Workflows, experimental findings, and their provenance: towards semantically rich linked data and method sharing for collaborative science

Paolo Missier

Paolo.Missier@ncl.ac.uk

Newcastle University, UK

Credible workshop

Sophia-Antipolis

October, 2012

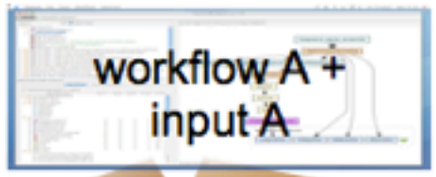
Prologue: DCC "REPRISE" workshop, 2009

MANCHESTER 1824

The University of Manchester

Full-fledged data-mediated collaborations

exp. A

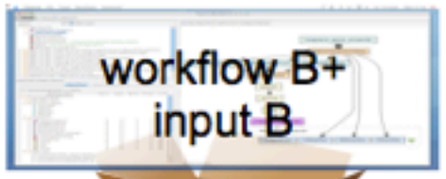


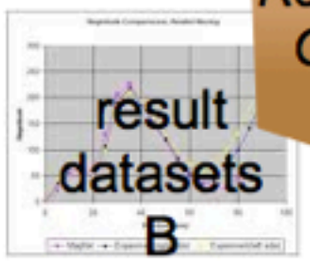
Research Object A

result provenance A

result A → input B

exp. B





Research Object B

result provenance B

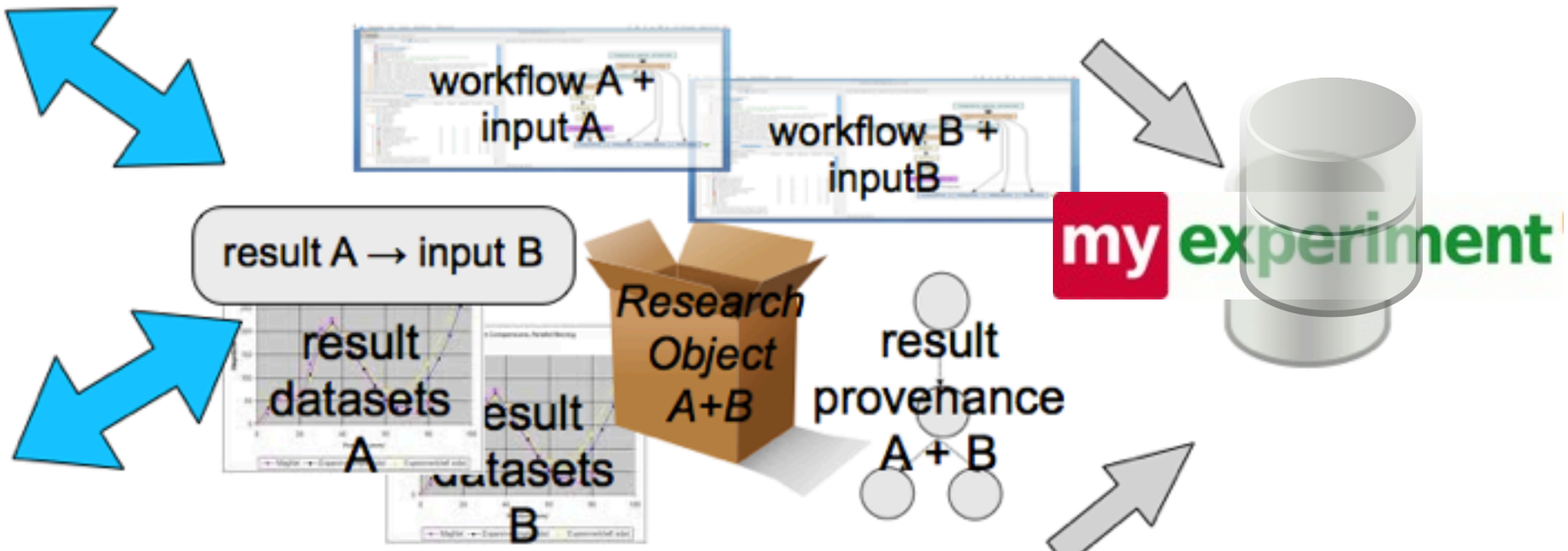
myexperiment

“Virtual experimental science” (DCC’09)

MANCHESTER 1824

The University of Manchester

Full-fledged data-mediated collaborations



Provenance composition accounts for implicit collaboration

[LMB+10]

Outline

1. Workflow and their provenance
   - Janus: workflows + provenance + semantics for Taverna
     - (+ LOD) [MLB+12, ZSM+11]
   - The PROV Semantic data model for provenance (ongoing) W3C
   - PROV-W: *unofficial* workflow extension (and semantic annotations)

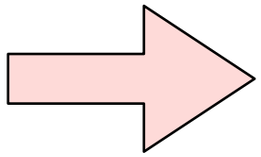
2. Packaging and sharing: data + methods + provenance
   - Research Objects in the Workflow 4Ever project
   - DataONE and Data Packages

1: Workflows and provenance

- The Janus provenance ontology for Taverna
- The W3C PROV ontology
- PROV-W

Example workflow (Taverna)

chr: 17
start: 28500000
end: 3000000



QTL → Ensembl Genes

Ensembl Gene → Uniprot Gene

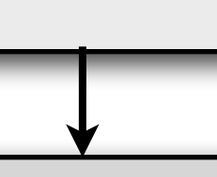
Ensembl Gene → Entrez Gene

Uniprot Gene → Kegg Gene

Entrez Gene → Kegg Gene

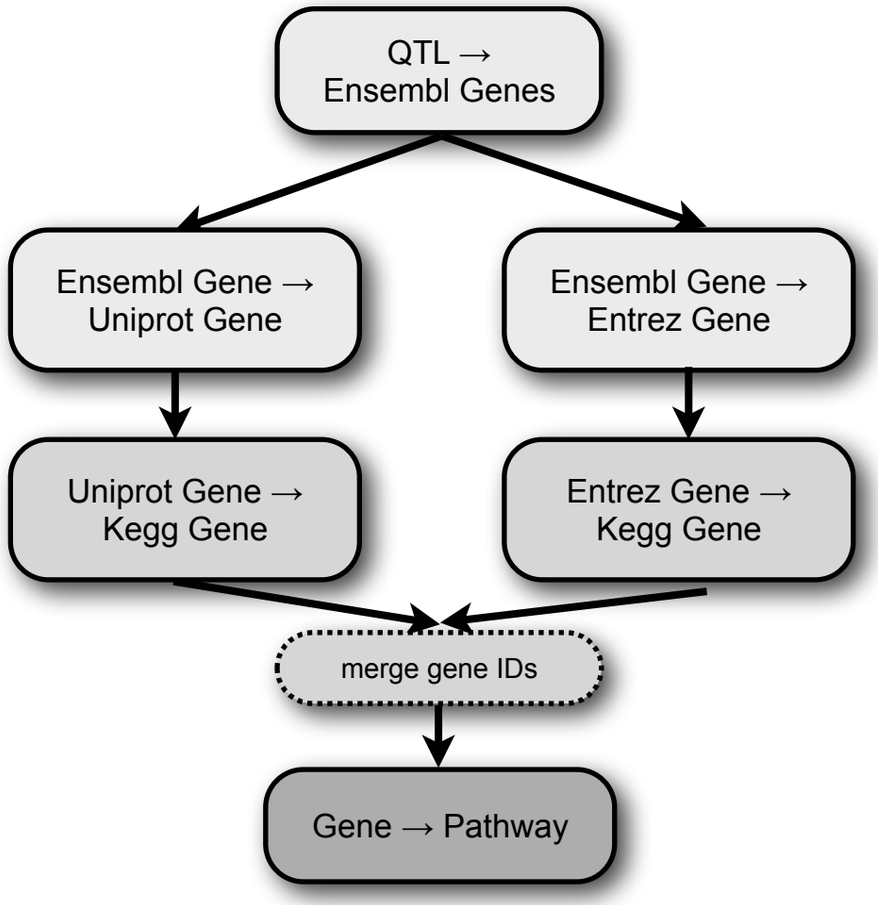
merge gene IDs

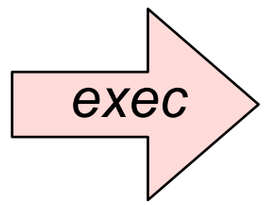
Gene → Pathway

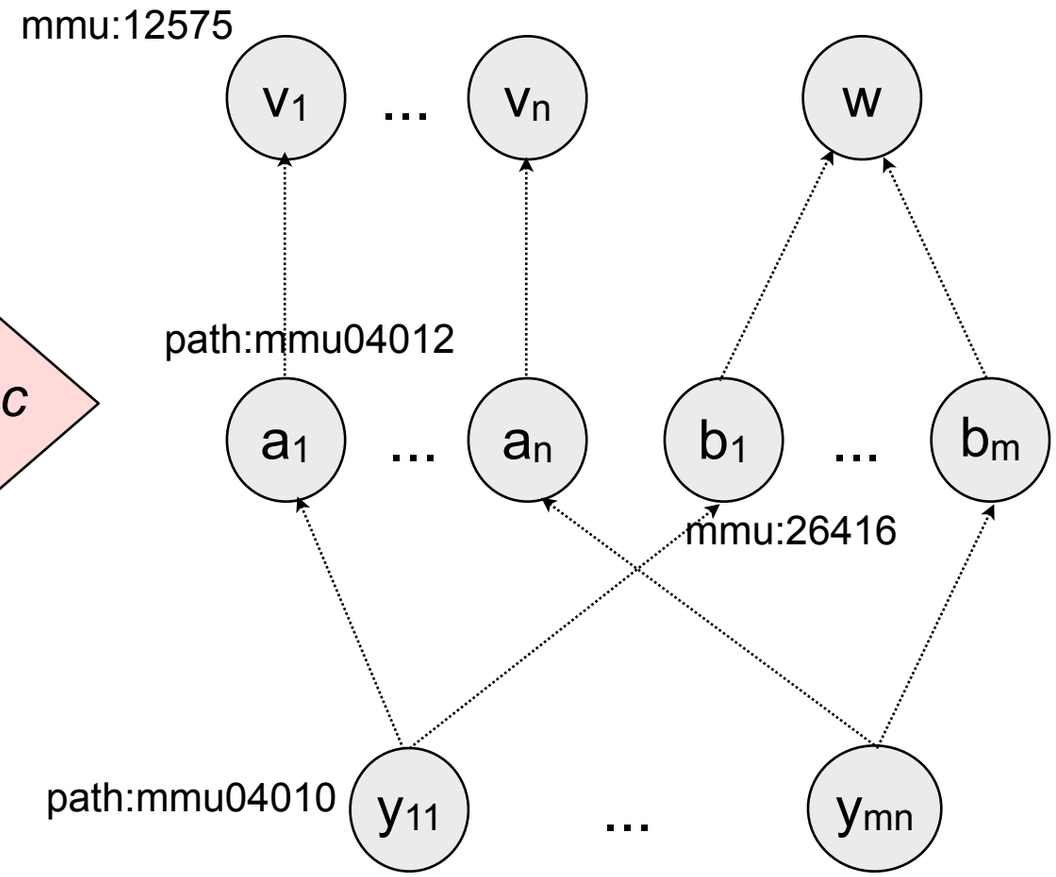


path:mmu04210 Apoptosis,
path:mmu04010 MAPK, ...

Baseline provenance of a workflow run





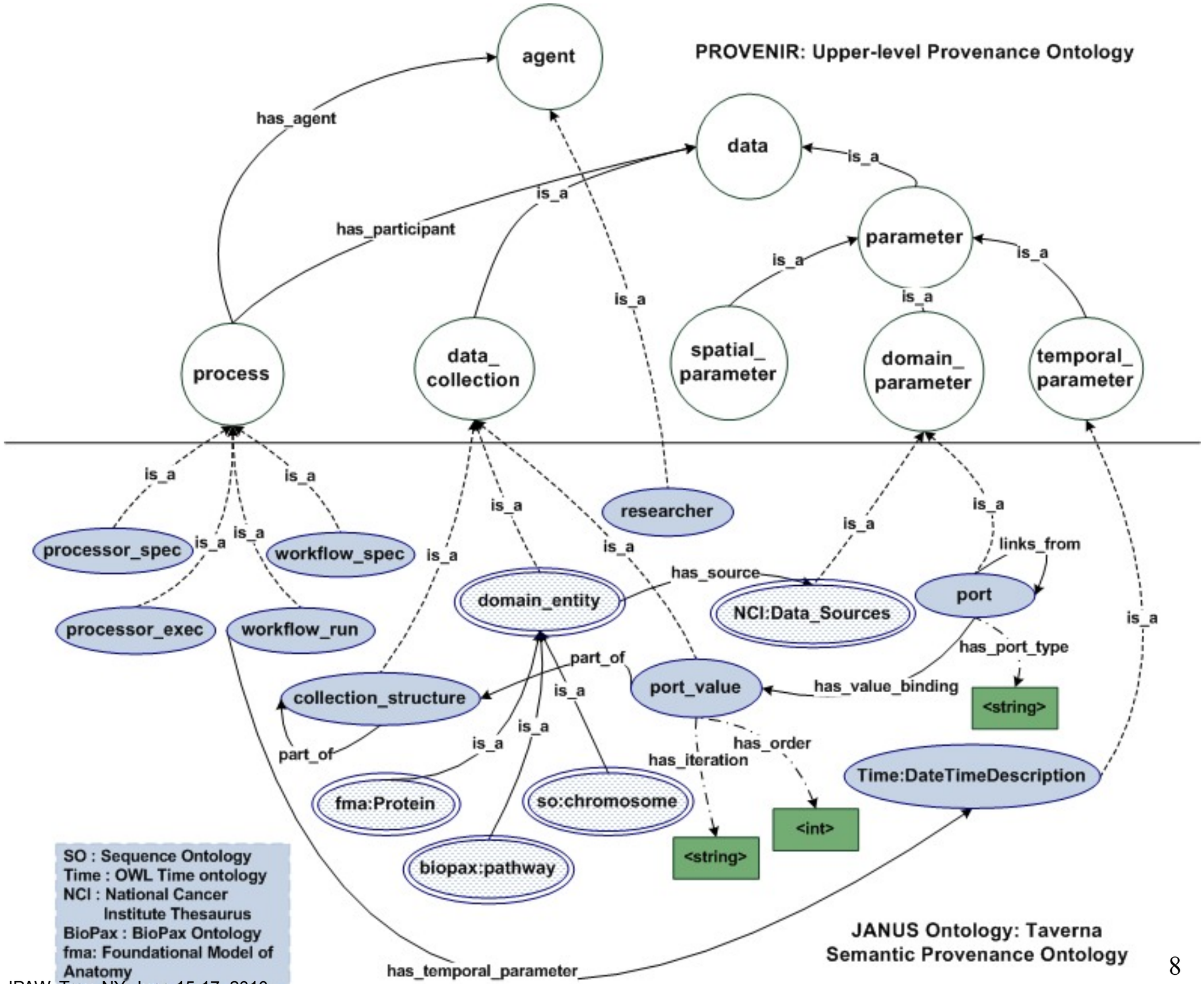


path:mmu04010→*derivedFrom*→mmu:26416

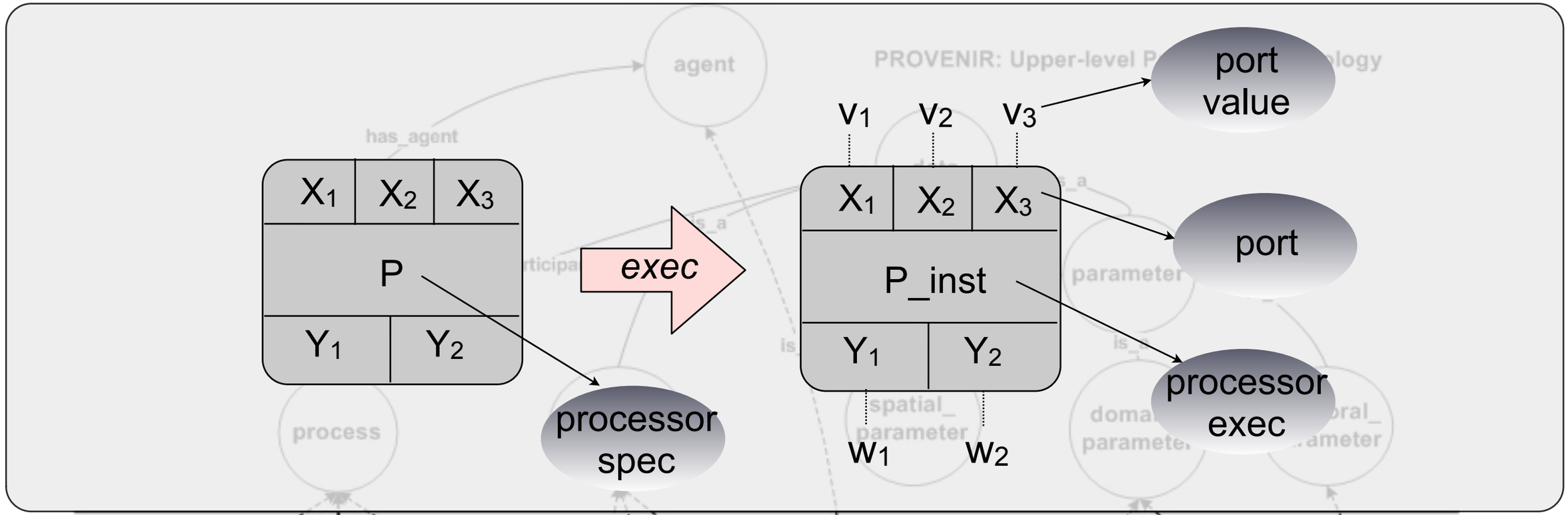
path:mmu04012→*derivedFrom*→mmu:12575

- The graph encodes all direct data dependency relations
- Baseline query model: compute paths amongst sets of nodes
  - Transitive closure over data dependency relations

