Structuring and linking of biological data guided by ontologies and the organizing principles of mathematical models

InOvive Meeting

December 15, 2022

People involved in the PhD

- **Olivier Inizan**, INRAE/LISN
- **Fatiha Saïs**, LISN (supervision)
- **Danaï Symeonidou**, INRAE (co-supervision)
- **Anne Goelzer**, INRAE (co-supervision)

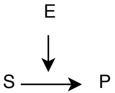
Context

System biology, whole cell modeling, BioSys group @INRAE

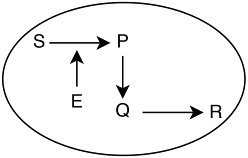
- A modeling approach that is a powerfull principle to understand the organisation of biological networks in bacteria
- Exploit this principle in the field of knowledge representation
- Molecular scale, entities are molecules

$H2O$, $Glucose$, DNA, $Proteins$, $Enzymes$, ... are molecules

In the cell molecules are transformed through **biochemical reactions** (BR)



One biochemical reaction



A network of reactions

Molecular Biology

Molecular knowledge is complex

- different types of molecules
- different types of processes
- different scales

Experimental data

- heterogeneous, different types
- fragmented, not always available

Molecular Biology

Molecular knowledge is complex

different types of molecules

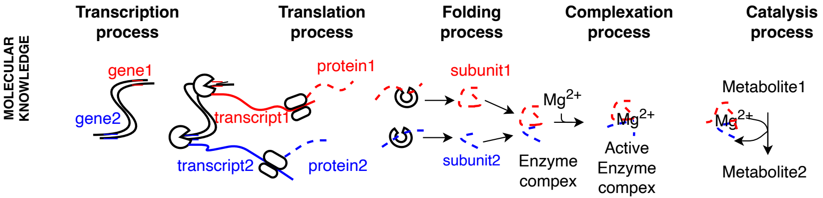
different types of processes

different scales

Experimental data

heterogeneous, different types

fragmented, not always available



Molecular Biology

Molecular knowledge is complex

different types of molecules

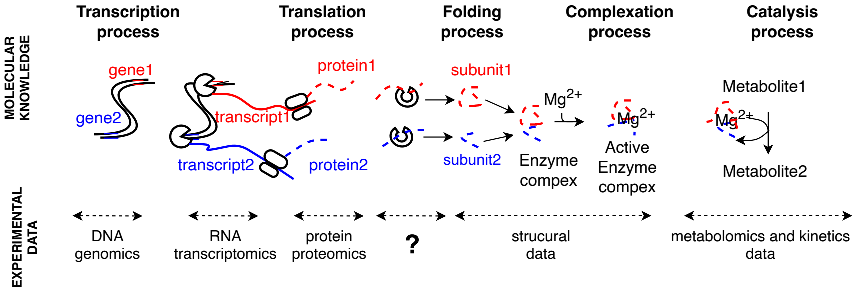
different types of processes

different scales

Experimental data are

heterogeneous, different types

fragmented, not always available



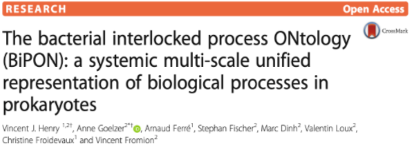
General questions

Taking into account the facts that (i) molecular biology provides complex knowledge and (ii) experimental data are heterogeneous and fragmented:

- Is it possible to link together these data and this knowledge?
- Is it possible to check the consistency of these data and knowledge?
- Is it possible to refine the links with fresh data and knowledge while keeping the consistency?

