Ontology based knowledge management of biomedical models and data

Pierre Grenon, EMBL-EBI, UK
pgrenon@ebi.ac.uk

CrEDIBLE, Session 2
16 Oct 2012

Context

Biomedical community specific efforts to mine data using knowledge.

VPH's RICORDO (ricordo.eu)

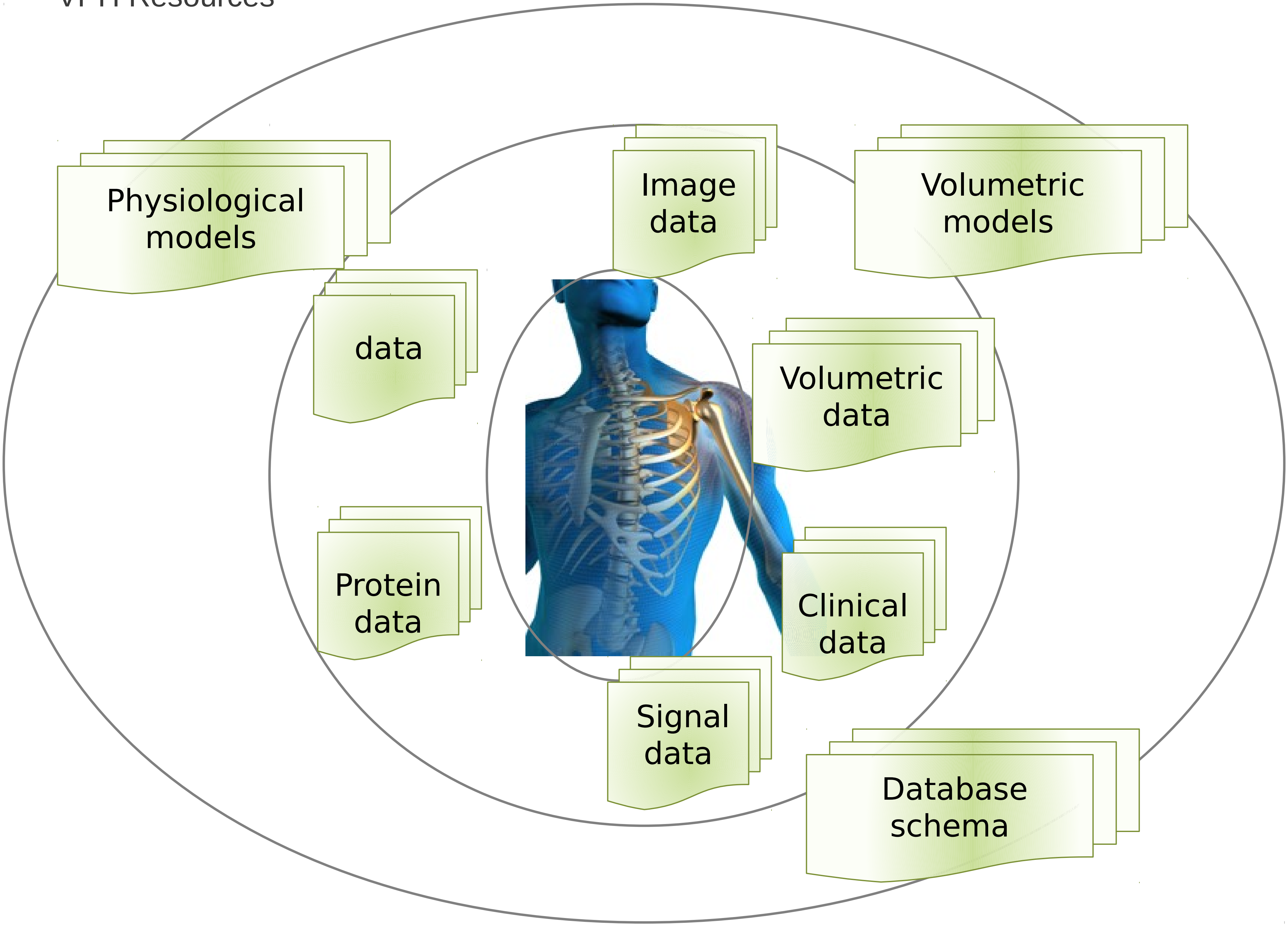
Cross-domain (any system modelling and clinical data)

Bridging community (modellers of different sorts and clinicians)

Standardised and reusable resources, tools and methods

Work pursued under IMI's DDMORE in the domain of drug and pharmacological modelling (ddmore.eu)

VPH Resources



Physiological models

Image data

Volumetric models

data

Volumetric data

Protein data

Clinical data

Signal data

Database schema

Different types of resources

Models

Data structures

Database schema

Different purposes and constitution

Physiological models

Geometric models

Volumetric data

Clinical data

Protein data

Image data

Different syntaxes

MathLab models

SBML models

FieldML models

Fortran models

CellML models

```
<reaction metaid="_230655" id="reaction7"
  name="deactivation of cyclin protease">
      <listOfReactants>
        <speciesReference species="X"/>
      </listOfReactants>
      <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
          ...
        </math>
        <listOfParameters>
          <parameter metaid="_961177" id="K4" value="0.005"/>
          <parameter metaid="_961180" id="V4" value="0.5"/>
        </listOfParameters>
      </kineticLaw>
    </reaction>
```

SBML

```
[Severity of A 1] [float] NULL,
[Severity of B 2] [float] NULL,
[Severity of C] [nvarchar](255) NULL,
[C gradient] [float] NULL,
```

Database Schema

```
namespace Model1
{ …
        enum ModelType {
                //! follows specs of volumetric models
                HEART_MODEL } ;

        enum SubpartType {
                LV_ENDO_SUBPART = 1, //1
                LV_EPI_SUBPART, //2
                LA_SUBPART, //3
                RV_SUBPART, //4
                RA_SUBPART, //5
            ... } ; ...  }; }
```

Volumetric model

```
<component cmeta:id="C" name="C">
  <variable units="dimensionless" name="X"/>
  <variable units="rate_constant" name="kd"/>
  <variable units="micromolar" name="Kd"/>
  ...
  <math xmlns="http://www.w3.org/1998/Math/MathML">
    ...
  </math>
</component>
```

CellML

Main Problems

Syntactic heterogeneity

Resources are maintained in distinct, disconnected languages and formats

Semantic opaqueness

Resources do not carry explicit, machine-processable formalisation of their biomedical interpretation

```
<reaction metaid="_230655" id="reaction7"
 name="deactivation of cyclin protease">
   <listOfReactants>
     <speciesReference species="X"/>
   </listOfReactants>
   <kineticLaw>
     <math xmlns="http://www.w3.org/1998/Math/MathML">
       ...
     </math>
     <listOfParameters>
       <parameter metaid="_961177" id="K4" value="0.005"/>
       <parameter metaid="_961180" id="V4" value="0.5"/>
     </listOfParameters>
   </kineticLaw>
  </reaction>
```

SBML

```
namespace Model1
{ …
        enum ModelType {
                //! follows specs of volumetric models
                HEART_MODEL } ;

        enum SubpartType {
                LV_ENDO_SUBPART = 1, //1
                LV_EPI_SUBPART, //2
                LA_SUBPART, //3
                RV_SUBPART, //4
                RA_SUBPART, //5
            ... } ; ...  }; }
```

Volumetric model

Similar syntactic serialisation

```
<component cmeta:id="C" name="C">
 <variable units="dimensionless" name="X"/>
 <variable units="rate_constant" name="kd"/>
 <variable units="micromolar" name="Kd"/>
 ...
 <math xmlns="http://www.w3.org/1998/Math/MathML">
   ...
 </math>
</component>
```

CellML

```
[Severity of MR (VC)] [float] NULL,
[Severity of MR (Area method)] [float] NULL,
[Severity of TR] [nvarchar](255) NULL,
[TR gradient] [float] NULL,
[PAP (mmHg)] [float] NULL,
[AV VTI] [float] NULL,
```

Database Schema

Problems

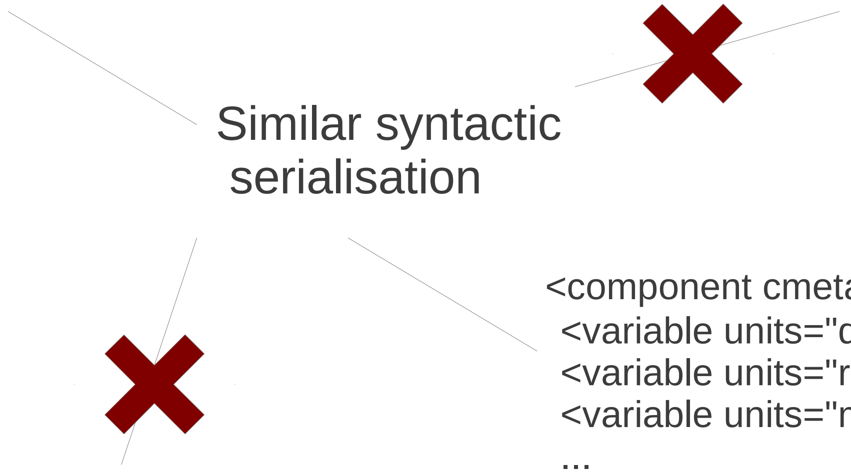
```
<reaction metaid="_230655" id="reaction7"
  name="deactivation of cyclin protease">
    <listOfReactants>
      <speciesReference species="X"/>
    </listOfReactants>
    <kineticLaw>
      <math xmlns="http://www.w3.org/1998/Math/MathML">
        ...
      </math>
      <listOfParameters>
        <parameter metaid="_961177" id="K4" value="0.005"/>
        <parameter metaid="_961180" id="V4" value="0.5"/>
      </listOfParameters>
    </kineticLaw>
  </reaction>
```

SBML

```
namespace Model1
{ …
        enum ModelType {
                //! follows specs of volumetric models
                HEART_MODEL } ;

        enum SubpartType {
                LV_ENDO_SUBPART = 1, //1
                LV_EPI_SUBPART, //2
                LA_SUBPART, //3
                RV_SUBPART, //4
                RA_SUBPART, //5
           ... } ; ...  }; }
```

Volumetric model



Similar syntactic serialisation

```
<component cmeta:id="C" name="C">
 <variable units="dimensionless" name="X"/>
 <variable units="rate_constant" name="kd"/>
 <variable units="micromolar" name="Kd"/>
 ...
 <math xmlns="http://www.w3.org/1998/Math/MathML">
   ...
 </math>
</component>
```

CellML

```
[Severity of MR (VC)] [float] NULL,
[Severity of MR (Area method)] [float] NULL,
[Severity of TR] [nvarchar](255) NULL,
[TR gradient] [float] NULL,
[PAP (mmHg)] [float] NULL,
[AV VTI] [float] NULL,
```

Database Schema

```
<reaction metaid="_230655" id="reaction7"
 name="deactivation of cyclin protease">
   <listOfReactants>
     <speciesReference species="X"/>
   </listOfReactants>
   <kineticLaw>
     <math xmlns="http://www.w3.org/1998/Math/MathML">
       ...
     </math>
     <listOfParameters>
       <parameter metaid="_961177" id="K4" value="0.005"/>
       <parameter metaid="_961180" id="V4" value="0.5"/>
     </listOfParameters>
   </kineticLaw>
 </reaction>
```

SBML

```
namespace Model1
{ …
        enum ModelType {
                //! follows specs of volumetric models
                HEART_MODEL } ;

        enum SubpartType {
                LV_ENDO_SUBPART = 1, //1
                LV_EPI_SUBPART, //2
                LA_SUBPART, //3
                RV_SUBPART, //4
                RA_SUBPART, //5
            ... } ; ...  }; }
```

Volumetric model

```
<component cmeta:id="C" name="C">
 <variable units="dimensionless" name="X"/>
 <variable units="rate_constant" name="kd"/>
 <variable units="micromolar" name="Kd"/>
 ...
 <math xmlns="http://www.w3.org/1998/Math/MathML">
   ...
 </math>
</component>
```

CellML

```
[Severity of MR (VC)] [float] NULL,
[Severity of MR (Area method)] [float] NULL,
[Severity of TR] [nvarchar](255) NULL,
[TR gradient] [float] NULL,
[PAP (mmHg)] [float] NULL,
[AV VTI] [float] NULL,
```

Database Schema

What are they about?
How do they relate?