Janus: a semantic provenance model

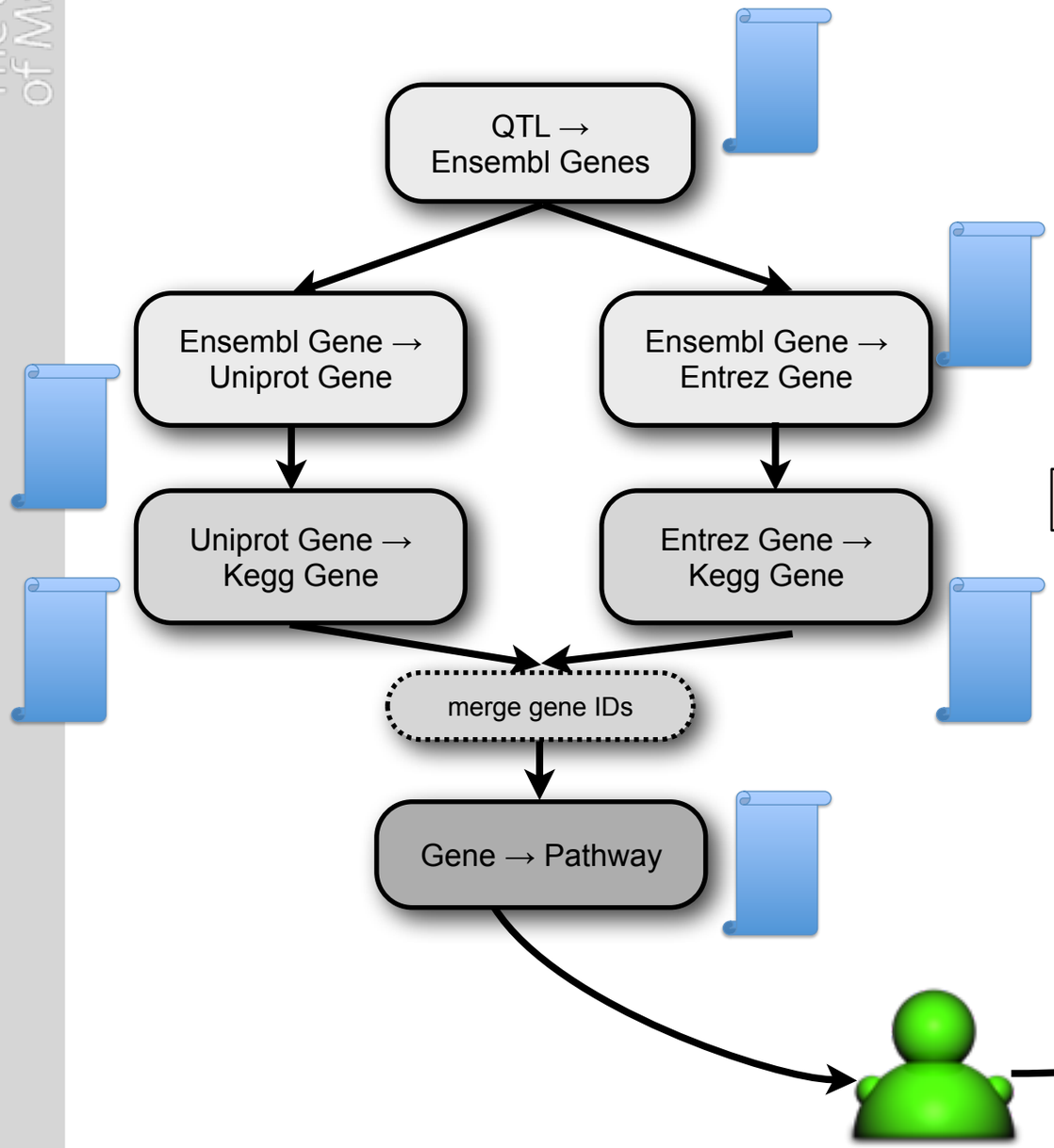


Janus: a semantic provenance model



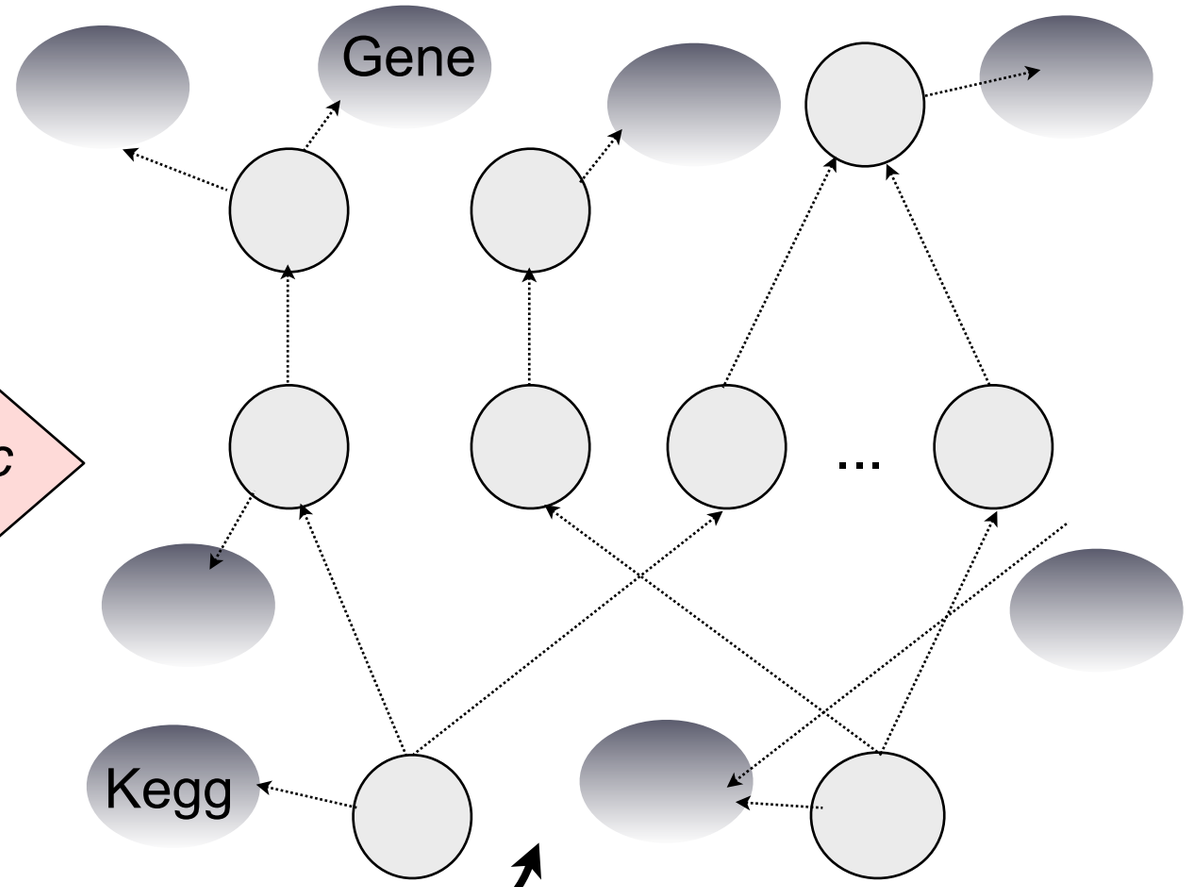
Annotated provenance

Annotated workflow



exec

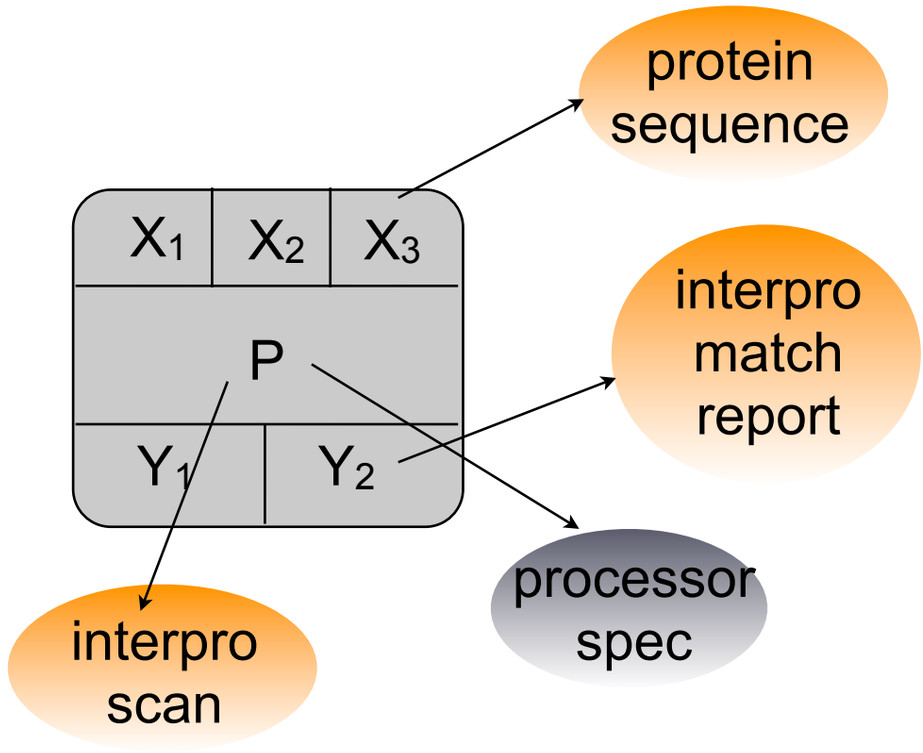
Annotated provenance graph



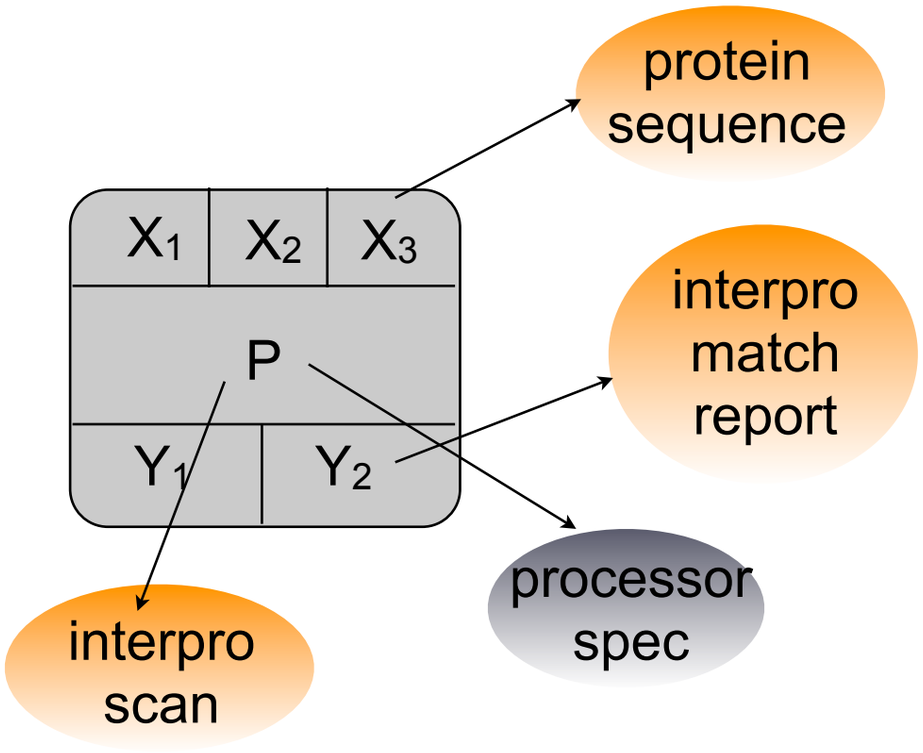
Find all genes within the input QTL region that are involved in a given KEGG pathway.

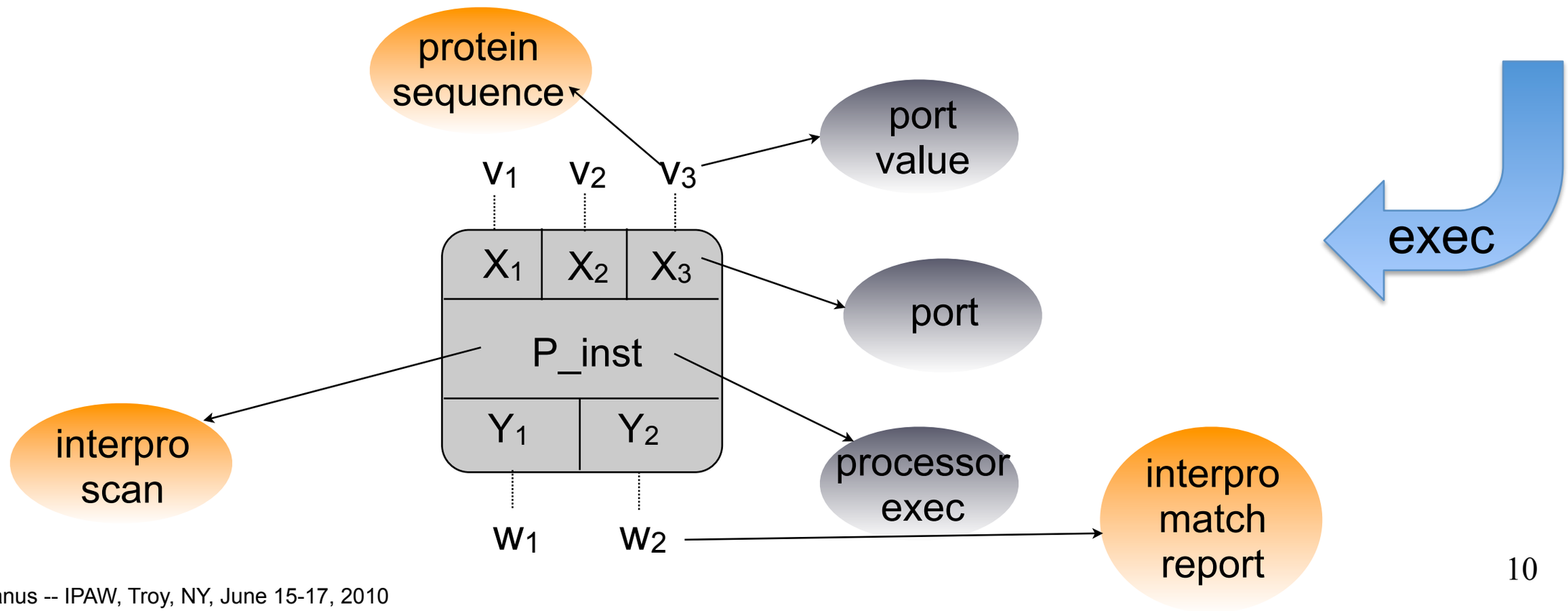
List relevant PubMed publications for the pathways listed in the result set.

