



# Spreadsheets to OWL with Populous

**Mikel Egaña Aranguren**

3205 School of Computer Science  
Universidad Politécnica de Madrid (UPM)  
28660 Boadilla del Monte  
Spain

Ontology Engineering Group (OEG)  
<http://www.oeg-upm.net>

[megana@fi.upm.es](mailto:megana@fi.upm.es)  
<http://mikeleganaaranguren.com>

**Robert Stevens**

Biohealth Informatics Group  
School of Computer Science  
University of Manchester  
Manchester  
UK

[Robert.steven@manchester.ac.uk](mailto:Robert.steven@manchester.ac.uk)

**Simon Jupp**

Functional Genomic Group  
European Bioinformatics Institute  
Wellcome Trust Genome Campus  
Hinxton  
UK

[jupp@ebi.ac.uk](mailto:jupp@ebi.ac.uk)



Engaging life scientists in data annotation and ontology population

Protégé and OWL are scary..

Need for simple “form filling” style of knowledge gathering and describing data - so we use spreadsheets.

Q1 How do we get people to annotate data in spreadsheets according to ontologies?

Q2 How do we transform those spreadsheets into sets of axioms?

Especially if one doesn't know OWL;

Hard to do complex patterns of axioms;

Hard to be consistent and conform to a style;

Hard to re-factor an ontology's content

Doing all this in bulk is tedious and error prone

Knowledge

All Eukaryotic Cells are either nucleated or anucleate, some cells are multinucleate

Ontologically

'Eukaryotic Cells' *has\_nucleation* some 'Nucleation'  
'Nucleation' *subClassOf* {mononucleate , binucleate , polynucleate , anucleate}

Differentia

'Eukaryotic Cells' *has\_nucleation* some 'Nucleation'  
'Nucleation' *subClassOf* {mononucleate , binucleate , polynucleate , anucleate}

Real Examples

'Eukaryotic Cells'	'Nucleation'
Mononuclear phagocyte	mononucleate
Flight Muscle cell	multinucleate
Red Blood cell	anucleate

Repetative pattern

**'Protein'**  
*has\_molecular\_function* some **'Molecular Function'**  
*is\_capable\_of* some **'Biological Process'**  
*located\_in* some **'Cellualr component'**

Axioms often added in regular ways

There are often patterns of axioms for a particular way of representation

There are also design patterns – standard well recognised solutions

Analogous to software patterns

Doing the same thing in the same way... it's a good thing

Want consistent axiom generation

Want to write axioms according to patterns

Separate knowledge gathering from axiom generation

Engage domain experts not experts in OWL and/or ontologies

Validate content to go into the ontology

Do all of this in a familiar environment i.e. spreadsheets

Spreadsheets are often used simply to organise data

Basic tabulation

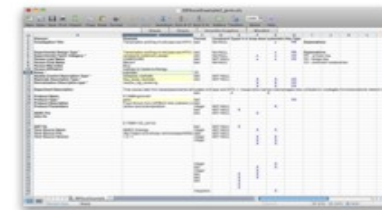
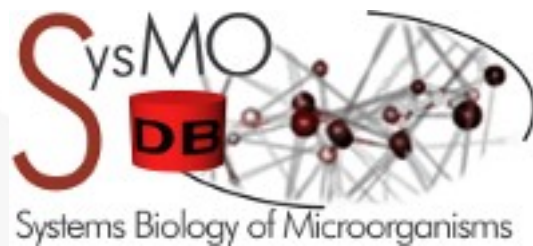
Saying the same kinds of things repeatedly

A very familiar environment

Want to capitalise on this...

<http://www.rightfield.org.uk>

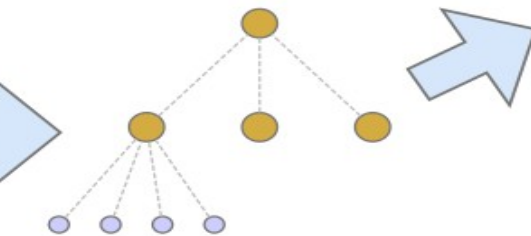
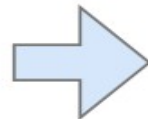
- Semantic Annotation by Stealth