Problems

Main Problems and Solutions

Syntactic heterogeneity

Resources are maintained in distinct, disconnected languages and formats

Use a unified, standardised knowledge representation framework

Semantic opaqueness

Resources do not carry explicit, machine-processable formalisation of their biomedical interpretation

Annotate resources with well defined, curated references (reference ontologies)

Syntactic heterogeneity is resolved by standardised, systematic representation

Unique representation language

Unique identifiers for objects

Description of models, databases, and their parts,

using a schema (of types and relationships)

Biomedical interpretation is part of metadata.

Syntactic heterogeneity is resolved by standardised, systematic representation

Unique representation language

Unique identifiers for objects

Description of models, databases, and their parts,

using a schema (of types and relationships)

Biomedical interpretation is part of metadata.

RDF is a prime candidate for machine processable metadata

Resource Description Framework

- W3C standard
- Widely used and supported
- Query language (SPARQL)
- Tools and implementation of stores

Syntactic heterogeneity is resolved by standardised, systematic representation

Unique representation language

Unique identifiers for objects

Description of models, databases, and their parts,

using a schema (of types and relationships)

Biomedical interpretation is part of metadata.

RDF is a prime candidate for machine processable metadata

Resource Description Framework

- W3C standard
- Widely used and supported
- Query language (SPARQL)
- Tools and implementation of stores

Syntactic integration puts everything on a par, making possible:

- Sharing tools and APIs
- Realising semantic clarity (representation of interpretation)

Semantic standardisation

- **Annotating** models, databases, and their parts
- Linking to standardised, shareable biomedical meanings
- Linking in a variety of ways
- Specification of these links **as machine-processable metadata**
- Thus expressing **biomedical interpretations**

Semantic standardisation

- **Annotating** models, databases, and their parts
- Linking to standardised, shareable biomedical meanings
- Linking in a variety of ways
- Specification of these links **as machine-processable metadata**
- Thus expressing **biomedical interpretations**

- Ontologies consist in a theory of terms and relationships
- They can be formalised
- Machine-processable formalisations facilitate automation and further computational tasks such as reasoning
- There are sustained efforts to produce standard, curated biomedical ontologies

Semantic standardisation

- **Annotating** models, databases, and their parts
- Linking to standardised, shareable biomedical meanings
- Linking in a variety of ways
- Specification of these links **as machine-processable metadata**
- Thus expressing **biomedical interpretations**

- Ontologies consist in a theory of terms and relationships
- They can be formalised
- Machine-processable formalisations facilitate automation and further computational tasks such as reasoning
- There are sustained efforts to produce standard, curated biomedical ontologies

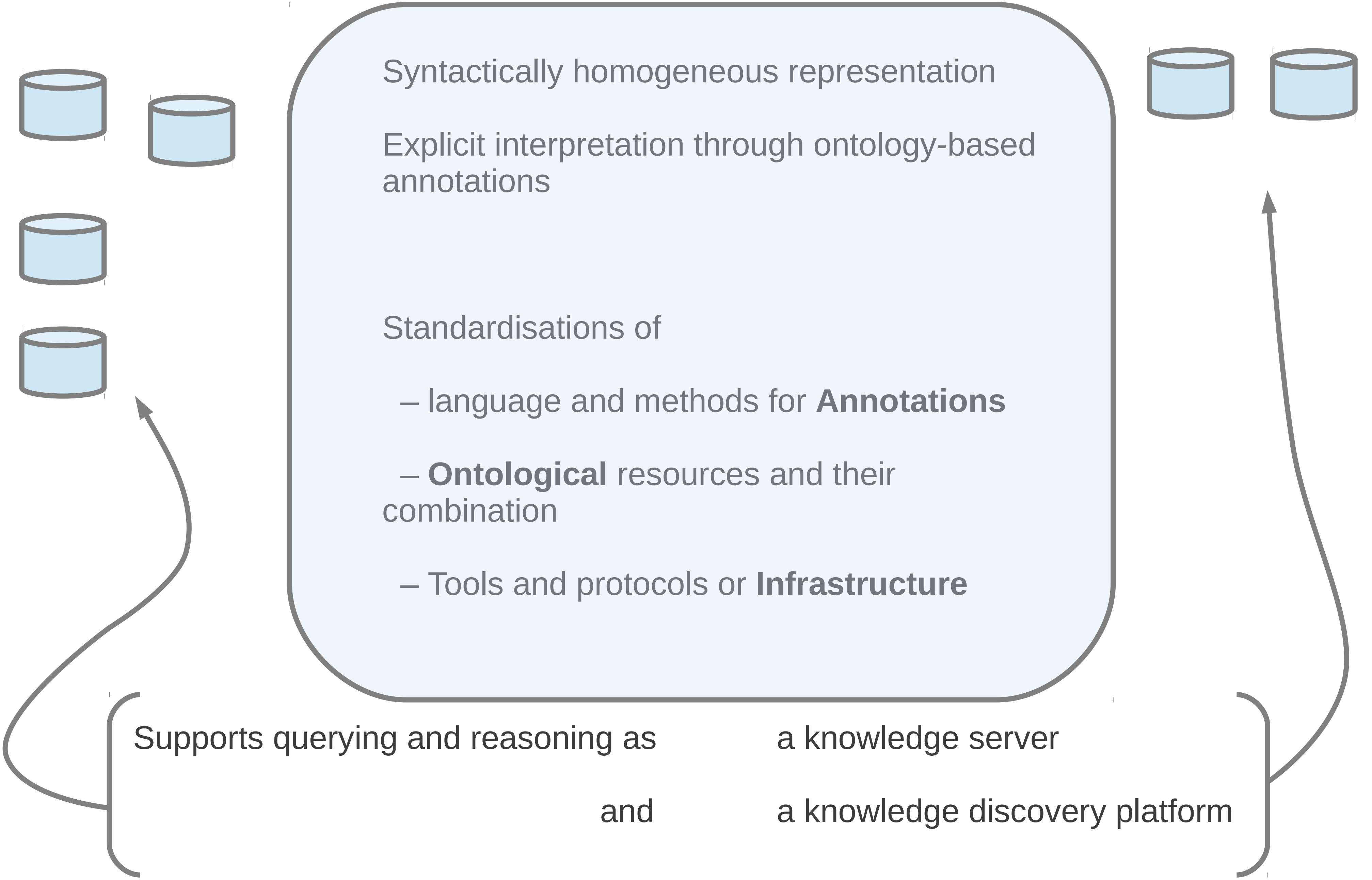
Ontologies and ontology-based annotations allow:

Semantic clarity (expression of biomedical interpretation)

Shared understanding through reuse and standardisation

Semantic interoperability through reuse and mappings

RICORDO's Knowledge Management Framework



Syntactically homogeneous representation

Explicit interpretation through ontology-based annotations

Standardisations of

- language and methods for **Annotations**
- **Ontological** resources and their combination
- Tools and protocols or **Infrastructure**

Supports querying and reasoning as a knowledge server

and a knowledge discovery platform

Representation and annotation

(Baseline
freetext annotation)

