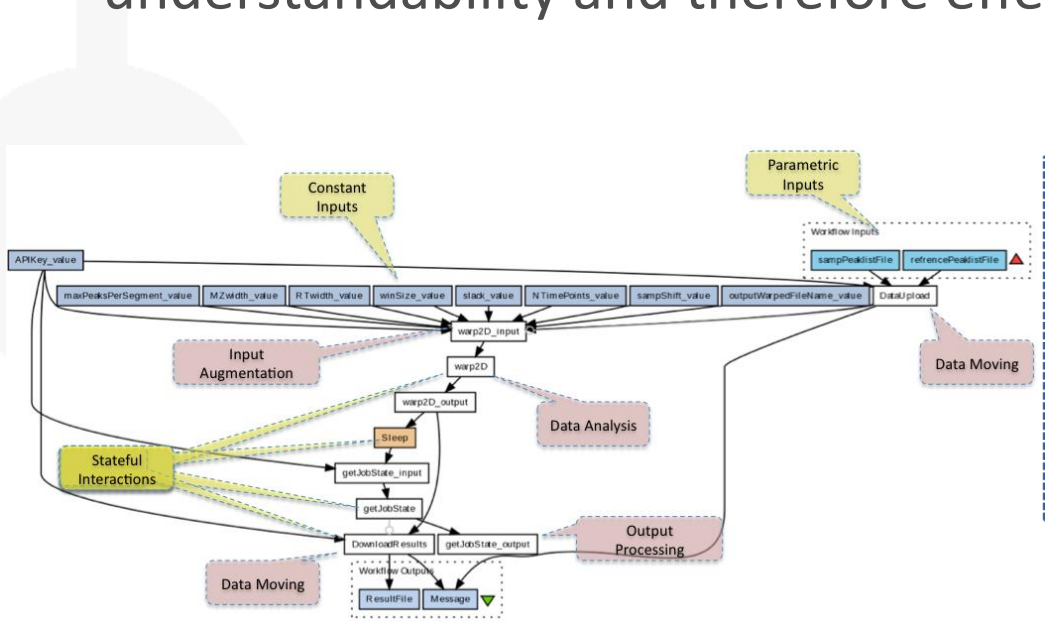
- Understandability and reuse

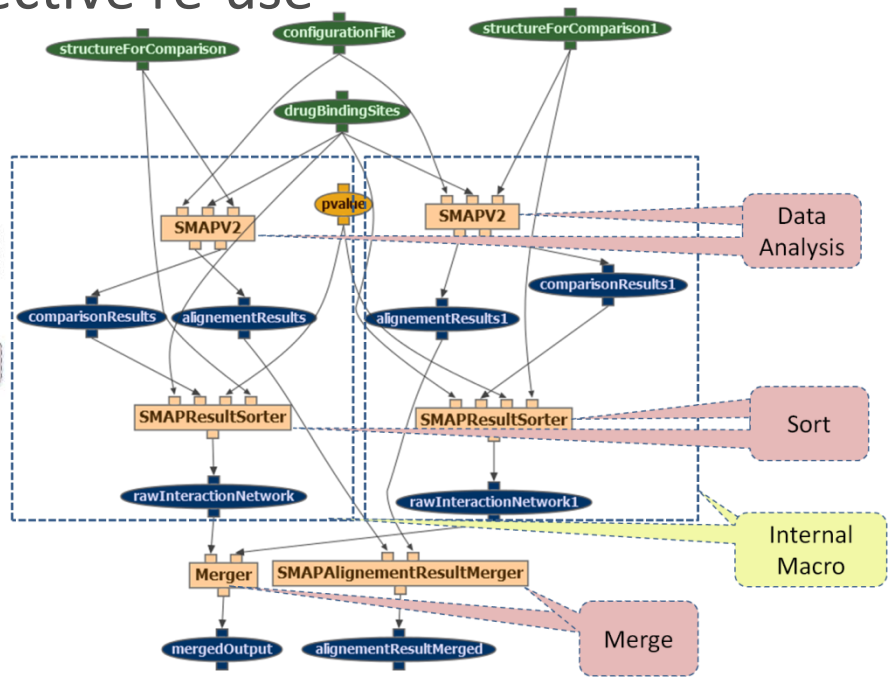


http://sensefinancial.com/wp-content/uploads/2012/02/contribution.jpg

- Reverse-engineer the set of current practices in workflow development through an analysis of empirical evidence

- Identify workflow abstractions that would facilitate understandability and therefore effective re-use





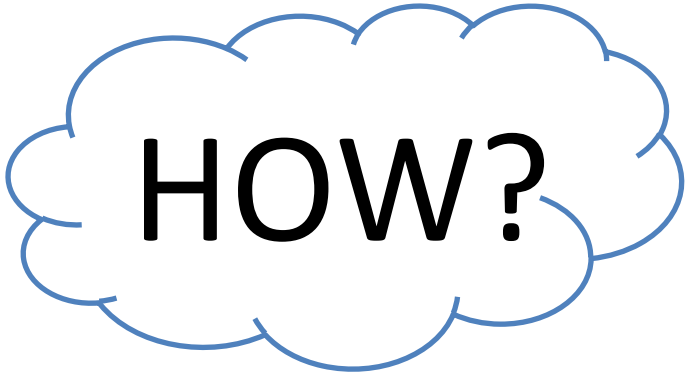
•Workflow motif: Domain independent conceptual abstraction on the workflow steps.

1. Data-oriented motifs: What kind of manipulations does the workflow have?

   •E.g.:
   - •Data retrieval
   - •Data preparation
   - • etc.

2. Workflow-oriented motifs: How does the workflow perform its operations?

   •E.g.:
   - •Stateful steps
   - •Stateless steps
   - •Human interactions
   - •etc.



Data-Oriented Motifs

- Data Retrieval
- Data Preparation
  - Format Transformation
  - Input Augmentation and Output Splitting
  - Data Organisation
- Data Analysis
- Data Curation/Cleaning
- Data Moving
- Data Visualisation