From structure annotations to values annotations



From structure annotations to values annotations

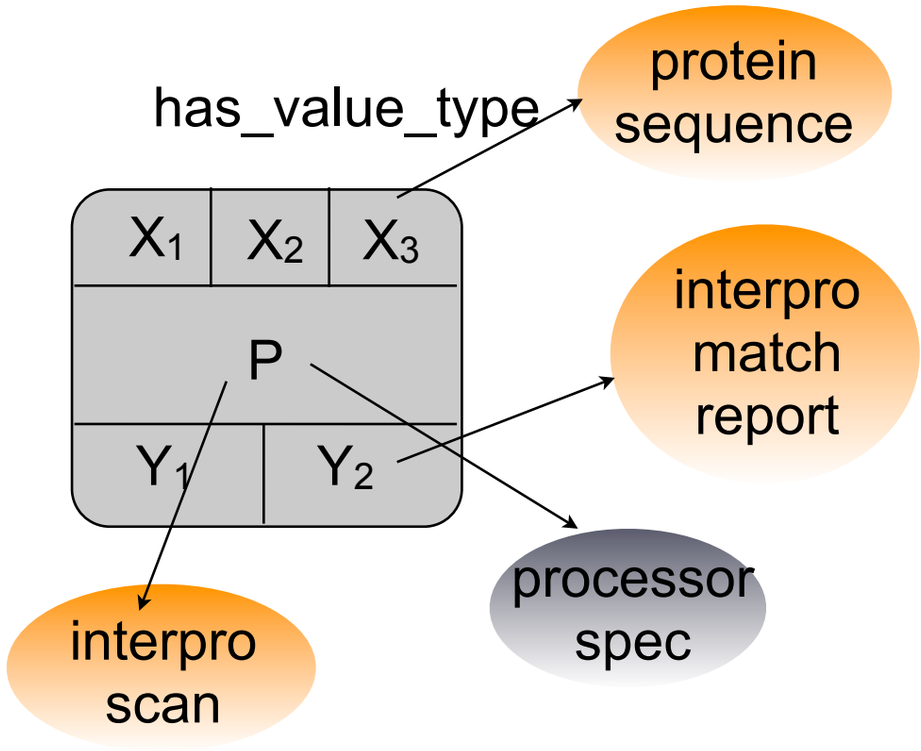




Annotations propagation rules

$$\frac{X\ \texttt{rdf:type}\ \texttt{Port} \qquad C = \{c\} \qquad X\ \texttt{has\_value\_type}\ c \qquad X\ \texttt{has\_value}\ v \qquad v\ \texttt{rdf:type}\ \texttt{PortValue}}{v\ \texttt{rdf:type}\ C}$$

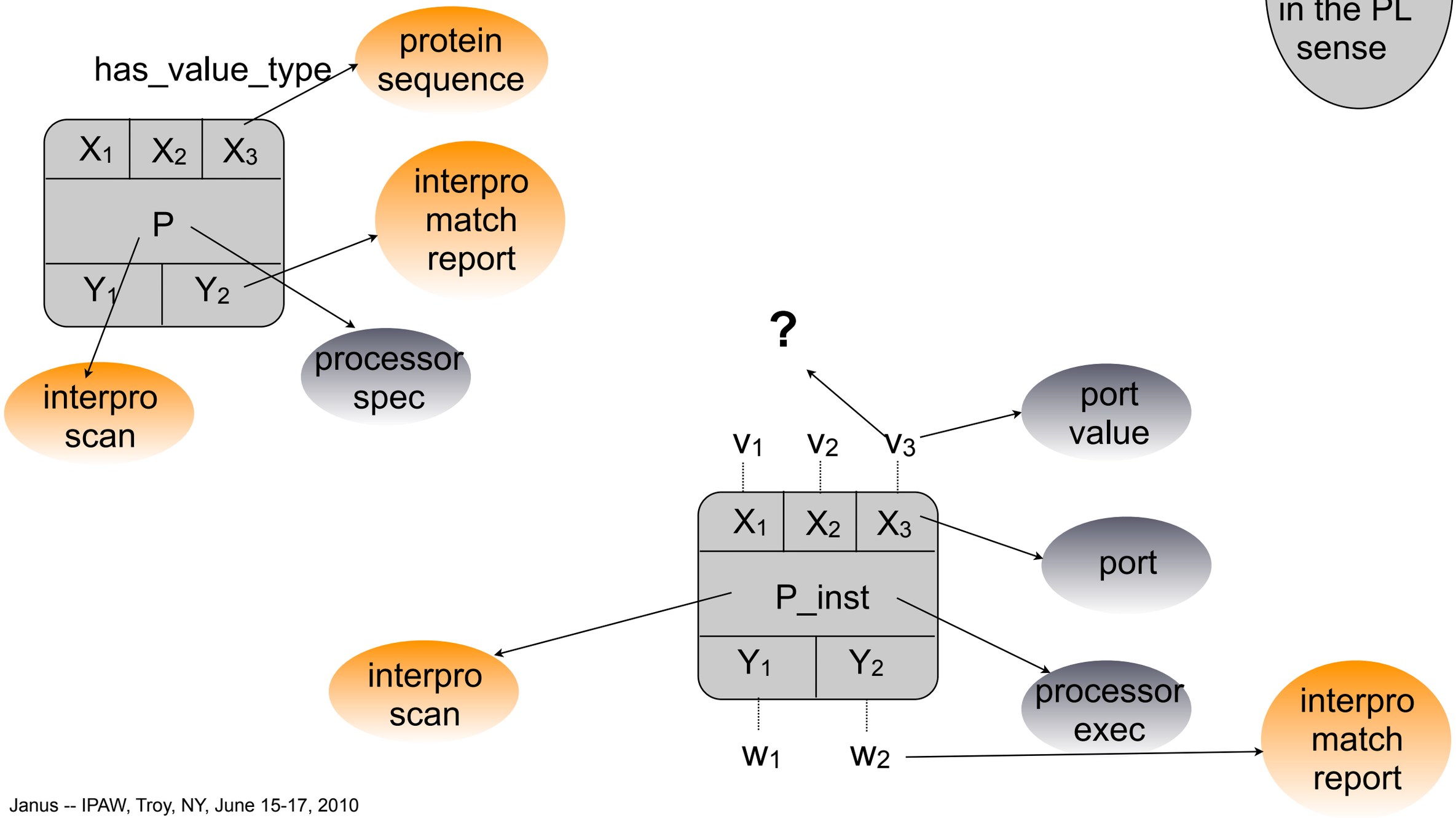
denotes data type in the PL sense



Annotations propagation rules

$$\frac{X\ \texttt{rdf:type}\ \texttt{Port} \qquad C = \{c\} \qquad X\ \texttt{has\_value\_type}\ c \qquad X\ \texttt{has\_value}\ v \qquad v\ \texttt{rdf:type}\ \texttt{PortValue}}{v\ \texttt{rdf:type}\ C}$$

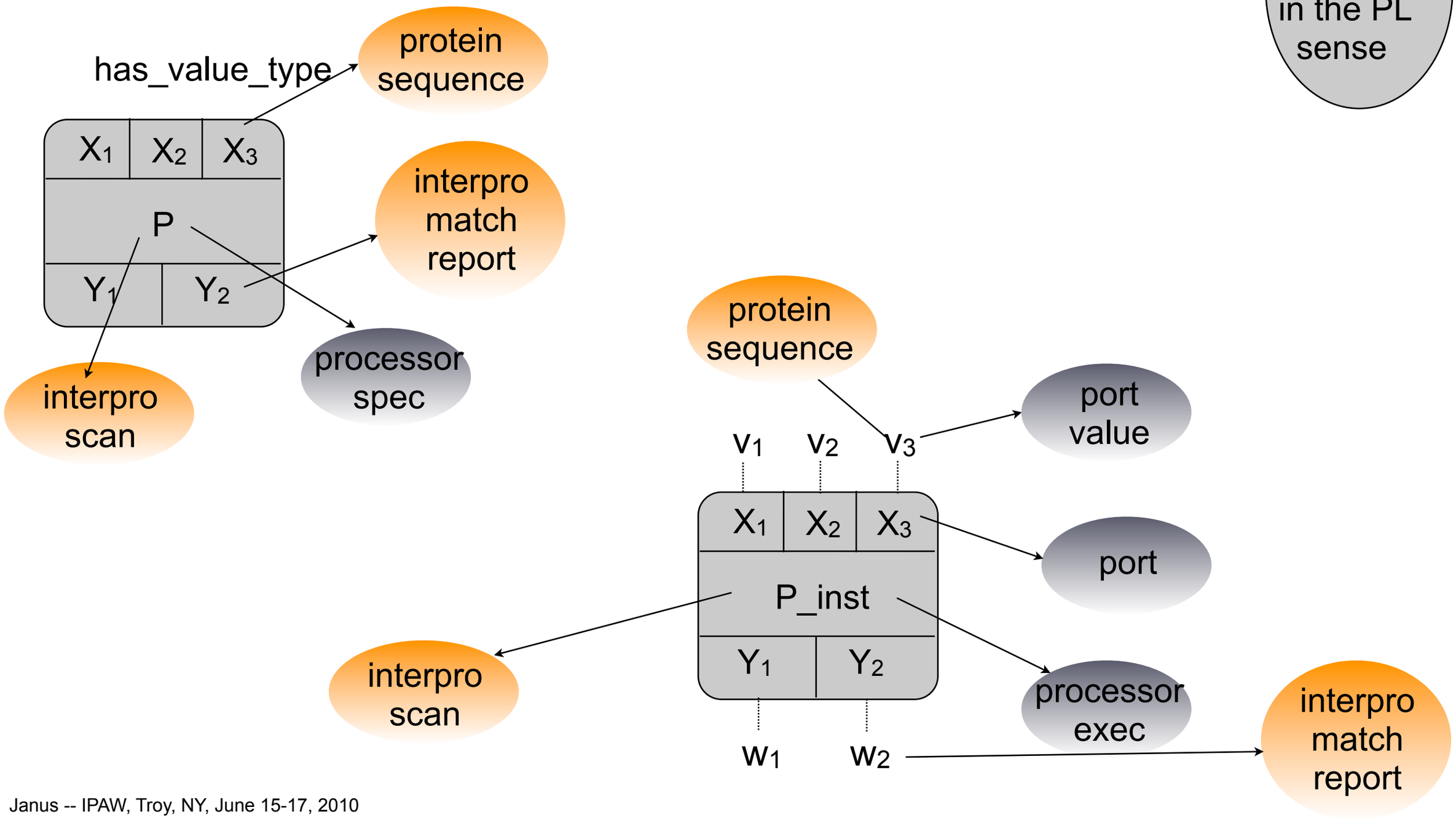
denotes data type in the PL sense



Annotations propagation rules

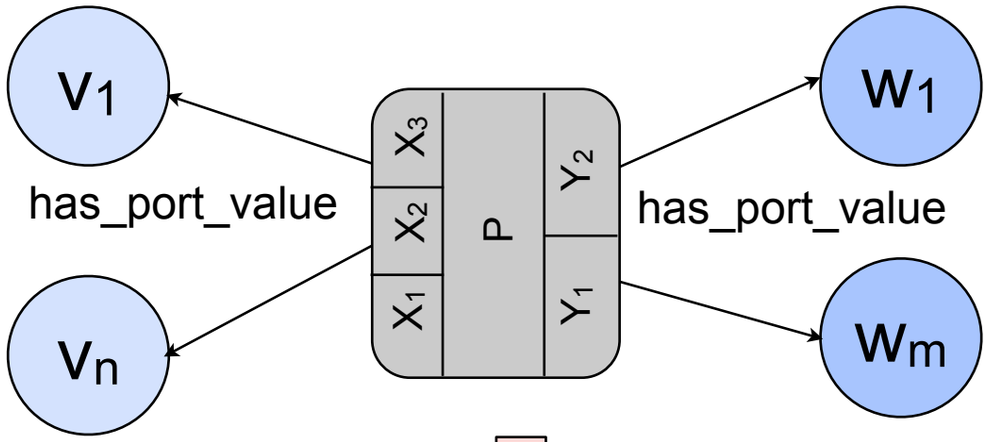
$$\frac{X\ \texttt{rdf:type}\ \texttt{Port} \qquad C = \{c\} \qquad X\ \texttt{has\_value\_type}\ c \qquad X\ \texttt{has\_value}\ v \qquad v\ \texttt{rdf:type}\ \texttt{PortValue}}{v\ \texttt{rdf:type}\ C}$$