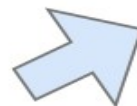
Terms Embedded into  
Excel Workbook



Ontology



"Portion" of ontology terms





# Excel validations

The screenshot shows an Excel spreadsheet titled 'IDFExcelExample2\_jerm.xls'. The spreadsheet has columns A through G. Column A contains 'Element' labels, column B contains 'Example' values, column C contains 'Format' (text, integer, date), column D contains 'Constraint' (Not NULL, NOT NULL), column E contains 'Typed in b drop-down automatic', column F contains 'drop-down automatic', and column G contains 'automatic'. A dropdown menu is open for the 'Roles' field in row 11, listing various roles. The roles listed are: array\_manufacturer, biomaterial\_provider, biosequence\_provider, consortium\_member, consultant, curator, data\_analyst, data\_coder, funder, hardware\_manufacturer, institution, investigator, software\_manufacturer, and submitter (checked). The spreadsheet also shows other rows with data and validation rules.

Element	Example	Format	Constraint	Typed in b drop-down automatic	drop-down automatic	automatic
Investigation Title	Transcription profiling of wild-type and ATF3 -	text	Not NULL			x
Experimental Design Type *	Transcription profiling of wild-type and ATF3 -	text	Not NULL			x
Experimental Factor Category *	compound_treatment_design	text	Not NULL		x	x
Person Last Name	COMPOUND	text	NOT NULL		x	x
Person First Name	Marzolf	text	Not NULL			
Person Mid Initial						
Person Affiliation	Institute for Systems Biology					
Roles *	submitter					
Quality Control Description Type *		text	NOT NULL			
Replicate Description Type *		text	NOT NULL		x	x
Normalization Description type *					x	x
Experiment Description	de-stimulated wild-type and ATF3 -/- mouse bone marrow macrophages	text		x		x
Protocol Name						
Protocol Type *		text				
Protocol Description	selection L1	text				
Protocol Parameters		integer	NOT NULL			x
SDRF File		text	NOT NULL	x		
data file		text	NOT NULL		x	
ADF file				x		
Term Source Name		integer	NOT NULL		x	x
Term Source File		date	NOT NULL			
Term Source Version		integer	NOT NULL		x	x
		integer			x	x
		integer			x	x
term source version	1.0.0	integer	NOT NULL		x	x
term source file	1.0.0	date	NOT NULL			
term source name	MGEC	integer	NOT NULL		x	x

# Populous workflow

The image shows a screenshot of the RightField software interface. The main window is titled "RightField" and contains a spreadsheet with columns labeled A through S and rows numbered 1 through 47. The spreadsheet is currently empty. To the right of the spreadsheet is a configuration panel with the following sections:

- HIERARCHY**: A search bar with a magnifying glass icon and a search button. Below it is a large empty box containing the text "No ontologies loaded".
- TYPE OF ALLOWED VALUES**: A list of radio buttons with the following options:
  - "Free text"
  - "Direct subclasses"
  - "Subclasses"
  - "Instances"
  - "Direct instances"
  - "All classes"Below the radio buttons is a dropdown menu labeled "Select ontology..." and an "Apply" button.
- ALLOWED VALUES**: A large empty box containing the text "Any".

At the bottom of the spreadsheet, there is a tab labeled "Sheet0".

The image shows a screenshot of the RightField application interface. The main window is a spreadsheet with columns labeled A through S and rows numbered 1 through 47. A modal dialog box is open in the center, titled "Load from file or directly from BioPortal". The dialog lists several ontologies, with "Cell type (OBO)" selected. Below the list is a "Filter:" input field and "Cancel" and "OK" buttons.

**RightField**

File Edit Tools Sheet

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47

A B C D E F G H I J K L M N O P Q R S

*Load from file or directly from BioPortal*

- C. elegans gross anatomy (OBO)
- C. elegans phenotype (OWL)
- Cancer Research and Management ACGT Master Ontology (OBO)
- Cardiac Electrophysiology Ontology (OWL-DL)
- Cell Behavior Ontology (null)
- Cell Cycle Ontology (OBO)
- Cell line ontology (OWL-DL)
- Cell Line Ontology (LEXGRID-XML)
- Cell line ontology (OWL)
- Cell type (OBO)**
- Cereal plant development (OBO)
- Cereal plant gross anatomy (OBO)
- Cereal plant trait (OBO)
- Chemical entities of biological interest (OWL)

