Using SPARQL queries to express integrity constraints in RDF graphs

Final internship report

Leandro Lovisolo

INRA SupAgro and INRIA GraphiK
Montpellier, France

February 25, 2016

Part I

Preliminaries

Problem statement

- We have an RDF graph where we store experimental data extracted from tables in scientific publications.
- The data extraction process is done semi-manually, thus it's very error-prone.
- Therefore, **we want to verify the integrity of the annotated data automatically.**

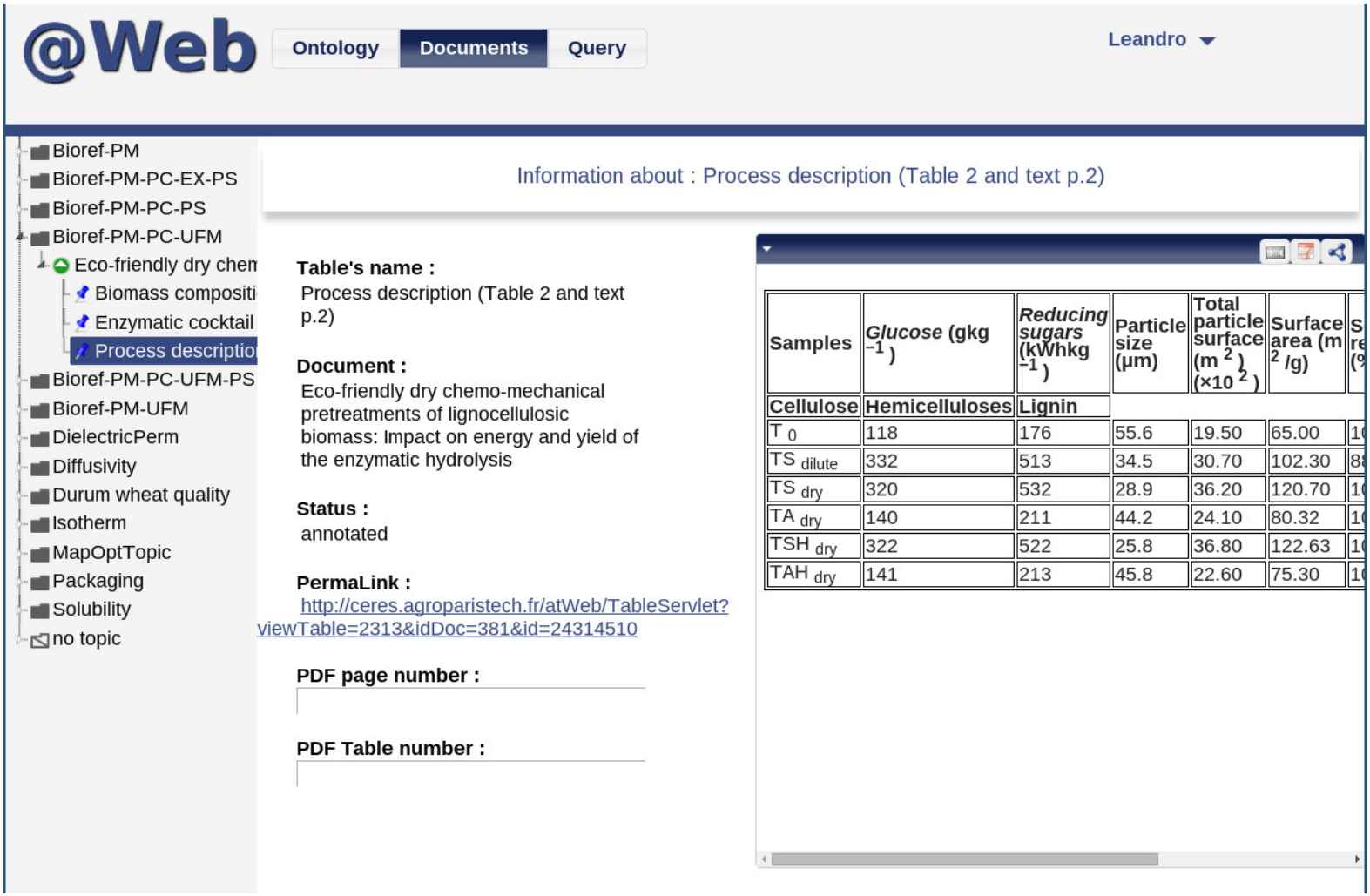
The **@Web** platform

Introduction

- A software platform used to annotate tables from scientific publications in heterogeneous formats (PDF files, Excel spreadsheets, etc.)
- Data is stored in an RDF graph following a predefined OWL ontology.
- Goal of my internship: add integrity constraint checking capabilities to the **@Web** platform.