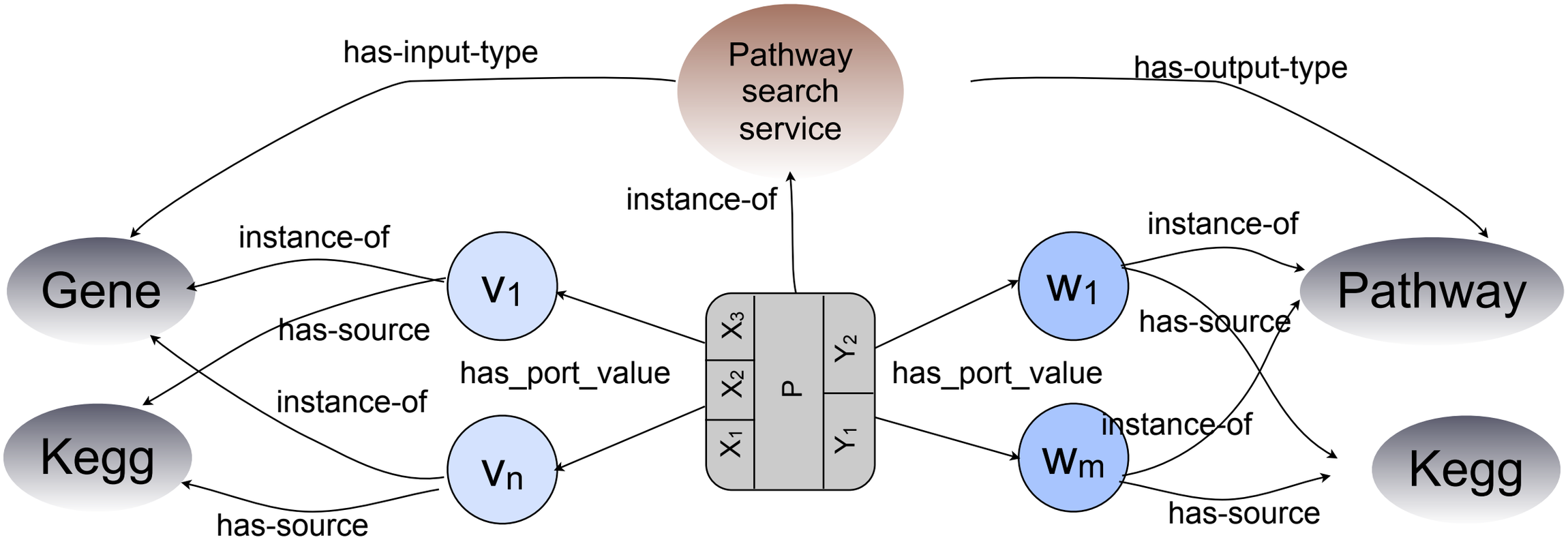
denotes data type in the PL sense



Annotations as semantic overlay

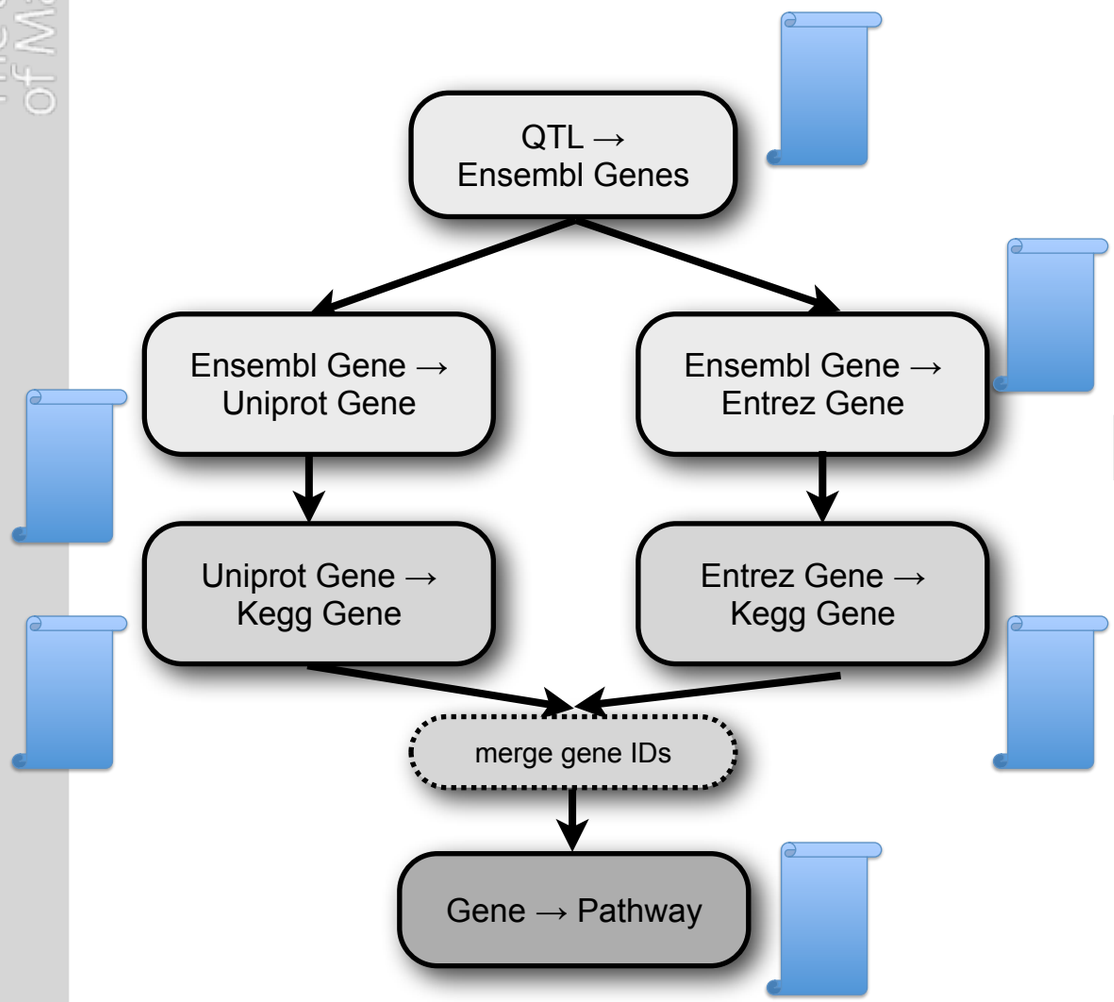
Provenance graph fragment





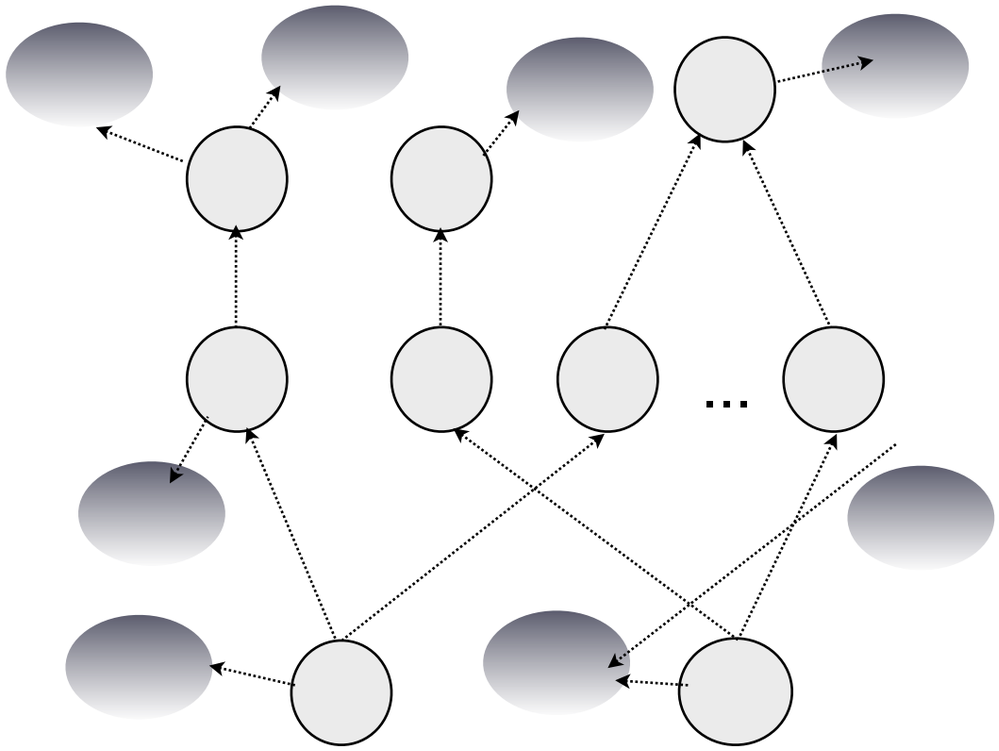
Extensions to Linked Data

Annotated workflow



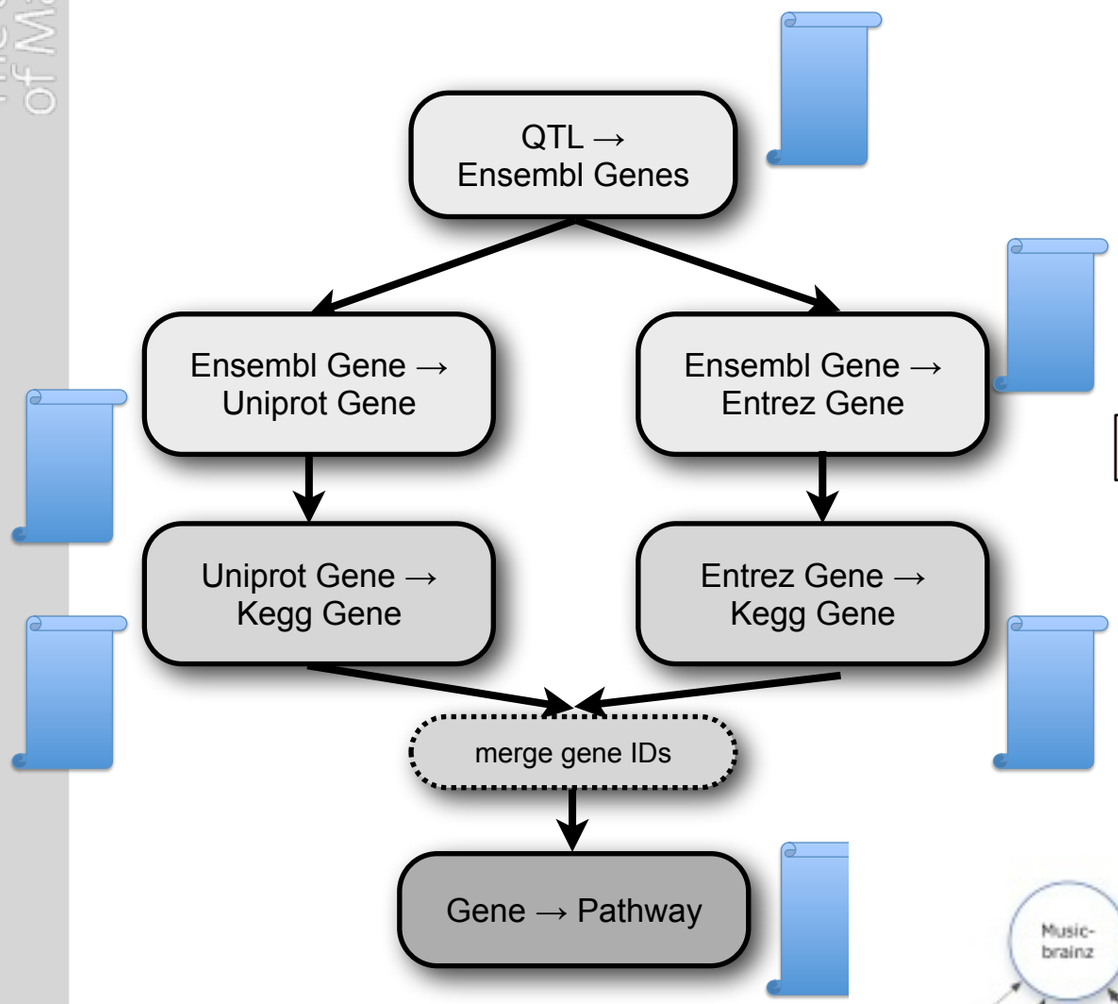
exec

Annotated provenance graph



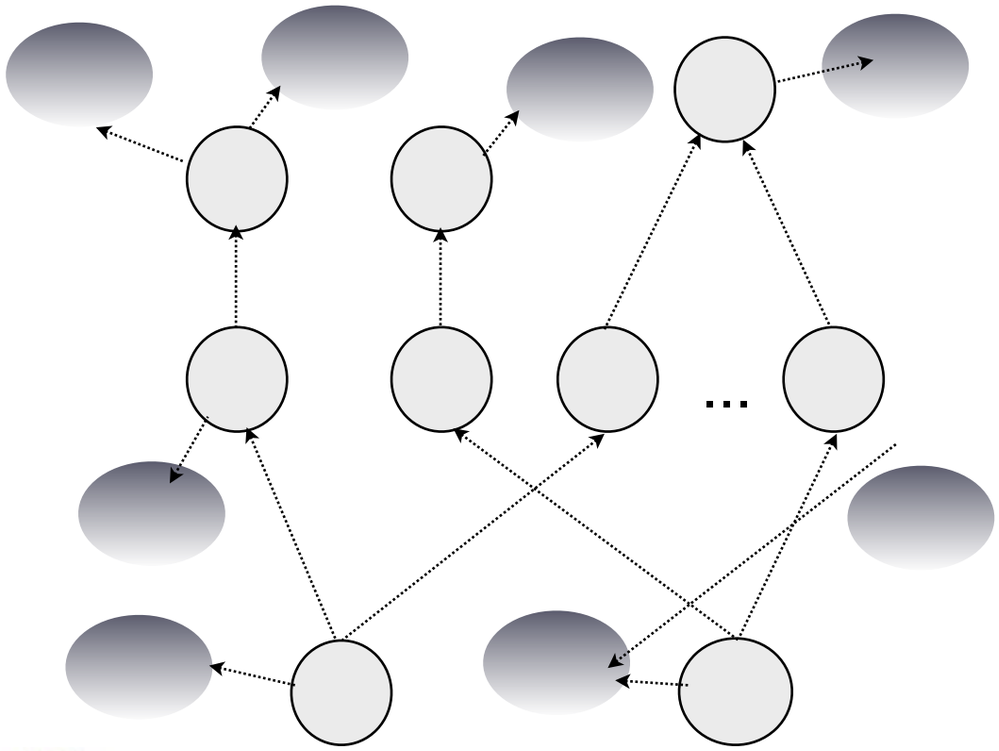
Extensions to Linked Data

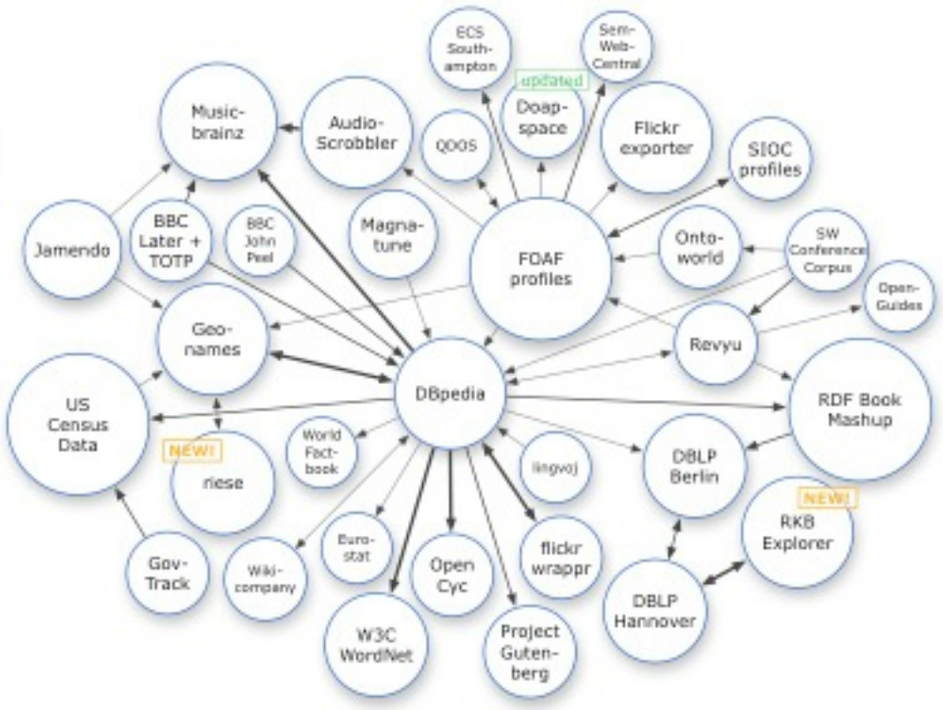
Annotated workflow



exec

Annotated provenance graph





- Publish
- I - Map IDs
- II - query

I - Mapping data values to LoD URIs

In our prototype we map data values to Bio2RDF as follows:

- IF $isType(d_i)$ == `Gene` AND $isSource(d_i)$ == `Entrez` THEN
  $uri(d_i)$ = http://bio2rdf.org/geneid: + $value(d_i)$ — Entrez Genes
- IF $isType(d_i)$ == `Gene` AND $isSource(d_i)$ == `UniProt` THEN
  $uri(d_i)$ = http://bio2rdf.org/uniprot: + $value(d_i)$ — Uniprot Genes
- IF $isType(d_i)$ == `Gene` AND $isSource(d_i)$ == `KEGG` THEN
  $uri(d_i)$ = http://bio2rdf.org/kegg: + $value(d_i)$ — KEGG Genes
- IF $isType(d_i)$ == `Pathway` AND $isSource(d_i)$ == `KEGG` THEN
  $uri(d_i)$ = http://bio2rdf.org/path: + $value(d_i)$ — KEGG Pathways

Linked Data Query example:

List relevant PubMed publications for the pathways listed in the workflow result set

The PROV ontology from the W3C

PROV Model Primer

W3C Working Draft 24 July 2012

This version:
http://www.w3.org/TR/2012/WD-prov-primer-20120724/
Latest published version:
http://www.w3.org/TR/prov-primer/
Latest editor's draft:
http://dvcs.w3.org/hg/prov/raw-file/default/primer/Primer.htr
Previous version:
http://www.w3.org/TR/2012/WD-prov-primer-20120503/
Editors:
Yolanda Gil, Information Sciences Institute, University of S
Simon Miles, King's College London, UK
Authors:
Khalid Belhajjame, University of Manchester
Helena Deus, Digital Enterprise Research Institute (DERI)
Daniel Garijo, Universidad Politécnica de Madrid
Graham Klyne, University of Oxford
Paolo Missier, Newcastle University
Stian Soiland-Reyes, University of Manchester
Stephan Zednik, Rensselaer Polytechnic Institute

PROV-O: The PROV Ontology

W3C Working Draft 24 July 2012

This version:
http://www.w3.org/TR/2012/WD-prov-o-20120724/
Latest published version:
http://www.w3.org/TR/prov-o/
Latest editor's draft:
https://dvcs.w3.org/hg/prov/raw-file/default/ontology/Overvi
Previous version:
http://www.w3.org/TR/2012/WD-prov-o-20120503/
Editors:
Timothy Lebo, Rensselaer Polytechnic Institute, USA
Satya Sahoo, Case Western Reserve University, USA
Deborah McGuinness, Rensselaer Polytechnic Institute, U
Authors:
(In alphabetical order)
Khalid Belhajjame, University of Manchester, UK
James Cheney, University of Edinburgh, UK
David Corsar, University of Aberdeen, UK
Daniel Garijo, Universidad Politécnica de Madrid, Spain
Stian Soiland-Reyes, University of Manchester, UK
Stephan Zednik, Rensselaer Polytechnic Institute, USA
Jun Zhao, University of Oxford, UK