Filter:

Cancel OK

**HIERARCHY**

Q

No ontologies loaded

**TYPE OF ALLOWED VALUES**

- "Free text"
- "Direct subclasses"
- "Subclasses"
- "Instances"
- "Direct instances"
- "All classes"

Select ontology... ▾

Apply

**ALLOWED VALUES**

Any

Sheet0

# Populous workflow

The image shows a screenshot of the RightField application. The main window is a spreadsheet with columns labeled A through S and rows numbered 1 through 47. A blue box with the text "Ontology browser" is overlaid on the spreadsheet. To the right of the spreadsheet is a panel titled "HIERARCHY" which contains a list of ontology terms with radio buttons next to them. A red circle highlights this list. Below the hierarchy list is a section titled "TYPE OF ALLOWED VALUES" with several radio button options: "Free text" (selected), "Direct subclasses", "Subclasses", "Instances", "Direct instances", and "All classes". There is also a "Select ontology..." dropdown menu and an "Apply" button. At the bottom of the panel is a section titled "ALLOWED VALUES" with a large white box containing the text "Any". The application title bar shows "RightField" and the menu bar includes "File", "Edit", "Tools", and "Sheet".

RightField

File Edit Tools Sheet

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47

A B C D E F G H I J K L M N O P Q R S

Ontology browser

HIERARCHY

- Thing
  - adrenal medulla cell
  - cell
  - DbXref
  - Definition
  - dental papilla cell
  - glial cell
  - hair papilla cell
  - mesangial cell
  - mesenchyme condensation cell
  - muscle precursor cell
  - neuroepithelial cell
  - neuron associated cell
  - neuron associated cell (sensory/vertebrate)

TYPE OF ALLOWED VALUES

- "Free text"
- "Direct subclasses"
- "Subclasses"
- "Instances"
- "Direct instances"
- "All classes"

Select ontology... ▾

Apply

ALLOWED VALUES

Any

Sheet0

The image shows a screenshot of the RightField application interface. The main window is a spreadsheet with columns A through R and rows 1 through 47. Column A is highlighted in light blue. A dialog box titled "TYPE OF ALLOWED VALUES" is open in the foreground, showing several radio button options: "Free text" (selected), "Direct subclasses", "Subclasses", "Instances", "Direct instances", and "All classes". Below these options is a "Select ontology..." dropdown menu and an "Apply" button. The background spreadsheet has three blue callout boxes with the following text:

- 1. *Select column*
- 2. *Select Class in Ontology*
- 3. *Select allowed values*

The application window title is "RightField" and the menu bar includes "File", "Edit", "Tools", and "Sheet". The status bar at the bottom shows "Sheet0".

# Populous workflow

The screenshot displays the Populous software interface. The main window is a spreadsheet with columns A, B, and C. The data in column A includes various cell types, and column B contains their corresponding morphological or functional terms. The right-hand panel features a search bar, a 'HIERARCHY' list with radio buttons, and two 'ALLOWED VALUES' lists. Overlaid on the spreadsheet are four callout boxes highlighting specific features: 'Label rendering', 'Tab completion', 'Syntax Highlighting', and 'Multi-value cells'.

	A	B
1	mononuclear phagocyte	mononucleate
2	anucleate erythrocyte	anucleate
3	flight muscle cell	multinucleate
4	garland cell	binucleate
5	Proximal tubule epithelial cell	mononucleate
6	receptor cell	
7	renal principal cell	
8	reticulocyte	
9	red sensitive photoreceptor cell	
10	regulatory B cell	
11	renal intercalated cell	
12	receptor cell (sensu Animalia)	
13	retinal rod cell	
14	retinal ganglion cell	
15	regulatory T cell	
16	renin secreting cell	
17	resident monocyte	
18	renal alpha-intercalated cell	
19	reticular cell	
20	retinal bipolar neuron	
21	retinal cone cell	
22	receptor cell	
23	renal principal cell	
24	reticulocyte	
25		
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		
36		
37		
38		
39		
40		
41		

