Janus:
from Workflows to Semantic Provenance and Linked Open Data

Paolo Missier
Carole Goble
University of Manchester, UK

Jun Zhao
University of Oxford, UK

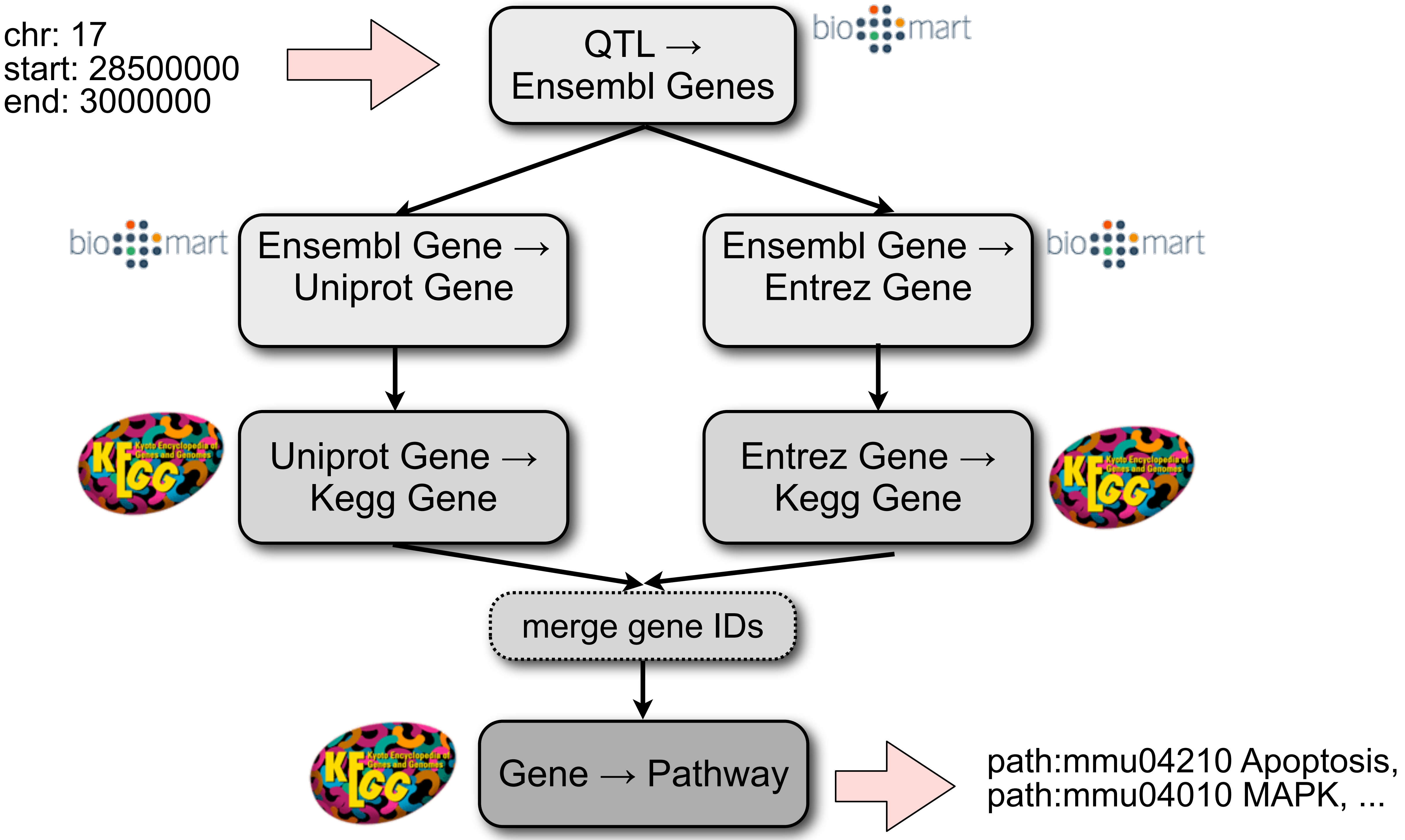
Satya S. Sahoo
Amit Sheth
Wright State University, USA

IPAW' 10
Troy, NY
June 15-16, 2010

Key ideas

- Janus:
  - a semantic provenance model with domain-specific extensions
  - designed around the Taverna workflow model
- From domain-agnostic provenance graphs
- To domain-aware graphs through explicit annotations
- From local provenance graphs and queries scoped to the graph
- To
  - Graphs published as Linked Data
  - Queries extended into the Web of Data

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

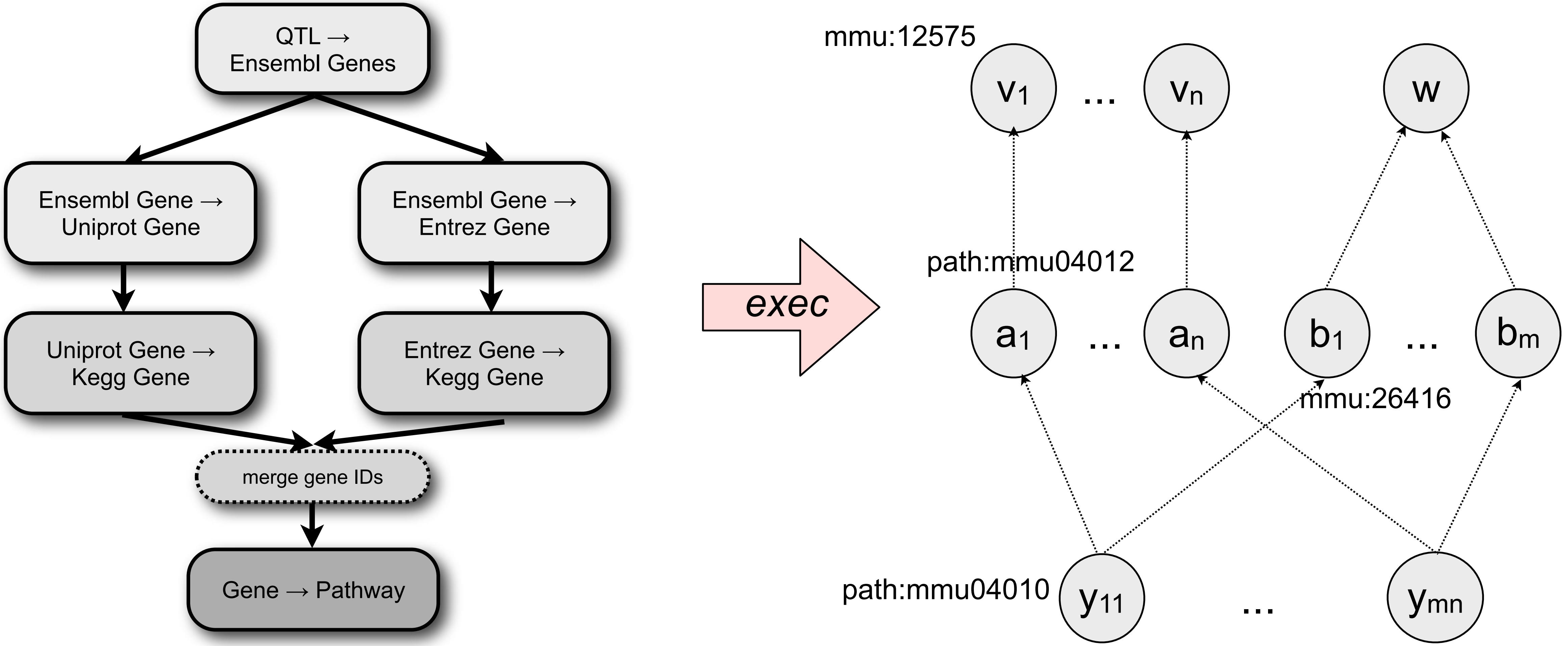
merge gene IDs

Gene → Pathway

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

biomart
biomart
biomart
KEGG Kyoto Encyclopedia of Genes and Genomes
KEGG Kyoto Encyclopedia of Genes and Genomes
KEGG Kyoto Encyclopedia of Genes and Genomes

Baseline provenance of a workflow run

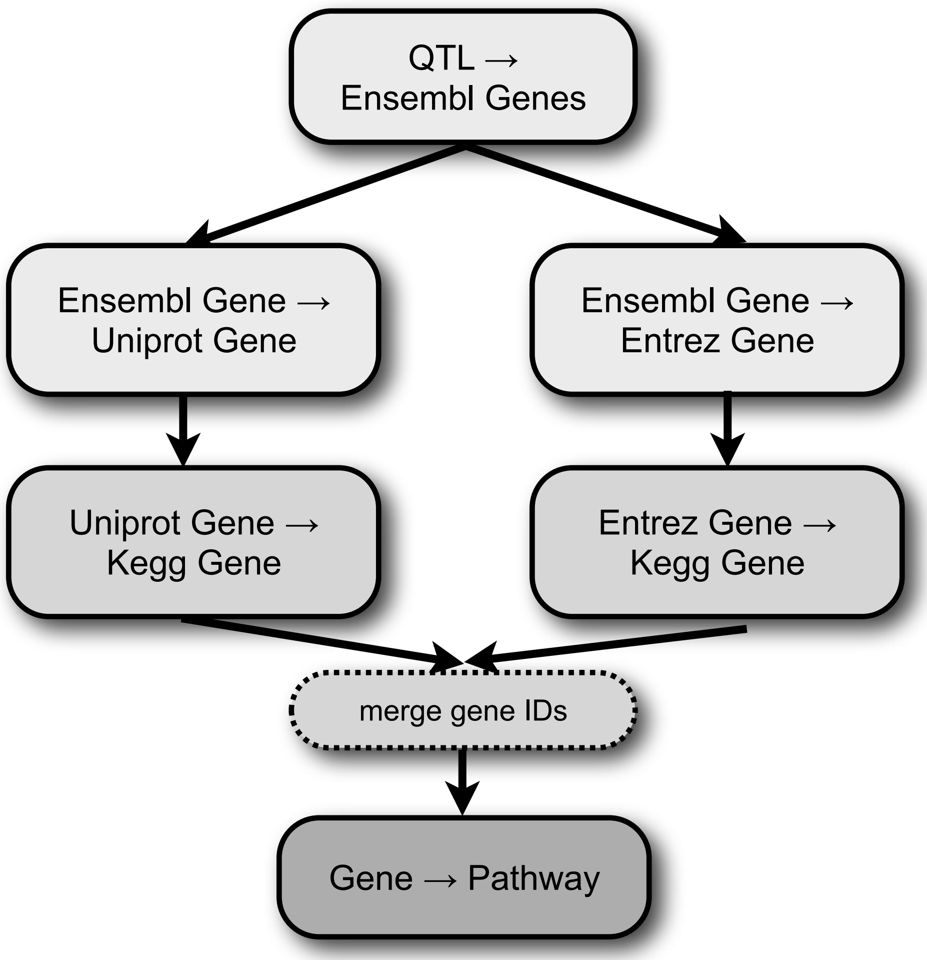


path:mmu04010→*derives_from*→mmu:26416

path:mmu04012→*derives_from*→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

Reference user questions



Q0: Find all intermediate and initial input values that contribute to the computation of a certain output value.