- Plus a growing catalogue of examples from group members:

http://www.w3.org/2011/prov/wiki/PROV_examples

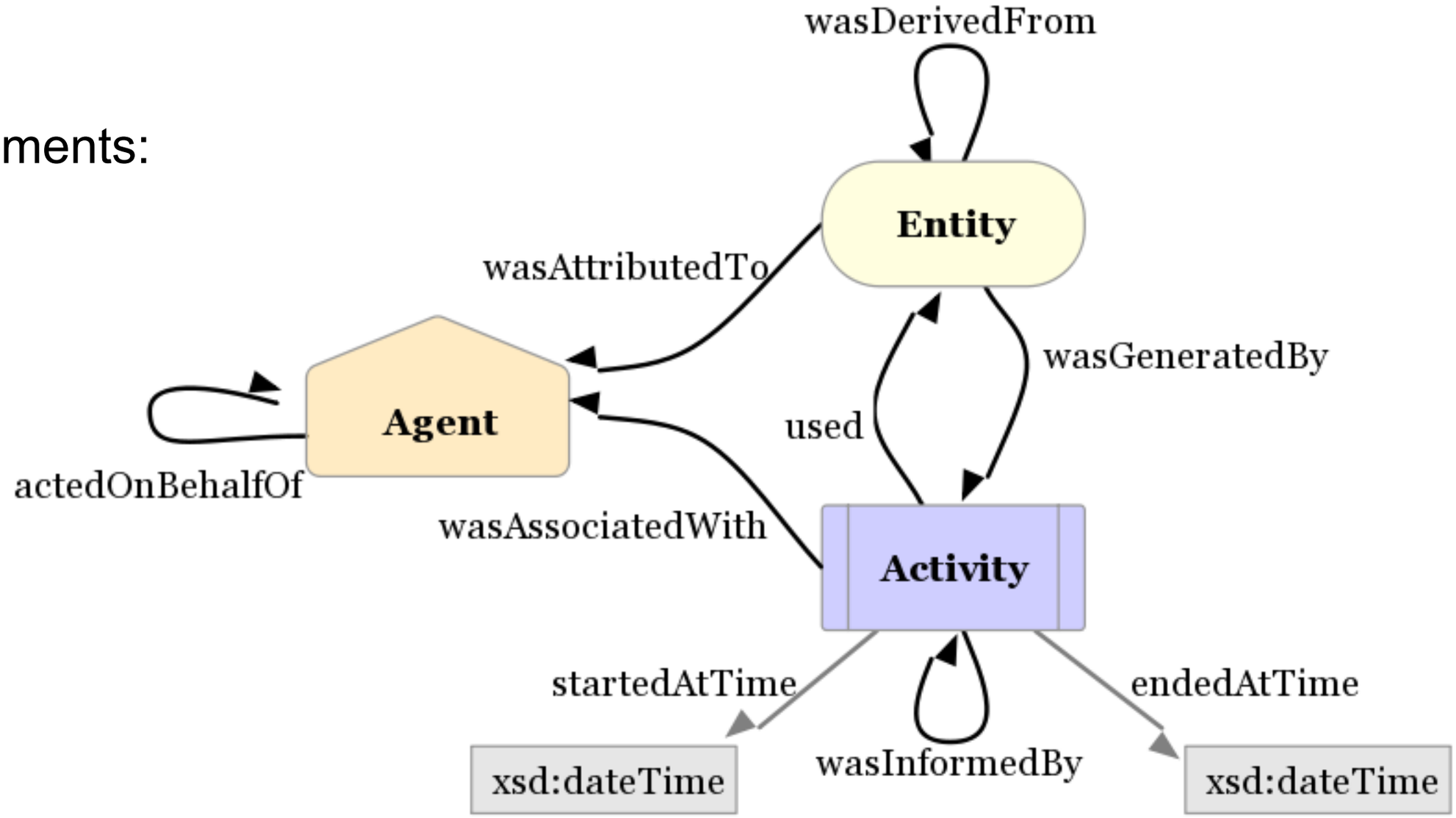
PROV Ontology (PROV-O)

- "PROV-O defines the normative OWL2 Web Ontology Language encoding of the PROV Data Model" [1]

http://dvcs.w3.org/hg/prov/raw-file/default/ontology/ProvenanceOntology.owl

state as of Oct., 2012: Last Call -- now closed to public comments

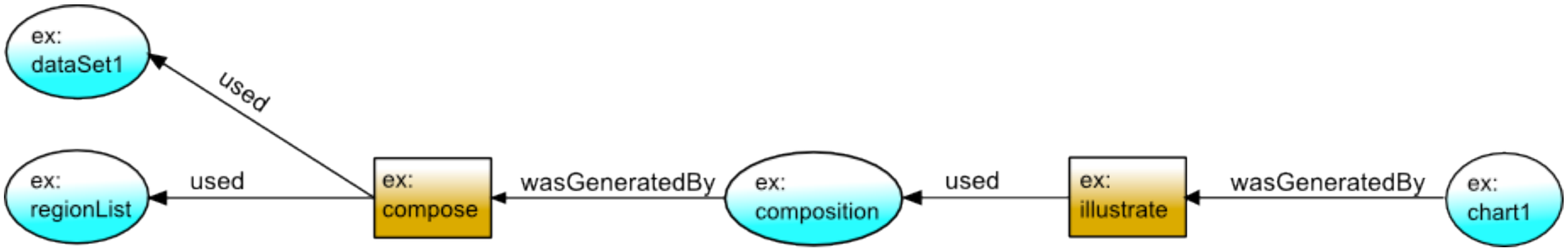
Core elements:



[1] Current version: http://www.w3.org/TR/prov-o/

PROV-O encoding: simple example

- Examples drawn from the PROV Primer document




```
PROV-N:
used(ex:compose, ex:dataSet1, -)
used(ex:compose, ex:regionList, -)
wasGeneratedBy(ex:composition, ex:compose, -)
used(ex:illustrate, ex:composition, -)
wasGeneratedBy(ex:chart1, ex:illustrate, -)
```

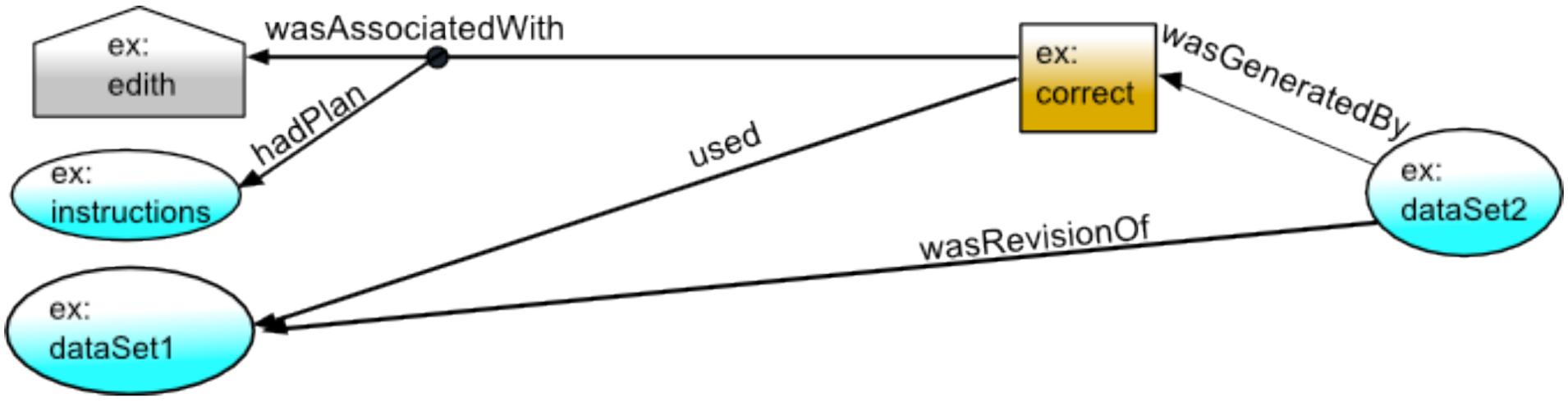
PROV-O (Turtle):

```
ex:compose       a prov:Activity;
                 prov:used              ex:dataSet1 ;
                 prov:used              ex:regionList .
ex:composition   a prov:Entity;
                 prov:wasGeneratedBy ex:compose .
ex:illustrate    a prov:Activity;
                 prov:used              ex:composition .
ex:chart1        a prov:Entity;
                 prov:wasGeneratedBy ex:illustrate .
```


[2] Current editor's draft: http://dvcs.w3.org/hg/prov/raw-file/default/primer/Primer.html

Plans in PROV-O

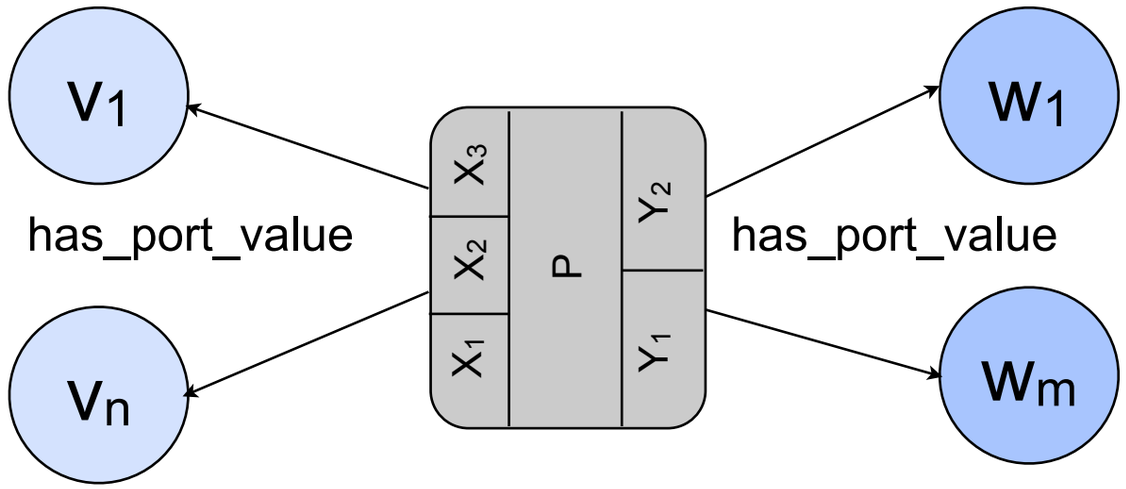
- PROV deliberately does not deal with program structure
  - workflow, processor, port

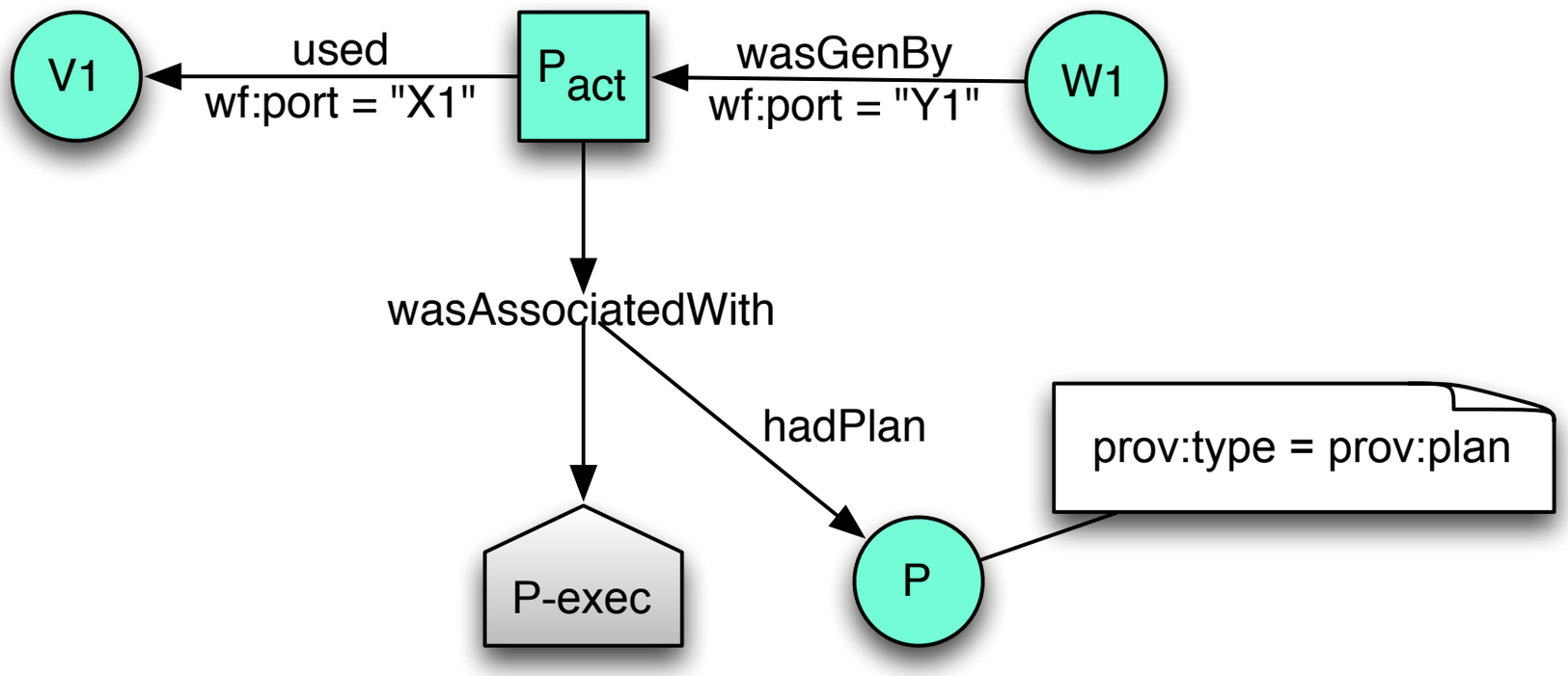



```
ex:correct       a prov:Activity; prov:used ex:dataSet1.
ex:edith         a prov:Agent, prov:Person .
ex:instructions a prov:Plan .

ex:correct prov:qualifiedAssociation [
                  a Association ;
                  prov:agent   ex:edith ;
                  prov:hadPlan ex:instructions ].
ex:dataSet2 prov:wasGeneratedBy ex:correct .
ex:dataSet2 prov:wasRevisionOf  ex:dataSet1 .
```

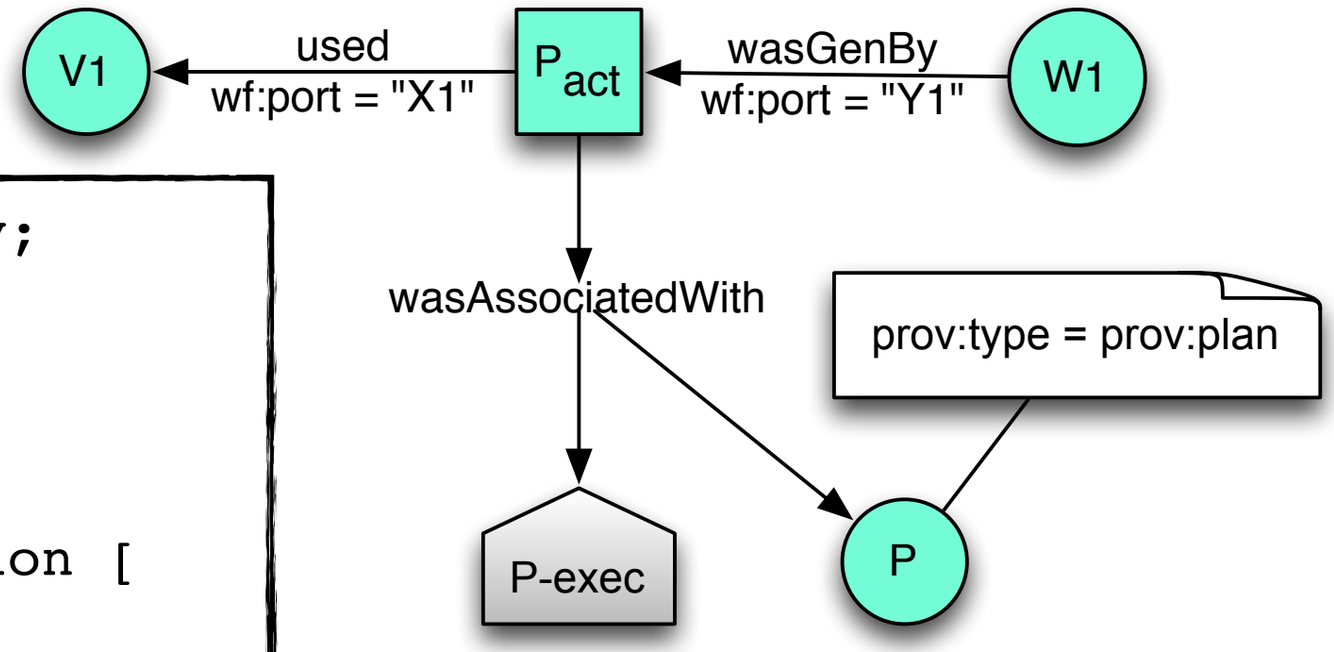
One possible rendering of the Janus example