**Label rendering**

**Tab completion**

**Syntax Highlighting**

**Multi-value cells**

CELLS: A6:A6

HIERARCHY

Thing

- abdominal tone
- abdominal tone value
- abiotic stress sensitivity
- abiotic stress sensitivity value
- absolute acceleration
- absolute acceleration value
- absolute activity
- absolute activity value
- absolute age
- absolute age value
- absolute alternation

ALLOWED VALUES

- "Free text"
- "Direct subclasses"
- "Subclasses"
- "Instances"
- "Direct instances"

ALLOWED VALUES

- B cell
- B lymphoblast
- B-1 B cell
- B-1a B cell
- B-1b B cell
- B-2 B cell
- Be cell
- Be1 Cell
- Be2 cell

Demo of Populous in action.

## Manchester OWL syntax

Class: CL:0003523

Annotation:  
rdfs:label 'Kidney Cell'

EquivalentTo:  
CL:0000000 and OBO\_REL:part\_of some MAO\_000629

Populous - /Users/simon/Documents/e-lico/kupo\_cells\_sept10\_Populous.xls

File Edit Insert

	A	B	C	D
1	Cell Term	cell label	part_of	participates_in
2	kidney cell	renal cell	kidney	
3	kidney epithelial cell		kidney epithelium	
4	renal tubule epithelial cell		renal tubule epithelium	
5	kidney glomerular epithelial cell		kidney glomerular epithelium	
6	renal tubule cell		renal tubule	
7	kidney cortex cell	renal cortex cell	kidney cortex	
8	renal cortex tubule cell		renal cortex tubule	
9	kidney medulla cell	renal medullary cell	kidney medulla	
10	kidney outer medulla cell	renal outer	outer renal medulla	
11	kidney inner medulla cell	renal inner medullary	inner renal medulla	
12	inner renal medulla loop of henle cell		inner renal medulla loop of henle	
13	juxtaglomerular complex cell	juxtaglomerular	juxtaglomerular complex	regulation of glomerular filtration, regulation of blood circulation
14	kidney blood vessel cell	renal blood vessel	kidney blood vessel	
15	kidney arterial blood vessel cell	renal arterial blood	kidney arterial blood vessel	
16	kidney capillary endothelial cell	renal capillary cell	kidney capillary, capillary endothelium	
17	kidney venous blood vessel cell	renal venous blood vessel cell	kidney venous blood vessel	
18	renal corpuscle cell	cell of the renal	renal corpuscle	
19	mesangial cell		mesangium	phagocytosis, extracellular matrix constituent
20	glomerular mesangial cell		glomerular mesangium	
21	extraglomerular mesangial cell		extraglomerular mesangium	
22	podocyte	visceral epithelial	glomerular visceral epithelium	glomerular filtration, regulation of glomerular anatomical structure arrangement
23	bowmans capsule epithelial cell	epithelial cell of the glomerular parietal	bowmans capsule	
24	parietal epithelial cell		glomerular parietal epithelium	
25	glomerular cell		glomerulus	
26	glomerular capillary endothelial cell	glomerular capillary	glomerular capillary endothelium	glomerular filtration, regulation of glomerular anatomical structure arrangement
27	renal afferent arteriole cell	afferent arteriole cell	afferent arteriole	regulation of glomerular filtration
28	renal afferent arteriole endothelial cell	afferent arteriole	afferent arteriole, arteriole endothelium	
29	juxtaglomerular cell	afferent arteriole	part of afferent arteriole forming	renin secretion into blood stream, detection
30	renal afferent arteriole smooth muscle	afferent arteriole	afferent arteriole, arteriole smooth	
31	renal efferent arteriole cell	efferent arteriole cell	afferent arteriole	regulation of glomerular filtration
32	renal efferent arteriole endothelial cell	efferent arteriole	afferent arteriole, arteriole endothelium	
33	renal efferent arteriole smooth muscle	efferent arteriole	afferent arteriole, arteriole smooth	
34	proximal tubule epithelial cell	PTEC, proximal	renal proximal tubule	renal sodium ion absorption, potassium ion
35	proximal convoluted tubule epithelial		proximal convoluted tubule	
36	proximal straight tubule cell		proximal straight tubule	
37	loop of henle epithelial cell	henle's loop	loop of henle	extracellular matrix constituent secretion
38	loop of henle ascending limb	henle's loop	loop of henle ascending limb	renal sodium ion absorption, potassium ion
39	loop of henle thick ascending limb	henle's loop thick	loop of henle ascending limb thick	
40	loop of henle thin ascending limb	henle's loop thin	loop of henle ascending limb thin	
41	loop of henle medullary thick	henle's loop	distal straight tubule premacula	