First Steps / State of art

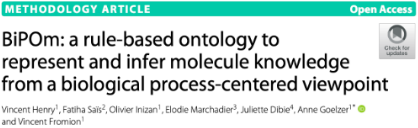
Formal representations of knowledge and data



RESEARCH Open Access

The bacterial interlocked process ONtology (BiPON): a systemic multi-scale unified representation of biological processes in prokaryotes

Vincent J. Henry [1,2†], Anne Goelzer [2*†], Arnaud Ferré [1], Stephan Fischer [2], Marc Dinh [2], Valentin Loux [2], Christine Froidevaux [1] and Vincent Fromion [2]



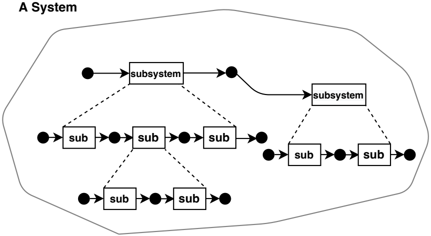
METHODOLOGY ARTICLE Open Access

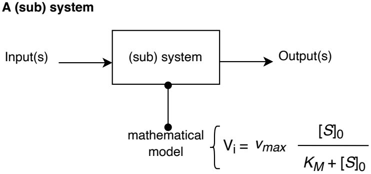
BiPOm: a rule-based ontology to represent and infer molecule knowledge from a biological process-centered viewpoint

Vincent Henry [1], Fatiha Saïs [2], Olivier Inizan [1], Elodie Marchadier [3], Juliette Dibie [4], Anne Goelzer [1*] and Vincent Fromion [1]

1. Represent the molecular knowledge
   - The ontologies BiPON and BiPOM
   - They use the systemic approach to tackle the complexity
2. Represent the experimental data
   - Organize the data (since they are heterogeneous and fragmented)
   - Link this representation with the representation of knowledge

The systemic approach

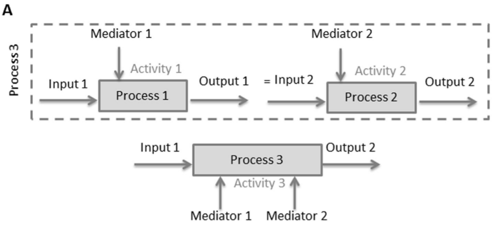




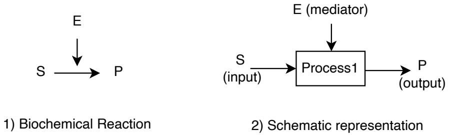
- Aim: simulate the behavior of a given system (for us the system is a bacteria)
- An interesting property: tackle the complexity of a given system
- How?: break down this system into connected sub-systems
- A sub-system: inputs, outputs and a function to fulfill
- The transformation of inputs to outputs is done through a mathematical model

BiPON/BiPOM and the systemic approach

- In BiPOM a sub-system is a Biological Process



- The **Biochemical Reaction** (BR) transforms the molecule S to the molecule P and is mediated by the molecule E:



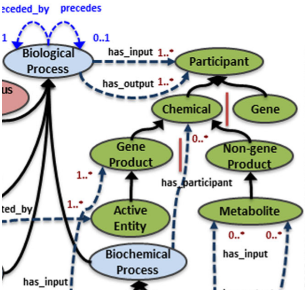
1) Biochemical Reaction

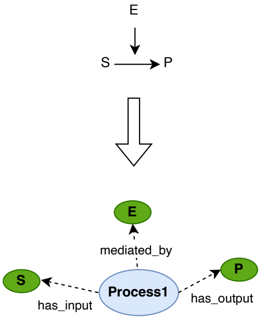
2) Schematic representation

BiPON/BiPOM

The limit of BiPON/BiPOM to represent experimental data

- Individuals that are inputs (resp. outputs or mediators) of a *BiologicalProcess* are always a molecule
- Ex: in BiPOM, individuals of class *Chemical*, *Gene*, (non)*GeneProduct*, *Metabolite* are molecules

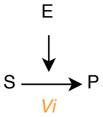




- Well suited to represent *BR* since by definition a *BR* transform molecules to another molecules
- But *BR* are not directly relied to experimental data

Mathematical models are relied to experimental data

Ex: the speed of $S \longrightarrow P$ is given by the following model

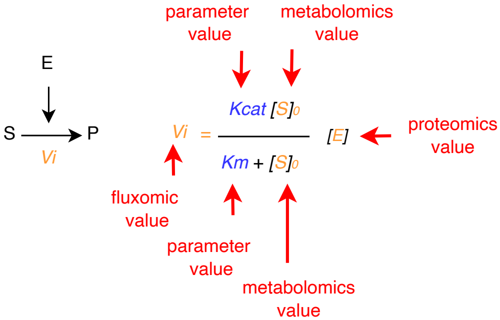


$$Vi = \frac{Kcat\,[S]_0}{Km + [S]_0}\,[E]$$

- A mathematical model: variables and parameters
- The values of variables and parameters represent (experimental) data
- The model organizes variables and parameters: $V_i = f(S_0, E, K_{cat}, K_m)$
    1. The inputs of the model are S_0, E, K_{cat}, K_m
    2. The output of the model is V_i