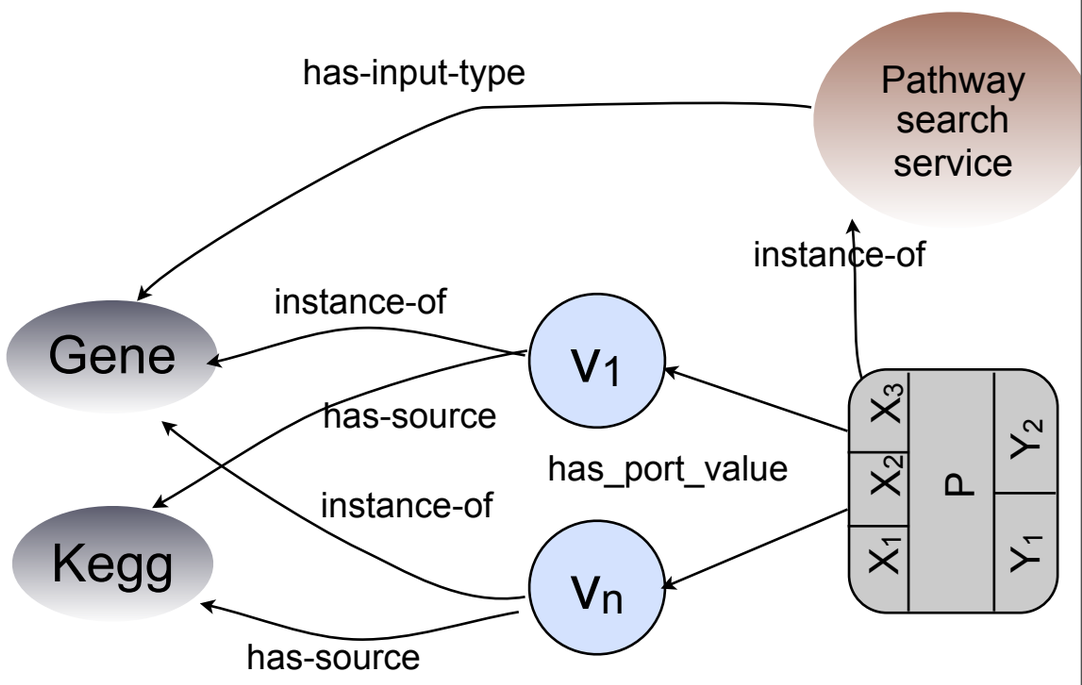




PROV-W: Workflow pattern with annotations

```
:P     a prov:Plan, prov:Entity;
       a PathwaySearchService.
:P_exec a prov:Agent.
:P_act a prov:Activity;
       prov:used  :V1;
       prov:qualifiedAssociation [
          a prov:Association;
          prov:agent    :P_exec;
          prov:hadPlan :P;  ];
        prov:qualifiedUsage [
          a prov:Usage ;
          prov:Entity :V1;
          wf:port “X1” ].
:V1    a prov:Entity;
       a :Gene; :hasSource :Kegg.
:W1    a prov:Entity;
       prov:qualifiedGeneration [
         a prov:Generation;
         prov:activity :P_act;
         wf:port “Y1”
       ];
       a :Pathway; :hasSource :Kegg.
```





The complete suite of PROV specifications

PROV-DM: The PROV Data Model

W3C Working Draft 24 July 2012

This version:
http://www.w3.org/TR/2012/WD-prov-dm-20120724/
Latest published version:
http://www.w3.org/TR/prov-dm/
Latest editor's draft:
http://dvcs.w3.org/hg/prov/raw-file/default/model/prov-dm.html
Previous version:
http://www.w3.org/TR/2012/WD-prov-dm-20120503/ (color-coded diffs)
Editors:
Luc Moreau, University of Southampton
Paolo Missier, Newcastle University
Contributors:
Khalid Belhajjame, University of Manchester
Reza B'Far, Oracle Corporation
James Cheney, University of Edinburgh
Sam Coppens, IBBT - Ghent University
Stephen Cresswell, legislation.gov.uk
Yolanda Gil, Invited Expert
Paul Groth, VU University of Amsterdam
Graham Klyne, University of Oxford
Timothy Lebo, Rensselaer Polytechnic Institute
Jim McCusker, Rensselaer Polytechnic Institute
Simon Miles, Invited Expert
James Myers, Rensselaer Polytechnic Institute
Satya Sahoo, Case Western Reserve University
Curt Tilmes, National Aeronautics and Space Administration

The complete suite of PROV specifications

PROV-DM: The PROV Data Model

W3C Working

This version:
http://www.w3
Latest published
http://www.w3
Latest editor's dra
http://dvcs.w3
Previous version:
http://www.w3
Editors:
Luc Moreau,
Paolo Missie
Contributors:
Khalid Belhaj
Reza B'Far, C
James Chene
Sam Coppen
Stephen Cres
Yolanda Gil,
Paul Groth, VU University of Amsterdam
Graham Klyne, University of Oxford
Timothy Lebo, Rensselaer Polytechnic Institute
Jim McCusker, Rensselaer Polytechnic Institute
Simon Miles, Invited Expert
James Myers, Rensselaer Polytechnic Institute
Satya Sahoo, Case Western Reserve University
Curt Tilmes, National Aeronautics and Space Administration

Constraints of the Provenance Data Model

W3C Working Draft 11 September 2012

This version:
http://www.w3.org/TR/2012/WD-prov-constraints-20120911/
Latest published version:
http://www.w3.org/TR/prov-constraints/
Latest editor's draft:
http://dvcs.w3.org/hg/prov/raw-file/default/model/prov-constraints.html
Previous version:
http://www.w3.org/TR/2012/WD-prov-constraints-20120503/ (color-coded diffs)
Editors:
James Cheney, University of Edinburgh
Paolo Missier, Newcastle University
Luc Moreau, University of Southampton
Author:
Tom De Nies, IBBT - Ghent University

The complete suite of PROV specifications

PROV-DM: The PROV Data Model

W3C Working

This version:
http://www.w3
Latest published
http://www.w3
Latest editor's dra
http://dvcs.w3
Previous version:
http://www.w3
Editors:
Luc Moreau,
Paolo Missie
Contributors:
Khalid Belhaj
Reza B'Far,
James Chene
Sam Coppen
Stephen Cres
Yolanda Gil,
Paul Groth, VU University
Graham Klyne, University o
Timothy Lebo, Rensselaer
Jim McCusker, Rensselaer
Simon Miles, Invited Expert
James Myers, Rensselaer Polytechnic Institute
Satya Sahoo, Case Western Reserve University
Curt Tilmes, National Aeronautics and Space Administration

Constraints of the Provenance Data Model

W3C W

This vers
http
Latest pu
http
Latest ed
http
Previous
http
Editors:
Jam
Pao
Luc
Author:
Tom

PROV-N: The Provenance Notation

W3C Working Draft 24 July 2012

This version:
http://www.w3.org/TR/2012/WD-prov-n-20120724/
Latest published version:
http://www.w3.org/TR/prov-n/
Latest editor's draft:
http://dvcs.w3.org/hg/prov/raw-file/default/model/prov-n.html
Previous version:
http://www.w3.org/TR/2012/WD-prov-n-20120503/ (color-coded diffs)
Editors:
Luc Moreau, University of Southampton
Paolo Missier, Newcastle University
Contributors:
James Cheney, University of Edinburgh
Stian Soiland-Reyes, University of Manchester

The complete suite of PROV specifications

PROV-DM: The PROV Data Model

W3C Working

This version:
http://www.w3
Latest published
http://www.w3
Latest editor's dra
http://dvcs.w3
Previous version:
http://www.w3
Editors:
Luc Moreau,
Paolo Missier
Contributors:
Khalid Belhaj
Reza B'Far,
James Chene
Sam Coppen
Stephen Cres
Yolanda Gil,
Paul Groth, VU University
Graham Klyne, University o
Timothy Lebo, Rensselaer
Jim McCusker, Rensselaer
Simon Miles, Invited Expert
James Myers, Rensselaer Polytechnic
Satya Sahoo, Case Western Reserve U
Curt Tilmes, National Aeronautics and

Constraints of the Provenance Data Model

W3C W

This vers
http
Latest pu
http
Latest ed
http
Previous
http
Editors:
Jam
Pao
Luc
Author:
Tom

PROV-N: The Provenance Notation

W3C Working Draft 24 July 2012

This ver
htt
Latest p
htt
Latest e
htt
Previou
htt
Editors:
Luc
Pa
Contribu
Jar
Sti

PROV-AQ: Provenance Access and Query

W3C Working Draft 19 June 2012

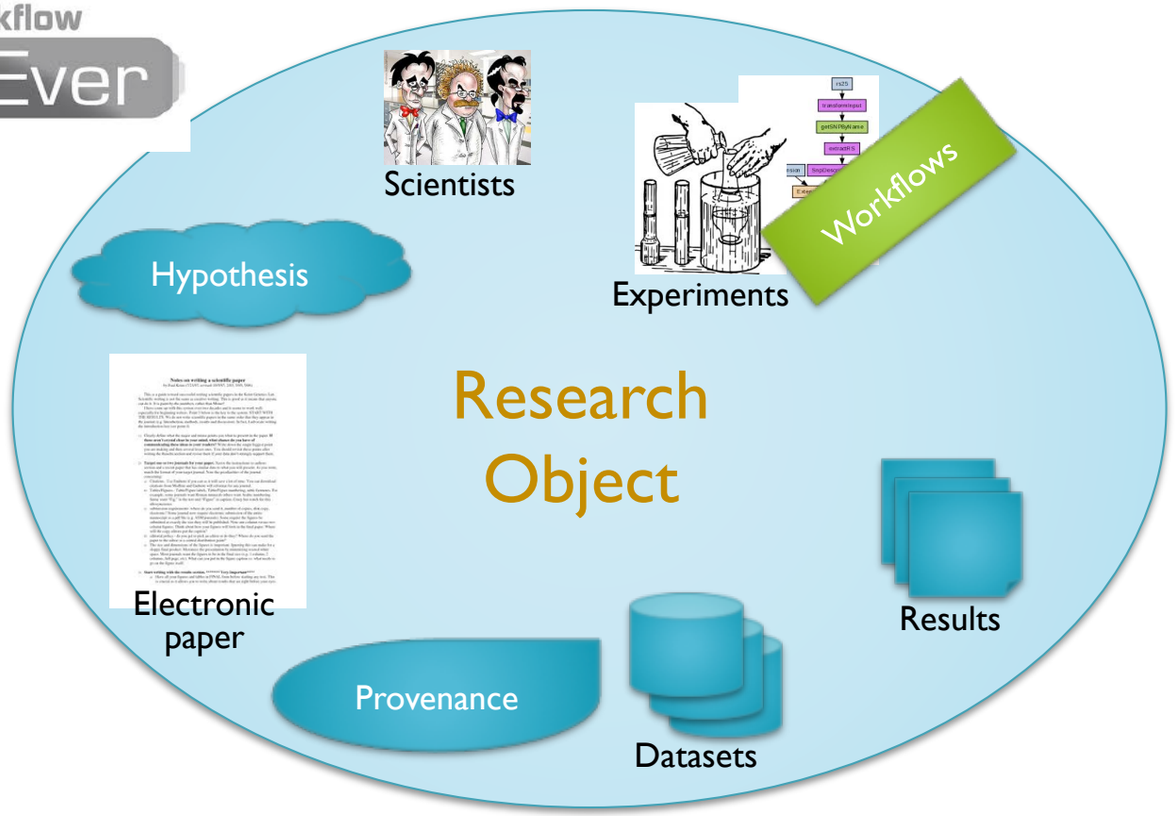
This version:
http://www.w3.org/TR/2012/WD-prov-aq-20120619/
Latest published version:
http://www.w3.org/TR/prov-aq/
Latest editor's draft:
http://dvcs.w3.org/hg/prov/raw-file/tip/paq/prov-aq.html
Previous version:
http://www.w3.org/TR/2012/WD-prov-aq-20120110/
Editors:
Graham Klyne, University of Oxford
Paul Groth, VU University Amsterdam
Authors:
Luc Moreau, University of Southampton
Olaf Hartig, Invited Expert
Yogesh Simmhan, Invited Expert
James Myers, Rensselaer Polytechnic Institute
Timothy Lebo, Rensselaer Polytechnic Institute
Khalid Belhajjame, University of Manchester
Simon Miles, Invited Expert

2: Packaging and sharing:
data + methods + provenance

- The Research Object model from Wf4Ever
  with: Khalid Belhajjame, University of Manchester

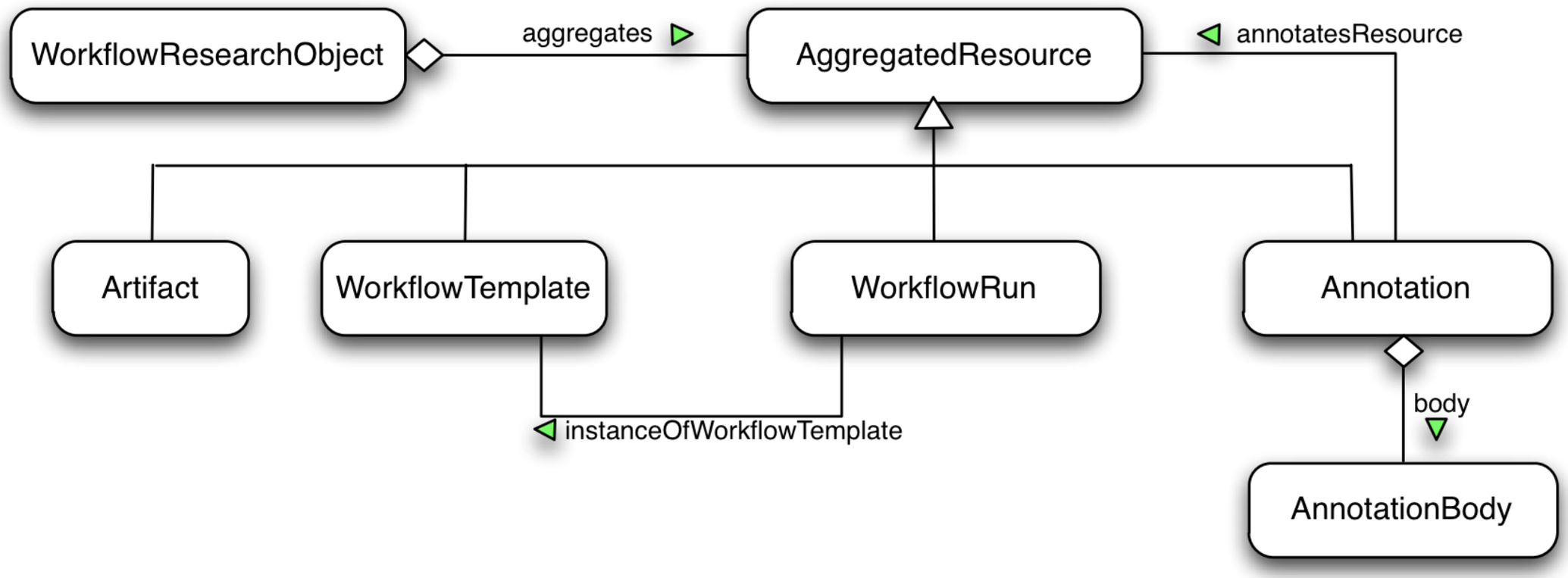
- The DataONE data preservation architecture