The **@Web** platform

Screenshot



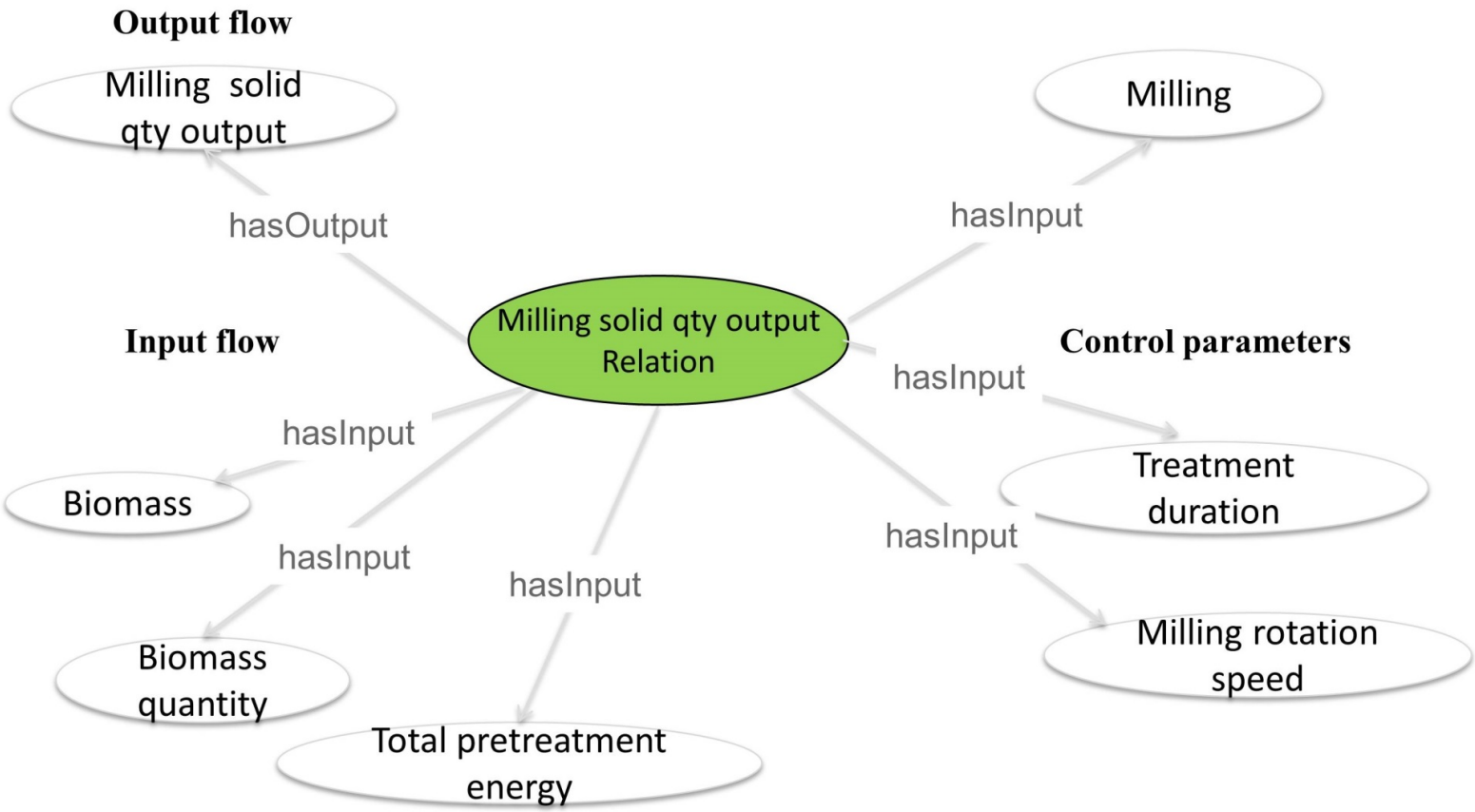
The **@Web** platform

n-ary relation pattern

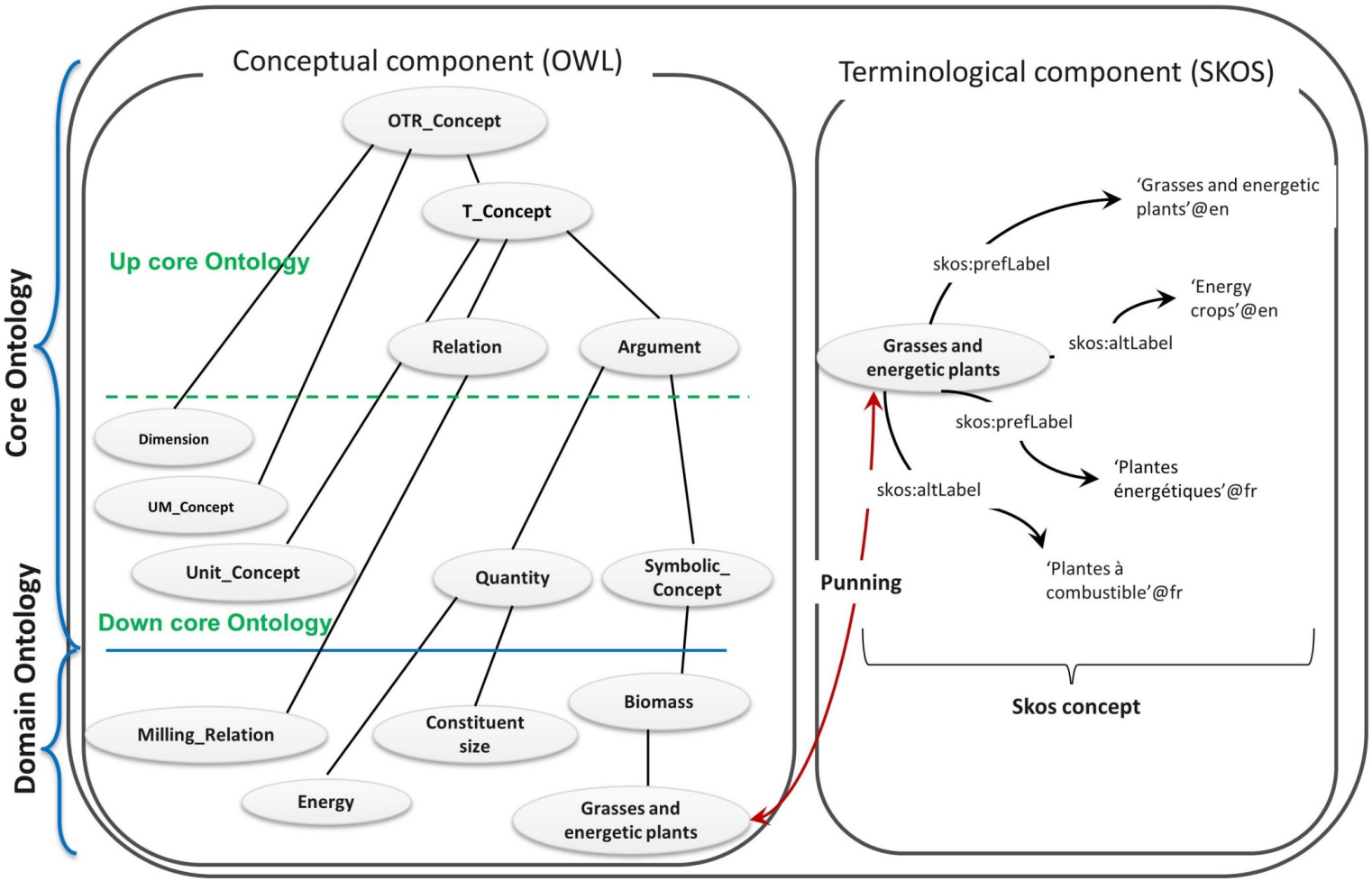
- We're trying to represent experiments composed of many inputs and a single output.
- An OWL ontology is created where OWL classes are defined for each kind of experiment we're interested in representing.
- Instances of each experiment class are connected to their respective input arguments and output argument via OWL object and data properties.
- We're thus defining a pattern for *n*-ary relations.