Model: A

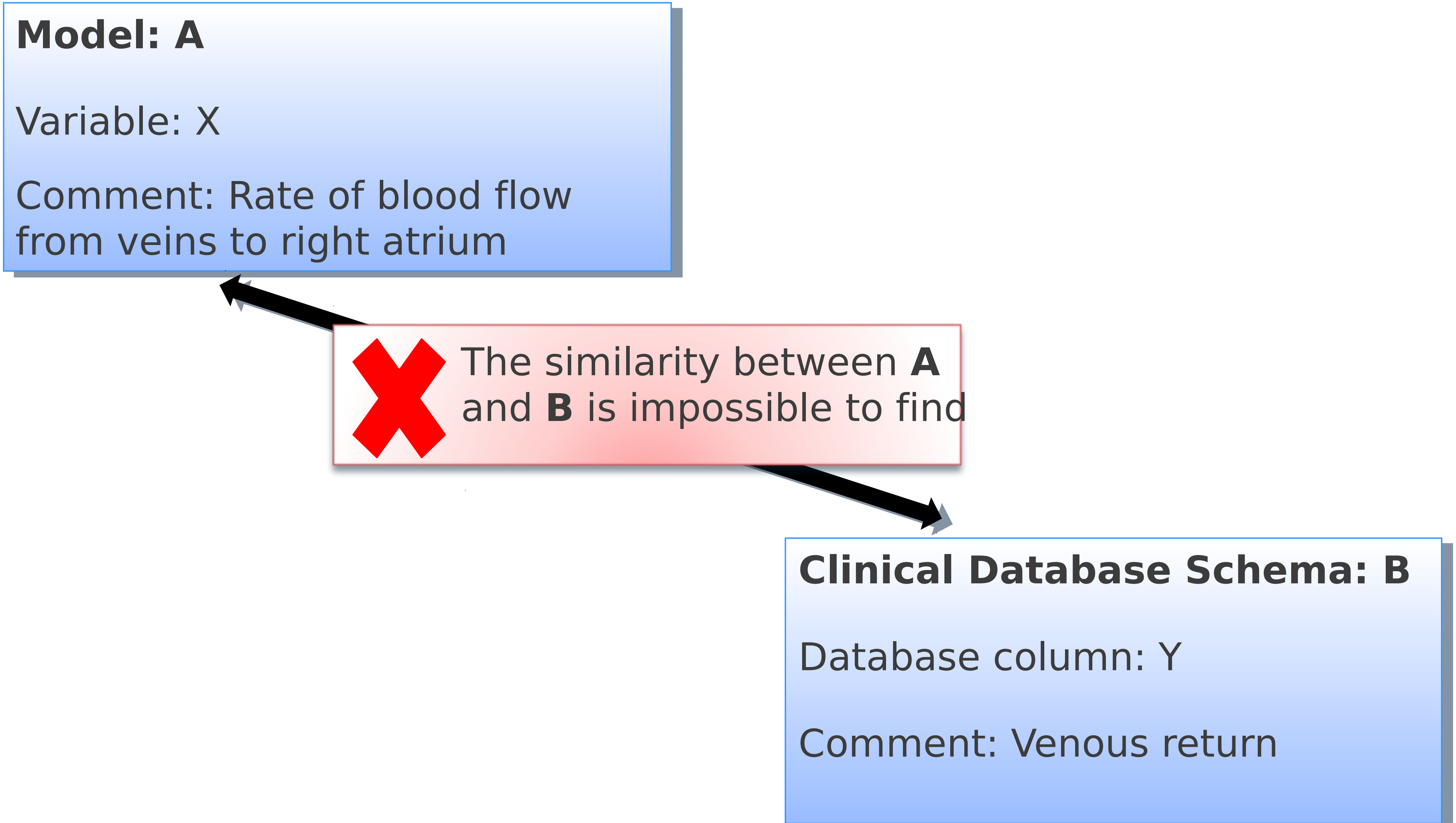
Variable: X

Comment: Rate of blood flow from veins to right atrium

Clinical Database Schema: B

Database column: Y

Comment: Venous return



Model: A

Variable: X

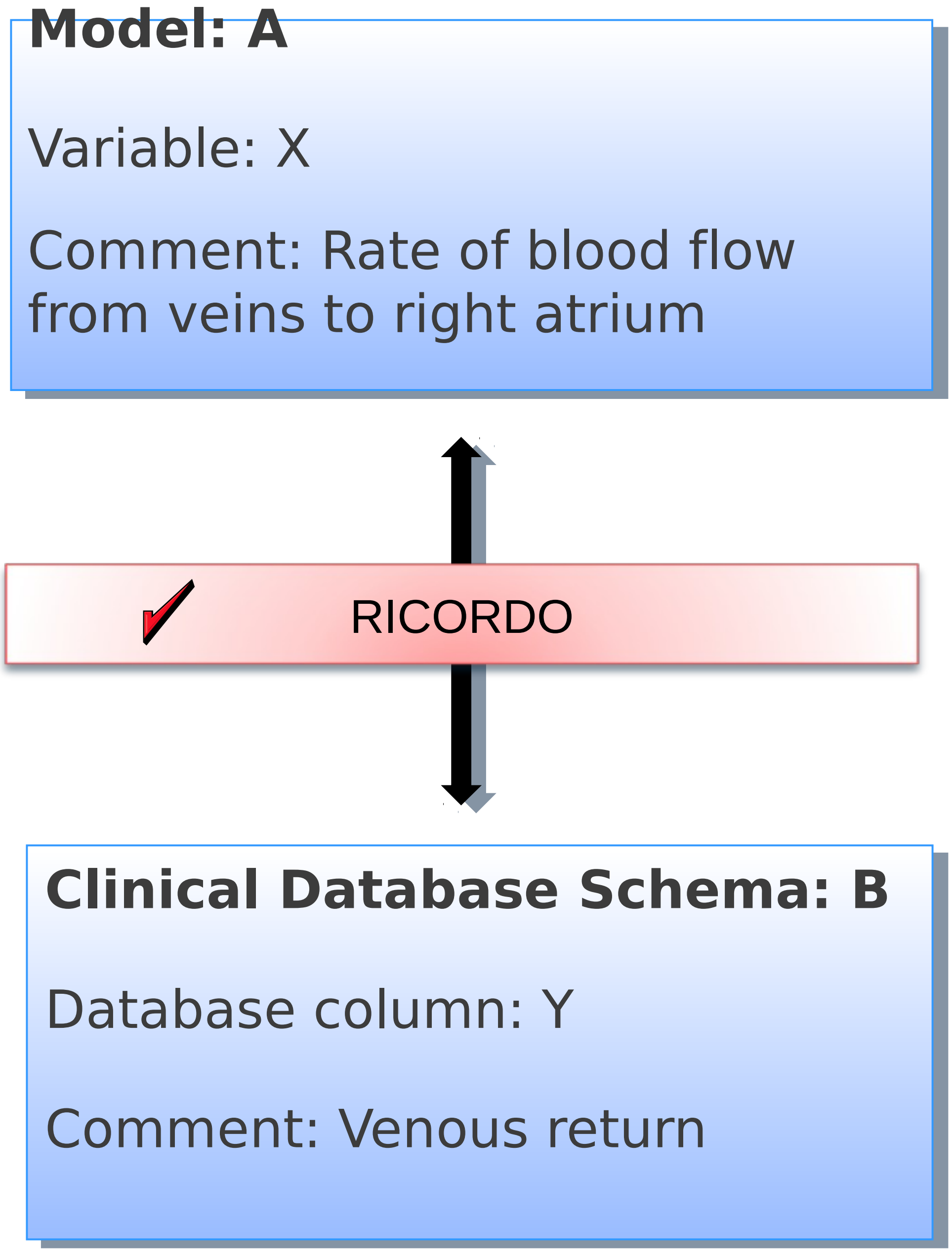
Comment: Rate of blood flow from veins to right atrium

The similarity between **A** and **B** is impossible to find

Clinical Database Schema: B

Database column: Y

Comment: Venous return



Model: A

Variable: X

Comment: Rate of blood flow from veins to right atrium

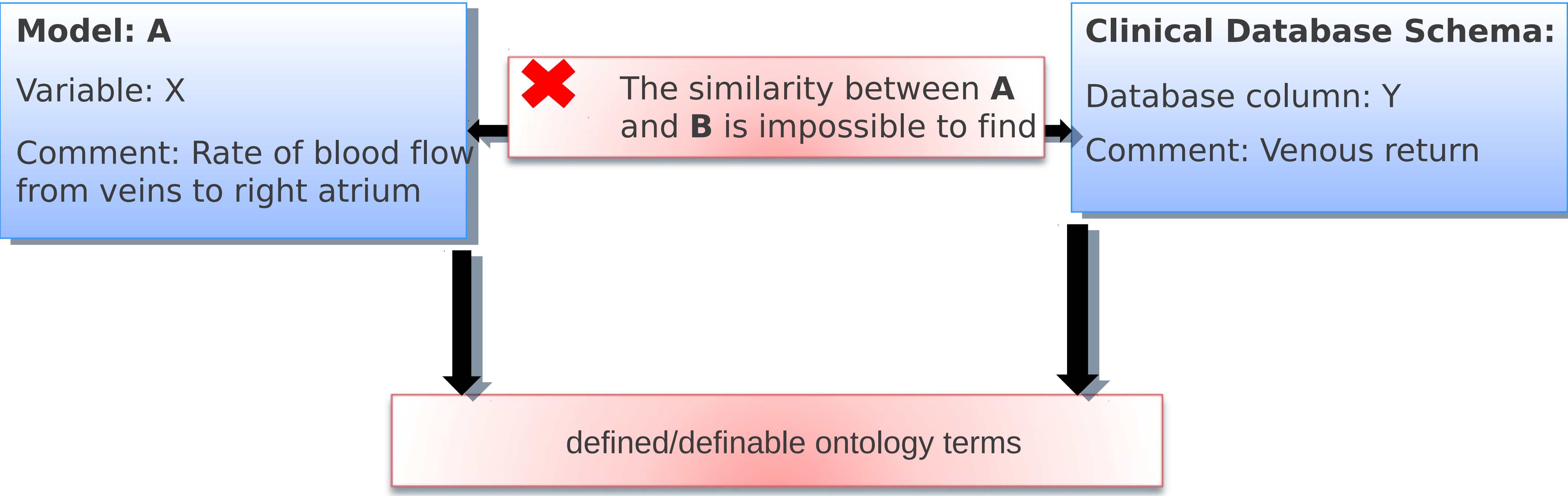
✓ RICORDO

Clinical Database Schema: B

Database column: Y

Comment: Venous return

Use of ontologies in machine processable annotations



Model: A

Variable: X

Comment: Rate of blood flow from veins to right atrium

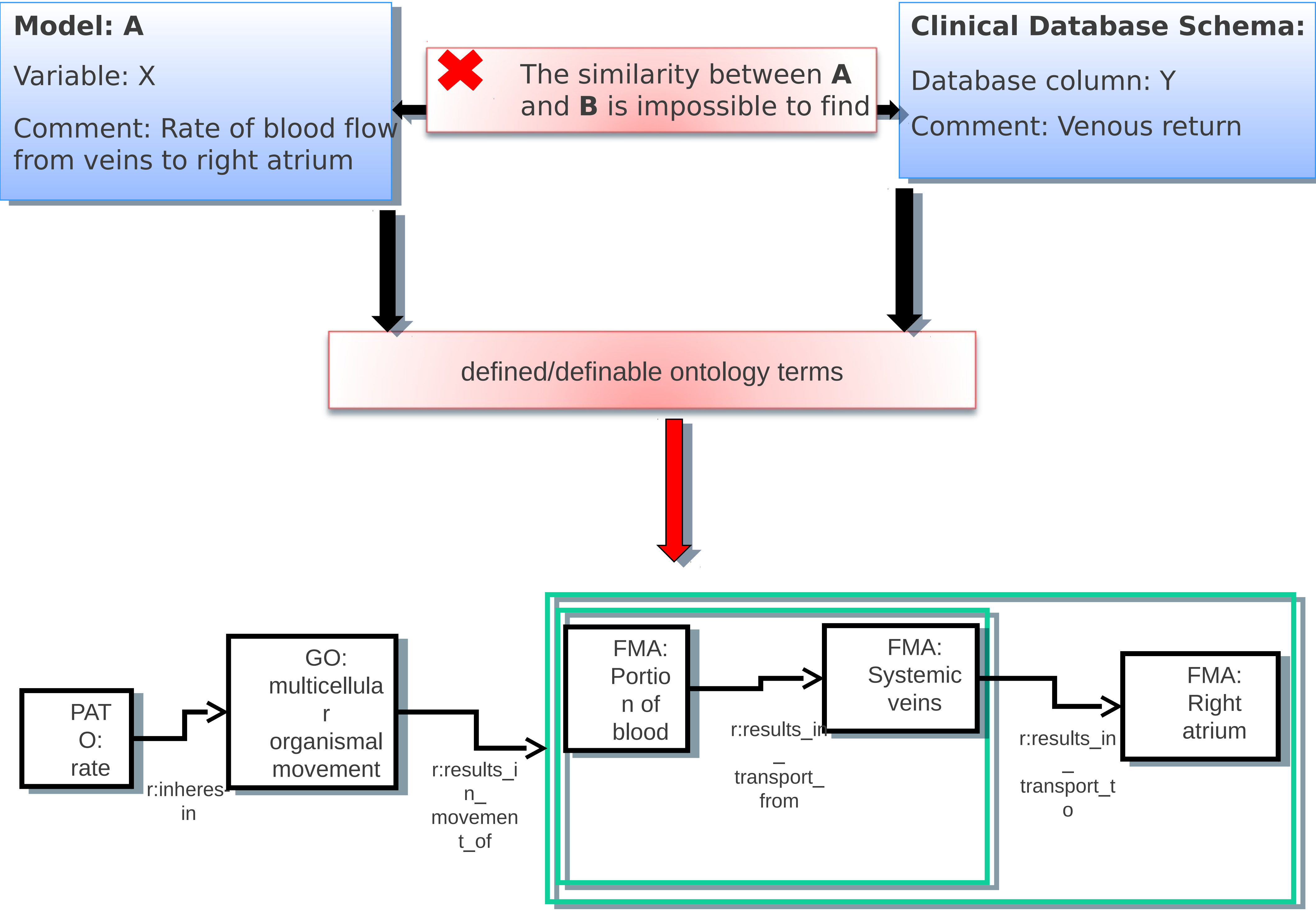
✖ The similarity between **A** and **B** is impossible to find

Clinical Database Schema:

Database column: Y

Comment: Venous return

defined/definable ontology terms



Model: A

Variable: X

Comment: Rate of blood flow from veins to right atrium

✖ The similarity between **A** and **B** is impossible to find

Clinical Database Schema:

Database column: Y

Comment: Venous return

defined/definable ontology terms

PATO: rate

r:inheres-in

GO: multicellular organismal movement

r:results_in_movement_of

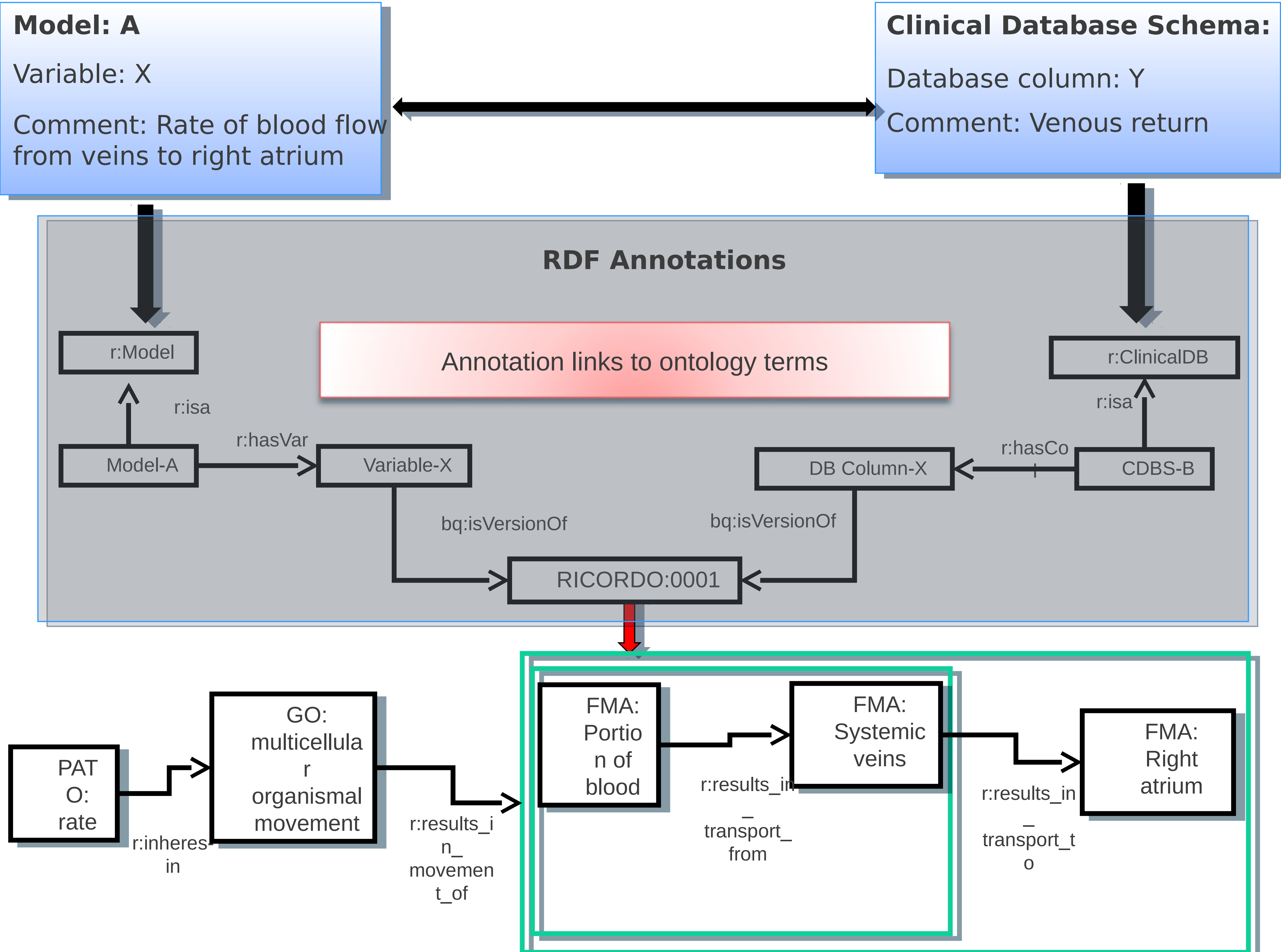
FMA: Portion of blood

r:results_in_transport_from

FMA: Systemic veins

r:results_in_transport_to

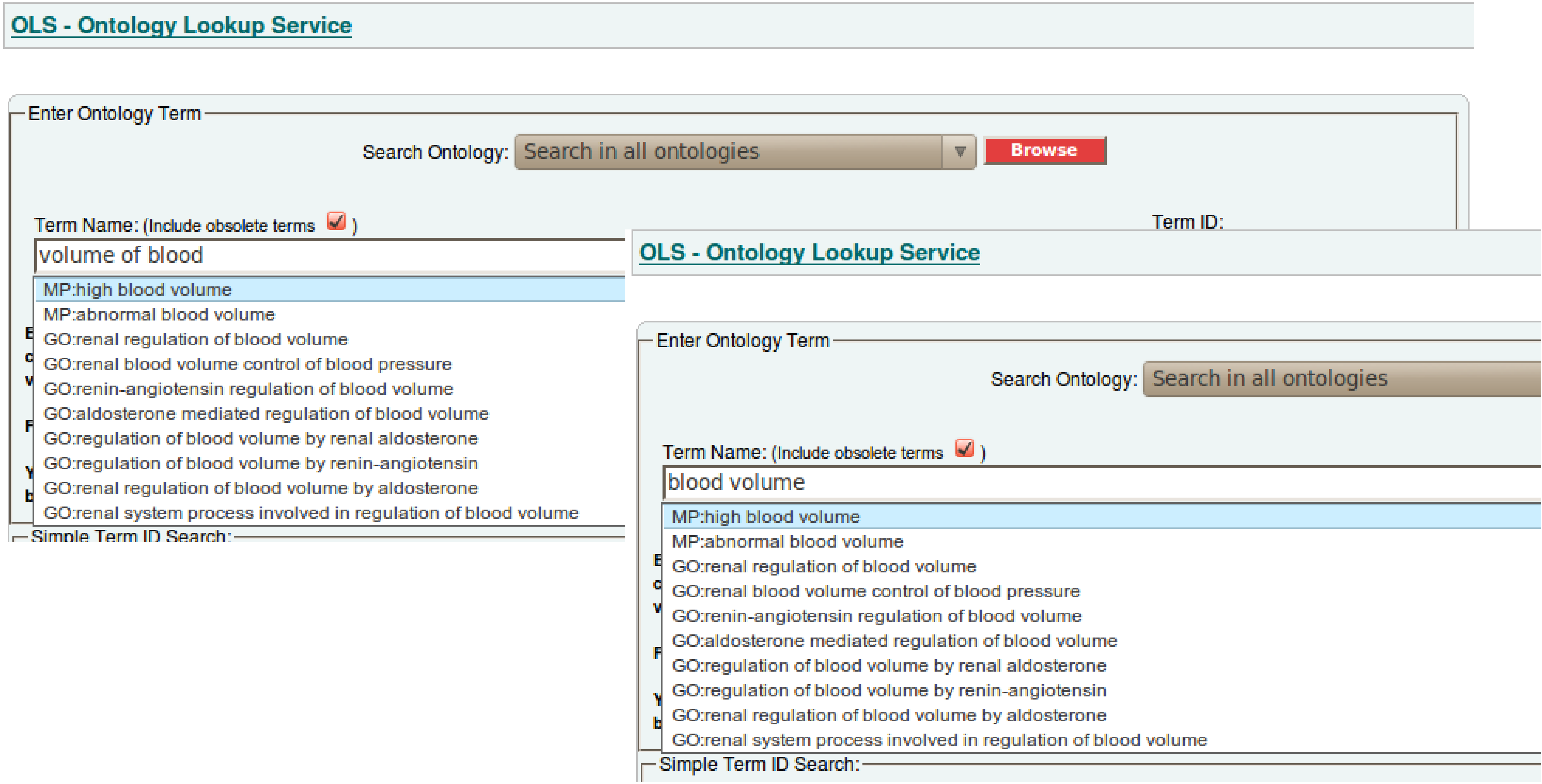
FMA: Right atrium


Model: A
Variable: X
Comment: Rate of blood flow from veins to right atrium
Clinical Database Schema:
Database column: Y
Comment: Venous return
RDF Annotations
r:Model
Annotation links to ontology terms
r:ClinicalDB
r:isa
r:isa
Model-A
r:hasVar
Variable-X
DB Column-X
r:hasCol
CDBS-B
bq:isVersionOf
bq:isVersionOf
RICORDO:0001
PATO: rate
r:inheres_in
GO: multicellular organismal movement
r:results_in_movement_of
FMA: Portion of blood
r:results_in_transport_from
FMA: Systemic veins
r:results_in_transport_to
FMA: Right atrium

From freetext to ontology

Volume of blood

- Volume_of_blood



OLS - Ontology Lookup Service

Enter Ontology Term

Search Ontology: Search in all ontologies

Term Name: (Include obsolete terms ☑)

Term ID:

volume of blood

- MP:high blood volume
- MP:abnormal blood volume
- GO:renal regulation of blood volume
- GO:renal blood volume control of blood pressure
- GO:renin-angiotensin regulation of blood volume
- GO:aldosterone mediated regulation of blood volume
- GO:regulation of blood volume by renal aldosterone
- GO:regulation of blood volume by renin-angiotensin
- GO:renal regulation of blood volume by aldosterone
- GO:renal system process involved in regulation of blood volume

Simple Term ID Search:

OLS - Ontology Lookup Service

Enter Ontology Term

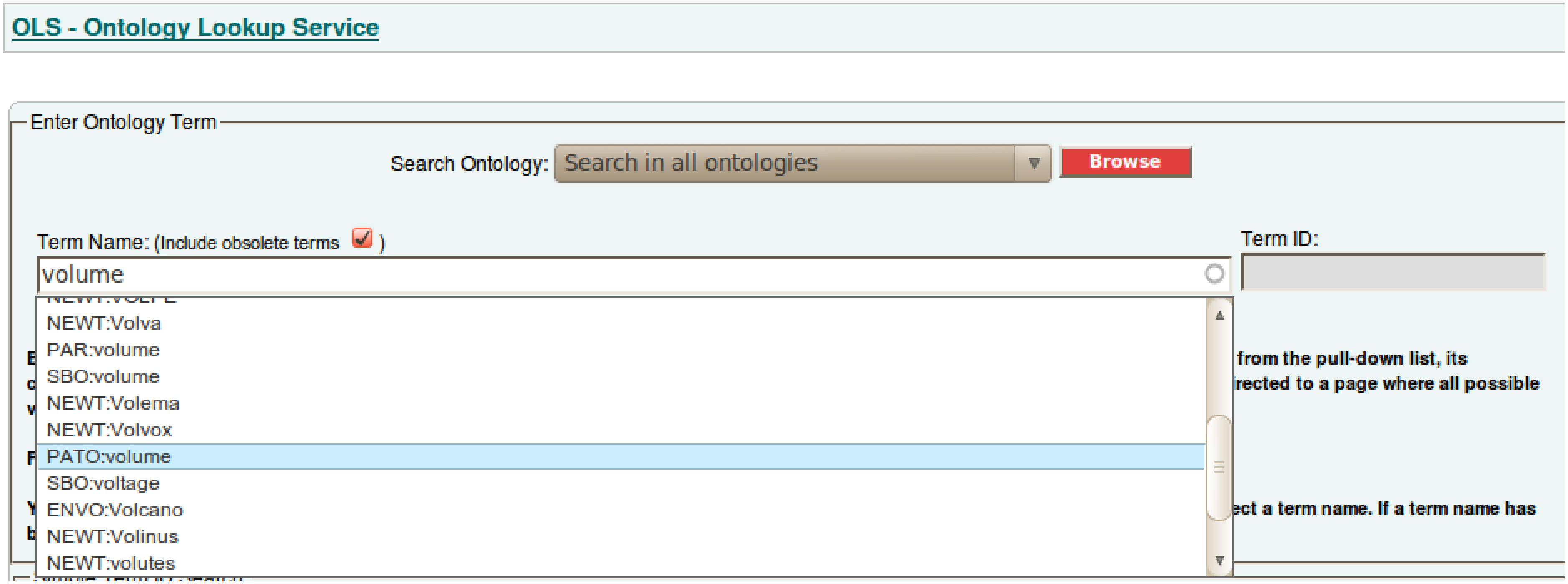
Search Ontology: Search in all ontologies

Term Name: (Include obsolete terms ☑)

blood volume

- MP:high blood volume
- MP:abnormal blood volume
- GO:renal regulation of blood volume
- GO:renal blood volume control of blood pressure
- GO:renin-angiotensin regulation of blood volume
- GO:aldosterone mediated regulation of blood volume
- GO:regulation of blood volume by renal aldosterone
- GO:regulation of blood volume by renin-angiotensin
- GO:renal regulation of blood volume by aldosterone
- GO:renal system process involved in regulation of blood volume

Simple Term ID Search:



OLS - Ontology Lookup Service

Enter Ontology Term

Search Ontology: Search in all ontologies

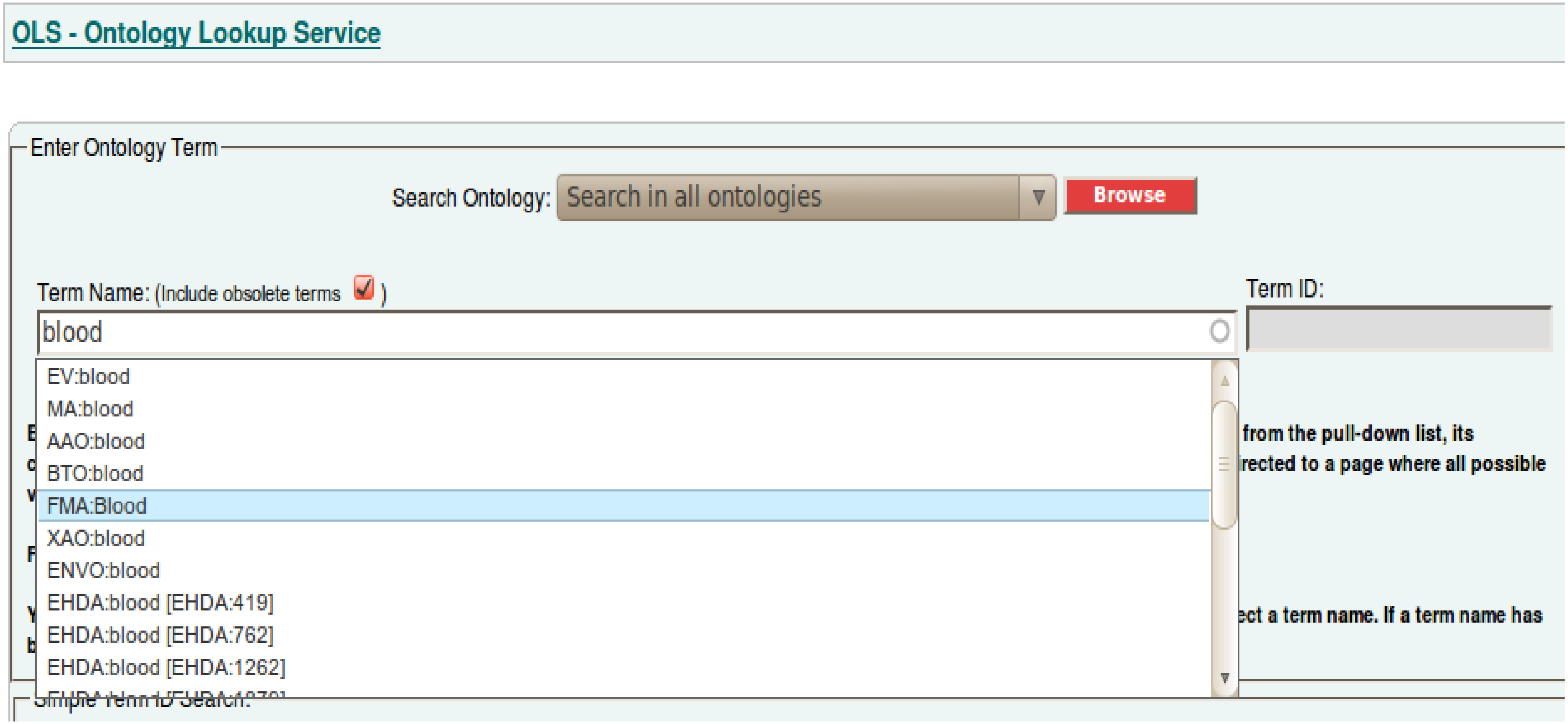
Term Name: (Include obsolete terms ☑) volume

Term ID:

from the pull-down list, its

rected to a page where all possible

ect a term name. If a term name has



OLS - Ontology Lookup Service

Enter Ontology Term

Search Ontology: Search in all ontologies

Term Name: (Include obsolete terms ☑)

blood

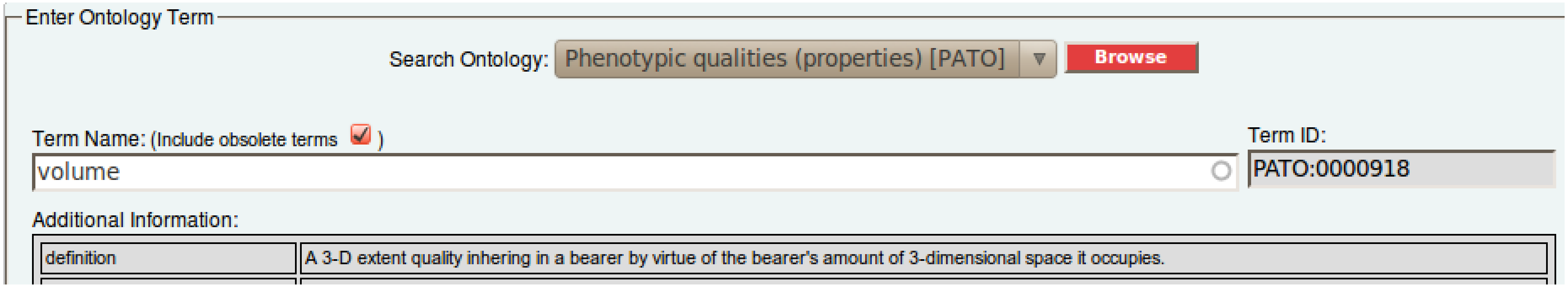
Term ID:

EV:blood
MA:blood
AAO:blood
BTO:blood
FMA:Blood
XAO:blood
ENVO:blood
EHDA:blood [EHDA:419]
EHDA:blood [EHDA:762]
EHDA:blood [EHDA:1262]

from the pull-down list, its
rected to a page where all possible

ect a term name. If a term name has

Simple Term ID Search:



Search Ontology: Phenotypic qualities (properties) [PATO]

Term Name: (Include obsolete terms ☑)

volume

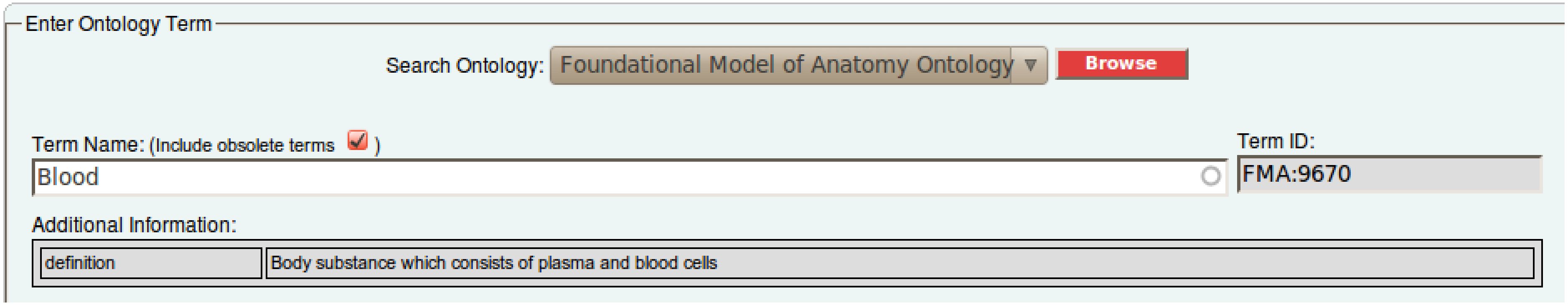
Term ID:

PATO:0000918

Additional Information:

definition	A 3-D extent quality inhering in a bearer by virtue of the bearer's amount of 3-dimensional space it occupies.

Volume_of_blood is a specialisation of PATO:volume



Search Ontology: Foundational Model of Anatomy Ontology

Term Name: (Include obsolete terms ☑)
Blood

Term ID:
FMA:9670

Additional Information:

definition	Body substance which consists of plasma and blood cells

An instance of Volume_of_blood

inheres (is a volume of)

an instance of a portion of blood, FMA:Blood

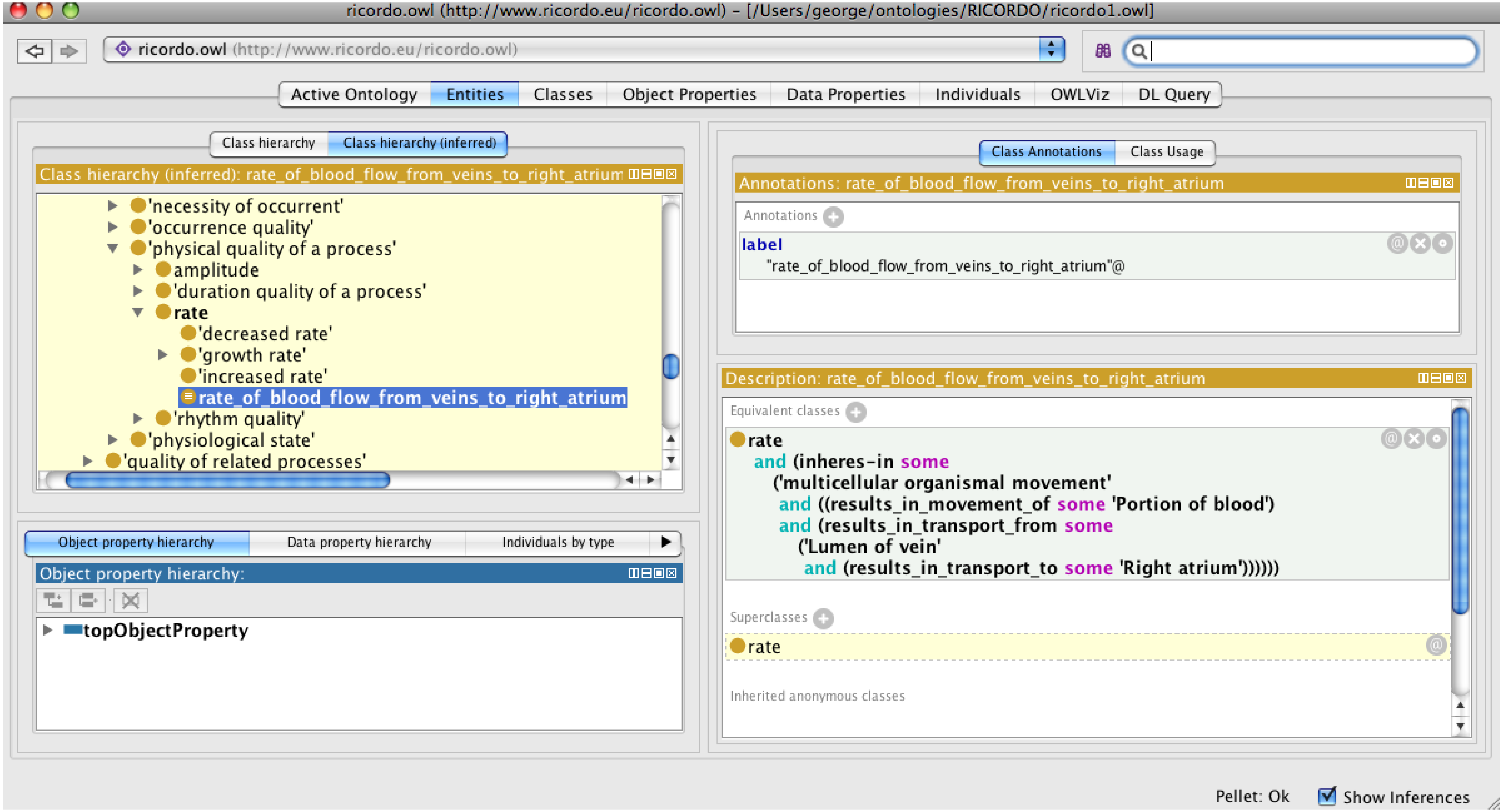
Volume_of_blood

- Volume_of_blood is a specialisation of Volume
- Volume_of_blood is a specialisation of the class of qualities of some portion of blood,

  that is to say, the class of entities that inhere in some portion of blood
- Formally:

  Volume and (inheres_in Some Blood)

Venous Return



Class hierarchy (inferred): rate_of_blood_flow_from_veins_to_right_atrium

- 'necessity of occurrent'
- 'occurrence quality'
- 'physical quality of a process'
  - amplitude
  - 'duration quality of a process'
  - rate
    - 'decreased rate'
    - 'growth rate'
    - 'increased rate'
    - rate_of_blood_flow_from_veins_to_right_atrium
  - 'rhythm quality'
- 'physiological state'
- 'quality of related processes'

Object property hierarchy:

- topObjectProperty

Annotations: rate_of_blood_flow_from_veins_to_right_atrium

Annotations

label
"rate_of_blood_flow_from_veins_to_right_atrium"@

Description: rate_of_blood_flow_from_veins_to_right_atrium

Equivalent classes

```
rate
  and (inheres-in some
     ('multicellular organismal movement'
      and ((results_in_movement_of some 'Portion of blood')
      and (results_in_transport_from some
         ('Lumen of vein'
          and (results_in_transport_to some 'Right atrium')))))
```