Mathematical models are relied to experimental data

The model organizes (experimental) data



$$Vi = \frac{Kcat\,[S]_0}{Km + [S]_0}\,[E]$$

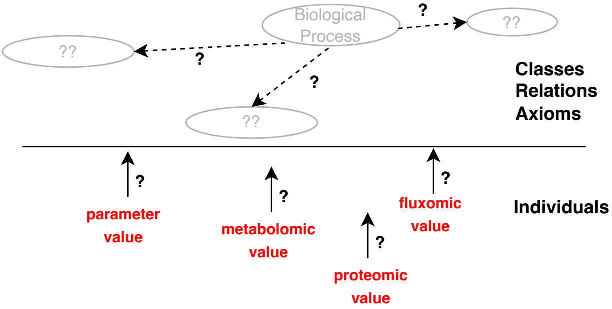
- The values of variables and parameters represent (experimental) data
- The model organizes these data: *fluxomics* = *f*(*proteomics*, *metabolomics*, *parameter_value*)
- BiPON/BiPOm: no classes to represent these values

Questions from a knowledge representation perspective

A new ontology for experimental data

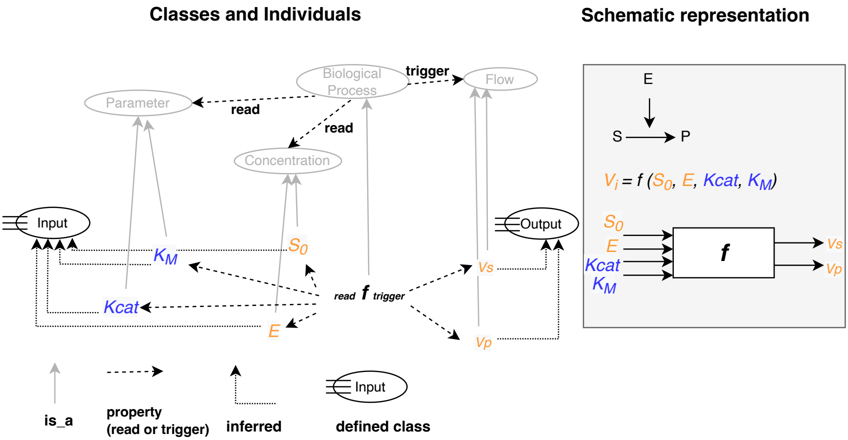
Through variables and parameters, models are data organizers:

1. Can we find classes that represent variables and parameters?
2. Can we organize these classes like variables and parameters are organized in the models?



The new ontology

Input and Output are inferred

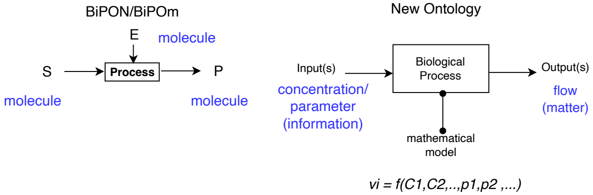


Classes and Individuals

Schematic representation

$$Input \equiv Concentration \sqcap \exists is\_read\_by.BiologicalProcess \sqcup Parameter \sqcap \exists is\_read\_by.BiologicalProcess$$
$$Output \equiv Flow \sqcap \exists triggered\_by.BiologicalProcess$$

Summary and discussion



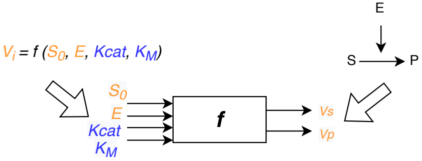
1. Since Inputs/Outputs are different classes, how to connect *BiologicalProcesses*? $S \longrightarrow P \longrightarrow Q \longrightarrow R$
   - The pool connects flows and concentration [1]
2. Generalization of Inputs/Outputs:
   - Inputs are the **information** that are the cause of the process (T°, pression, etc ...)
   - Outputs are **matter** exchange between the process and the env