Research Objects

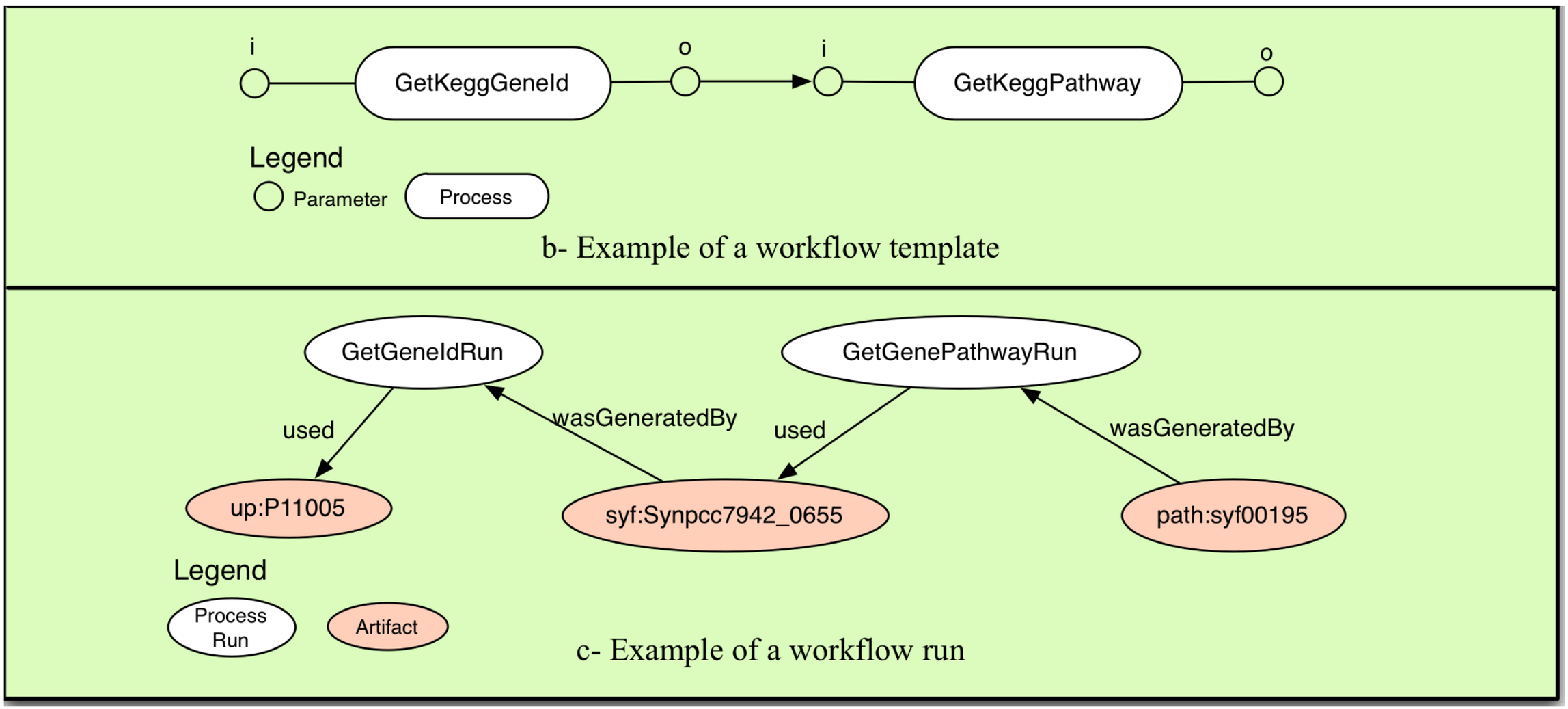


RO model specification:
http://wf4ever.github.com/ro/

RO primer:
http://wf4ever.github.com/ro-primer/

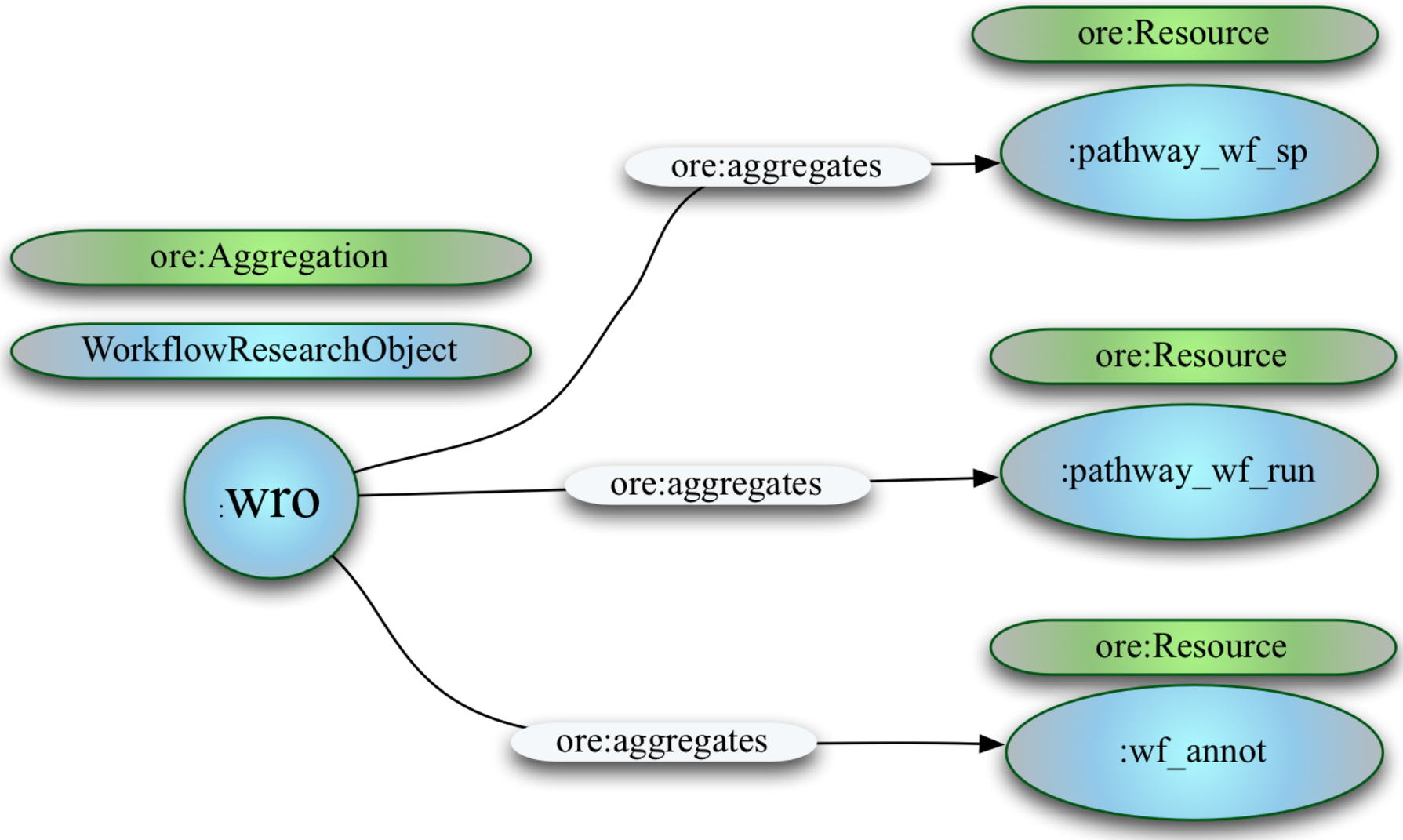


Example

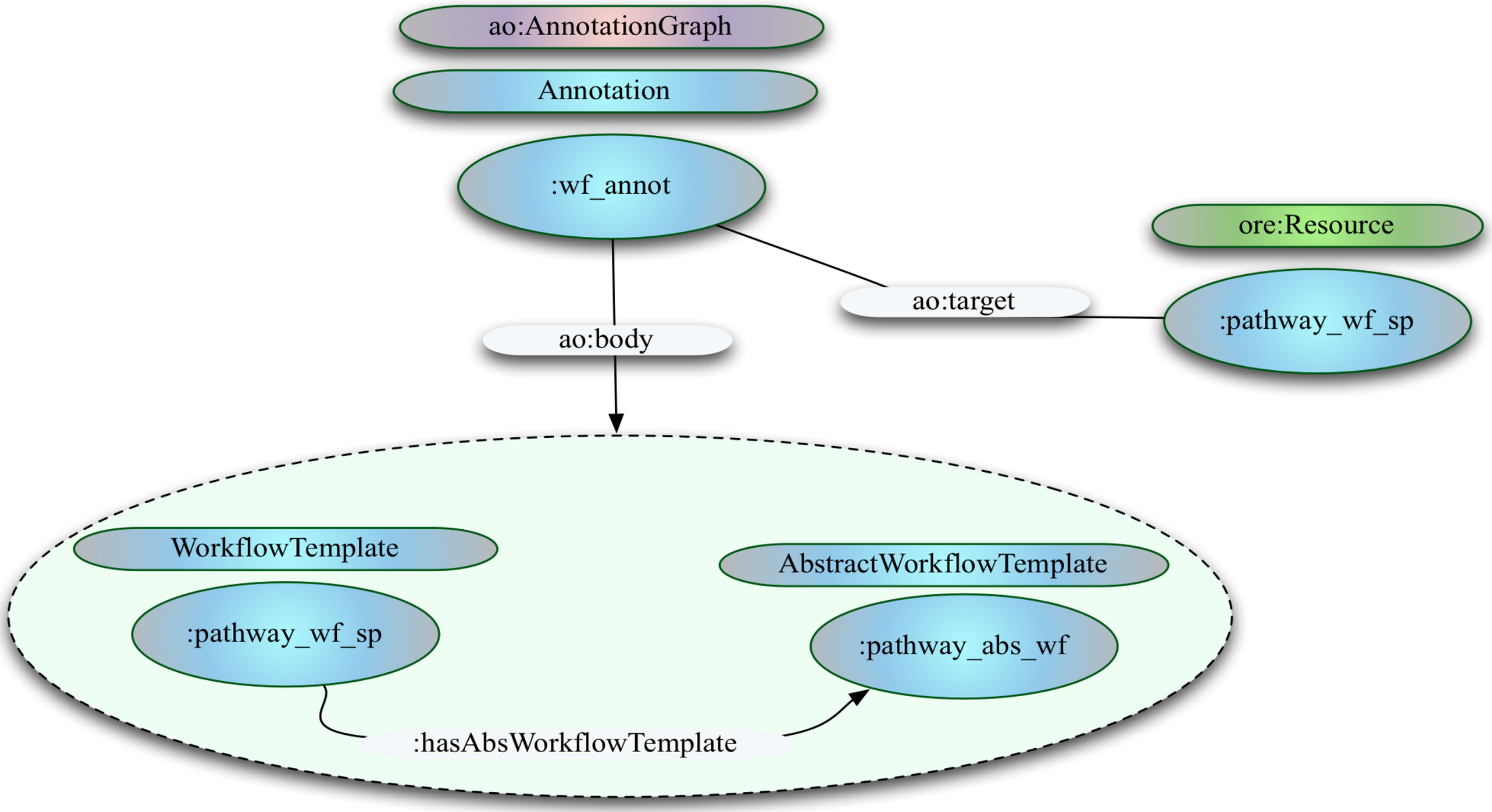

i
GetKeggGeneId
o
i
GetKeggPathway
o
Legend
Parameter
Process
b- Example of a workflow template
GetGeneIdRun
GetGenePathwayRun
used
wasGeneratedBy
used
wasGeneratedBy
up:P11005
syf:Synpcc7942_0655
path:syf00195
Legend
Process Run
Artifact
c- Example of a workflow run

Research Objects as ORE resources

- ORE = Object Exchange and Reuse
  - a small vocabulary and patterns for modelling generic “aggregation”