Q1. Find all those genes within the input QTL region that are involved in a given KEGG pathway.

Q2: Find all Uniprot-sourced genes

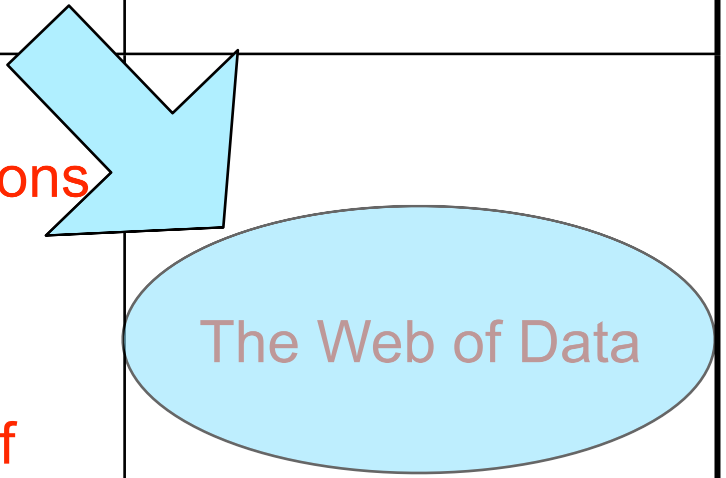
Q3: Find all Entrez genes that encode proteins involved in ATP binding (go: 0005524).

Q4: List relevant PubMed publications for the pathways listed in the result set.

Query types and model capabilities

	Query formulation effort	Annotation requirements	Query Scope
Q0	- Requires knowledge of process structure and data values - Graphical query constructor may be available	No annotations required	Single run graph or Multi-run graphs
Q1 Q2	Use of domain terms facilitates query formulation	Requires domain annotations on workflow tasks and on data values	Single run graph or Multi-run graphs
Q3 Q4	- Use of domain terms facilitates query formulation. - Can be integrated with browsers for LoD sources	- Requires domain annotations on workflow tasks and on data values - Relies on completeness of Linked Data Sources	The Web of Data

Query types and model capabilities

	Query formulation effort	Annotation requirements	Query Scope
Q0	- Requires knowledge of process structure and data values - Graphical query constructor may be available	No annotations required	Single run graph or Multi-run graphs
Q1 Q2	Use of domain terms facilitates query formulation	Requires domain annotations on workflow tasks and on data values	Single run graph or Multi-run graphs
Q3 Q4	- Use of domain terms facilitates query formulation. - Can be integrated with browsers for LoD sources	- Requires domain annotations on workflow tasks and on data values - Relies on completeness of Linked Data Sources	 The Web of Data

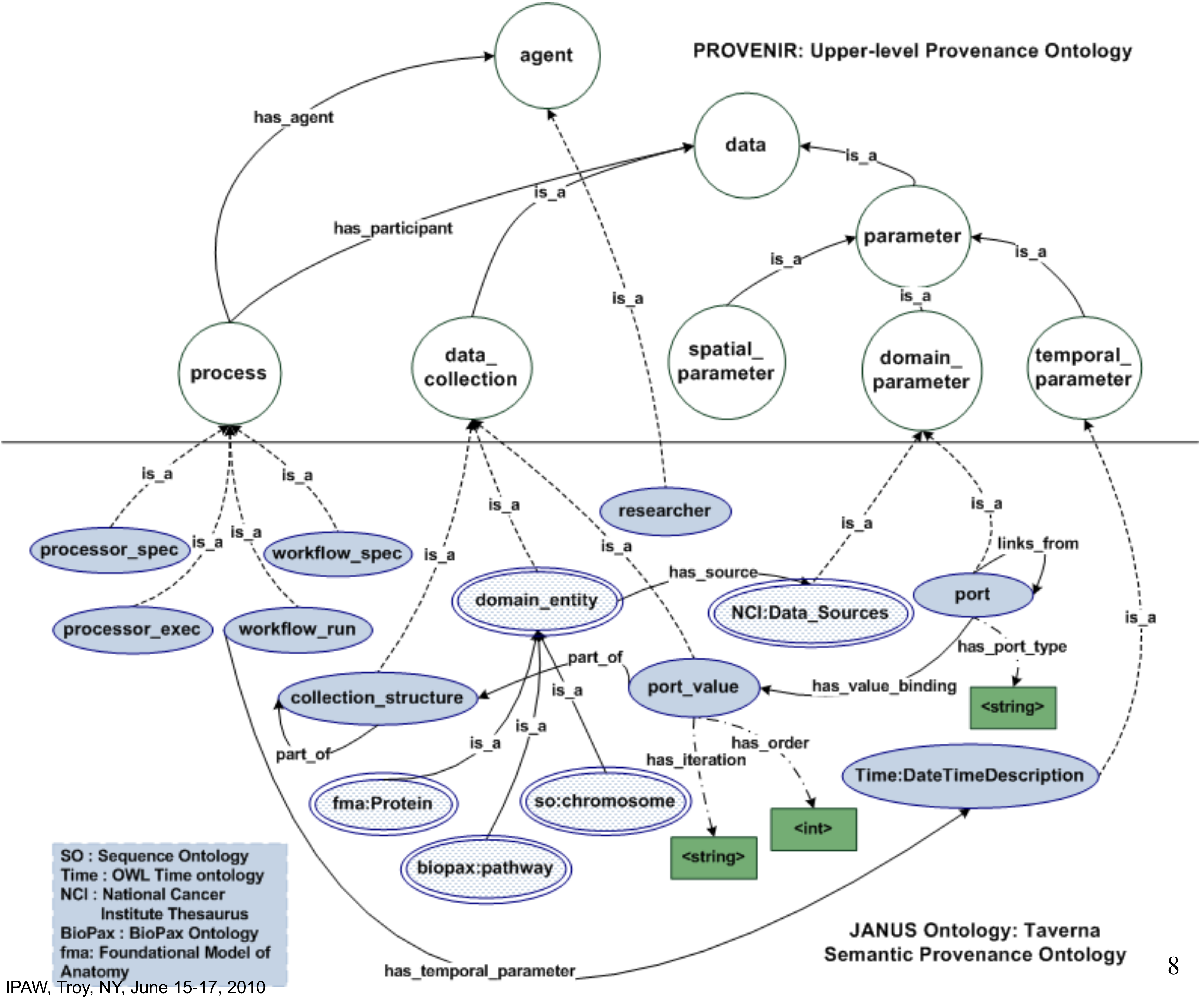
Janus: baseline model of provenance

- The semantic provenance model is an OWL ontology
  - defined for domain-agnostic provenance graphs
  - naturally extensible to domain concepts

- extends the Provenir upper ontology [*]
  - Itself an extension of the Basic Formal Ontology (BFO)
    - abstract concepts include *data*, *process*, and *agent*
  - Provenir adds 11 types of relationships:
    - partonomy relations
    - temporal information
    - precedence
    - causal relationships
    - ...

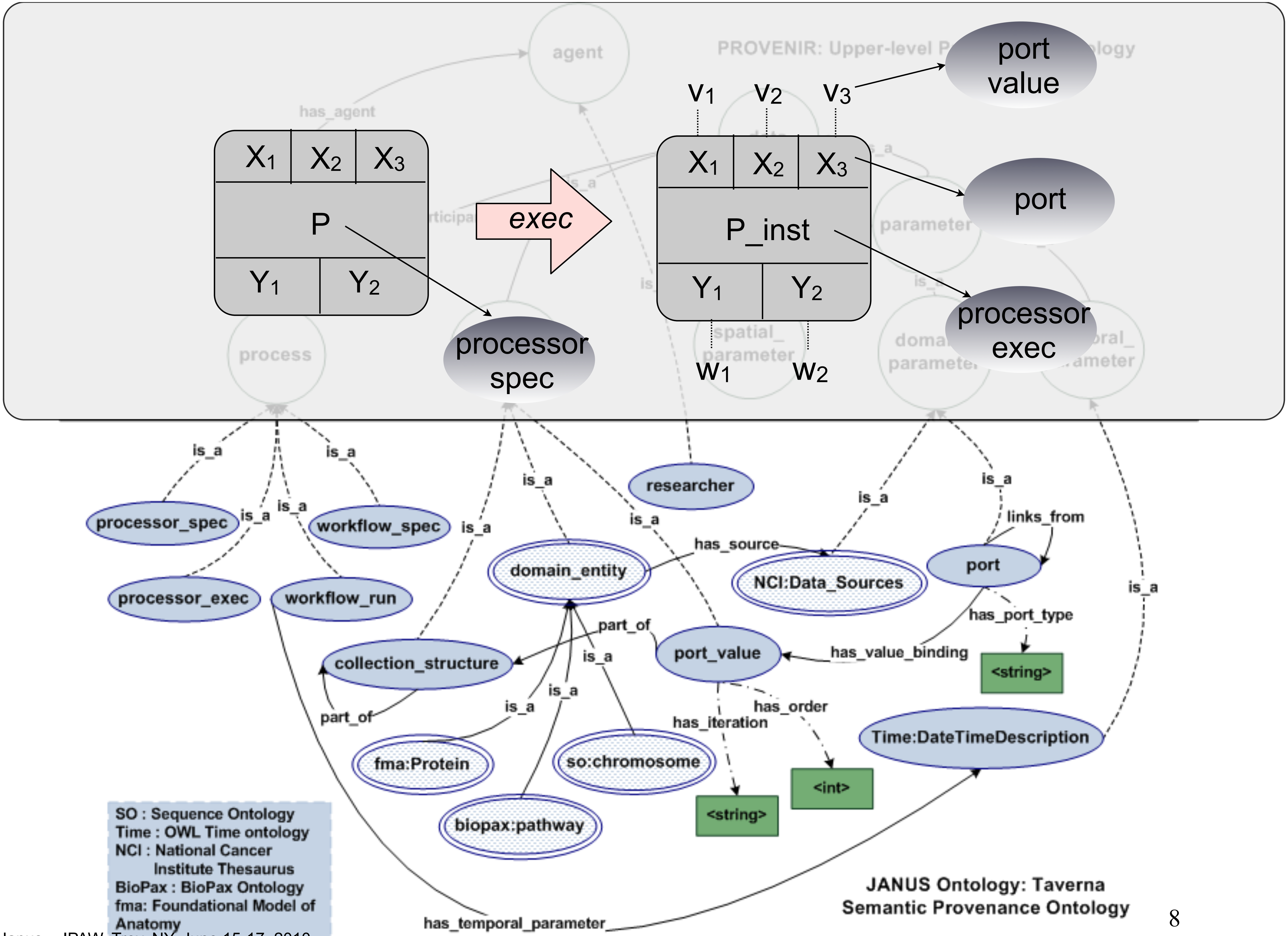
[*] S. Sahoo and A. Sheth. Provenir ontology: Towards a Framework for eScience Provenance Management, Knoesis Center Tech Report, 2009.