[1] Inizan O. et al. 2021, An ontology to structure biological data: the contribution of mathematical models

Discussion

In the example $S \longrightarrow P$:

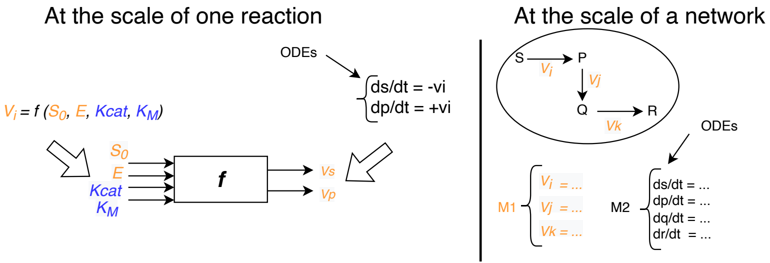
1. Where does the information come from?
2. Where does the matter exchange come from?



- The speed v_i contains the information, the inputs
- The *BR* contains matter exchange, the outputs

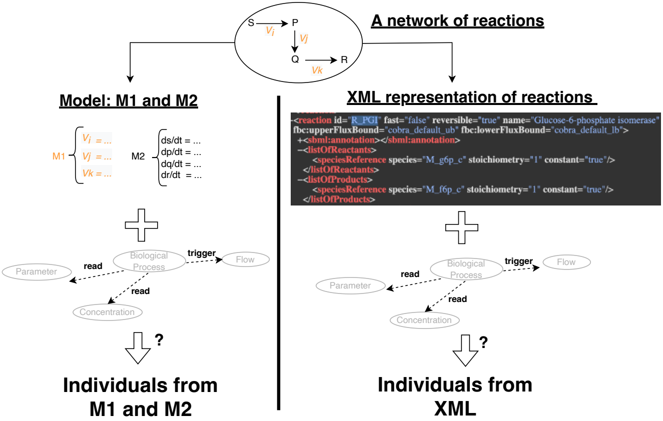
Discussion

1. Where does the information come from?
2. Where does the matter exchange come from?



- The matter exchange is also avalaible in another type of models: Ordinary Differential Equations (ODEs)
- A stricking fact: 2 sets of math expressions ($M1$ and $M2$) contains all the information we need to populate the ontology

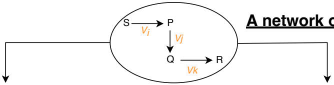
Questions



```
-<reaction id="R_PGI" fast="false" reversible="true" name="Glucose-6-phosphate isomerase"
fbc:upperFluxBound="cobra_default_ub" fbc:lowerFluxBound="cobra_default_lb">
  +<sbml:annotation></sbml:annotation>
  −<listOfReactants>
      <speciesReference species="M_g6p_c" stoichiometry="1" constant="true"/>
   </listOfReactants>
  −<listOfProducts>
      <speciesReference species="M_f6p_c" stoichiometry="1" constant="true"/>
   </listOfProducts>
```

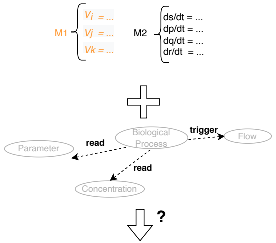
1. Given only *M*1 and *M*2 which model the dynamic of a network of reactions, can we populate the ontology?
2. Conversely, can we populate the ontology with only *BR* (in XML)?
3. What is the motivation to populate the ontology with 2 different sources?