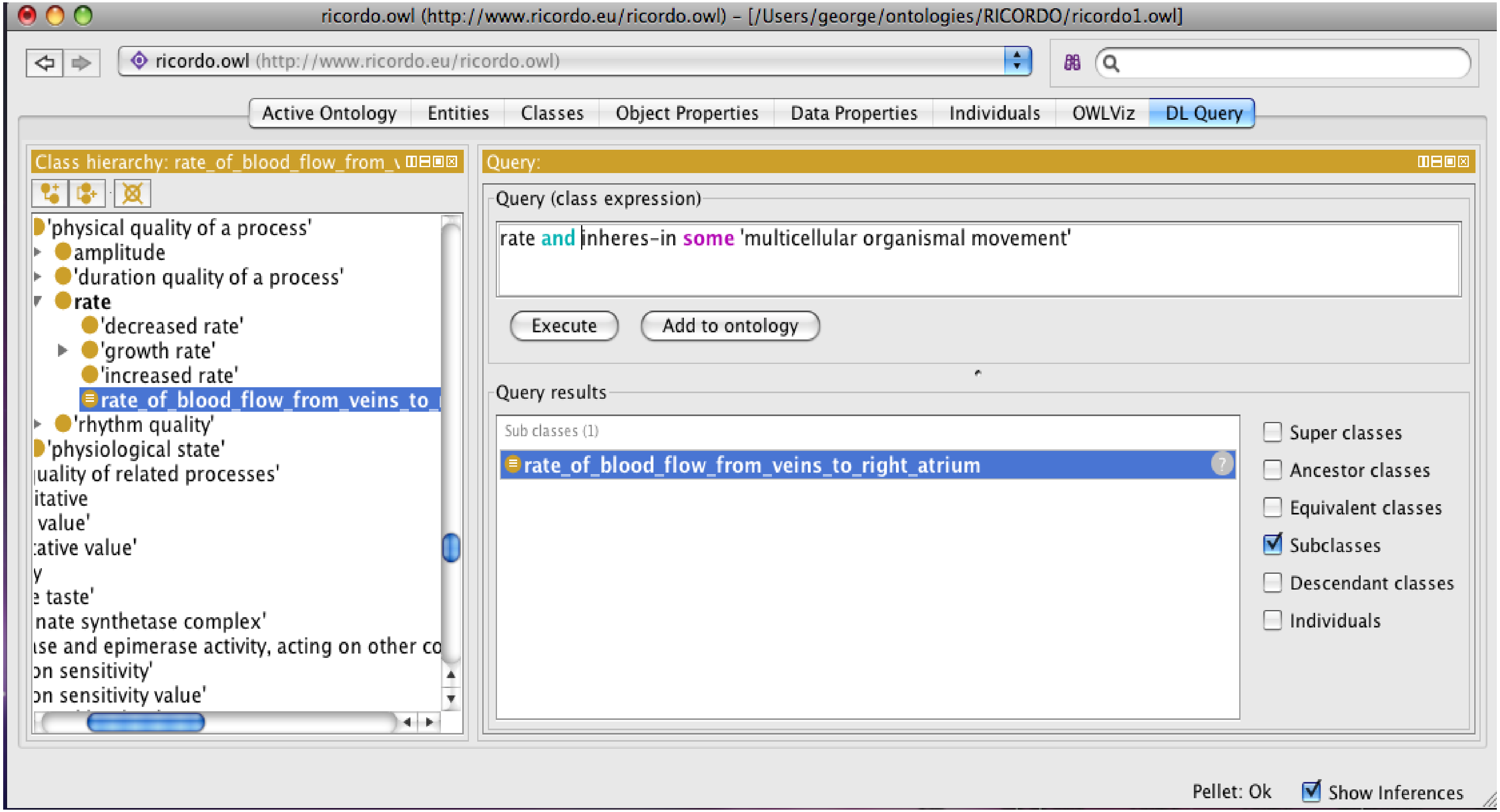
Superclasses

rate

Inherited anonymous classes

☑ Show Inferences

Reasoning



Query (class expression)

rate and inheres-in some 'multicellular organismal movement'

Query results

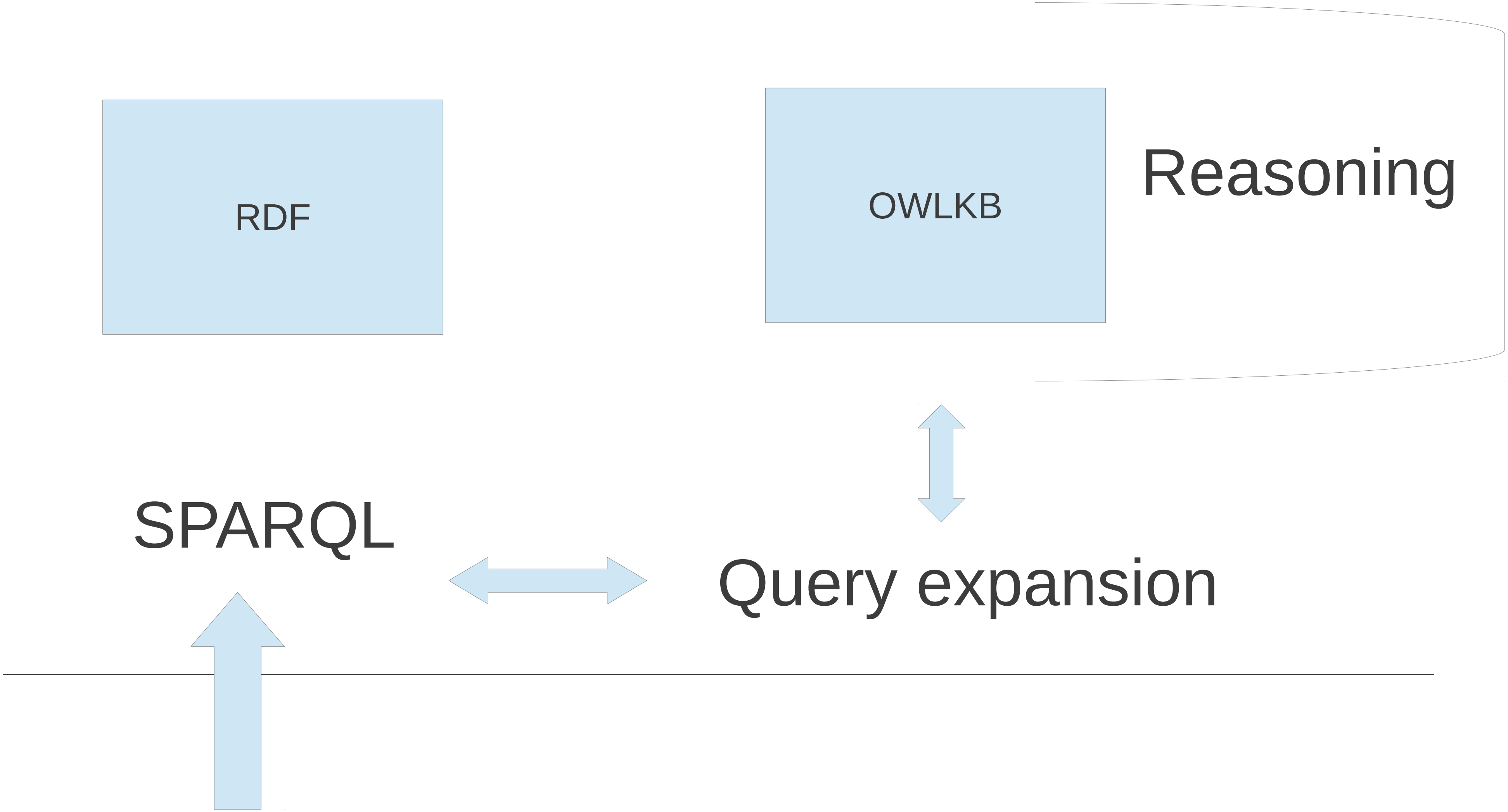
Sub classes (1)

rate_of_blood_flow_from_veins_to_right_atrium

☐ Super classes
☐ Ancestor classes
☐ Equivalent classes
☑ Subclasses
☐ Descendant classes
☐ Individuals

Pellet: Ok ☑ Show Inferences

Tools and Querying Workflow



RDF

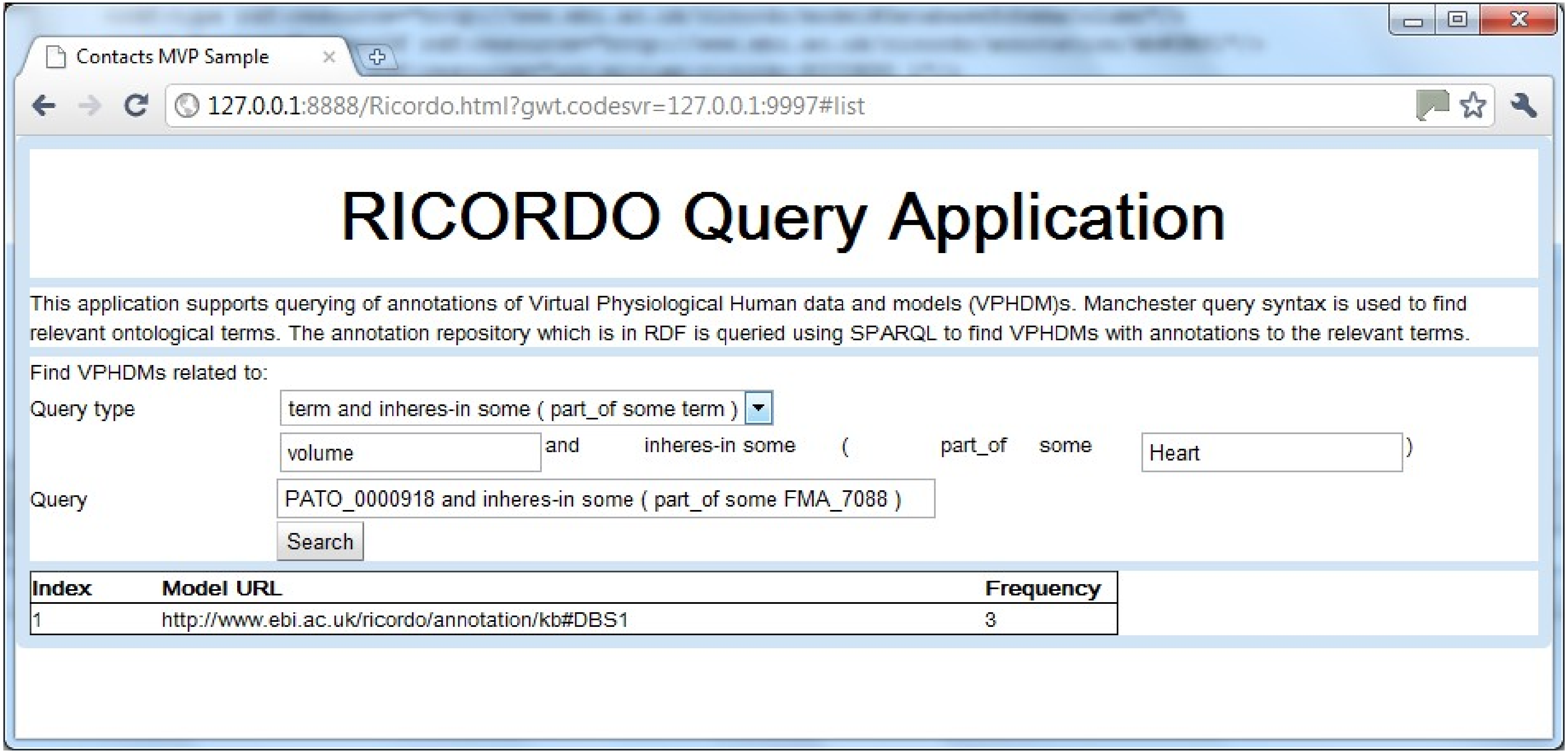
OWLKB

Reasoning

SPARQL

Query expansion

Searching with reasoning



RICORDO Query Application

This application supports querying of annotations of Virtual Physiological Human data and models (VPHDM)s. Manchester query syntax is used to find relevant ontological terms. The annotation repository which is in RDF is queried using SPARQL to find VPHDMs with annotations to the relevant terms.