Janus structure



SO : Sequence Ontology
Time : OWL Time ontology
NCI : National Cancer Institute Thesaurus
BioPax : BioPax Ontology
fma: Foundational Model of Anatomy

Janus structure



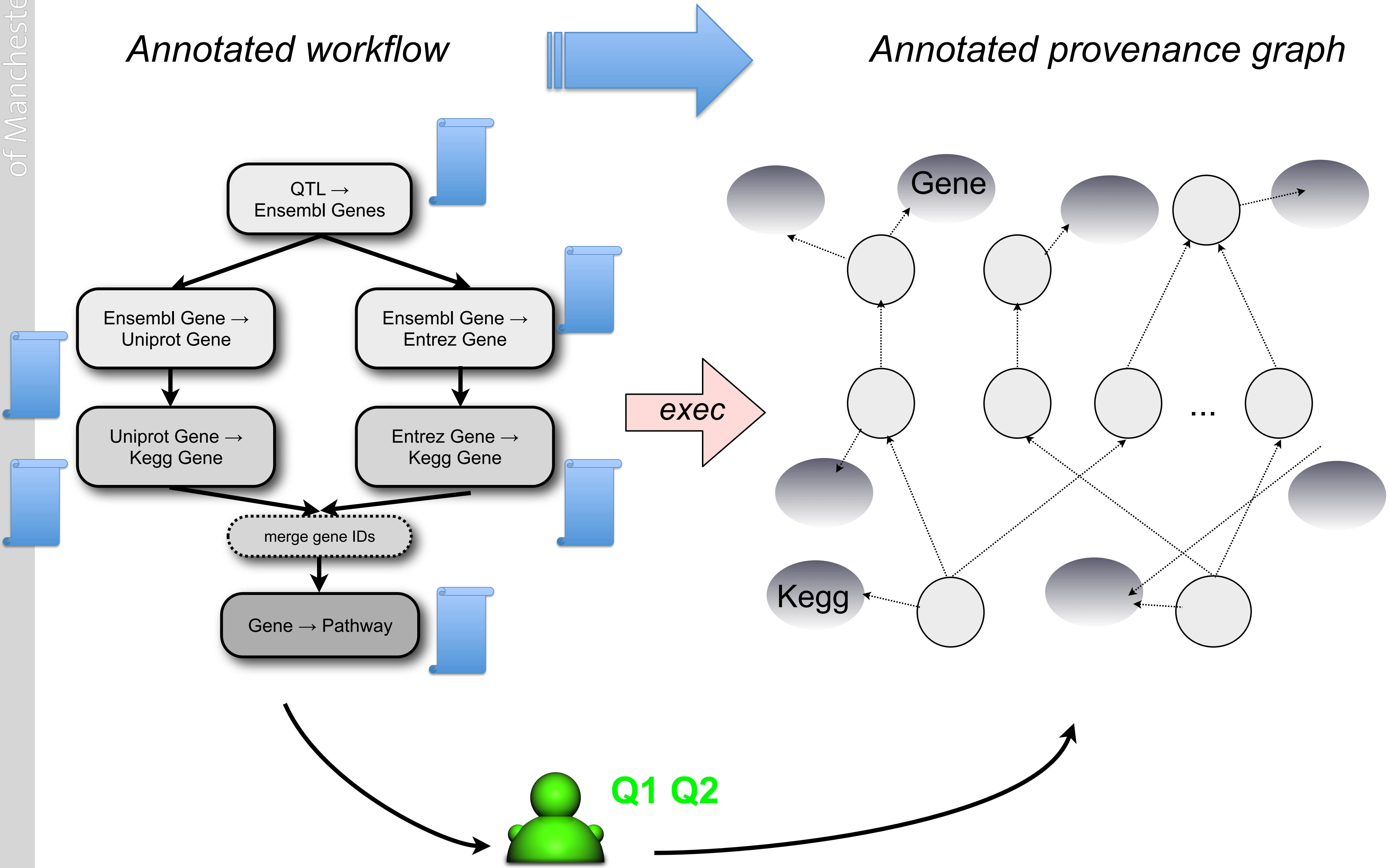
Example Janus domain-agnostic fragment

```
<rdf:Description rdf:about="http://purl.org/net/taverna/janus/remove_Nulls">
<janus:has_execution rdf:resource="http://purl.org/net/taverna/janus/remove_Nulls"/>
<knoesis:has_parameter rdf:resource="http://purl.org/net/taverna/janus/remove_Nulls/output"/>
<knoesis:has_parameter rdf:resource="http://purl.org/net/taverna/janus/remove_Nulls/input"/>
<obo:part_of rdf:resource="http://purl.org/net/taverna/janus/e589d90b-01f2-4de6-..."/>
<rdf:type rdf:resource="http://purl.org/net/taverna/janus#processor_spec"/>


<rdf:Description rdf:about="http://purl.org/net/taverna/janus/remove_Nulls/input">
<janus:has_value_binding rdf:resource="http://purl.org/net/taverna/janus/test1625"/>
<janus:links_from rdf:resource="http://purl.org/net/taverna/janus/merge_entrez_genes/
concatenated"/>
<janus:is_processor_input rdf:datatype="http://www.w3.org/2001/
XMLSchema#boolean">true</janus:is_processor_input>
<rdf:type rdf:resource="http://purl.org/net/taverna/janus#port"/>


<rdf:Description rdf:about="http://purl.org/net/taverna/janus/test1625">
<janus:has_iteration rdf:datatype="http://www.w3.org/2001/XMLSchema#string">[]</
janus:has_iteration>
<rdf:type rdf:resource="http://purl.org/net/taverna/janus#port_value"/>
</rdf:Description>
```