Questions



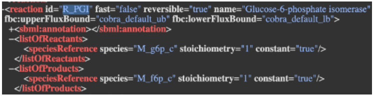
A network of reactions

Model: M1 and M2

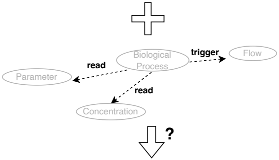


Individuals from M1 and M2

XML representation of reactions



```
-<reaction id="R_PGI" fast="false" reversible="true" name="Glucose-6-phosphate isomerase"
fbc:upperFluxBound="cobra_default_ub" fbc:lowerFluxBound="cobra_default_lb">
  +<sbml:annotation></sbml:annotation>
  -<listOfReactants>
      <speciesReference species="M_g6p_c" stoichiometry="1" constant="true"/>
   </listOfReactants>
  -<listOfProducts>
      <speciesReference species="M_f6p_c" stoichiometry="1" constant="true"/>
   </listOfProducts>
```



Individuals from XML

1. Given only *M*1 and *M*2 which models the dynamic of a network of reactions, can we populate the ontology?
2. Conversely, can we populate the ontology with only *BR* (in XML)?
3. What is the motivation to populate the ontology with different sources?

General questions

What is the motivation to populate the ontology with different sources?

Taking into account the facts that (i) molecular biology provides complex knowledge and (ii) experimental data are heterogeneous and fragmented:

- Is it possible to link together these data and this knowledge?
- Is it possible to check the consistency of these data and knowledge?
- Is it possible to refine the links with fresh data and knowledge while keeping the consistency?

Since the information of different sources are represented with the same classes and relations, a linkage can be considered.

Perspectives

1. Consider a linkage and (possibly) evaluate the consistency: we need two sources of informations represented with the ontology
2. Refine the links with another source of information

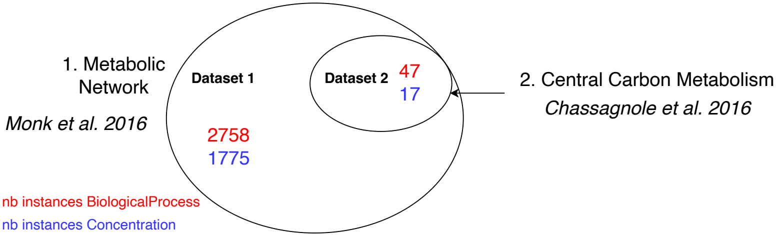
Perspectives

Brief description of 2 examples of sources of informations

- XML representation of reactions, Source 1 *Monk et al. 2016*
- M1 and M2, Source 2 *Chassagnole et al. 2002*

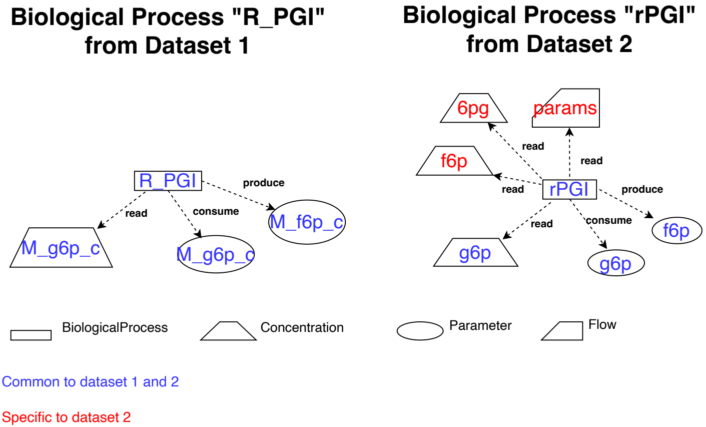
1. **Source 1** describes the reactions of the entire metabolic network of the bacteria *E.coli*.
2. **Source 2** is a model that study the dynamic of a sub-network of the metabolic network of the bacteria *E.coli*.

The ontology has been populated with each source, we have obtained 2 datasets:


1. Metabolic Network
Monk et al. 2016
Dataset 1
2758
1775
Dataset 2
47
17
2. Central Carbon Metabolism
Chassagnole et al. 2016
nb instances BiologicalProcess
nb instances Concentration