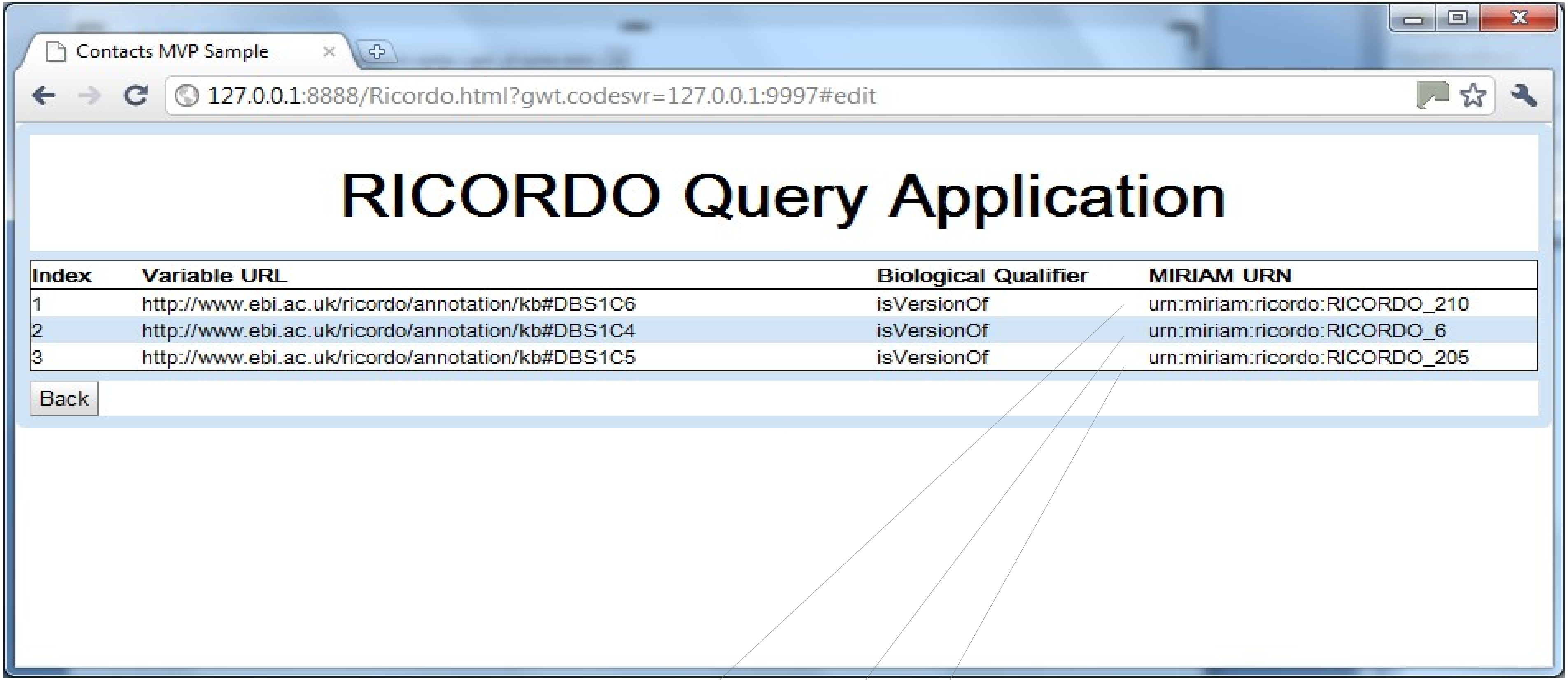
Find VPHDMs related to:

Query type: term and inheres-in some (part_of some term)

volume and inheres-in some (part_of some Heart)

Query: PATO_0000918 and inheres-in some (part_of some FMA_7088)

Index	Model URL	Frequency
1	http://www.ebi.ac.uk/ricordo/annotation/kb#DBS1	3



RICORDO Query Application

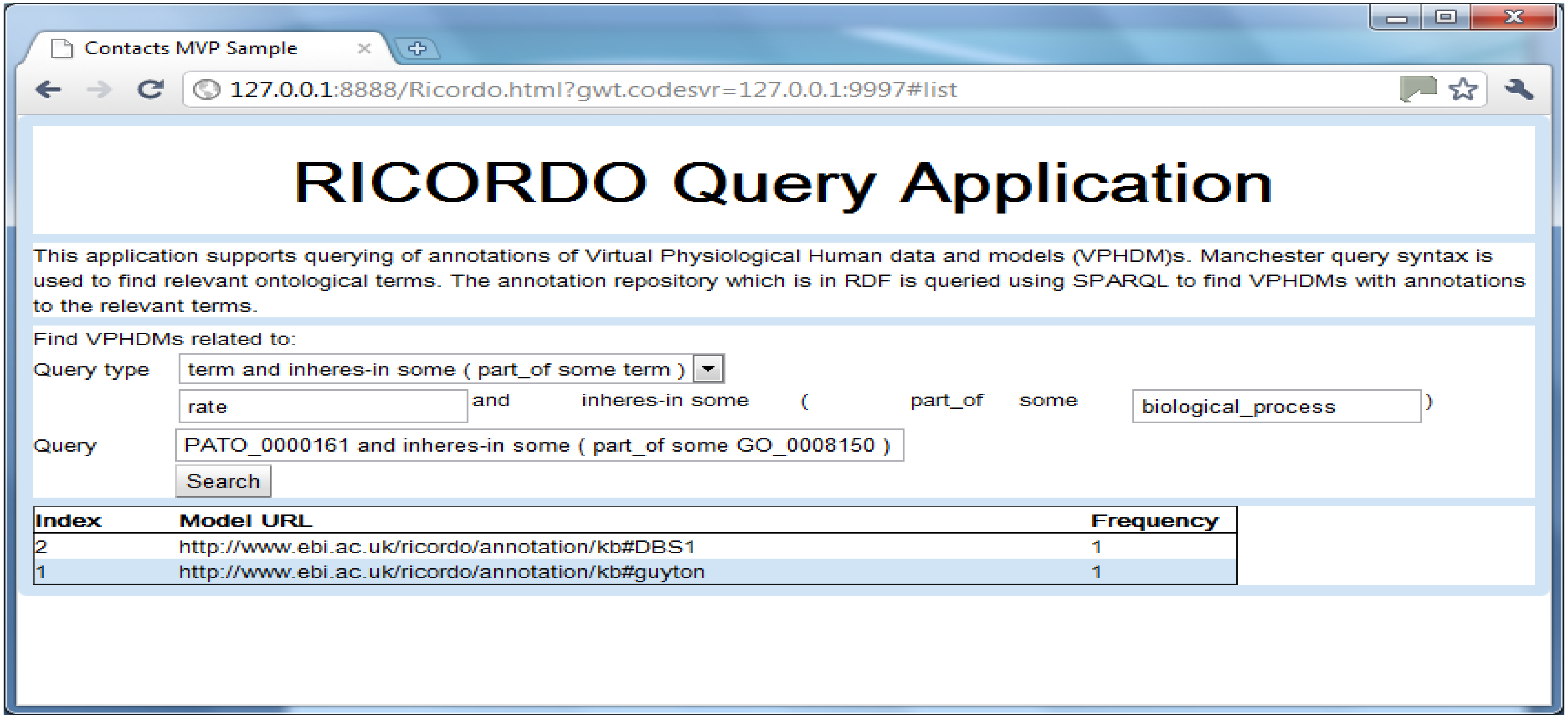
Index	Variable URL	Biological Qualifier	MIRIAM URN
1	http://www.ebi.ac.uk/ricordo/annotation/kb#DBS1C6	isVersionOf	urn:miriam:ricordo:RICORDO_210
2	http://www.ebi.ac.uk/ricordo/annotation/kb#DBS1C4	isVersionOf	urn:miriam:ricordo:RICORDO_6
3	http://www.ebi.ac.uk/ricordo/annotation/kb#DBS1C5	isVersionOf	urn:miriam:ricordo:RICORDO_205

Volume of the right ventricle in diastole

Volume of the left ventricle at end of diastole

Volume of the left ventricle at end of systole

Rates



RICORDO Query Application

This application supports querying of annotations of Virtual Physiological Human data and models (VPHDM)s. Manchester query syntax is used to find relevant ontological terms. The annotation repository which is in RDF is queried using SPARQL to find VPHDMs with annotations to the relevant terms.

Find VPHDMs related to:

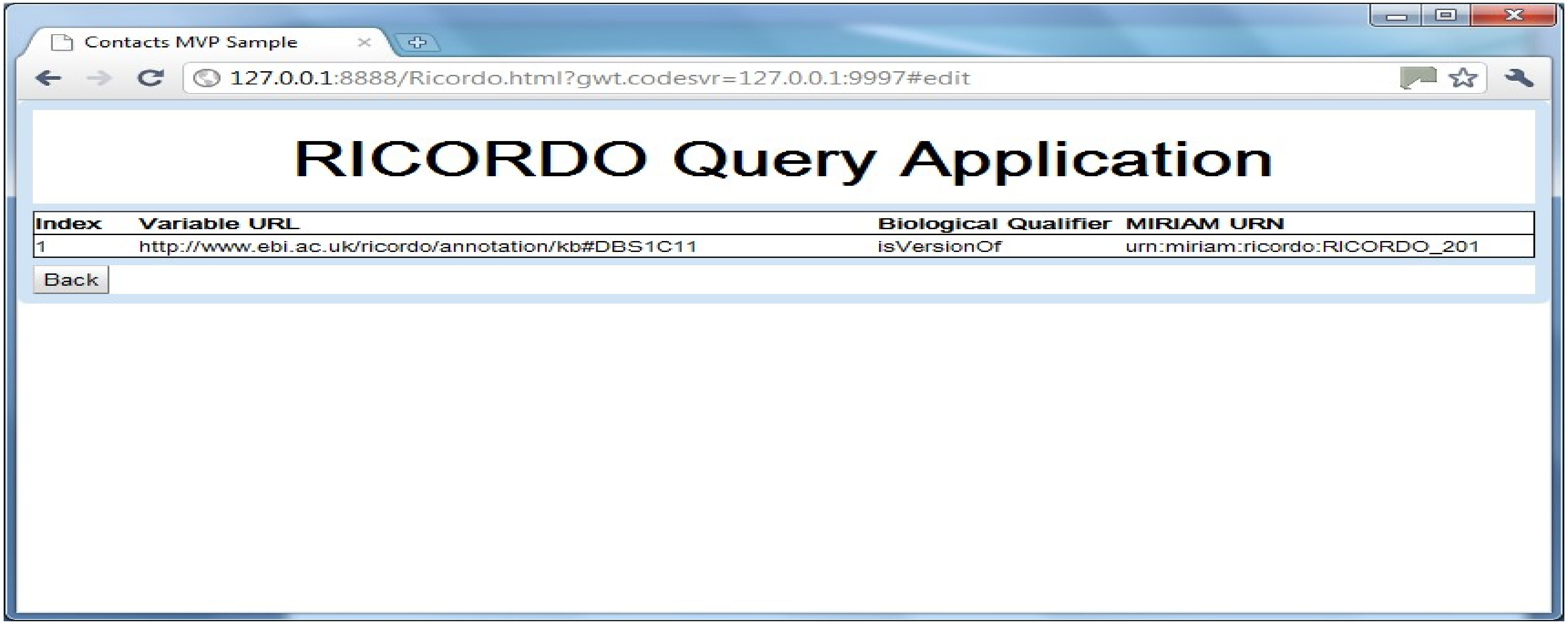
Query type: term and inheres-in some (part_of some term)

rate and inheres-in some (part_of some biological_process)