Annotated provenance



Annotated workflow

Annotated provenance graph

QTL → Ensembl Genes

Ensembl Gene → Uniprot Gene

Ensembl Gene → Entrez Gene

Uniprot Gene → Kegg Gene

Entrez Gene → Kegg Gene

merge gene IDs

Gene → Pathway

exec

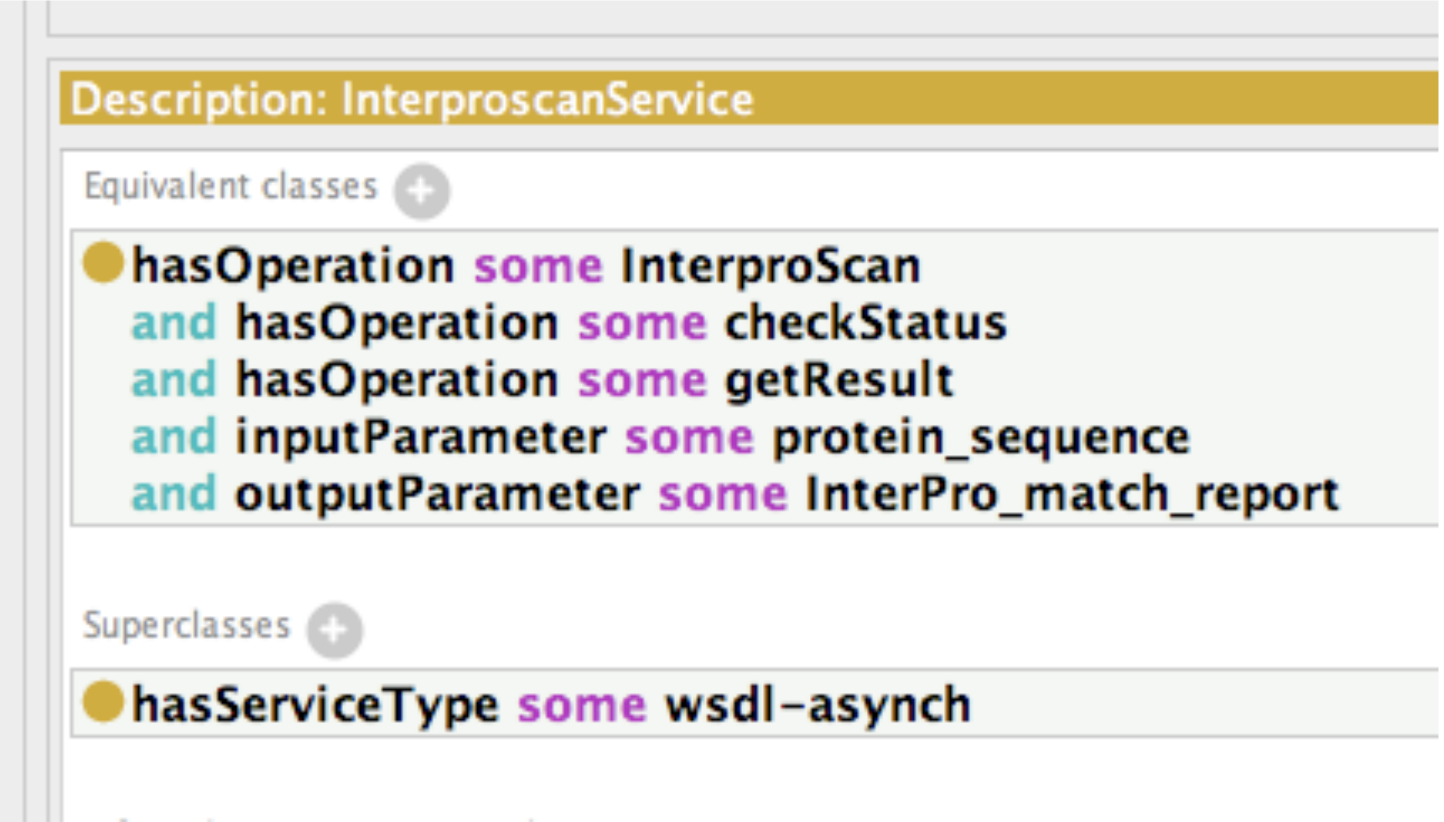
Gene

Kegg

…

Q1 Q2

Desiderata: from structure to values annotations



Description: InterproscanService

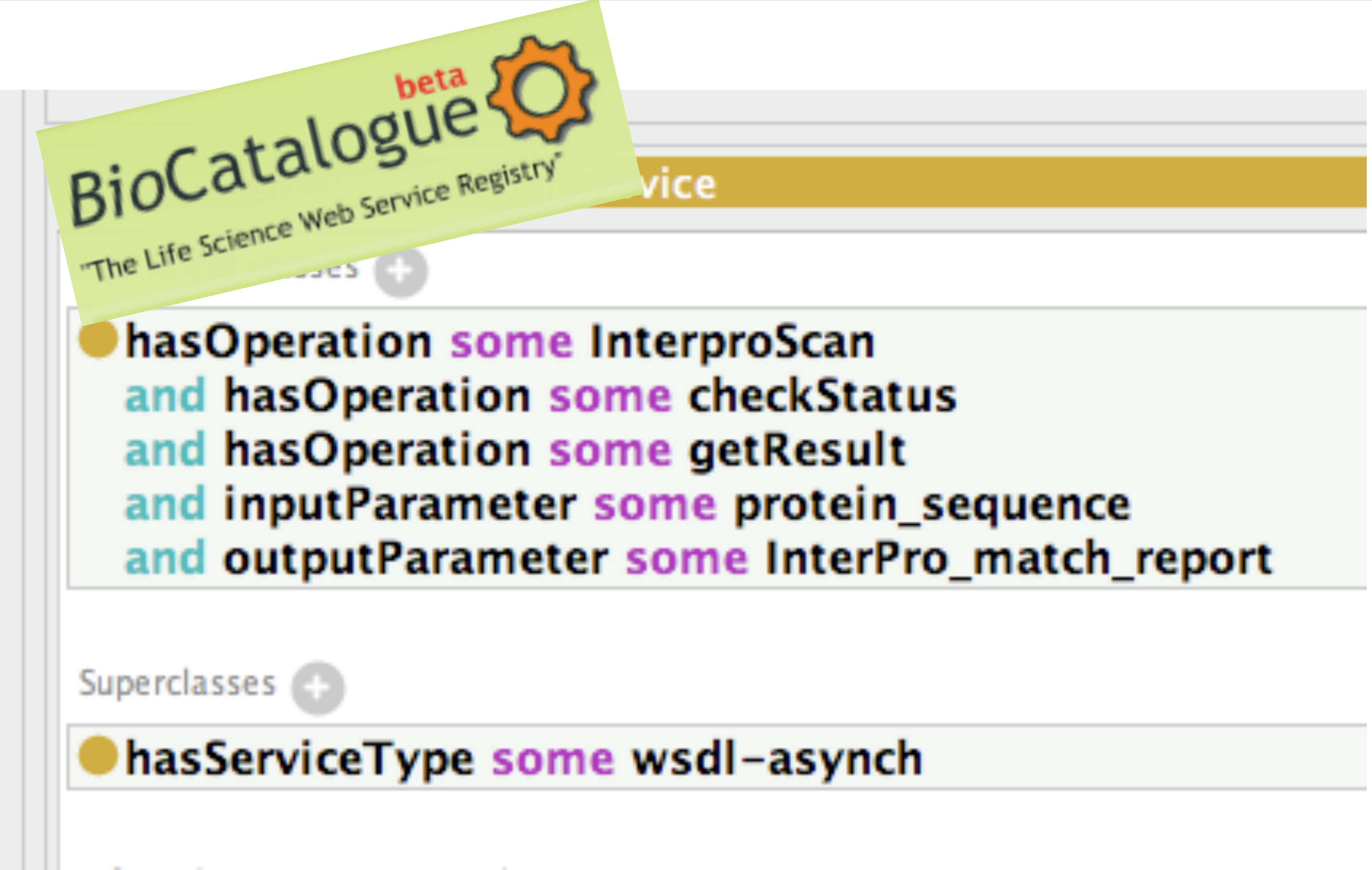
Equivalent classes

● hasOperation some InterproScan
and hasOperation some checkStatus
and hasOperation some getResult
and inputParameter some protein_sequence
and outputParameter some InterPro_match_report

Superclasses

● hasServiceType some wsdl-asynch

Desiderata: from structure to values annotations

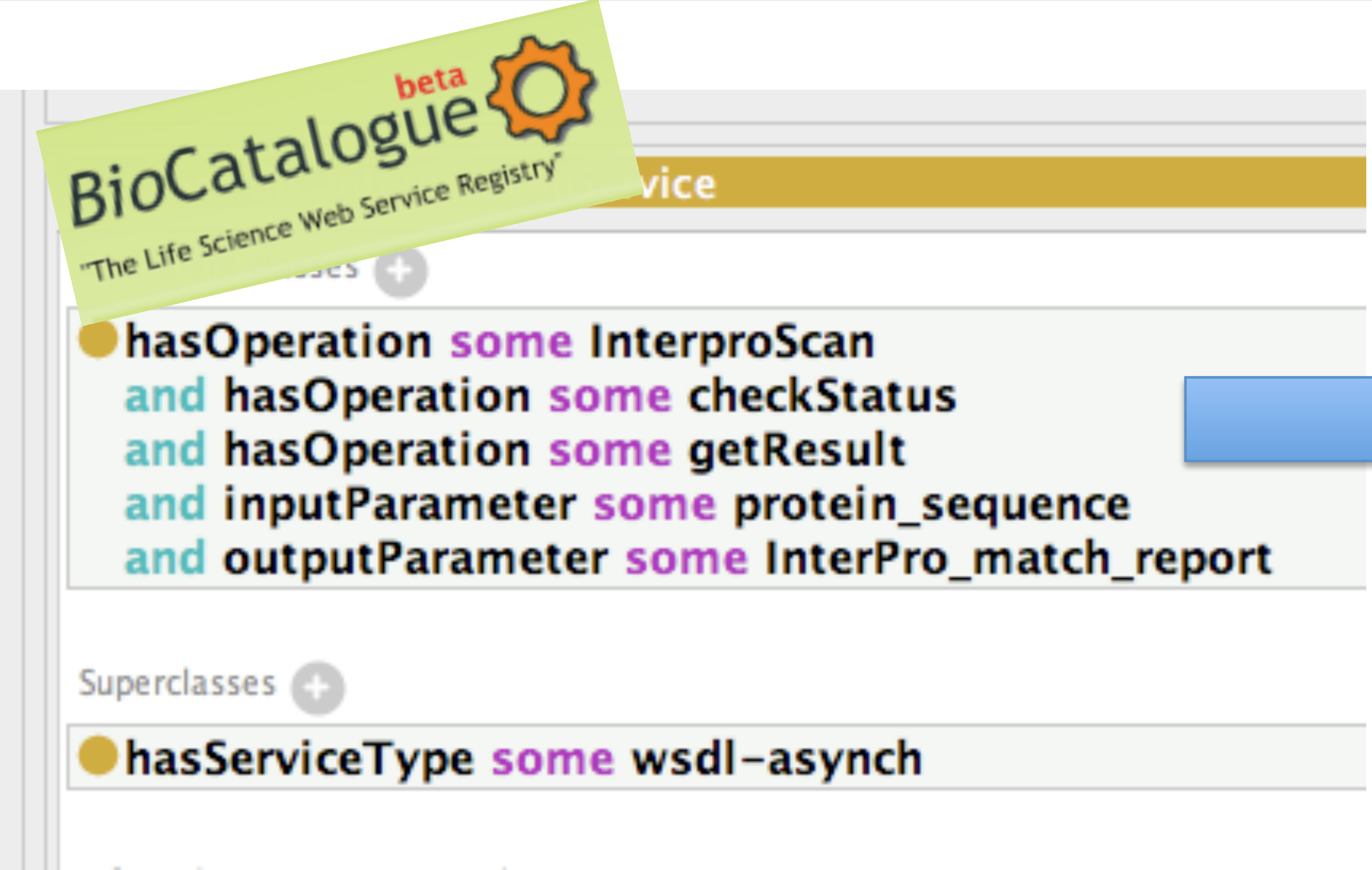


```
hasOperation some InterproScan
and hasOperation some checkStatus
and hasOperation some getResult
and inputParameter some protein_sequence
and outputParameter some InterPro_match_report
```

Superclasses

```
hasServiceType some wsdl-asynch
```

Desiderata: from structure to values annotations



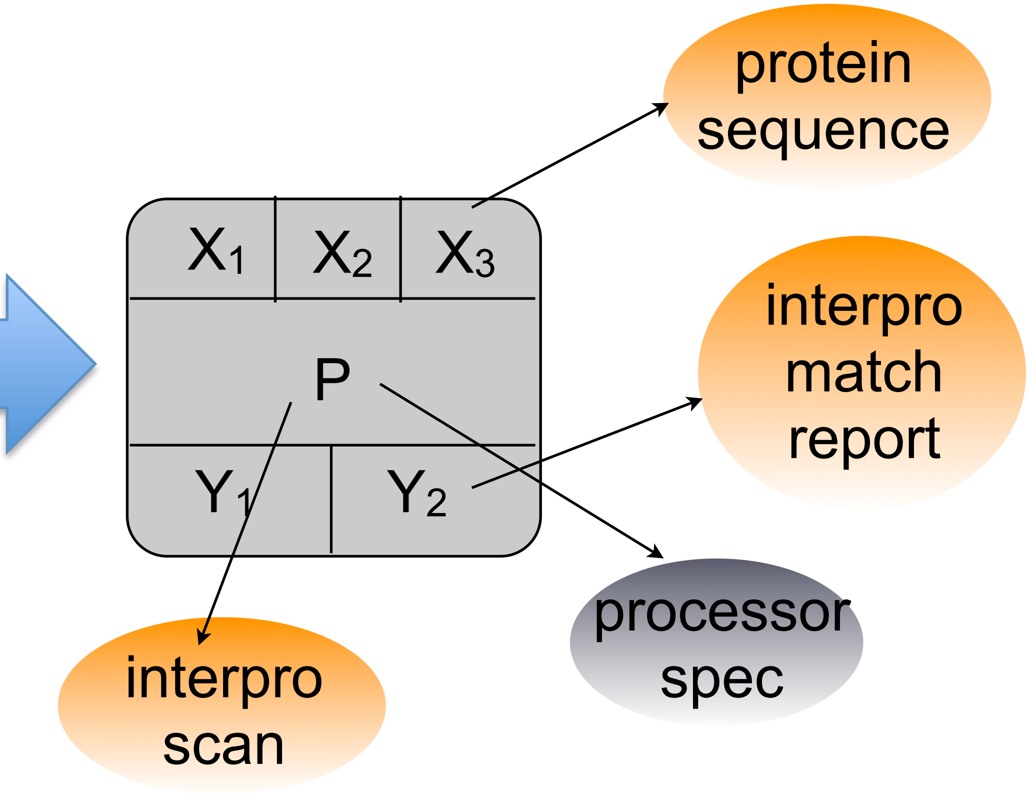
BioCatalogue beta

"The Life Science Web Service Registry"