Workflow-Oriented Motifs

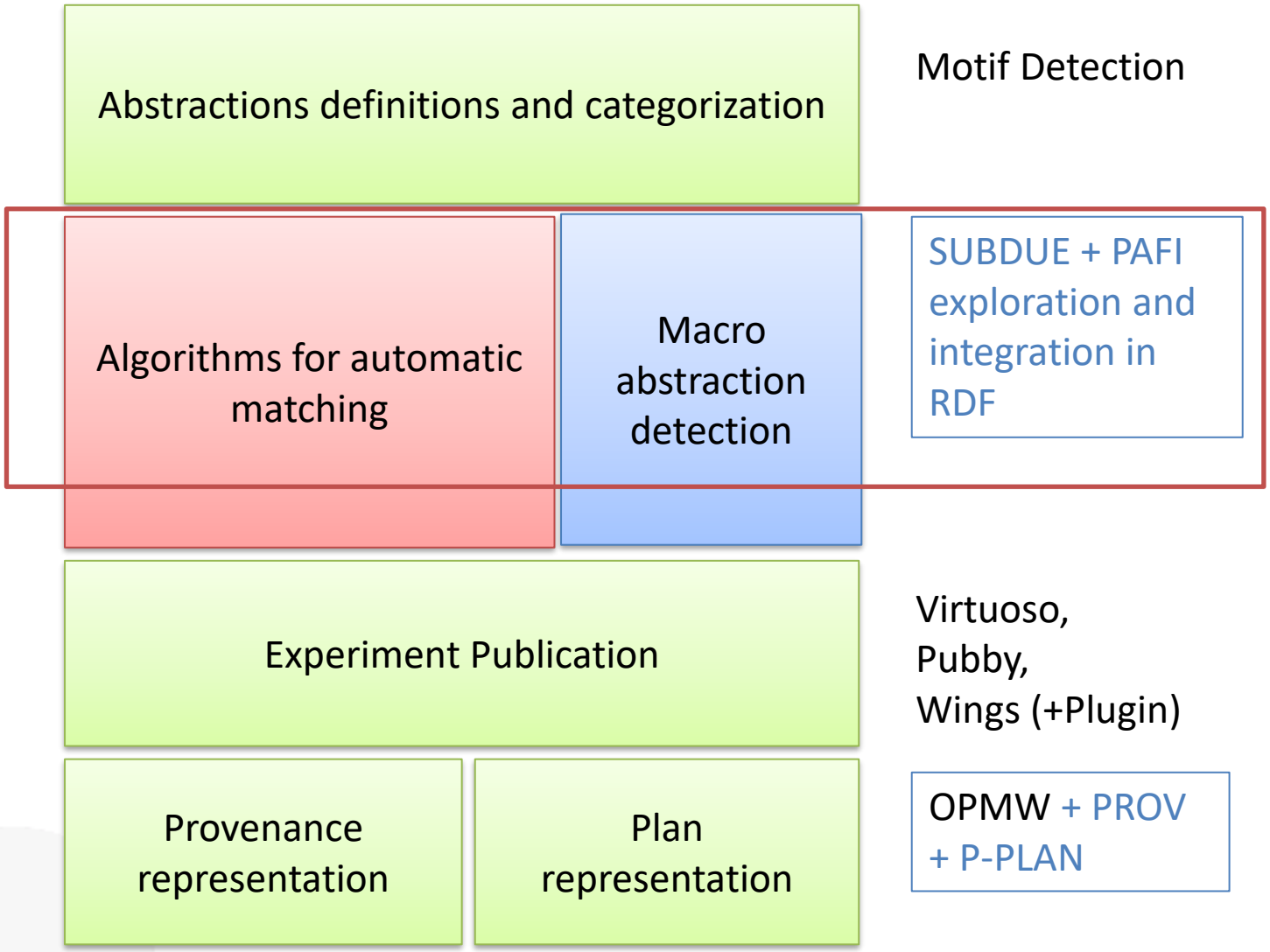
- Intra-Workflow Motifs
  - Stateful (Asynchronous) Invocations
  - Stateless (Synchronous) Invocations
  - Internal Macros
  - Human Interactions
- Inter-Workflow Motifs
  - Atomic Workflows
  - Composite Workflows
  - Workflow Overloading

Ontology Purl: **http://purl.org/net/wf-motifs**

Detecting common fragments among scientific workflows (macro motifs)

Detecting common scientific workflow fragments using execution provenance. Garijo, D.; Corcho, O.; and Gil, Y. In *Proceedings of of the seventh international conference on Knowledge capture*, page 33-40, Banff, 2013. ACM.

Summary: Work done at ISI



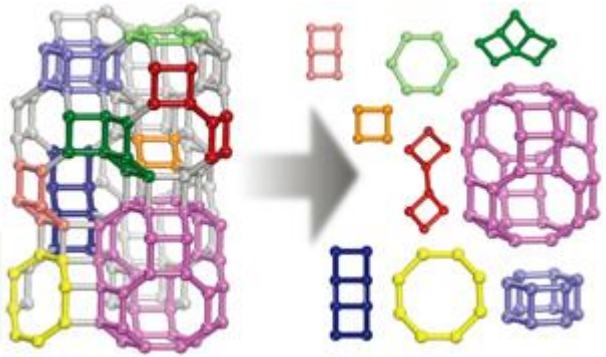
Problem statement:

Given a ***repository of workflow templates (either abstract or specific) or workflow execution traces, what are the workflow fragments I can deduce from it?***

Useful for:

- Systems like Taverna and Wings: (Many templates, little annotation to relate them)
  - Finding relationships between workflows and sub-workflows.
    - Most used fragments, most executed, etc.

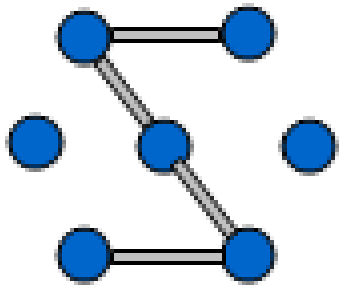
- Systems like GenePattern and Galaxy: (Many runs, nearly no templates published)
  - Proposing new templates with the popular fragments.



Challenges: Common workflow fragment detection

- Given a collection of workflows, which are the most common fragments?
  - Common sub-graphs among the collection
    - Sub-graph isomorphism (NP-complete)

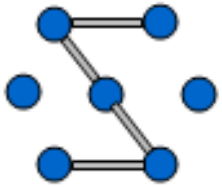
- We use the SUBDUE algorithm **[Holder et al 1994]** (hierachical clustering)
  - Graph Grammar learning
    - The rules of the grammar are the workflow fragments