In the perspective to link biological processes

The process named "PGI" in each dataset



Biological Process "R_PGI" from Dataset 1

Biological Process "rPGI" from Dataset 2

We want to investigate contextual identity links

Conclusion

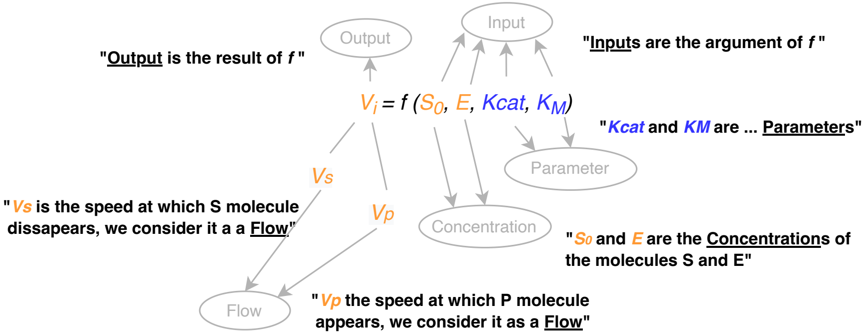
- Contributions:
  1. An ontology that was originally designed to structure data but that would allow us to consider the linkage between different sources of informations
  2. A RDF vocabulary for the representation and the querying of complex math expressions: the ODEs
- Short term perspective: investigate contextual identity links with the aid of datasets produced by us
- Middle term perspective, consider a new source of information that presents interesting characteristics:
  1. Another modelisation paradigm: constraints based instead of ODEs
  2. Covers a wider range biological processes

Thank you for your attention

Backup slides

A deeper look at the models

The experts help us to design the classes



$V_i = f\,(S_0, E, Kcat, K_M)$

Output

Input

Parameter

Concentration

Flow

Vs

Vp

"Output is the result of *f* "

"Inputs are the argument of *f* "

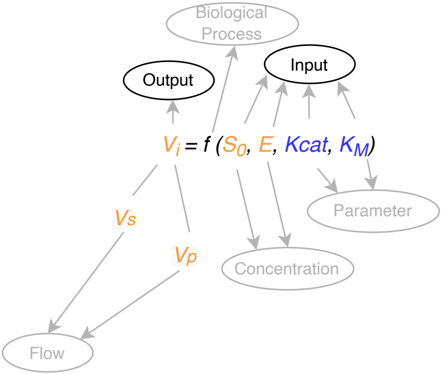
"*Kcat* and *KM* are ... Parameters"

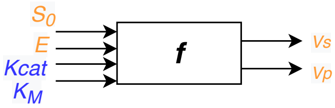
"*Vs* is the speed at which S molecule dissapears, we consider it a a Flow"

"*S0* and *E* are the Concentrations of the molecules S and E"

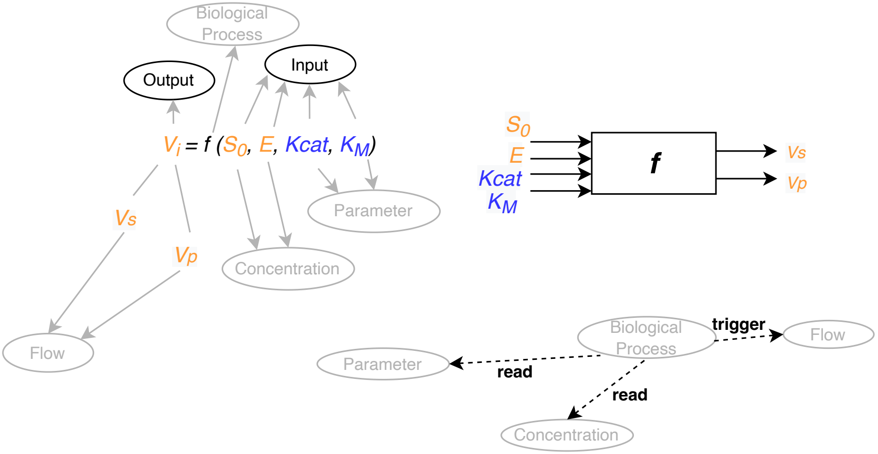
"*Vp* the speed at which P molecule appears, we consider it as a Flow"