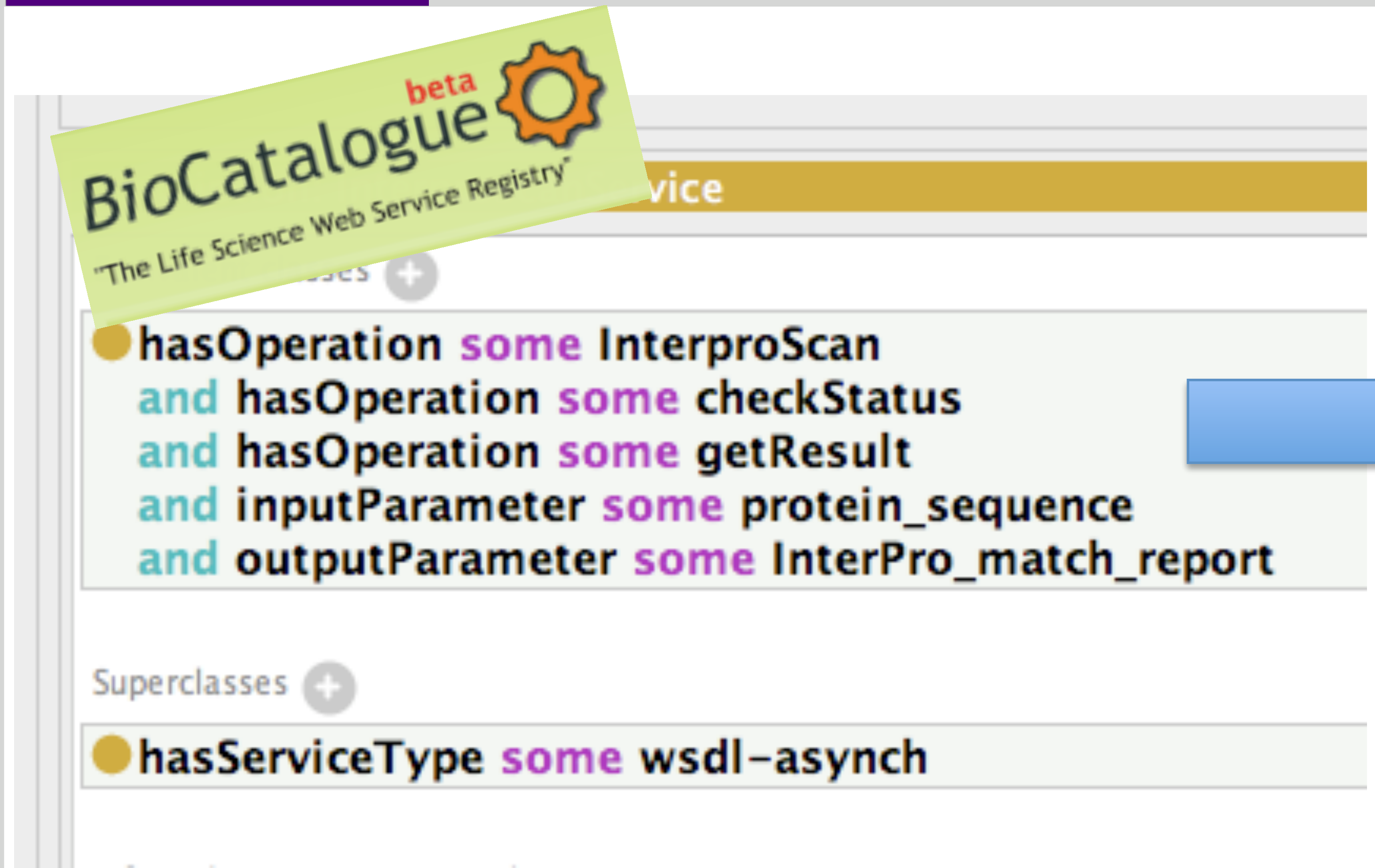
hasOperation some InterproScan
and hasOperation some checkStatus
and hasOperation some getResult
and inputParameter some protein_sequence
and outputParameter some InterPro_match_report

Superclasses

hasServiceType some wsdl-asynch



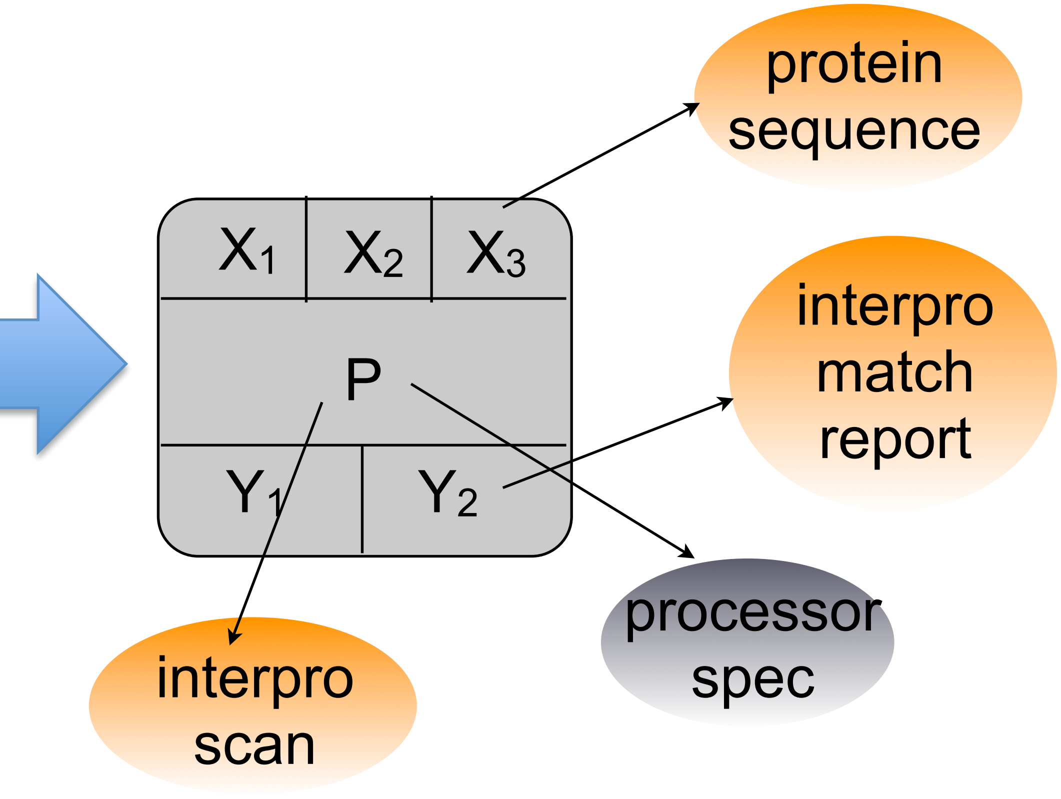
Desiderata: from structure to values annotations

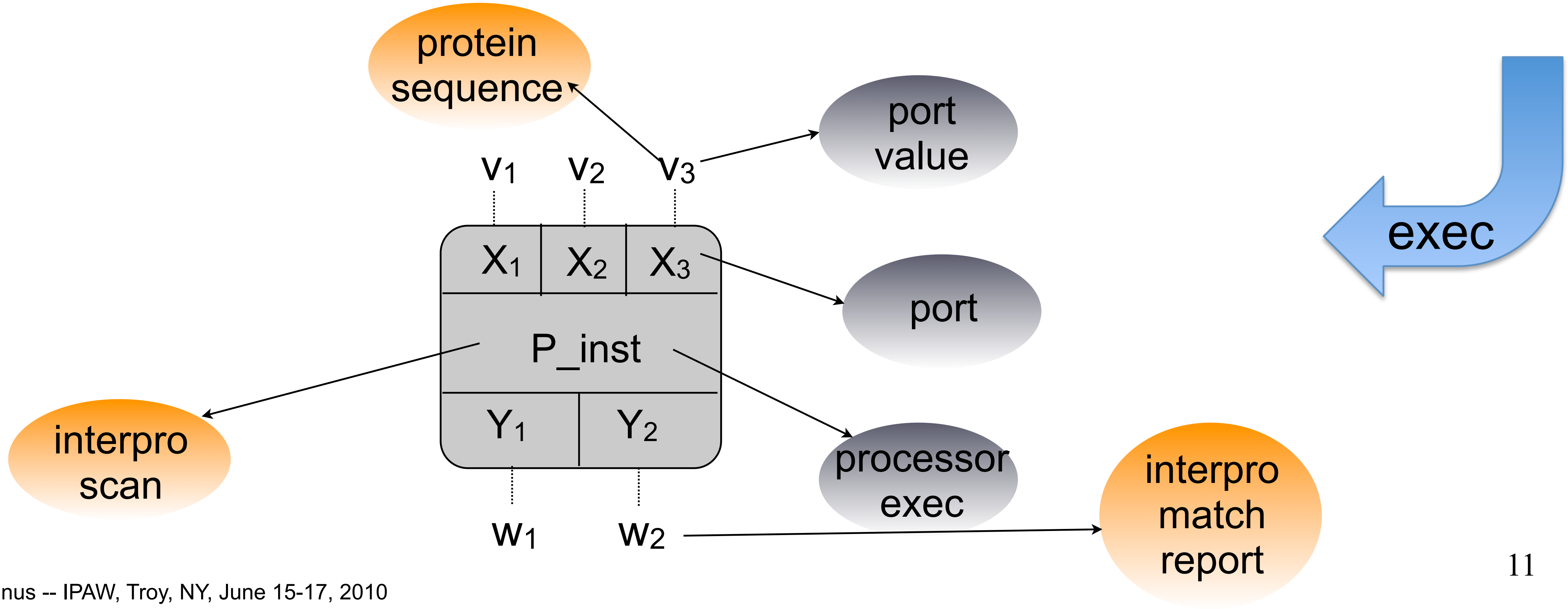


hasOperation some InterproScan
and hasOperation some checkStatus
and hasOperation some getResult
and inputParameter some protein_sequence
and outputParameter some InterPro_match_report