CELLS: B17:B17

HIERARCHY

Thing

- abdominal tone
- abdominal tone value
- abiotic stress sensitivity
- abiotic stress sensitivity value
- absolute acceleration
- absolute acceleration value
- absolute activity
- absolute activity value
- absolute age
- absolute age value
- absolute alternation

TYPE OF ALLOWED VALUES

- "Free text"
- "Direct subclasses"
- "Subclasses"
- "Instances"
- "Direct instances"

ALLOWED VALUES

Any

Sheet1



Ontology Pre Processor Language ([oppl.sf.net](http://oppl.sf.net))

Scripting language to automate the manipulation of OWL ontologies

Apply pre-defined very complex OWL modelling automatically

Based in Manchester OWL Syntax

## OPPL script

```
Variable declaration,  
Variable declaration,  
...  
SELECT  
Query,  
Query,  
...  
WHERE  
Constraint,  
Constraint,  
...  
BEGIN  
ADD/REMOVE Axiom,  
ADD/REMOVE Axiom,  
...  
END;
```

## OPPL script

```
Variable declaration,  
Variable declaration,  
...  
SELECT  
Query,  
Query,  
...  
WHERE  
Constraint,  
Constraint,  
...  
BEGIN  
ADD/REMOVE Axiom,  
ADD/REMOVE Axiom,  
...  
END;
```

## Populous OPPL script

```
Variable declaration,  
Variable declaration,  
...  
SELECT  
Query,  
Query,  
...  
WHERE  
Constraint,  
Constraint,  
...  
BEGIN  
ADD/REMOVE Axiom,  
ADD/REMOVE Axiom,  
...  
END;
```

## OPPL script

```

Variable declaration,
Variable declaration,
...
SELECT
Query,
Query,
...
WHERE
Constraint,
Constraint,
...
BEGIN
ADD/REMOVE Axiom,
ADD/REMOVE Axiom,
...
END;

```

## Populous OPPL script

```

Variable declaration,
Variable declaration,
...
SELECT
Query,
Query,
...
WHERE
Constraint,
Constraint,
...
BEGIN
ADD/REMOVE Axiom,
ADD/REMOVE Axiom,
...
END;

```

```

?cell:CLASS,
?parent:CLASS

BEGIN
ADD ?cell subClassOf ?parent

END;
```

Variable

Variable type

CLASS  
CONSTANT  
OBJECTPROPERTY  
DATAPROPERTY  
ANNOTATIONPROPERTY  
INDIVIDUAL

```
?cell:CLASS,  
?parent:CLASS  
BEGIN  
ADD ?cell subClassOf ?parent  
END;
```

OWL expression: Manchester OWL syntax + variables

OWL expression  
=  
createIntersection  
createUnion  
(?var.VALUES)

```
?cell:CLASS,  
?anatomyPart:CLASS,  
?partOfRestriction:CLASS = part_of some ?anatomyPart,  
?anatomyIntersection:CLASS = createIntersection(?  
partOfRestriction.VALUES)  
BEGIN  
ADD ?cell equivalentTo CL_0000000 and  
?anatomyIntersection  
END;
```

## OPPL builder

The screenshot displays the OPPL Builder interface, which is used for constructing and executing OPPL scripts. The interface is divided into several sections:

- DL Query**: Contains tabs for Rules, OPPL, OPPL Macros, and Patterns.
- OPPL Builder**: The main workspace for building the query.
  - Input Variables**: Lists variables like `?target:CLASS` and `?origin:CLASS`.
  - Generated Variables**: A section for defining variables generated by the query.
  - SELECT**: A query editor showing the query: `(?target) SubClassOf ((develops_from some ((?origin))))`.
  - WHERE**: A section for defining conditions.
  - Actions**: A list of actions to be performed, such as `Add (?target) SubClassOf ((develops_from only ((?origin))))`.
- Evaluate**: A button to execute the query.
- When removing consider Active Ontology Imported Closure**: A checkbox option.
- Affected axioms: 15891**: A list of axioms affected by the query, including `Add (CL_0002111) SubClassOf ((develops_from only ...))`.
- Bindings: 15891**: A table showing the results of the query, with columns for `?origin` and `?target`.
 