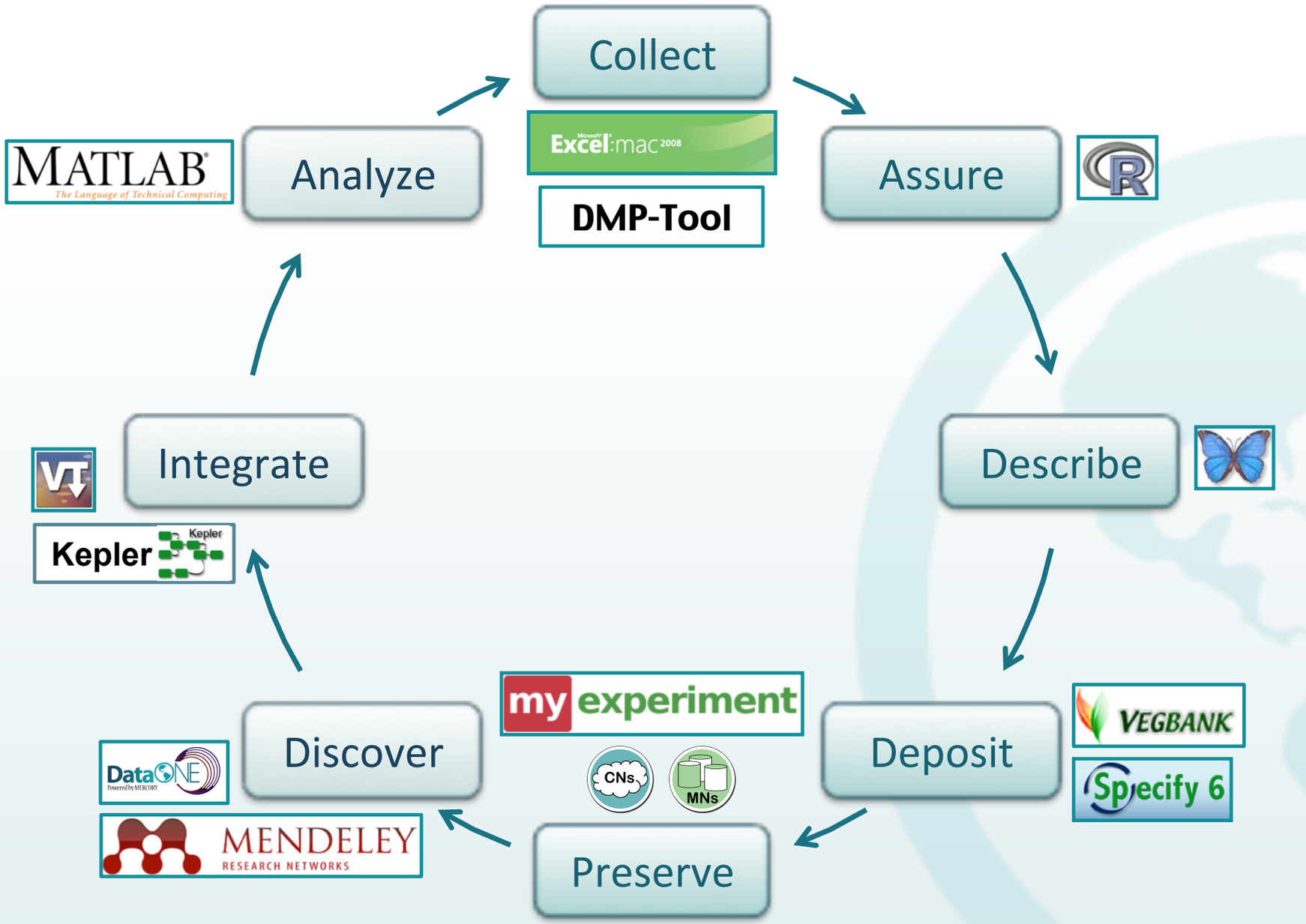
Annotations



Acknowledgement

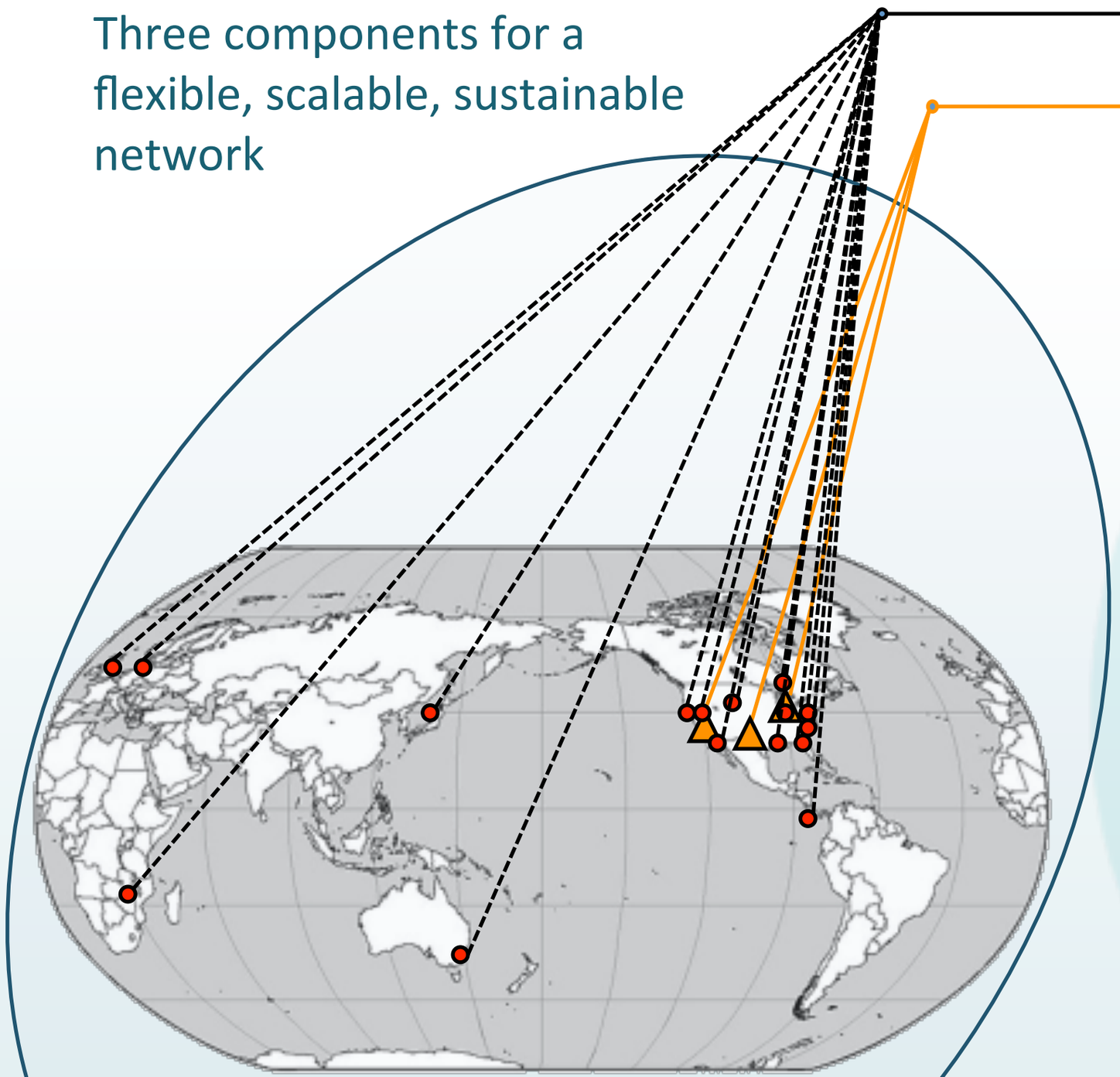
The University of Manchester
MANCHESTER 1824
iSOCO enabling the networked economy
LUMC
SUPERCOMPUTING POZNAN NETWORKING
UNIVERSITY OF OXFORD
INSTITUTO de ASTROFISICA de ANDALUCIA CSIC I·A·A
UNIVERSIDAD POLITECNICA MADRID UPM

Data Life Cycle Tool Support



Components

Three components for a flexible, scalable, sustainable network



Member Nodes

Coordinating Nodes

Investigator Toolkit

MENDELEY RESEARCH NETWORKS

VEGBANK

Specify 6

Java

my experiment

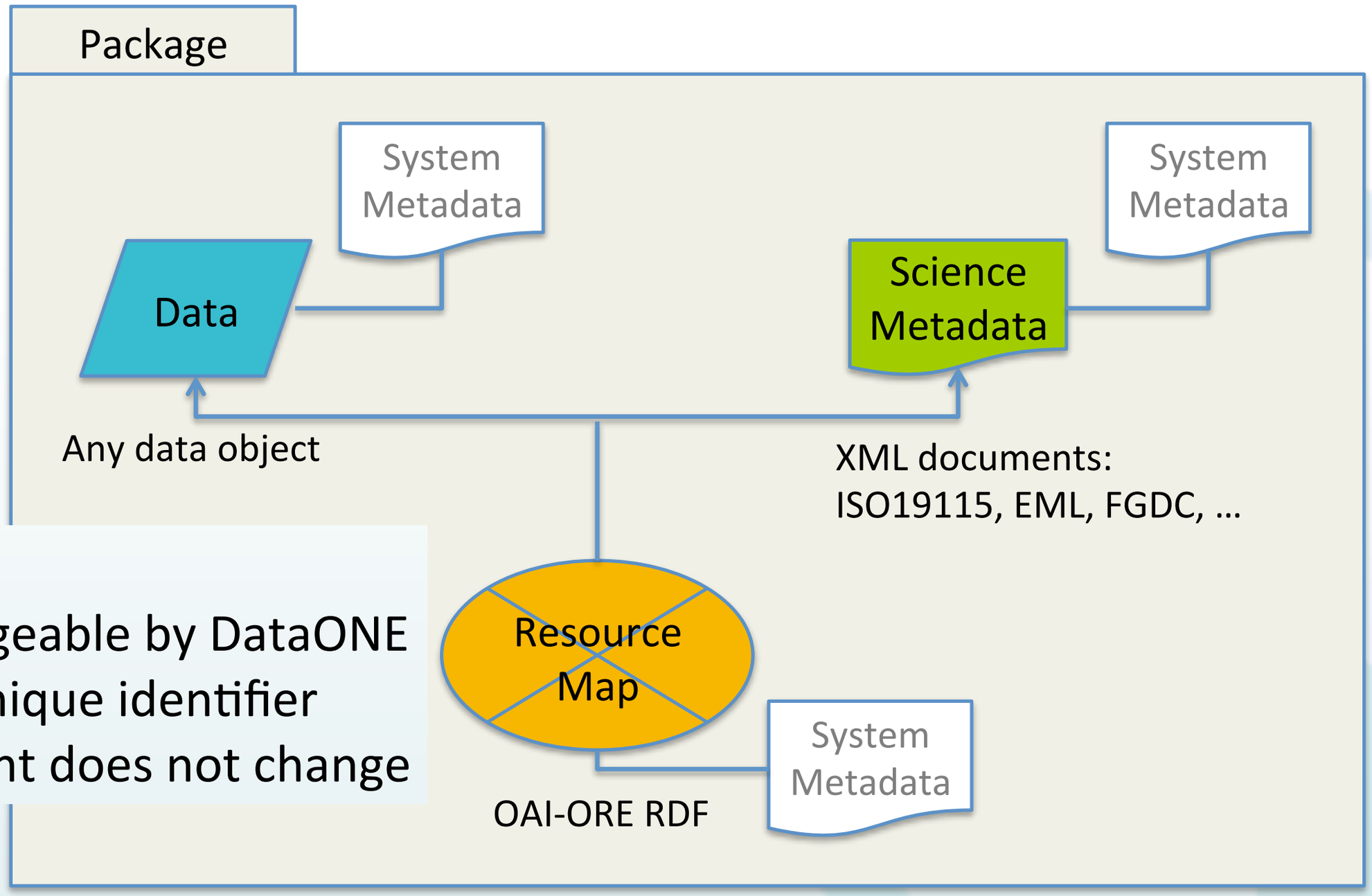
Excel:mac 2008

zotero

MATLAB The Language of Technical Computing

Kepler

Data Model

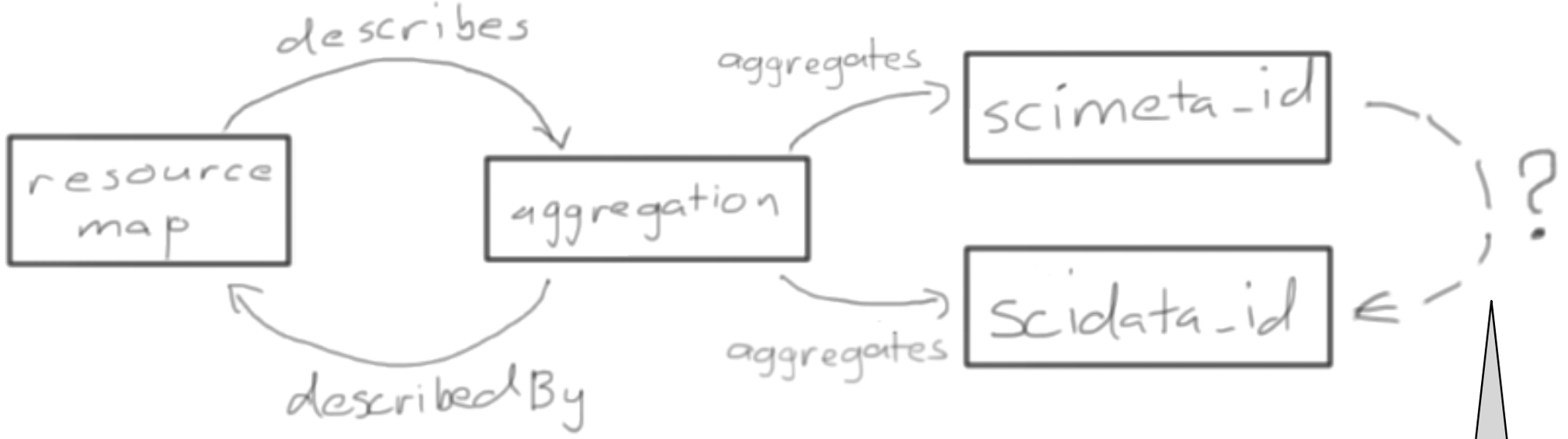


Granule:

- Manageable by DataONE
- Has unique identifier
- Content does not change

Data packages in DataONE

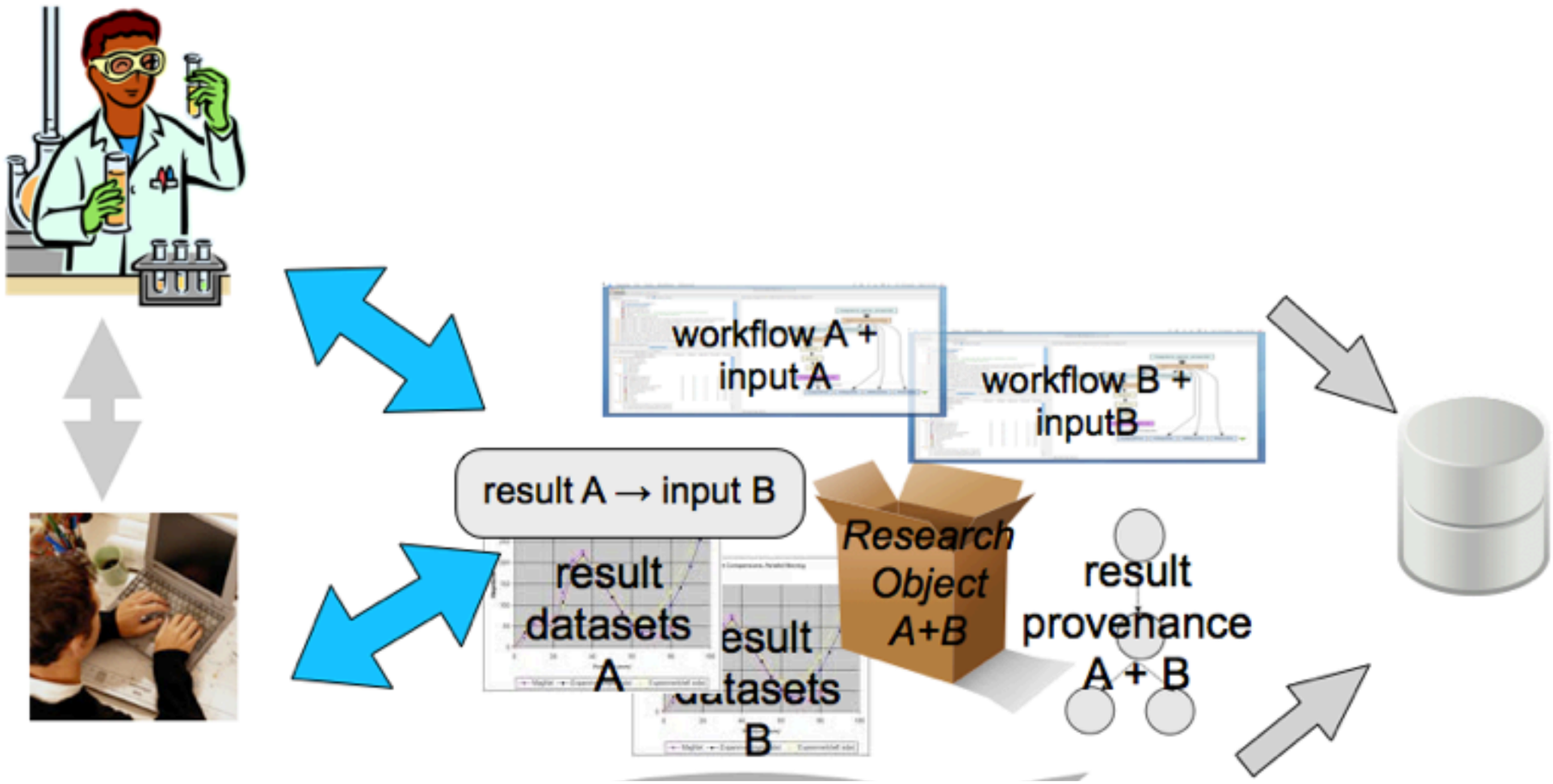
Package Content Associations Using OAI-ORE



http://mule1.dataone.org/ArchitectureDocs-current/design/DataPackage.html

Summary: Putting it all together

- Research objects fit well with DataONE packages
- Workflow and provenance fit well with Research Objects
- PROV for provenance
- *PROV-W* for workflow and provenance
- Semantic annotations fit well with PROV-W



References

[MLB+10] Missier, Paolo, Bertram Ludascher, Shawn Bowers, Manish Kumar Anand, Ilkay Altintas, Saumen Dey, Anandarup Sarkar, Biva Shrestha, and Carole Goble. "**Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science.**" In Proc.s 5th Workshop on Workflows in Support of Large-Scale Science (WORKS), 2010.

[MLB+12] Missier, Paolo, Bertram Ludascher, Shawn Bowers, Ilkay Altintas, Saumen Dey, and Michael Agun. "**Golden Trail: Retrieving the Data History That Matters from a Comprehensive Provenance Repository**." International Journal of Digital Curation 7, no. 1 (2012). http://www.dcc.ac.uk/events/idcc11.

[MSZ+10] Missier, Paolo, Satya S Sahoo, Jun Zhao, Amit Sheth, and Carole Goble. "***Janus: From Workflows to Semantic Provenance and Linked Open Data.***" In **Procs. IPAW 2010**. Troy, NY, 2010. http://www.springerlink.com/content/am3551t4q4614r47/.

[ZSM+11] Zhao, Jun, Satya S Sahoo, Paolo Missier, Amit Sheth, and Carole Goble. "***Extending Semantic Provenance into the Web of Data***." **IEEE Internet Computing** 15, no. 1 (2011): 40–48. http://doi.ieeecomputersociety.org/10.1109/MIC.2011.7.