Superclasses

hasServiceType some wsdl-asynch



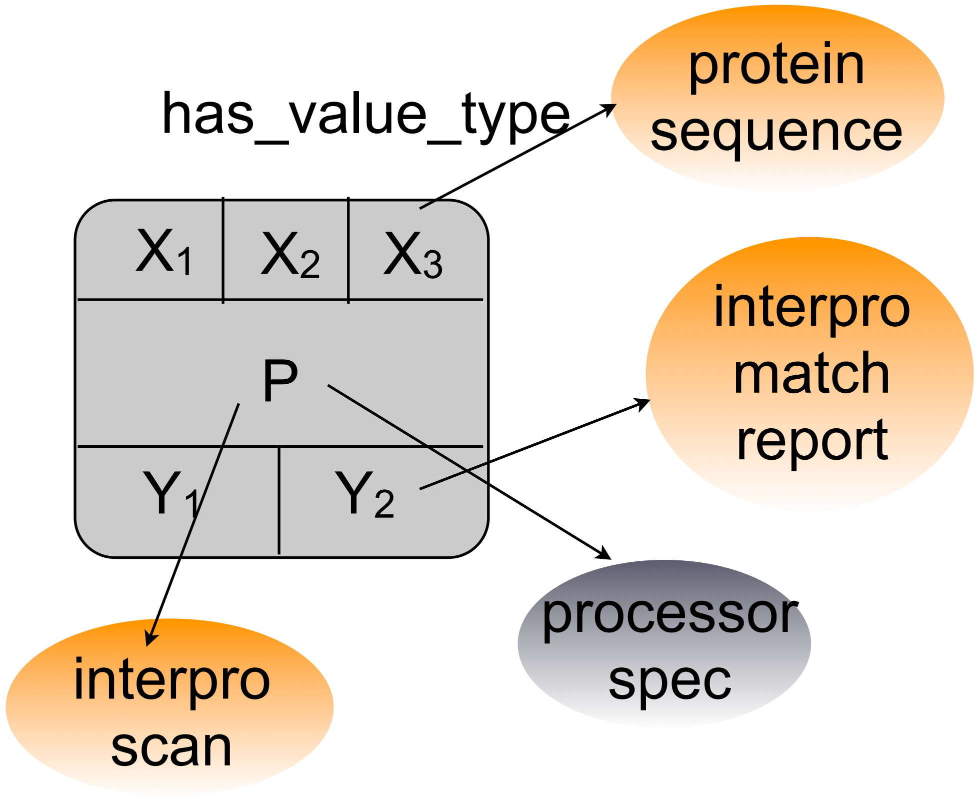


MANCHESTER 1824

Annotations propagation rules

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

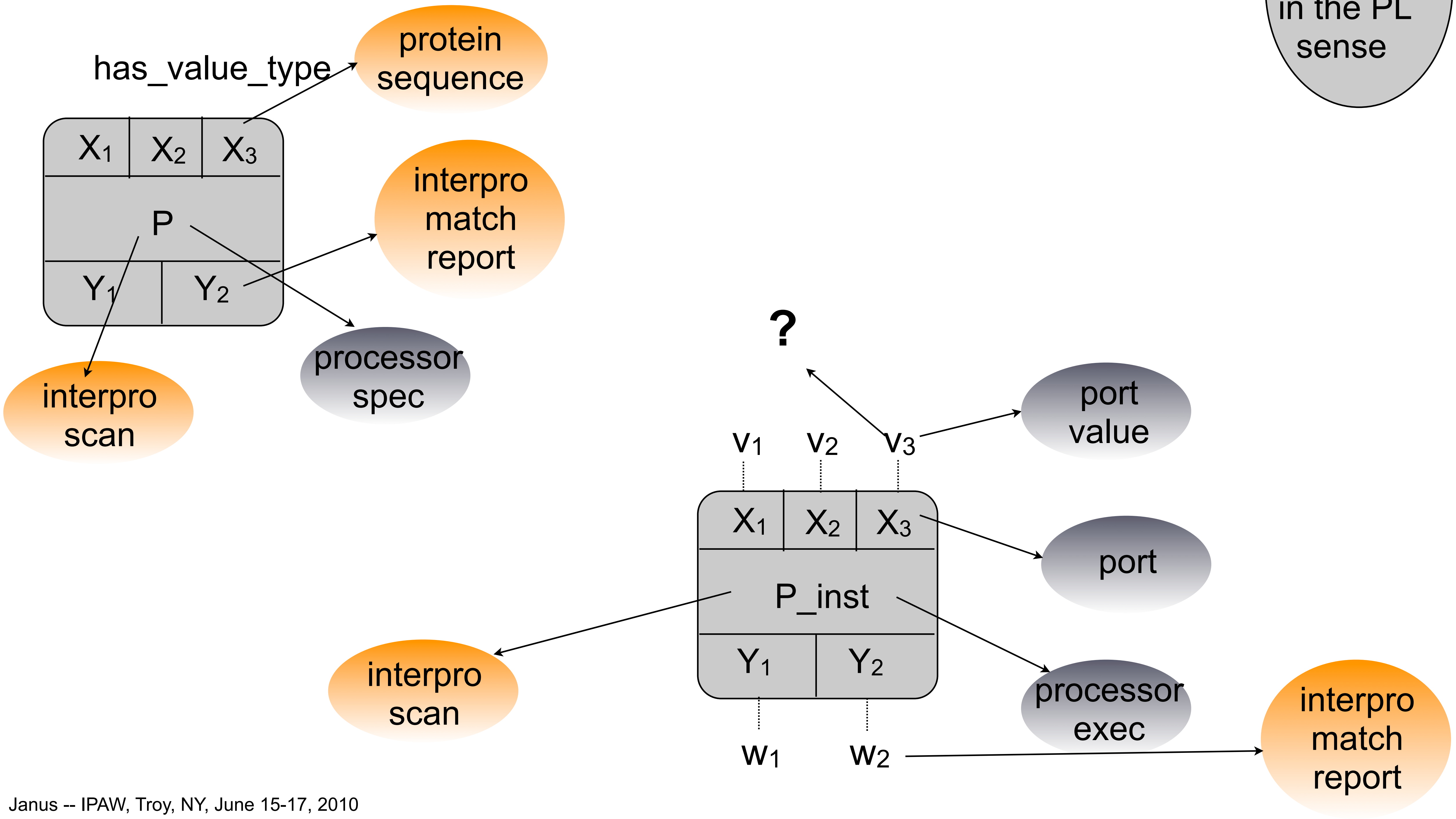
denotes data type in the PL sense



Annotations propagation rules

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

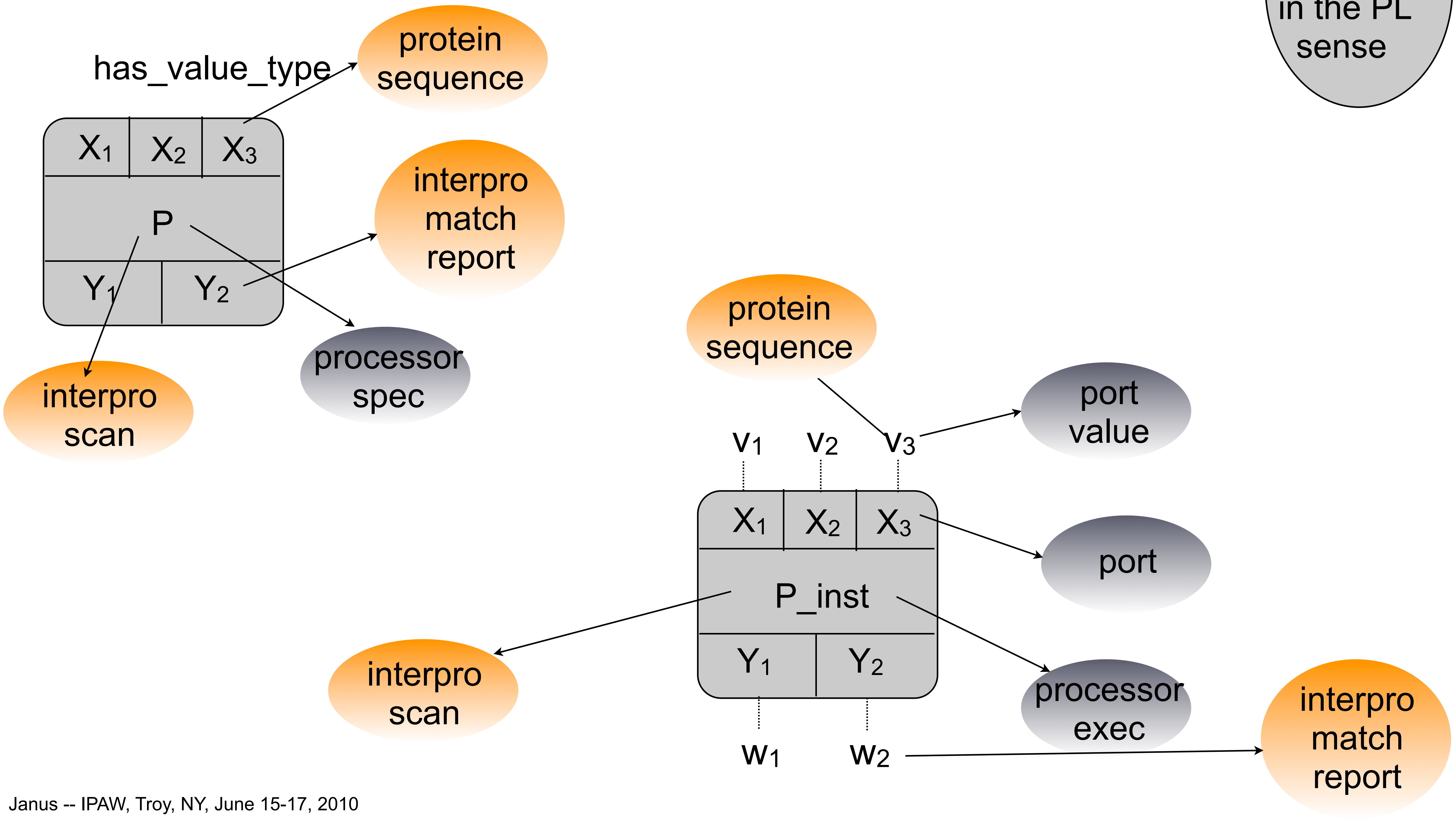
denotes data type in the PL sense



Annotations propagation rules

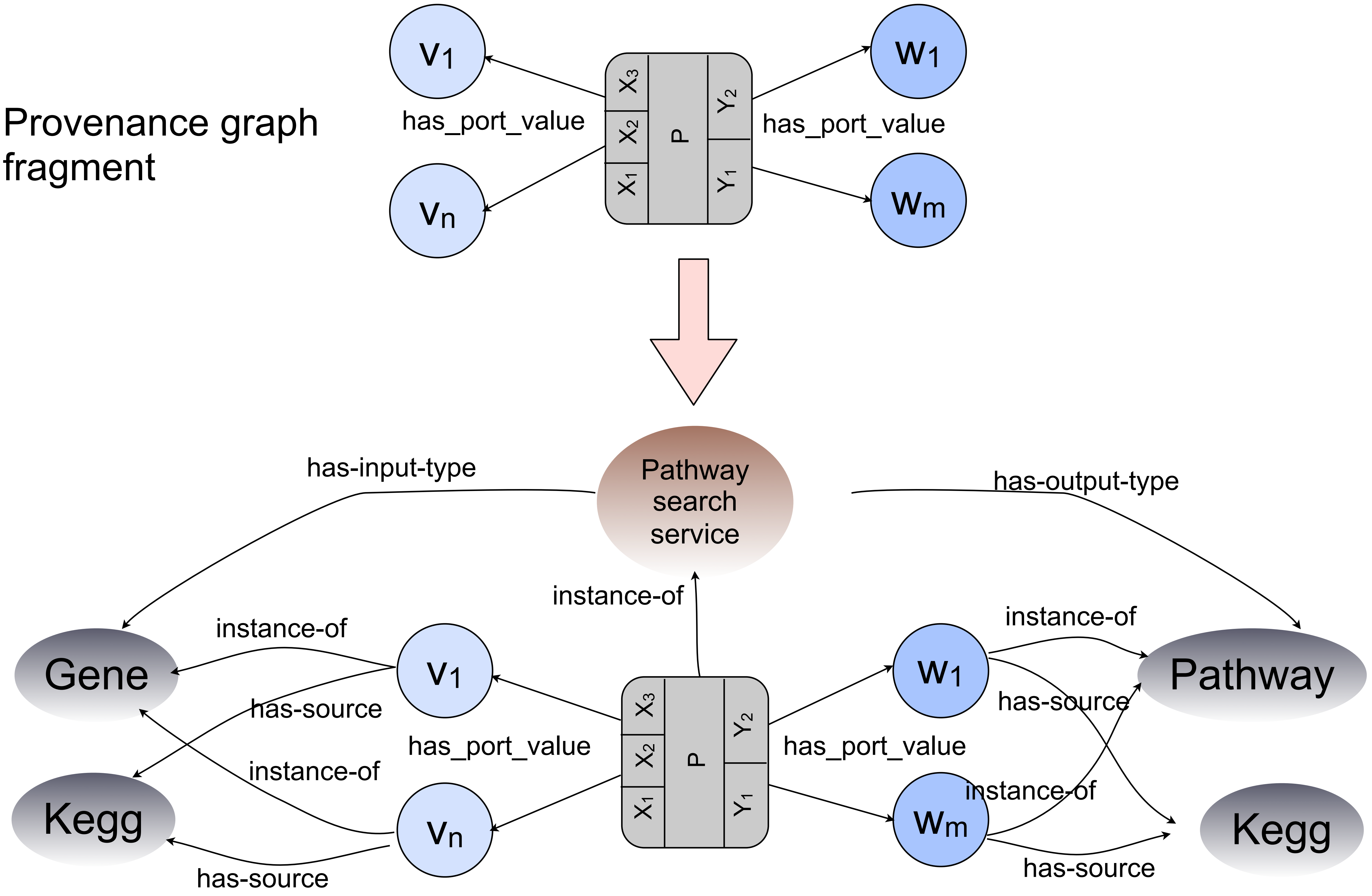
$$\frac{X\ \texttt{rdf:type Port} \qquad C=\{c\} \qquad X\ \texttt{has\_value\_type}\ c \qquad X\ \texttt{has\_value}\ v \qquad v\ \texttt{rdf:type PortValue}}{v\ \texttt{rdf:type}\ C}$$