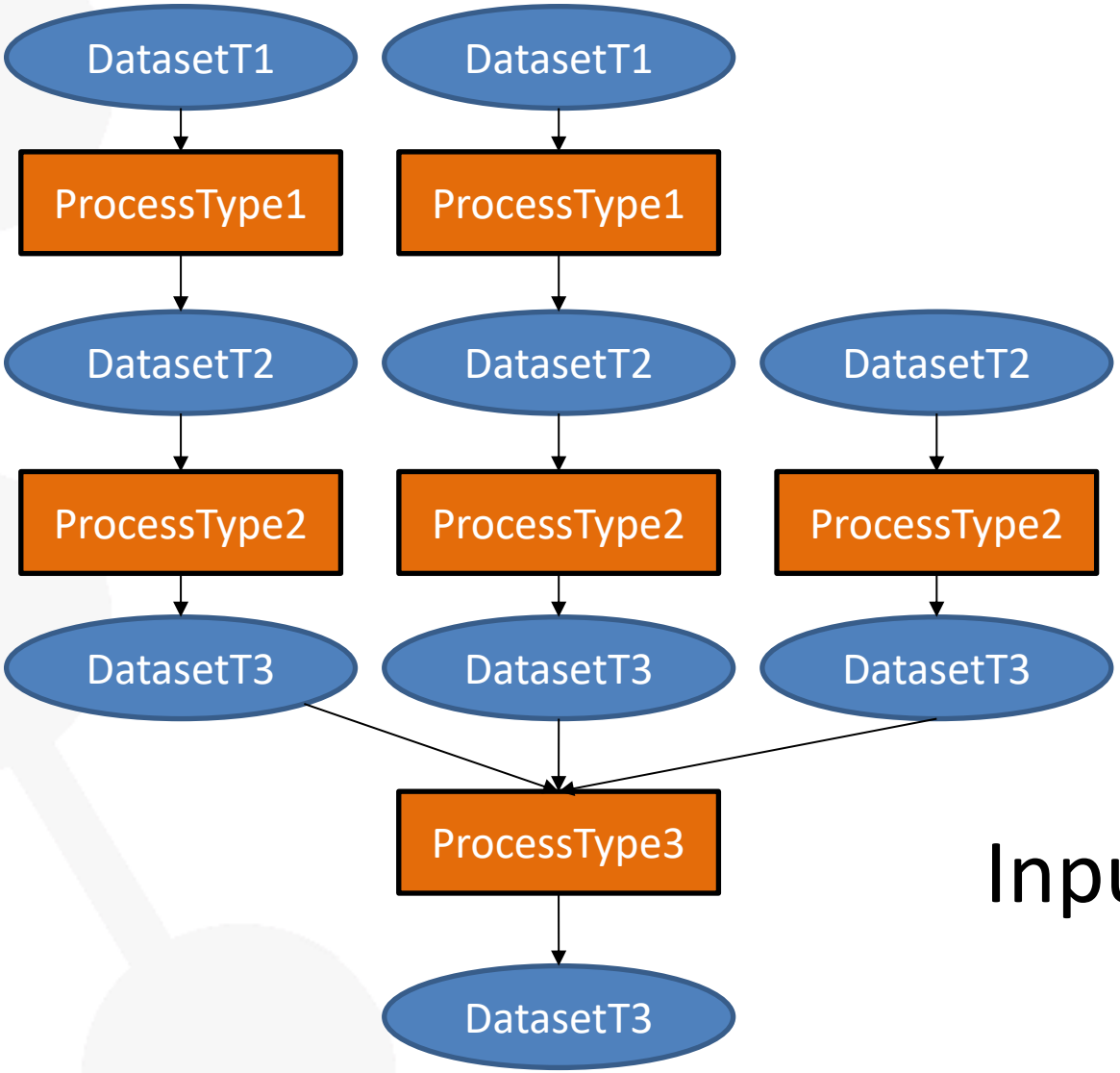


  - Graph based hierarchical clustering
    - Each cluster corresponds to a workflow fragment
  - Iterative algorithm with two measures for compressing the graph:
    - Minimum Description Length (MDL)
    - Size

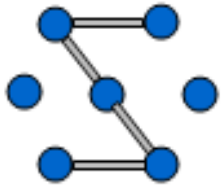
- Current tests with PAFI (http://glaros.dtc.umn.edu/gkhome/pafi/overview) ongoing.

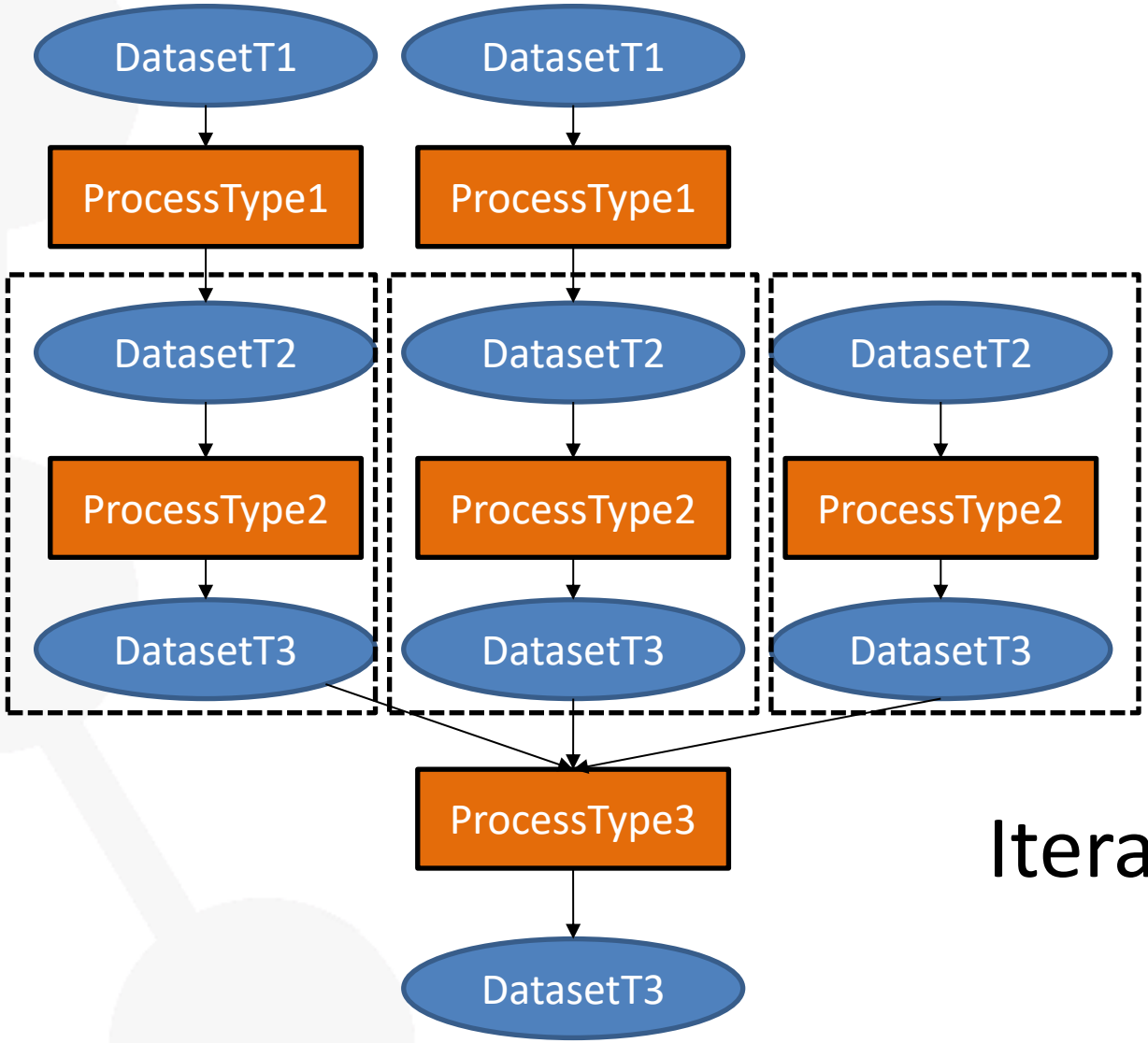
[Holder et al 1994]: **Substructure Discovery in the SUBDUE System** L. B. Holder, D. J. Cook, and S. Djoko. AAAI Workshop on Knowledge Discovery, pages 169-180, 1994.