?origin	?target
CL 0000000	CL 0000005
CL 0000000	CL 0000006
CL 0000000	CL 0000008
CL 0000000	CL 0000011
CL 0000000	CL 0000017
CL 0000000	CL 0000018
CL 0000000	CL 0000019
CL 0000000	CL 0000024
CL 0000000	CL 0000025
CL 0000000	CL 0000026
CL 0000000	CL 0000027
CL 0000000	CL 0000028
CL 0000000	CL 0000029
- Instantiated axioms**: A section for displaying instantiated axioms.
- Execute**: A button to execute the query.

## OPPL text editor

The screenshot shows the OPPL text editor interface. At the top, there are tabs for "DL Query", "Rules", "OPPL", "OPPL Macros", and "Patterns". The "OPPL" tab is active. Below the tabs, there are two sub-tabs: "OPPL Builder" and "OPPL Text Editor". The "OPPL Text Editor" is selected and contains the following script:

```
?target:CLASS,  
?origin:CLASS  
SELECT  
?target SubClassOf develops_from some ?origin  
BEGIN  
ADD ?target SubClassOf develops_from only ?origin  
END;
```

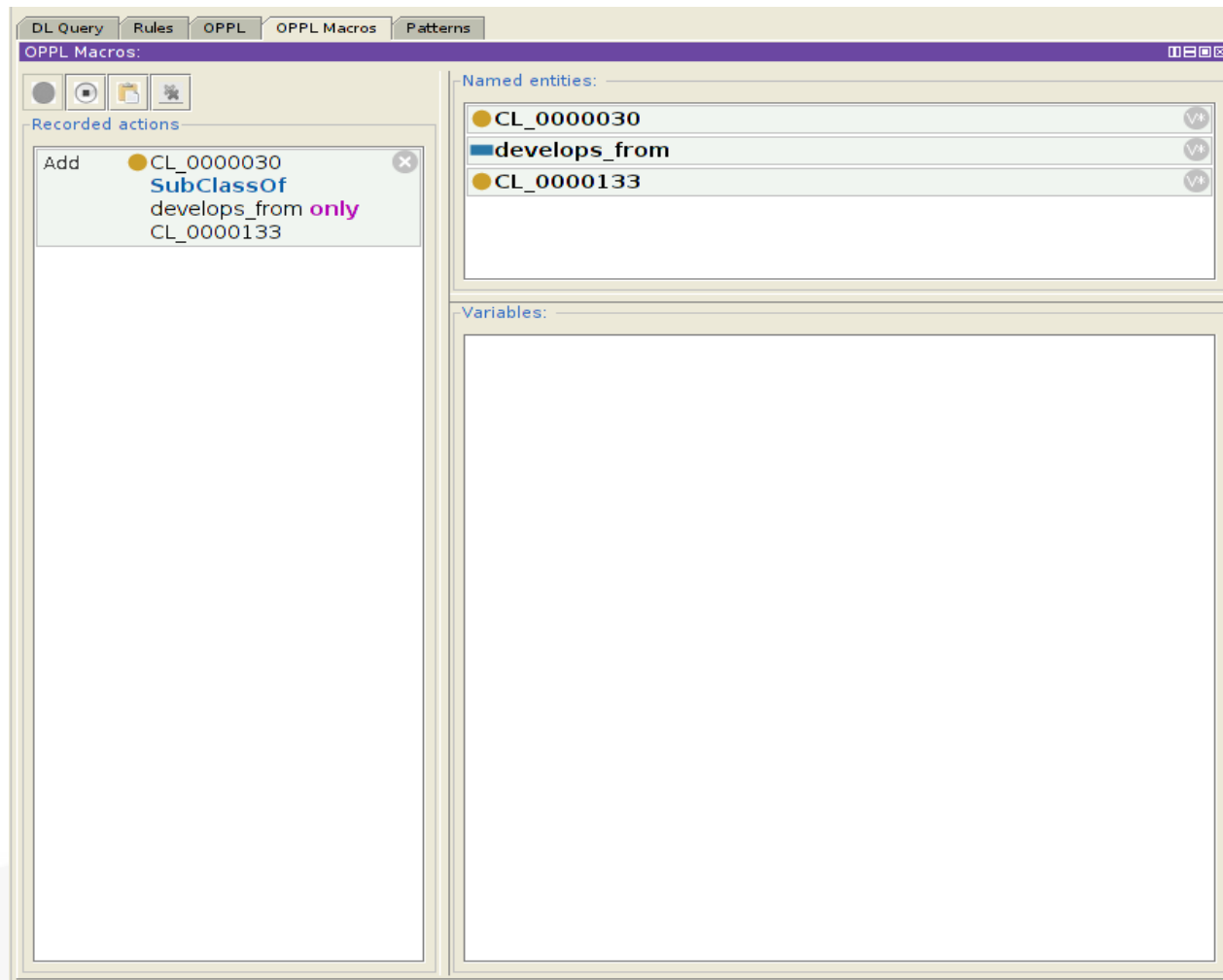
Below the editor is an "Evaluate" button. Underneath, there is a checkbox labeled "When removing consider Active Ontology Imported Closure". Below this, there are three panels: "Affected axioms:", "Bindings:", and "Instantiated axioms:". The "Bindings:" panel contains a table with two columns: "?origin" and "?target".