The **@Web** platform

Example *n*-ary relation



@Web ontology



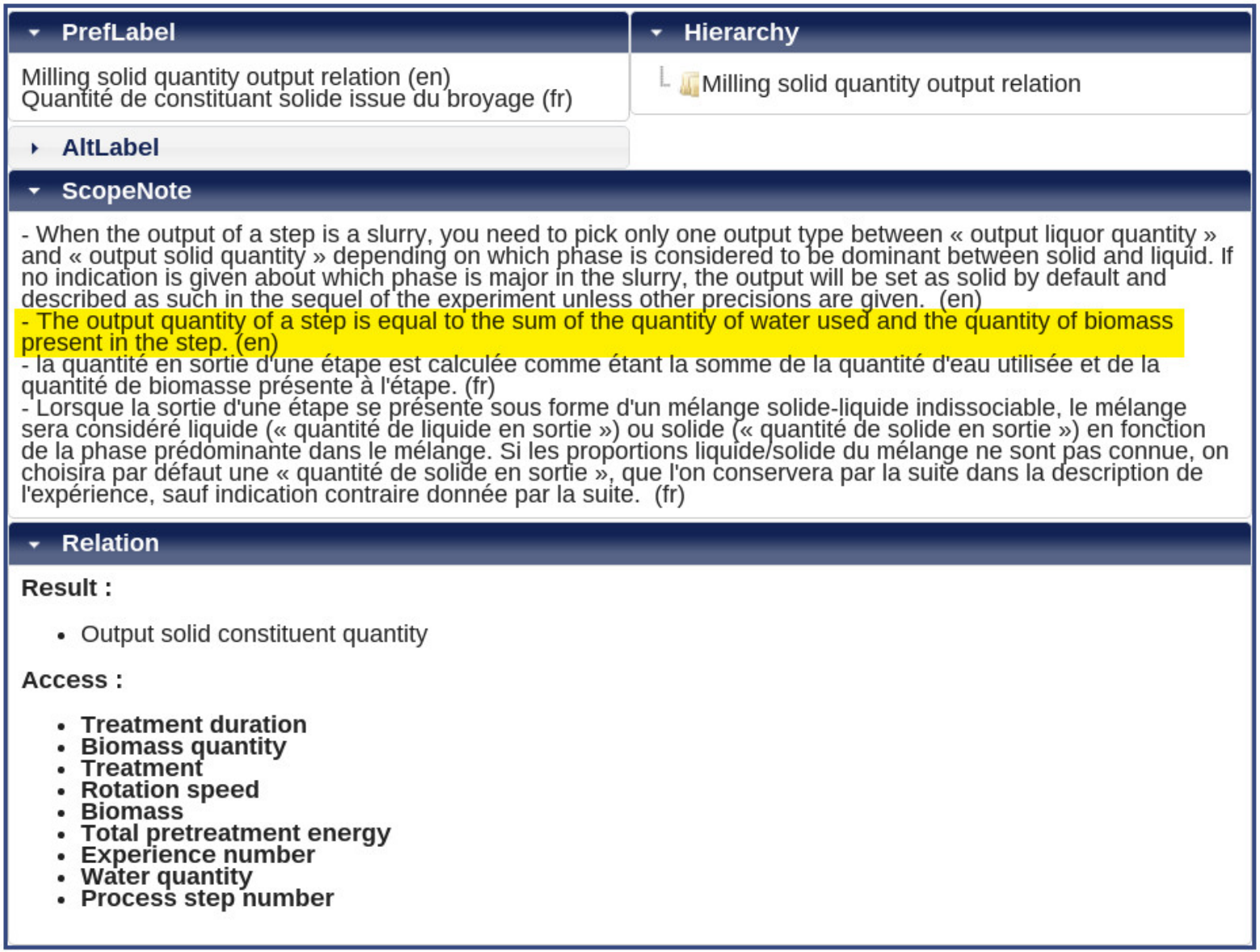
Annotated tables

Screenshot

n°	Output solid constituent size Unit : mm	Treatment	Experience number Unit : 1	Process step number Unit : 1	Biomass	Biomass quantity Unit : g	Total pretreatment energy Unit : kW.h.kg-1	Water quantity Unit : l	Rotation speed Unit : min-1	Treatment duration Unit : min	Output solid constituent quantity Unit : g	Temperature Unit : oC	Output liquor quantity Unit : l	Salt	Salt quantity Unit : g
1	3.000e+0	Cutting milling	0.000e+0	1.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]	0.000e+0	[-inf ; inf]	[-inf ; inf]					
2		Drying	0.000e+0	2.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]			[-inf ; inf]	[-inf ; inf]	6.000e+1			
3		Wet disk milling	0.000e+0	3.000e+0	Rice straw	1.000e+3	[-inf ; inf]	2.000e+1	[-inf ; inf]	[-inf ; inf]	1.000e+3	[1.800e+1 ; 2.400e+1]	0.000e+0	Salt	0.000e+0
4		Washing and centrifugation	0.000e+0	4.000e+0	Rice straw	1.000e+3	[-inf ; inf]	0.000e+0	9.000e+3	1.000e+1	1.000e+3	[1.800e+1 ; 2.400e+1]	2.000e+1	Salt	0.000e+0
5		Enzymatic hydrolysis treatment	0.000e+0	5.000e+0	Rice straw	[4.000e-2 ; 6.000e-2]			[-inf ; inf]	4.320e+3	[3.400e-2 ; 5.000e-2]	4.500e+1			