DatasetT1
DatasetT1
ProcessType1
ProcessType1
DatasetT2
DatasetT2
DatasetT2
ProcessType2
ProcessType2
ProcessType2
DatasetT3
DatasetT3
DatasetT3
ProcessType3
DatasetT3

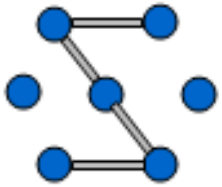
Input Graph

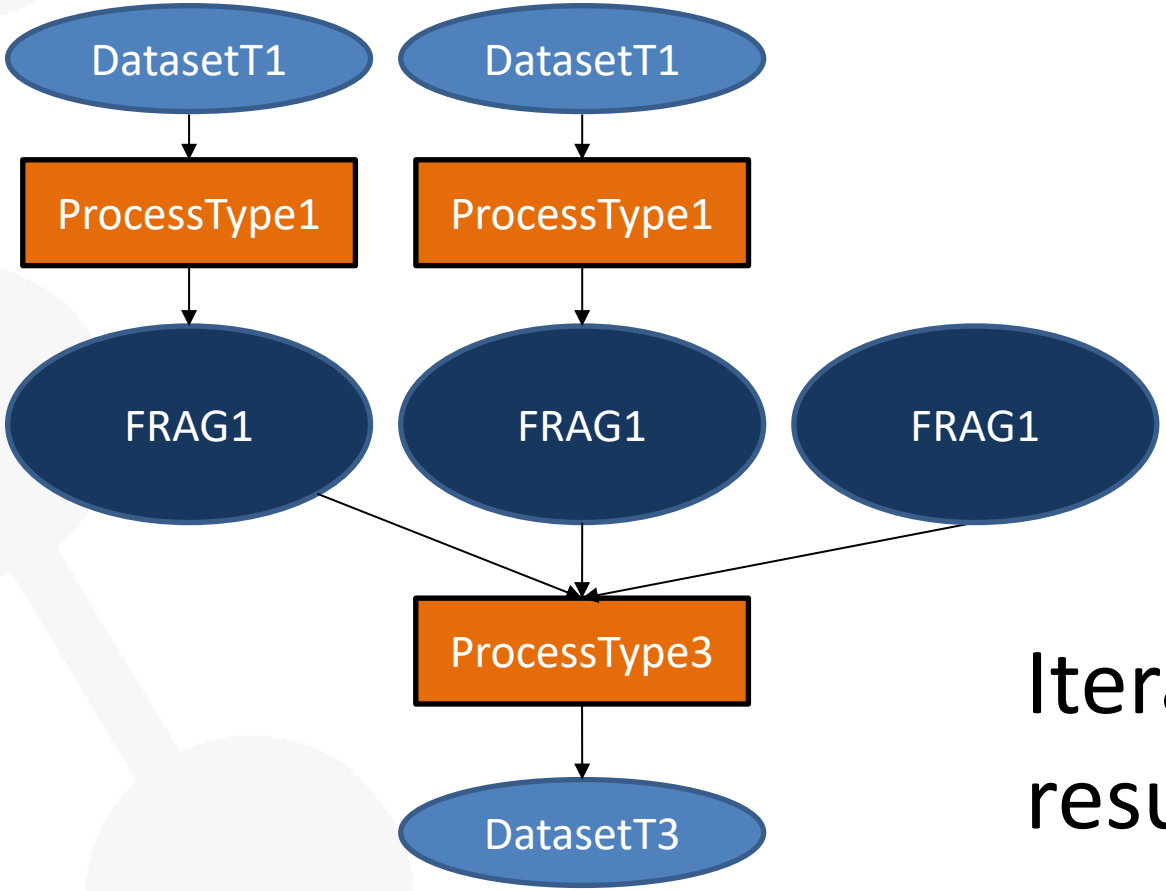




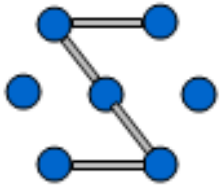
Fragment1

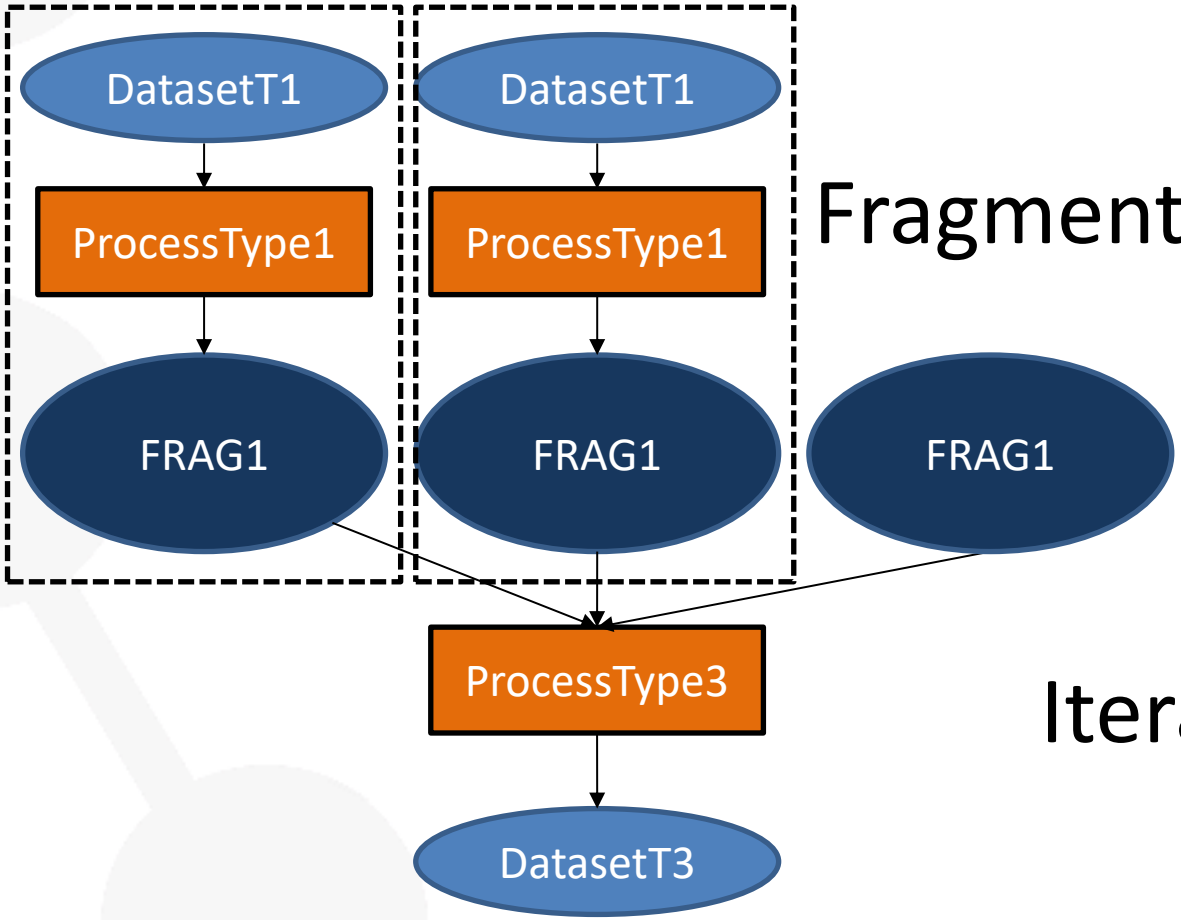
Iteration 1

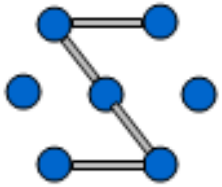


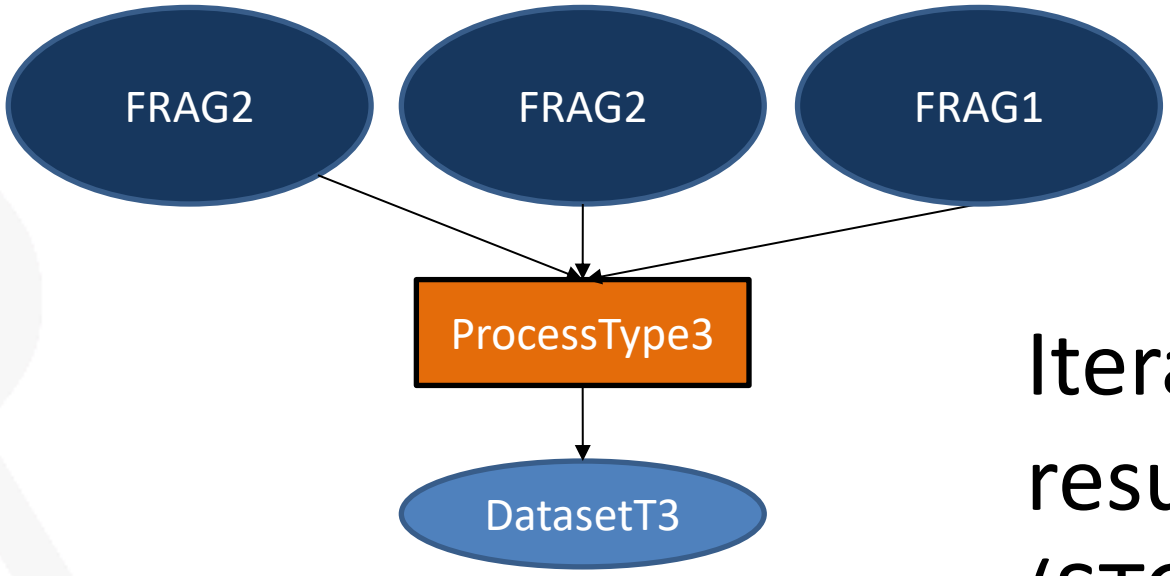


Iteration 1 result

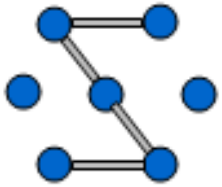



DatasetT1
DatasetT1
ProcessType1
ProcessType1
Fragment2
FRAG1
FRAG1
FRAG1
ProcessType3
Iteration 2
DatasetT3



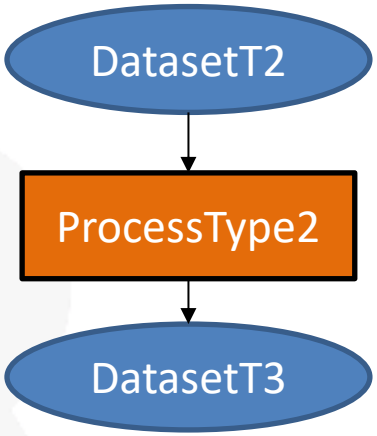


Iteration 2 result (STOP)

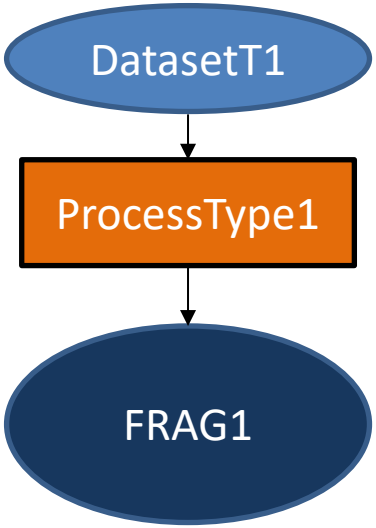


Results:

Fragment 1 (FRAG1) :



Fragment 2 (FRAG2):