?origin	?target
---------	---------

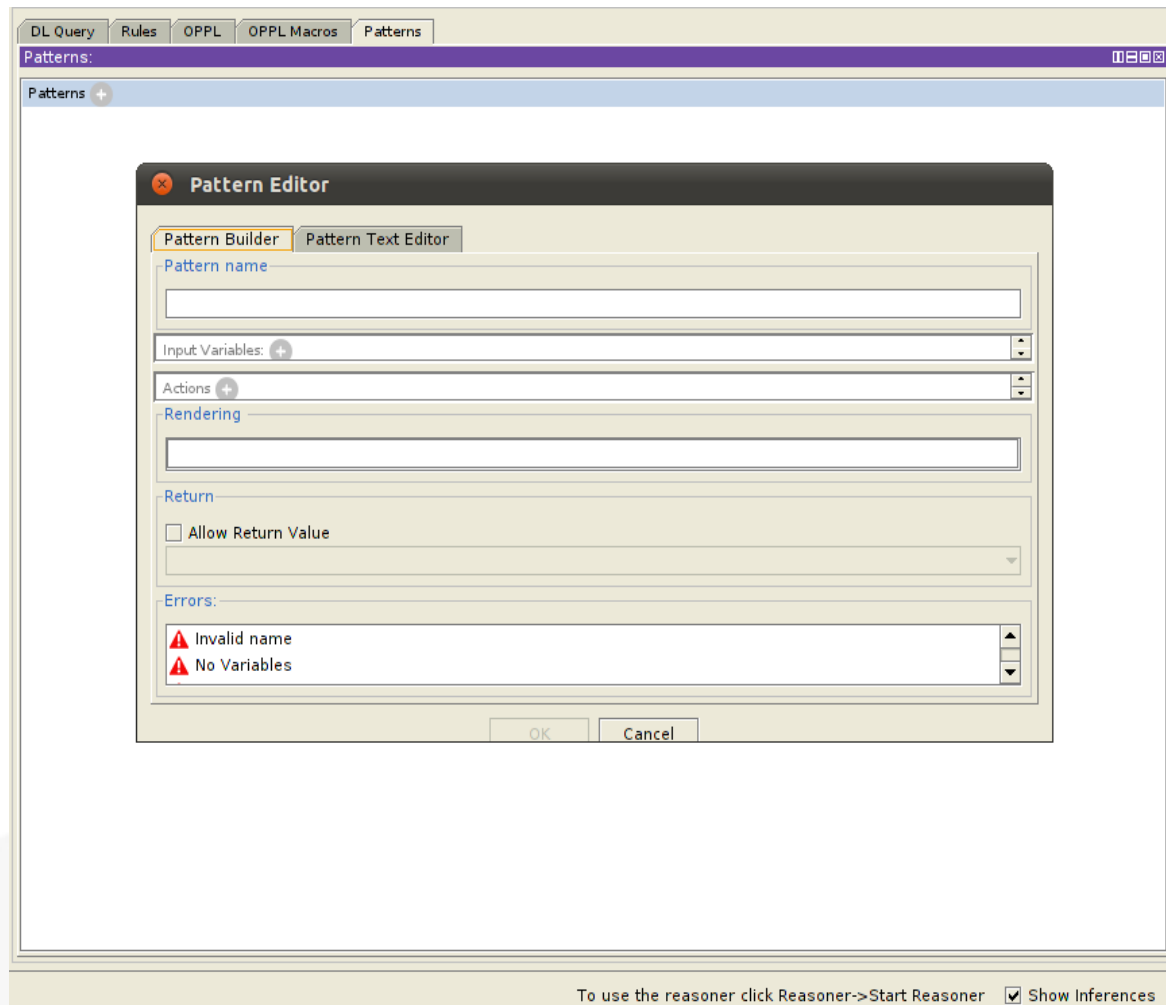
At the bottom of the interface is an "Execute" button.