denotes data type in the PL sense



Annotations as semantic overlay

Provenance graph fragment

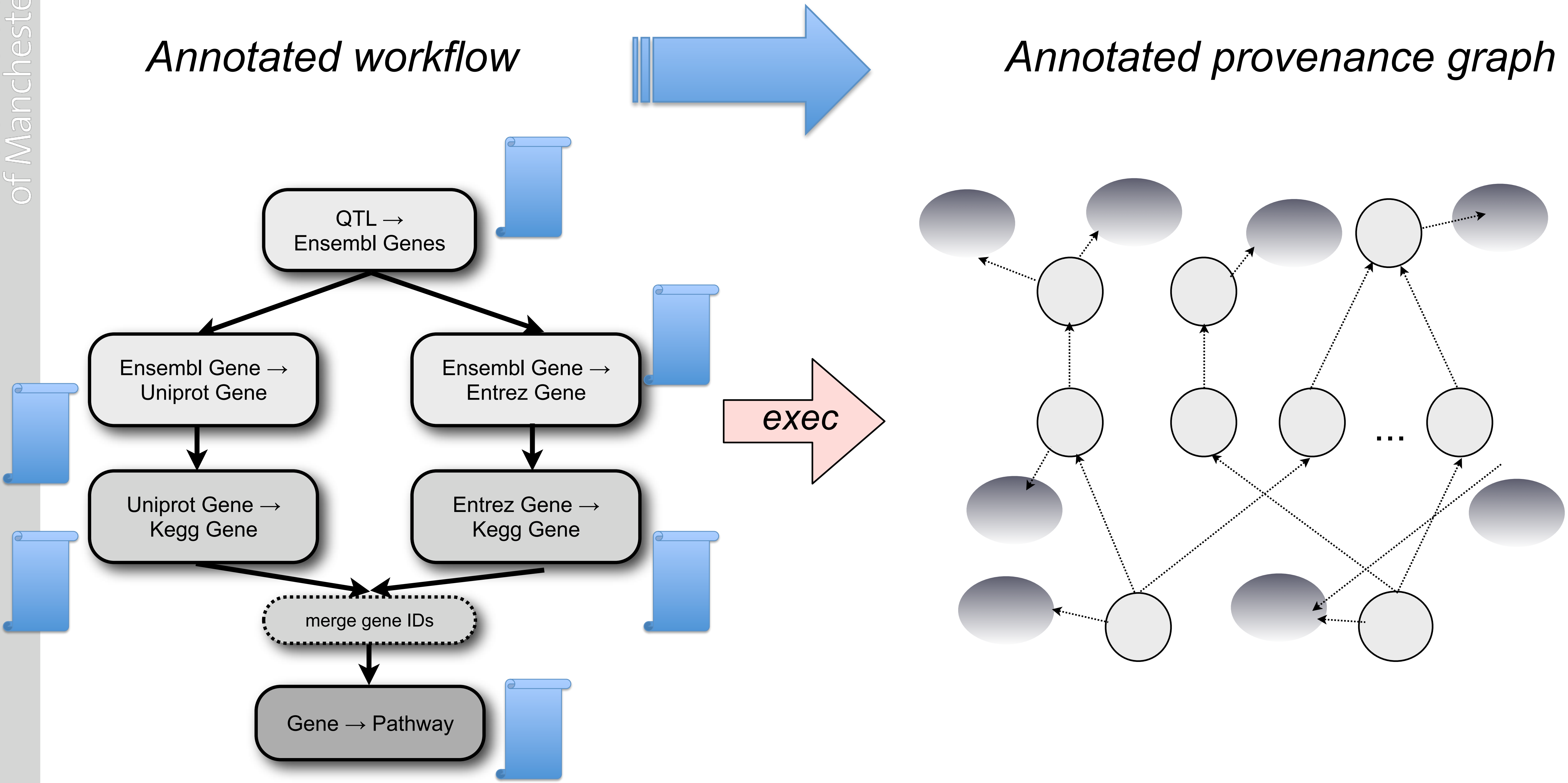


Example Janus domain-aware fragment

```
<rdf:Description rdf:about="http://purl.org/net/taverna/janus/test1625">
<janus:has_iteration>[]</janus:has_iteration>
<rdf:type rdf:resource="http://purl.org/net/taverna/janus#port_value"/>
<rdf:type rdf:resource="http://purl.org/obo/owl/sequence#gene"/>
<janus:has_source rdf:resource="http://purl.org/net/taverna/janus#KEGG"/>
</rdf:Description>
```