The function f is a process





The ontology



Biological Process

Output

Input

$V_i = f\,(S_0, E, Kcat, K_M)$

Vs

Vp

Parameter

Concentration

Flow

S_0

E

$Kcat$

K_M

f

Vs

Vp

Biological Process

trigger

Flow

Parameter

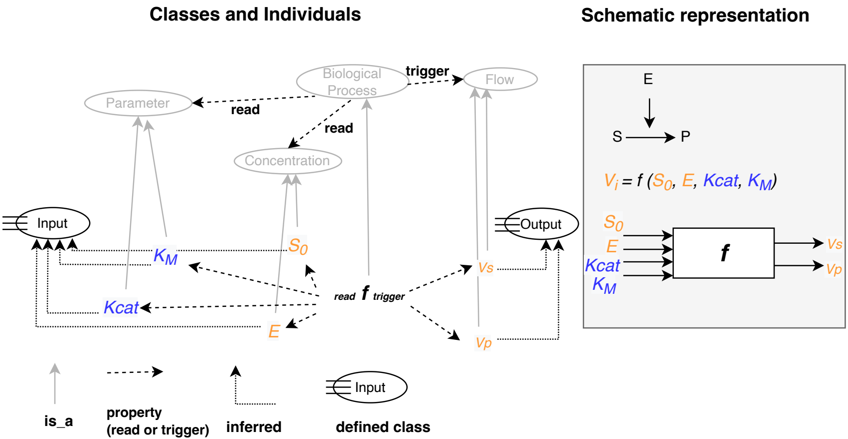
read

read

Concentration

The new ontology

Input and Output are inferred



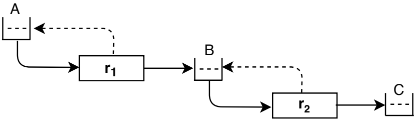
$Input \equiv Concentration \sqcap \exists is\_read\_by.BiologicalProcess \sqcup Parameter \sqcap \exists is\_read\_by.BiologicalProcess$
$Output \equiv Flow \sqcap \exists triggered\_by.BiologicalProcess$

RDF Vocabulary

Problem: query math expressions to extract biological processes, concentrations, flows

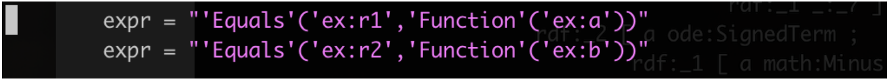
$$M_1 = \{r_1 = f_1(a), r_2 = f_2(b)\}$$

$$M_2 = \left\{ \begin{array}{rcccc} \text{Evolution} & = & \text{Production} & - & \text{Degradation} \\ \dfrac{da}{dt} & = & & - & r_1 \\ \dfrac{db}{dt} & = & r_1 & - & r_2 \\ \dfrac{dc}{dt} & = & r_2 & & \end{array} \right.$$

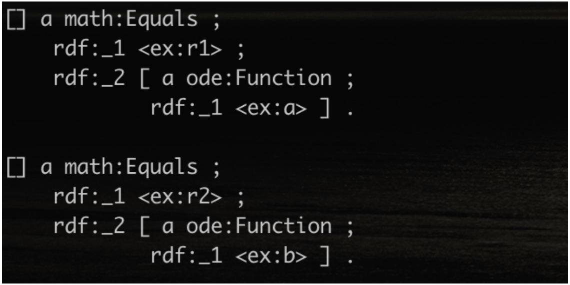


RDF Vocabulary

$$M_1 = \{r_1 = f_1(a), r_2 = f_2(b)\}$$



```
    expr = "'Equals'('ex:r1','Function'('ex:a'))"
    expr = "'Equals'('ex:r2','Function'('ex:b'))"
```

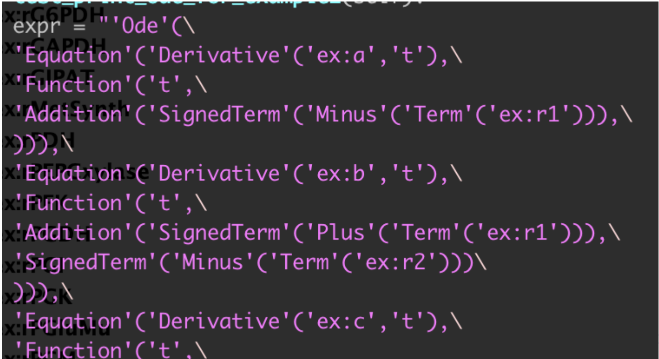


```
[] a math:Equals ;
   rdf:_1 <ex:r1> ;
   rdf:_2 [ a ode:Function ;
            rdf:_1 <ex:a> ] .

[] a math:Equals ;
   rdf:_1 <ex:r2> ;
   rdf:_2 [ a ode:Function ;
            rdf:_1 <ex:b> ] .
```