## OPPL macros



## OPPL patterns



OPPL publications:

<http://oppl2.sourceforge.net/documentation.html>

OPPL documentation:

[http://oppl2.sourceforge.net/oppl\\_documentation.html](http://oppl2.sourceforge.net/oppl_documentation.html)

OPPL patterns:

[http://oppl2.sourceforge.net/patterns\\_documentation.html](http://oppl2.sourceforge.net/patterns_documentation.html)

OPPL Manual: <http://oppl2.sourceforge.net/manual.pdf>

OPPL sample scripts:

<http://oppl2.sourceforge.net/taggedexamples/>

## Demo 2 -converting spreadsheets to OWL using OPPL

Ontological annotation by stealth

Real biological data + high quality meta-data

Development of a Kidney and Urinary Pathway Knowledge Base

Element	Example	Format	Constraint	Typed in b	drop-down	automatic	Key Type	Explanations
Investigation Title	Transcription profiling of wild-type and ATF3 -	text	Not NULL			x	PK	
Experimental Design Type *	Transcription profiling of wild-type and ATF3 -	text	Not NULL				PK	Explanations
Experimental Factor Category *	compound_treatment_design	text	Not NULL		x		CV	PK - primary k
Person Last Name	COMPOUND	text	Not NULL		x		CV	FK - foreign ke
Person First Name	Marzolf	text	Not NULL			x		CV - controlled
Person Mid Initial	Bruz							
Person Affiliation	Institute for Systems Biology							
Roles *	curator							
Quality Control Description Type *	biological_replicate	text	NOT NULL					
Replicate Description Type *	Dye_swap_replicate	text	NOT NULL		X		CV	
Normalization Description type *	median_log_centering				X		CV	
Experiment Description	Time course data from lipopolysaccharide-stimulated wild-type and ATF3 -/- mouse bone marrow macrophages were collected to inv			X		X		
Protocol Name	P-TABM-glichrist1							
Protocol Type *	grow	text						
Protocol Description	Flush femurs from C57BL/6 mice (Jackson La	text						CV
Protocol Parameters	carbon source,temperature	integer	NOT NULL				X	
SDRF File		text	NOT NULL	X				
data file		text	NOT NULL		X			
ADF file	E-TABM-102_sdrf.txt							
Term Source Name	MGED Ontology	integer	NOT NULL	X		X		
Term Source File	http://mged.sourceforge.net/ontologies/MGED	date	NOT NULL					
Term Source Version	1.3.1.1	integer	NOT NULL		X		X	
		integer			X		X	
		text			X		X	
		integer			X			

## Demo 3 - Experiment template for data annotation

## RightField

Matthew Horridge, Katy Wolstencroft, Stuart Owen,  
Carole Goble

## Populous

Simon Jupp, Robert Stevens funded by e-LICO  
EU-FP7 Collaborative Project (2009-2012) Theme ICT-4.4:  
Intelligent Content and Semantics  
and NIH funded NCBO driving biological project program

Mikel Egaña Aranguren is funded by the Marie Curie Cofund  
programme (FP7)

OPPL 2 is maintained by Luigi Iannone