this is rule-defined, too

Extensions to Linked Data



Annotated workflow

Annotated provenance graph

QTL → Ensembl Genes

Ensembl Gene → Uniprot Gene

Ensembl Gene → Entrez Gene

Uniprot Gene → Kegg Gene

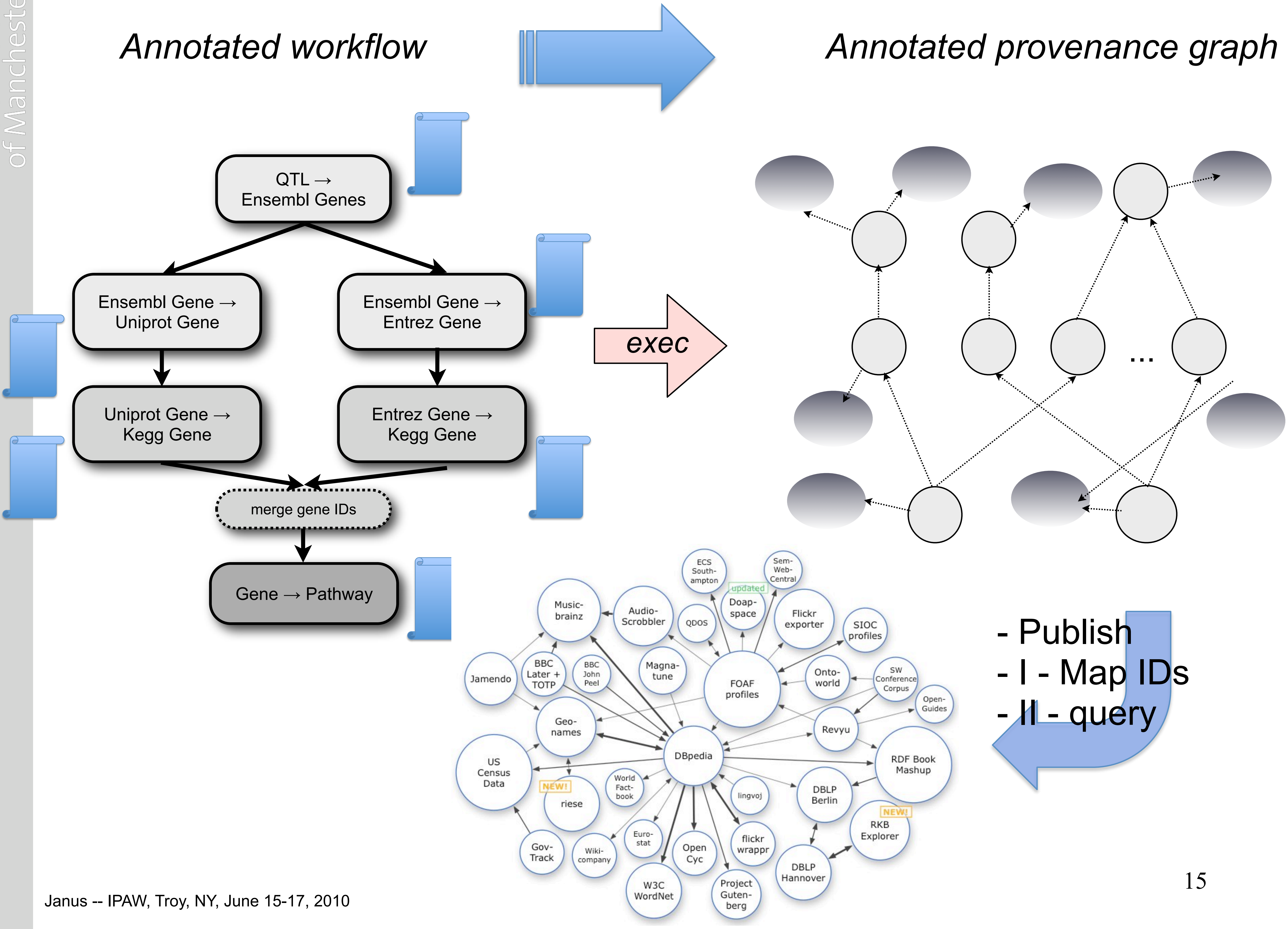
Entrez Gene → Kegg Gene

merge gene IDs

Gene → Pathway

exec

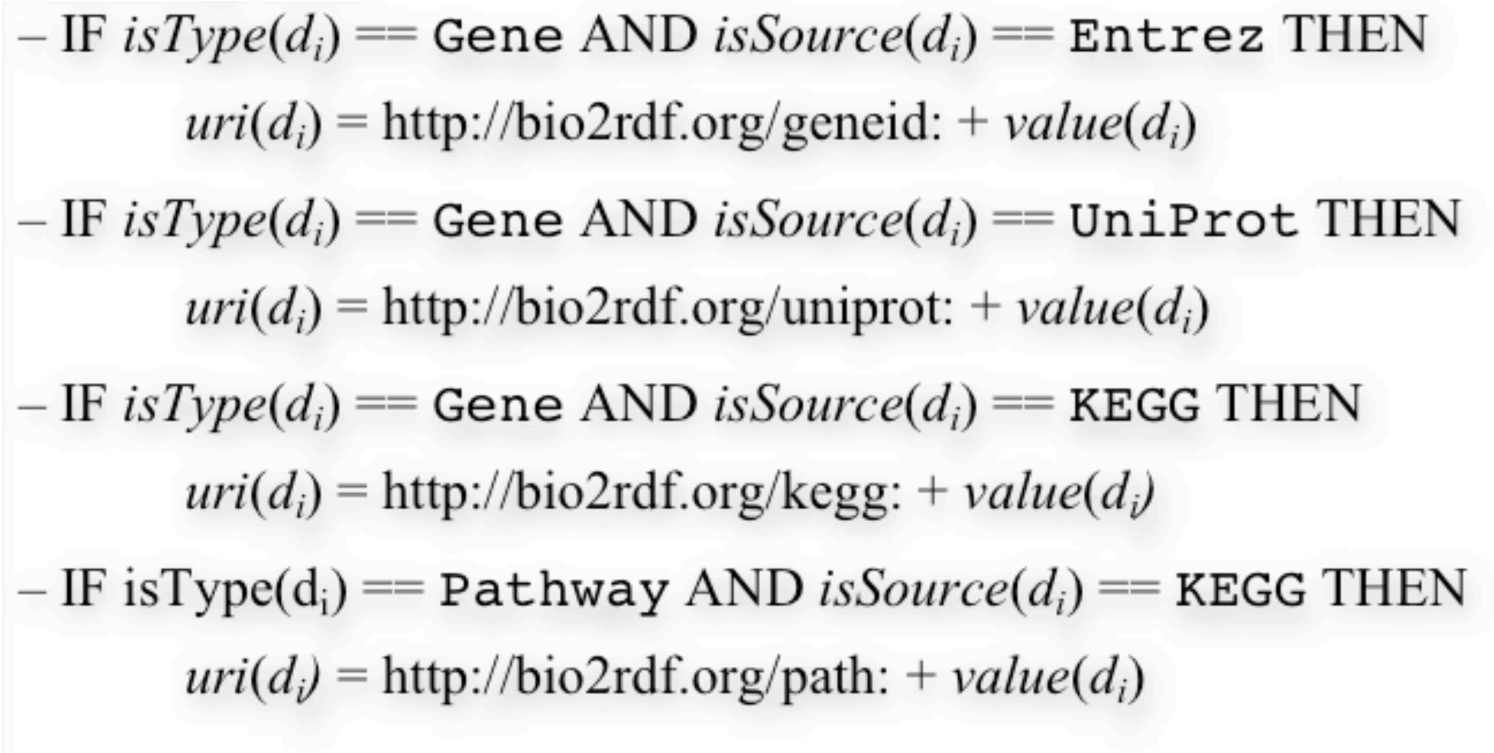
Extensions to Linked Data



- 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

```
<rdf:Description rdf:about="http://purl.org/net/taverna/janus/create_report/entrezGeneId">
   <janus:has_value_binding rdf:resource="http://purl.org/net/taverna/janus/test18"/>

<rdf:Description rdf:about="http://purl.org/net/taverna/janus/test18">
   <rdf:type rdf:resource="http://purl.org/net/taverna/janus#port_value"/>
   <rdfs:comment>11835</rdfs:comment>
   <rdf:type rdf:resource="http://purl.org/obo/owl/sequence#gene"/>
  <janus:has_source rdf:resource="http://purl.org/net/taverna/janus#entrez_gene"/>

<rdf:Description rdf:about="http://purl.org/net/taverna/janus/test18">
   <rdfs:seeAlso rdf:resource="http://bio2rdf.org/geneid:11835"/>
```

II - extended queries in the LoD setting

Strategy:

- use the SQUIN LoD query engine to query multiple “Web of Data” sources
    - only Bio2RDF in our case
- combine graph patterns on local provenance with conditions on remote LoD graphs

Q5: Find all Entrez genes that encode proteins involved in ATP binding (GO:0005524).

```
PREFIX uniprot: <http://purl.uniprot.org/core/>
PREFIX : <http://www.taverna.org.uk/janus#>
SELECT distinct ?entrezgene
WHERE {
?protein uniprot:classifiedWith <http://bio2rdf.org/go:0005524> .
?entrezgene <http://bio2rdf.org/bio2rdf_resource:xPath> ?protein .
?gene rdfs:seeAlso ?entrezgene
?gene rdf:type :port_gene
?gene :has_source :entrez_gene . }
```

Bio2RDF

local provenance graph

Current status

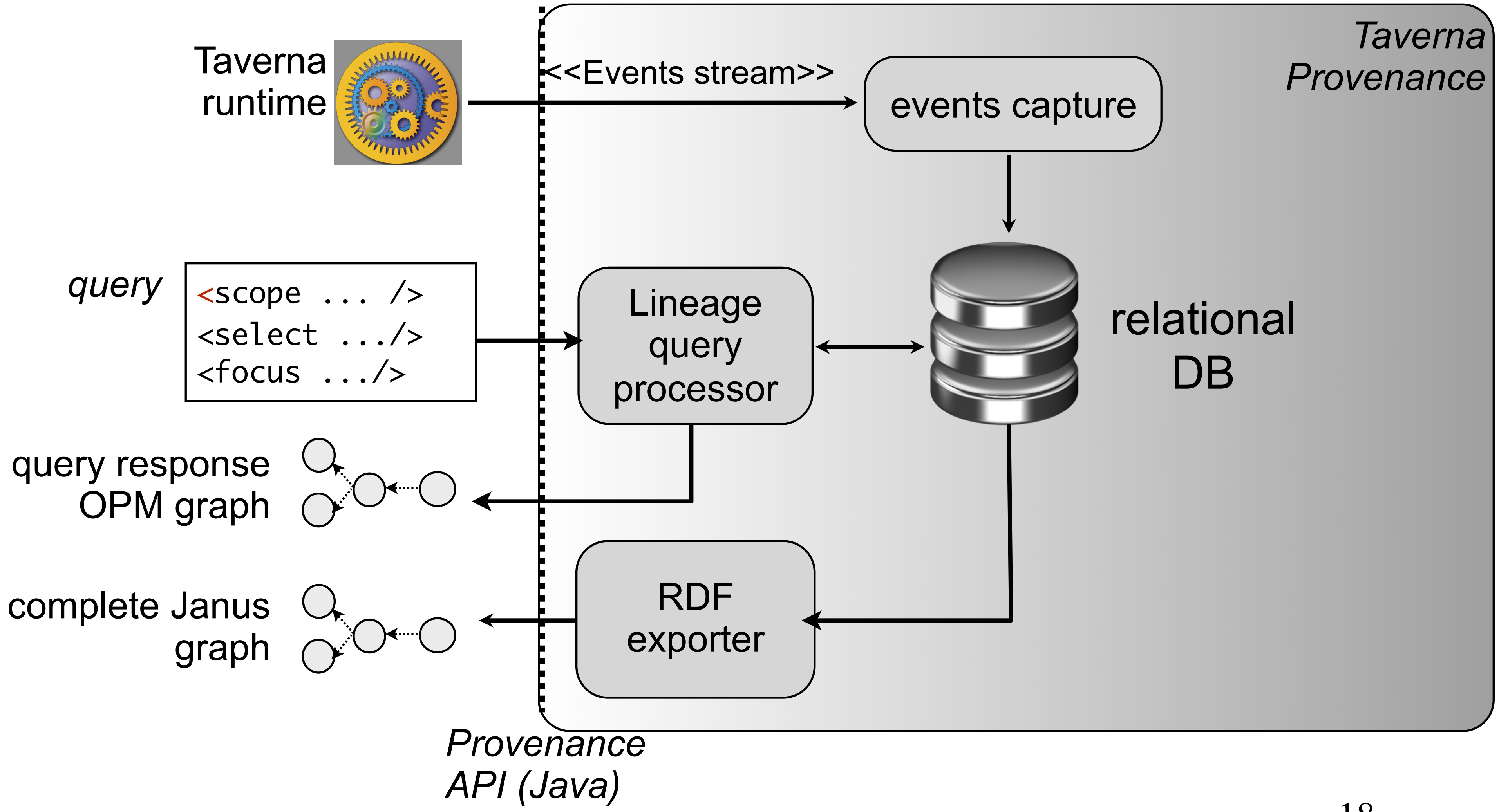
Current Taverna provenance architecture:

Lab prototype

- "Export as..." Janus RDF
- currently only queried using SPARQL
- manually published
- manually annotated

Production

- "native" (relational) graphs
- simple, efficient query language on native provenance



```
<scope ... />
<select .../>
<focus .../>
```

Summary, and moving forward

- Janus: a semantic model for workflow provenance
  - OWL ontology, extension of Provenir
  - should include attribution + system level provenance
  - alignment with OPM?
- Domain-aware graphs through annotations:
  - automatically propagated from workflow annotations when possible
  - but in practice no real workflows are annotated
- LoD integration:
  - powerful provenance publishing and query broadening
  - mapping rules currently limited
  - no completeness guarantee -- all joins are outer joins!