Query: PATO_0000161 and inheres-in some (part_of some GO_0008150)

Search

Index	Model URL	Frequency
2	http://www.ebi.ac.uk/ricordo/annotation/kb#DBS1	1
1	http://www.ebi.ac.uk/ricordo/annotation/kb#guyton	1



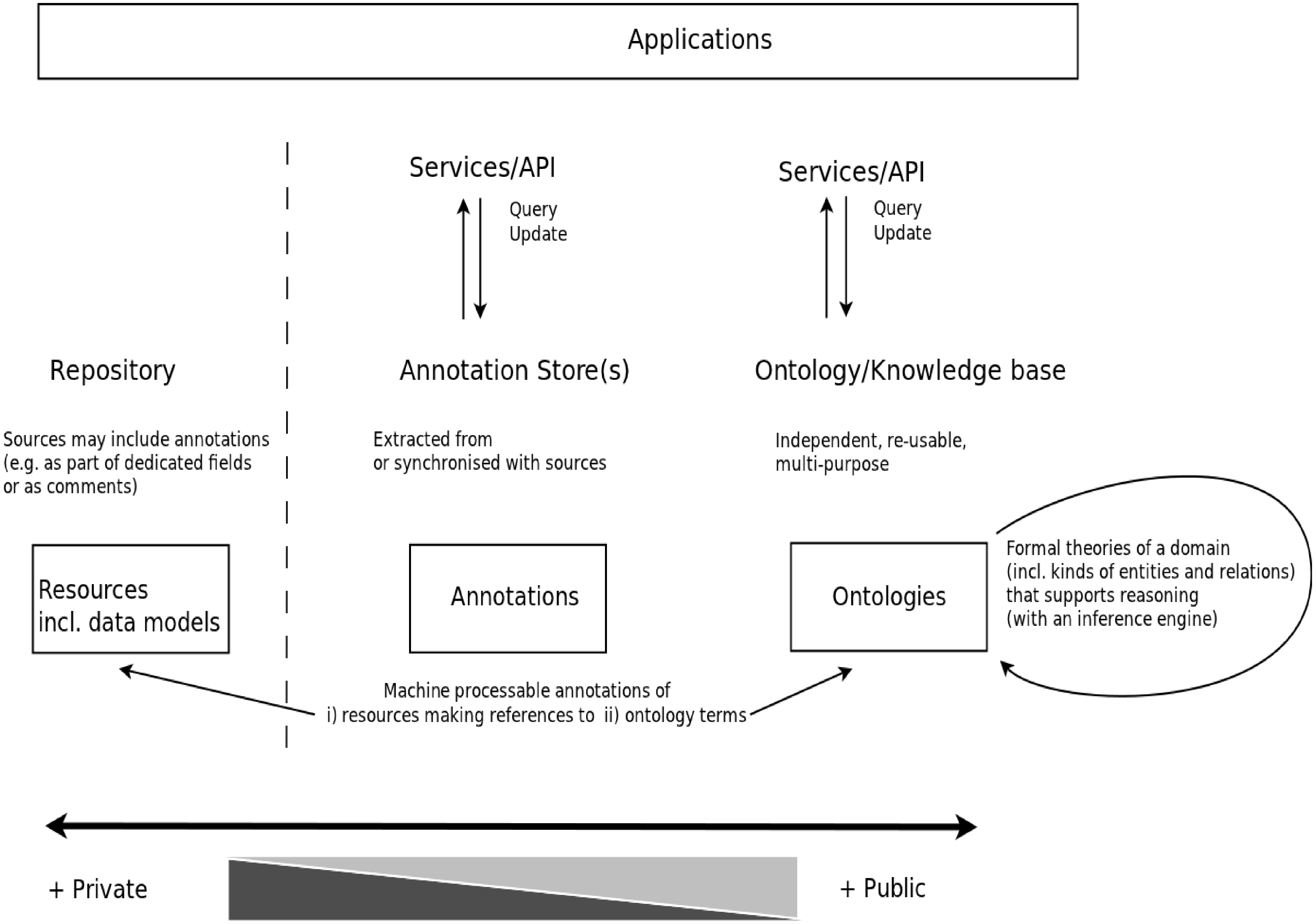
RICORDO Query Application

Index	Variable URL	Biological Qualifier	MIRIAM URN
1	http://www.ebi.ac.uk/ricordo/annotation/kb#DBS1C11	isVersionOf	urn:miriam:ricordo:RICORDO_201

Rate of blood flow from systemic veins to right atrium

a.k.a. Venous return

RICORDO Framework


Applications
Services/API
Query
Update
Services/API
Query
Update
Repository
Annotation Store(s)
Ontology/Knowledge base
Sources may include annotations (e.g. as part of dedicated fields or as comments)
Extracted from or synchronised with sources
Independent, re-usable, multi-purpose
Resources incl. data models
Annotations
Ontologies
Formal theories of a domain (incl. kinds of entities and relations) that supports reasoning (with an inference engine)
Machine processable annotations of
i) resources making references to ii) ontology terms
+ Private
+ Public

Main annotation strategy

- 1) Custodians of (VPH)DMRs generate annotations in compliance with RICORDO guidelines
- 2) RICORDO maintains a reference store of voluntarily submitted and publicly available annotations and of composites they use
- 3) Guidelines include reference implementation for tools and compliance criteria (relating to both annotations and composites instrumental to them)

RICORDO Software packages

Locally deployable solutions for:

Web services access to the **RDF store**
Core development site: https://github.com/sarala/ricordo-rdfstore

Knowledge base creation (ontology content is configurable) and support reasoning;
Core development site: https://github.com/sarala/ricordo-owlkb

Web services access to the knowledge base
Core development site: https://github.com/sarala/ricordo-owlkb-ws

Deployable
Reusable
Configurable

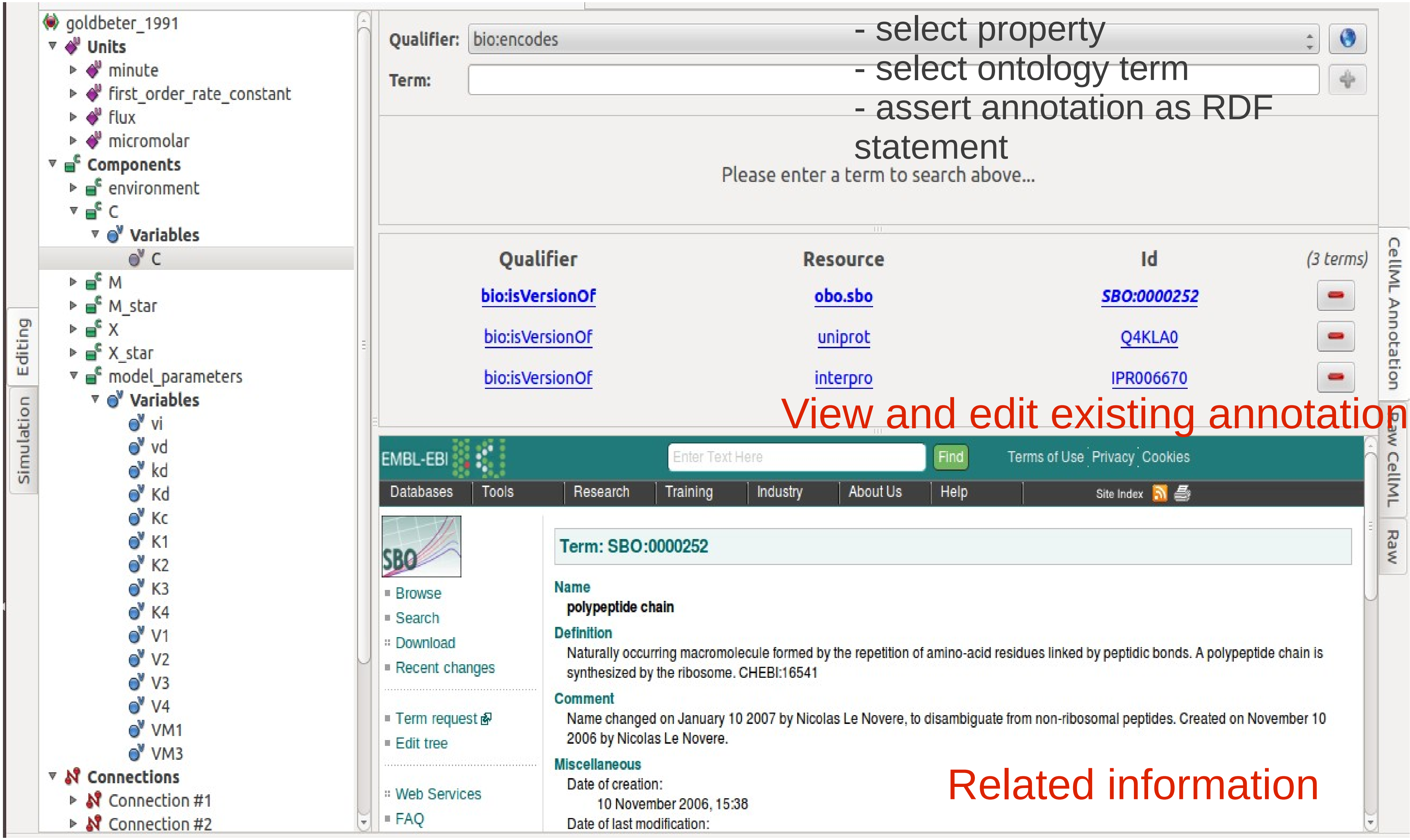
Documentation: https://sites.google.com/site/ricordotoolset/

Credit and contact: Sarala Wimalaratne, sarala@ebi.ac.uk

OpenCOR (CellML editor)

Browse model
and select element

Add annotations:
- select property
- select ontology term
- assert annotation as RDF statement



Qualifier: bio:encodes

Term:

Please enter a term to search above...

Qualifier	Resource	Id	(3 terms)
bio:isVersionOf	obo.sbo	*SBO:0000252*	➖
bio:isVersionOf	uniprot	Q4KLA0	➖
bio:isVersionOf	interpro	IPR006670	➖

View and edit existing annotations

Term: SBO:0000252

Name
polypeptide chain

Definition
Naturally occurring macromolecule formed by the repetition of amino-acid residues linked by peptidic bonds. A polypeptide chain is synthesized by the ribosome. CHEBI:16541

Comment
Name changed on January 10 2007 by Nicolas Le Novere, to disambiguate from non-ribosomal peptides. Created on November 10 2006 by Nicolas Le Novere.

Miscellaneous
Date of creation:
10 November 2006, 15:38
Date of last modification:

Related information

Open Questions

- Distribution/federation in infrastructure
- Operationalisation of knowledge

Further facts

- Multiplicity of sources
- physical separation
- different access policies and rights

- different motivations and goals (finding relevant data vs composing new models)

- Solution is deployable, local systems in addition to centralised reference store and knolwedge base

Infrastructures

Model 1

- Centralised repository of annotation
- Centralised knowledge base
- Integrated querying

Model 2

- Multiple repositories
- Multiple knowledge bases
- Configurable querying (select metadata repositores, select ontologies etc)

Issues

- Federation and distributed query
- Data annotation curation
- Versioning and update
- Right managements
- Provenance and trust
- Concurrent access
- A la carte configuration

Operational integration

- Semantic integration = standardised, consistent use of knowledge resources (ontologies) and annotation mechanisms resulting in a unified knowledge representation
- Supports search

- However, goals include not only finding data, models and other resources, also do something with these
- Requires access and ability to retrieve resources
- Requires ability to transform, configure so as to be used in software applications (e.g. model simulation)
- Role of ontologies and knowledge representation in data and model specification and application workflows

Summary

- RICORDO looks for a standardised approach to ontology-based knowledge management of biomedical resources
- Conceptual approach is relatively clear and mainstream ontological and knowledge engineering
- Implementation meets a varied and open ended set of use cases with constraints which are not all technical (e.g. proprietary information)
- Central model as reference implementation
- Real-life deployment involves local, distributed solutions
- Role of ontology and knowledge representation in operational integration