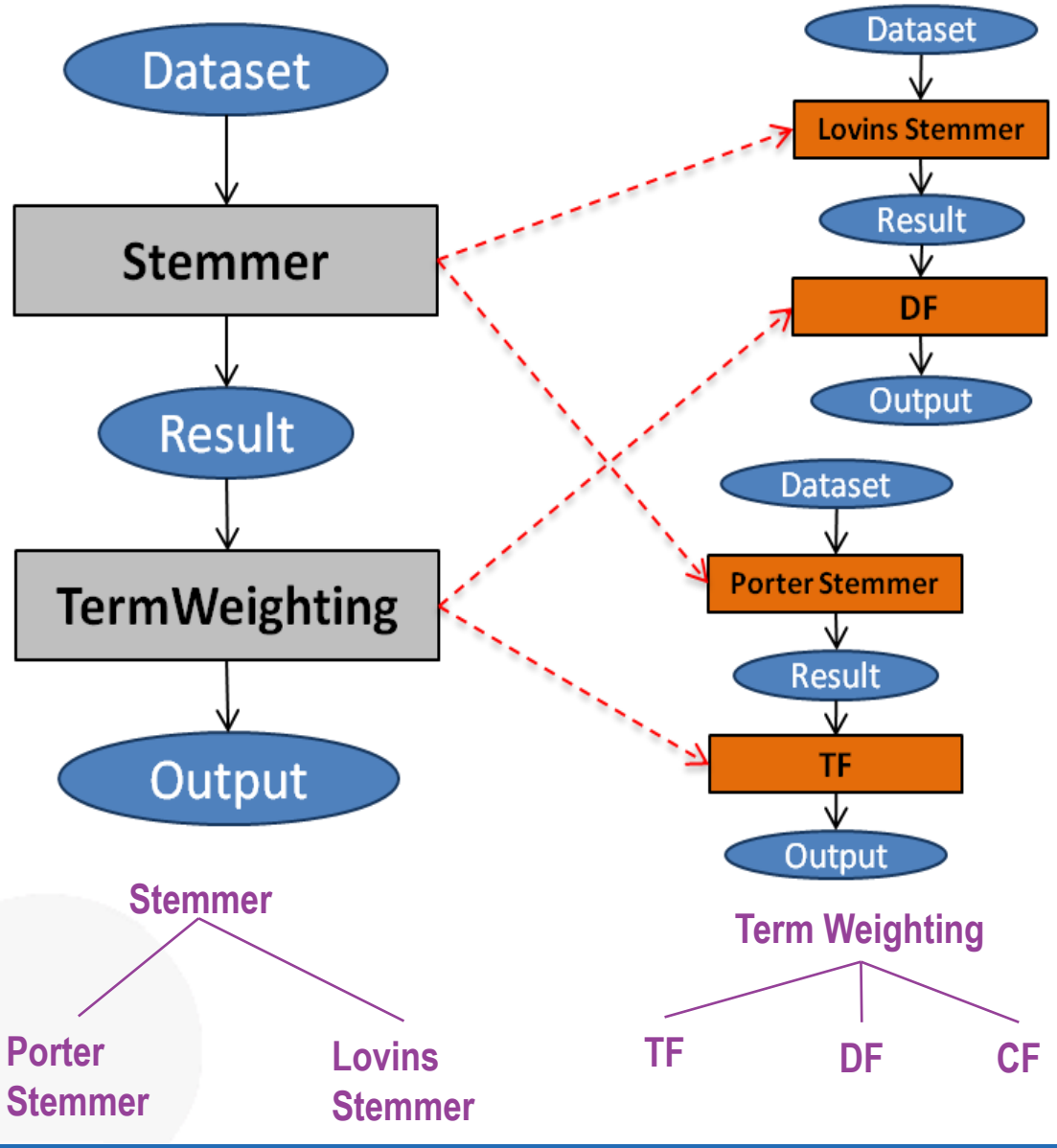


Occurrences:

3 times

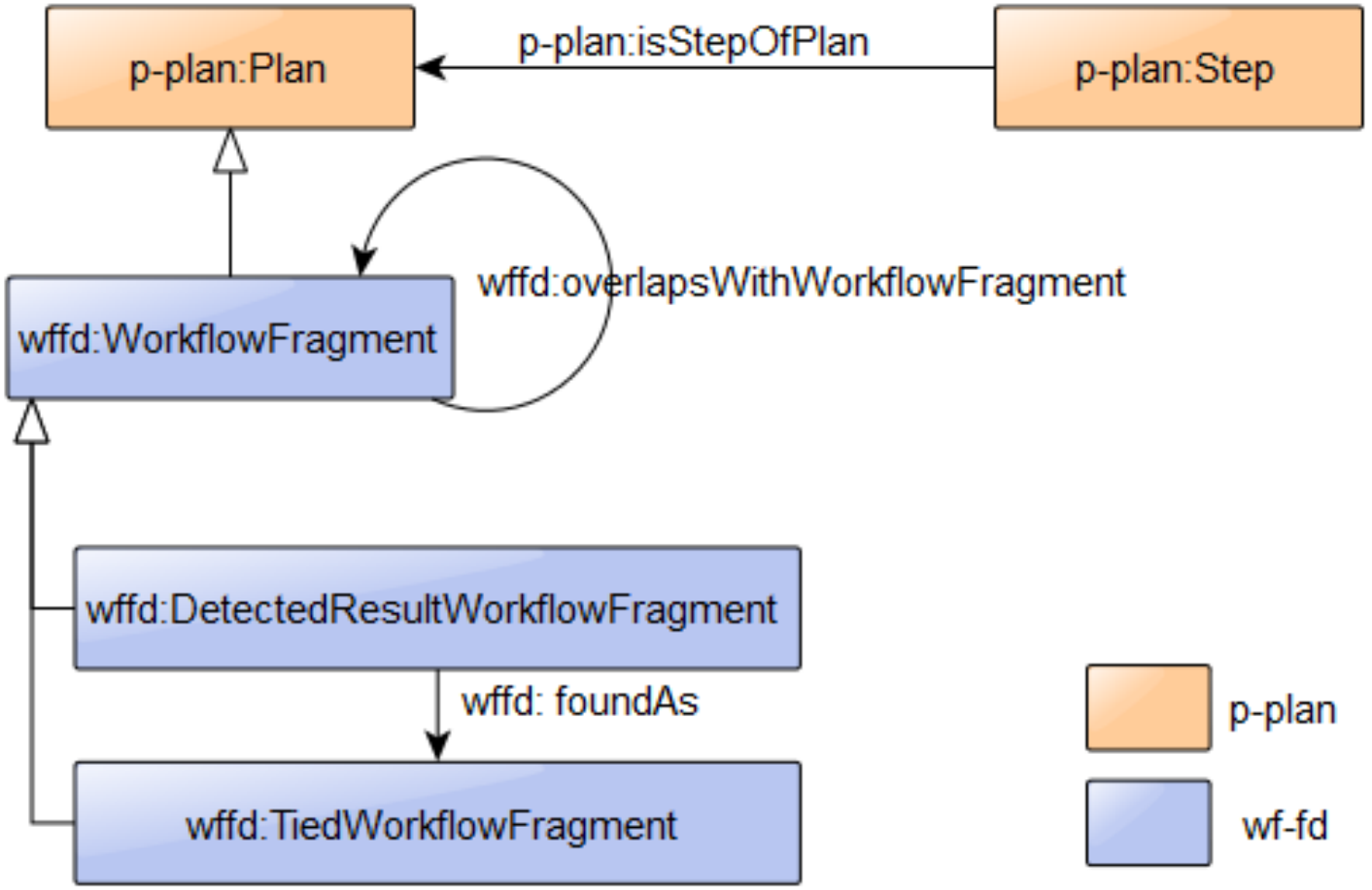
2 times

Challenges: Generalization of workflows

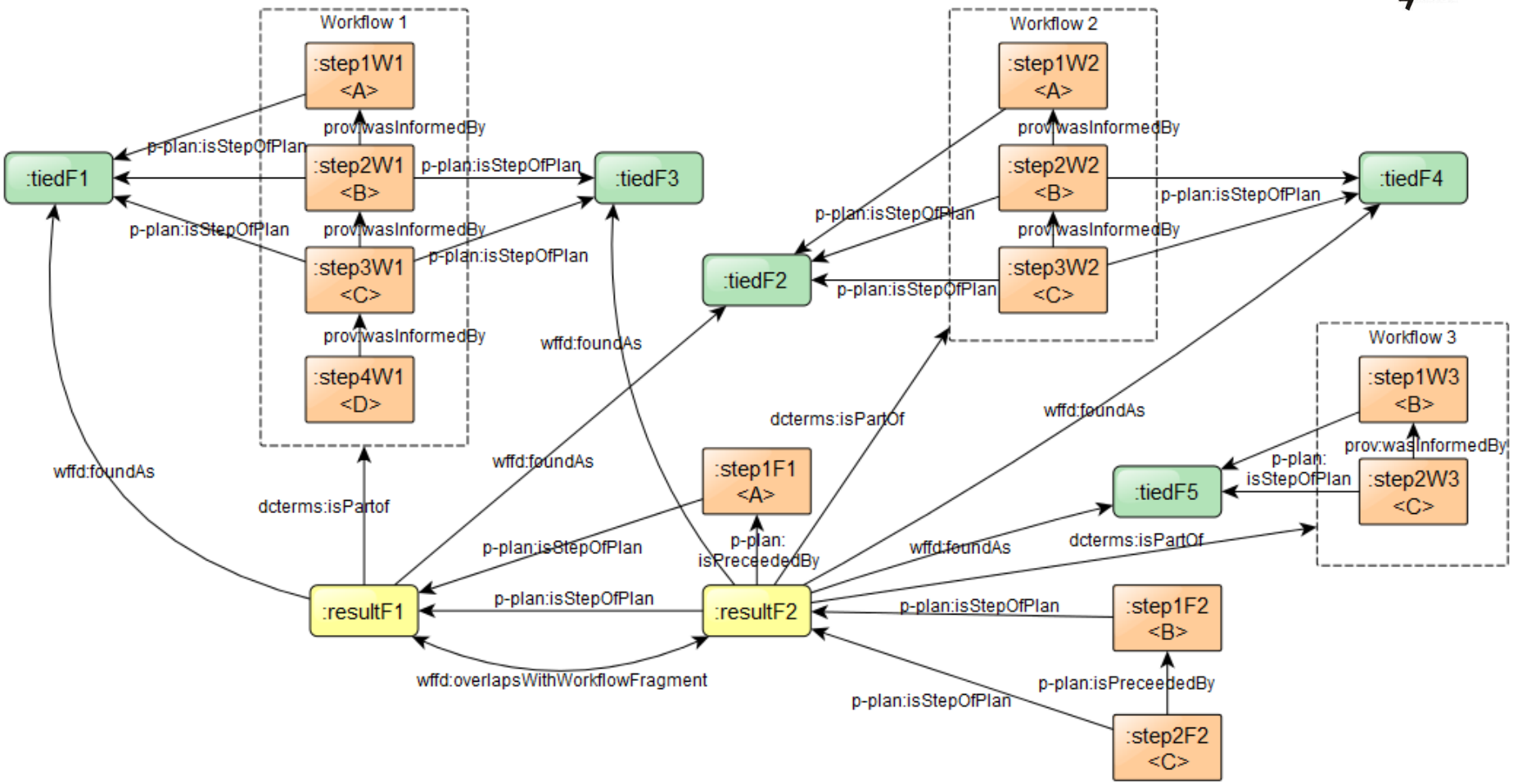


Exporting the fragment results: Wf-FD model

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



Exporting the fragment results: Wf-FD model

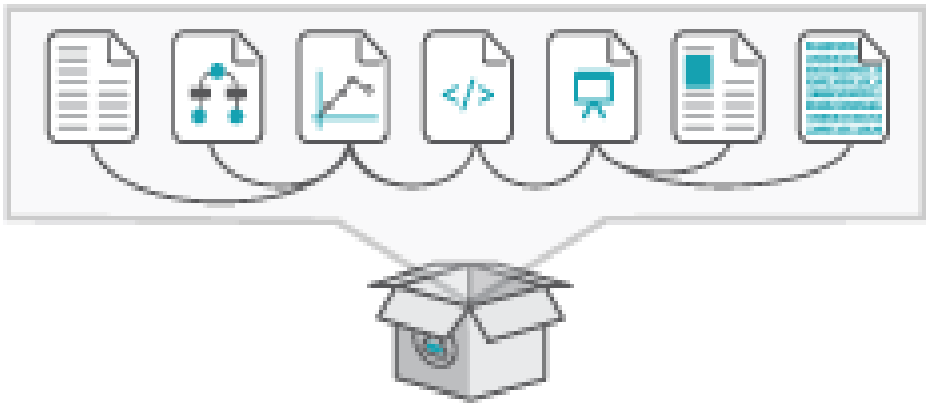


Research Objects

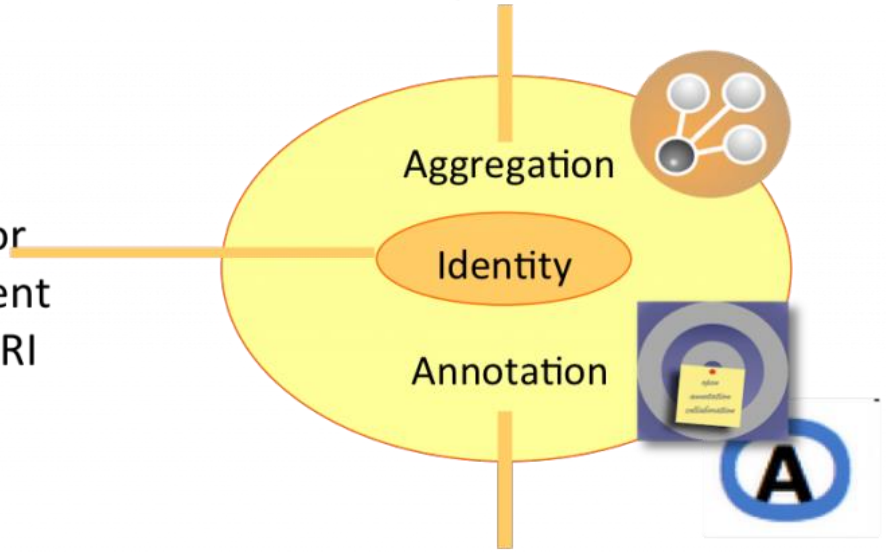
Workflow-Centric Research Objects: First Class Citizens in Scholarly Discourse. Belhajjame, K.; Corcho, O.; Garijo, D.; Zhao, J.; Missier, P.; Newman, D.; Palma, R.; Bechhofer, S.; Garcıa, E.; Manuel, .G. J.; Klyne, G.; Page, K.; Roos, M.; Ruiz, J. E.; Soiland-Reyes, S.; Verdes-Montenegro, L.; De Roure, D.; and Goble, C. In *Proceedings of the Second International Conference on the Future of Scholarly Communication and Scientific Publishing Sepublica2012*, page 1-12, Hersonissos, 2012

•Aggregation of resources that bundles together the contents of a research work:

- •Data
- •Experiments
- •Examples
- •Bibliography
- •Annotations
- •Provenance
- •ROs
- •Etc.