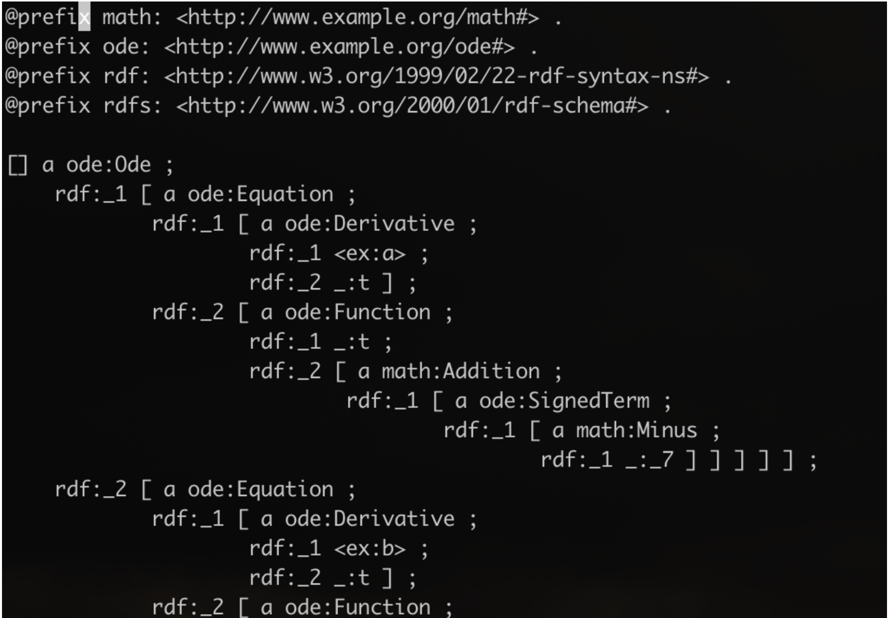
RDF Vocabulary

$$M_2 = \left\{ \begin{array}{rcccc} \text{Evolution} & = & \text{Production} & - & \text{Degradation} \\ \frac{da}{dt} & = & & - & r_1 \\ \frac{db}{dt} & = & r_1 & - & r_2 \\ \frac{dc}{dt} & = & r_2 & & \end{array} \right.$$



```
test_print_ode_for_example2(self):
expr = "'Ode'(\
'Equation'('Derivative'('ex:a','t'),\
'Function'('t',\
'Addition'('SignedTerm'('Minus'('Term'('ex:r1'))),\
))),\
'Equation'('Derivative'('ex:b','t'),\
'Function'('t',\
'Addition'('SignedTerm'('Plus'('Term'('ex:r1'))),\
'SignedTerm'('Minus'('Term'('ex:r2')))\
))),\
'Equation'('Derivative'('ex:c','t'),\
'Function'('t',\
```

RDF Vocabulary



```
@prefix math: <http://www.example.org/math#> .
@prefix ode: <http://www.example.org/ode#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

[] a ode:Ode ;
    rdf:_1 [ a ode:Equation ;
            rdf:_1 [ a ode:Derivative ;
                    rdf:_1 <ex:a> ;
                    rdf:_2 _:t ] ;
            rdf:_2 [ a ode:Function ;
                    rdf:_1 _:t ;
                    rdf:_2 [ a math:Addition ;
                            rdf:_1 [ a ode:SignedTerm ;
                                    rdf:_1 [ a math:Minus ;
                                            rdf:_1 _:_7 ] ] ] ] ] ;
    rdf:_2 [ a ode:Equation ;
            rdf:_1 [ a ode:Derivative ;
                    rdf:_1 <ex:b> ;
                    rdf:_2 _:t ] ;
            rdf:_2 [ a ode:Function ;
```