OAI-ORE
http://www.openarchives.org/ore/
Structuring and Bundling descriptions and components.



Standard id for each component
ORCID, DOI, URI

Annotation Data Ontology (AO) + Open Annotation
http://code.google.com/p/annotation-ontology/
A generic, domain-neutral annotation framework
http://www.openannotation.org/spec/core/

- Tool support
- Interoperability

What can you find in a Research Object? A real example

myexperiment

About | Mailing List | Publications

Logout | Give us Feedback | Invite

Home | Users | Groups | Workflows | Files | Packs | Services

All | Search

Home > Packs > GWAS to pathway

Pack: GWAS to pathway

Created at: 07/02/13 @ 08:32:47

| Tags (0) | Featured in Packs (0) | Favourited By (0) | Comments (0) |

Title: GWAS to pathway

Research object: http://sandbox.wf4ever-project.org/rodl/ROs/Pack384/

Description

This pack is for a workflow that finds KEGG pathways for genes from a GWAS.

Uploaded by

- Stian Soiland-Reyes (last uploaded on 2012-12-24 18:40)
- Khalid Belhajjame (last uploaded on 2012-11-02 18:40)

Authors

- Kristina Hettne (last authored on 2012-12-24 18:40)
- Marco Roos (last authored on 2012-10-15 11:24)

Contributors

Creator

Marco Roos

6 items in this pack

Navigate RO

- root
  - biblio/
    - produced/
    - used/
  - config/
    - scripts/
    - setup/
    - software/
    - web services/

New/Upload

Pack | GO

Stian Soiland-Reyes

My Profile [edit]
My Messages (3)
My Memberships
My History
My News

3 new messages
RE: hehe..
RE: Hello
RE: testing

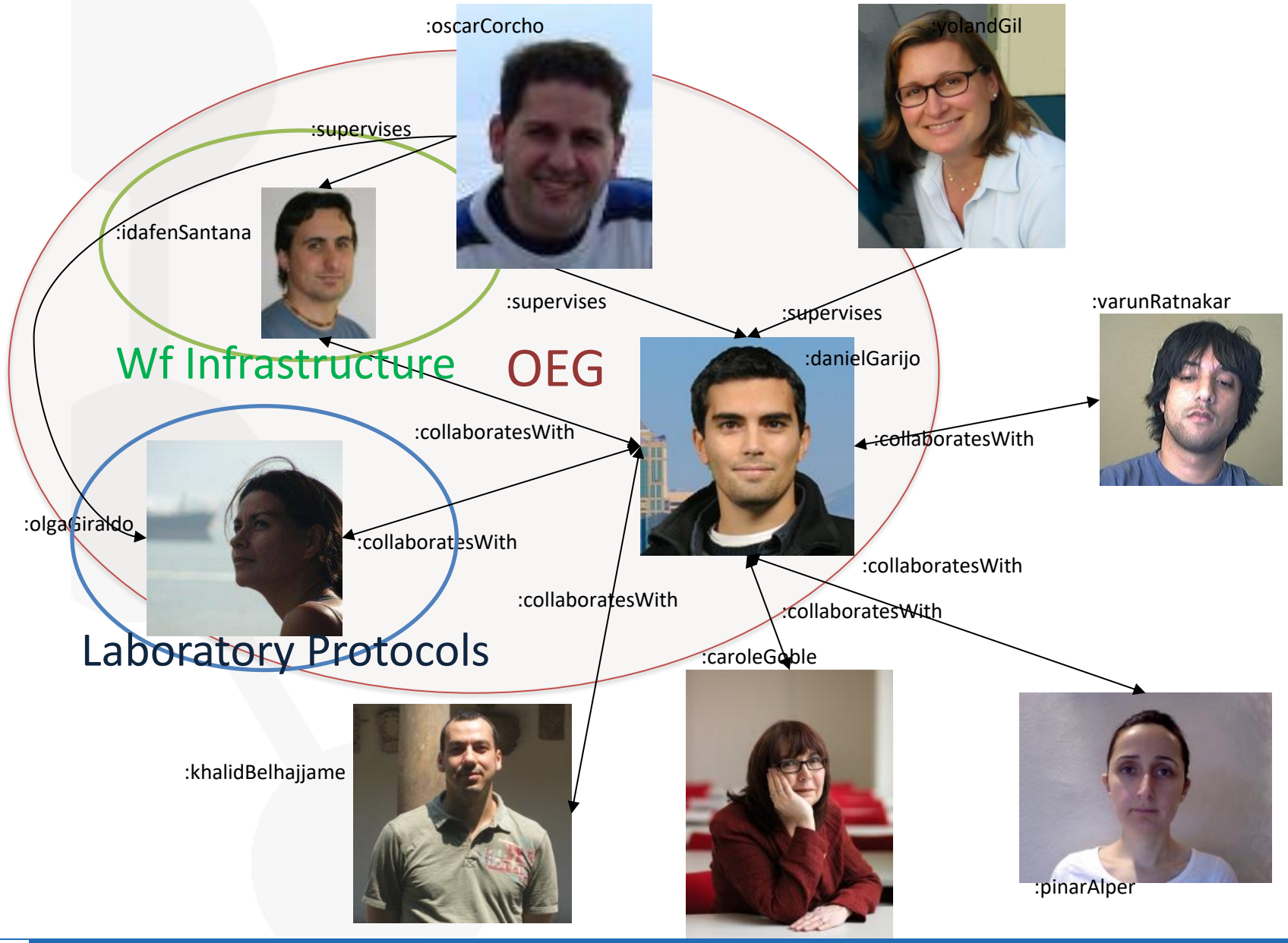
1 new friendship request
Raul Palma

My Stuff

21 Friends | 5 Groups | 47 Workflows | 17 Packs

- Finish integrating some other graph algorithms
  - PAFI (almost done)
  - Parsemis
- Test other workflow systems
  - LONI (from ISI)
  - GenePattern/Galaxy?
- Build a corpus for evaluation
  - Gold standard to test the different results.
- Perform evaluation
  - Analyze which is the best way to characterize the fragments
  - Presentation of the results
- Write paper(s)

Thanks !


:oscarCorcho
:yolandGil
:supervises
:idafenSantana
:supervises
:supervises
:varunRatnakar
Wf Infrastructure
OEG
:danielGarijo
:collaboratesWith
:collaboratesWith
:olgaGiraldo
:collaboratesWith
:collaboratesWith
:collaboratesWith
:collaboratesWith
Laboratory Protocols
:caroleGoble
:khalidBelhajjame
:pinarAlper

From Scientific Workflows to Research Objects: Publication and Abstraction of Scientific Experiments

Daniel Garijo Verdejo

Supervisors: Oscar Corcho, Yolanda Gil

Ontology Engineering Group
Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid

Date: 27/02/2014