6	3.000e+0	Cutting milling	1.000e+0	1.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]	0.000e+0	[-inf ; inf]	[-inf ; inf]					
7		Drying	1.000e+0	2.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]			[-inf ; inf]	[-inf ; inf]	6.000e+1			
8		Hot water treatment	1.000e+0	3.000e+0	Rice straw	1.000e+3	[-inf ; inf]	1.000e+1	0.000e+0	6.000e+1	1.000e+3	1.210e+2	0.000e+0	Salt	0.000e+0
9		Wet disk milling	1.000e+0	4.000e+0	Rice straw	1.000e+3	[-inf ; inf]	1.000e+1	[-inf ; inf]	[-inf ; inf]	1.000e+3	[1.800e+1 ; 2.400e+1]	0.000e+0	Salt	0.000e+0
10		Washing and centrifugation	1.000e+0	5.000e+0	Rice straw	1.000e+3	[-inf ; inf]	0.000e+0	9.000e+3	1.000e+1	1.000e+3	[1.800e+1 ; 2.400e+1]	2.000e+1	Salt	0.000e+0
11		Enzymatic hydrolysis treatment	1.000e+0	6.000e+0	Rice straw	[4.000e-2 ; 6.000e-2]			[-inf ; inf]	4.320e+3	[3.000e-2 ; 4.500e-2]	4.500e+1			
12	3.000e+0	Cutting milling	2.000e+0	1.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]	0.000e+0	[-inf ; inf]	[-inf ; inf]					
13		Drying	2.000e+0	2.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]			[-inf ; inf]	[-inf ; inf]	6.000e+1			
14		Hot water treatment	2.000e+0	3.000e+0	Rice straw	1.000e+3	[-inf ; inf]	1.000e+1	0.000e+0	6.000e+1	1.000e+3	1.350e+2	0.000e+0	Salt	0.000e+0
15		Wet disk milling	2.000e+0	4.000e+0	Rice straw	1.000e+3	[-inf ; inf]	1.000e+1	[-inf ; inf]	[-inf ; inf]	1.000e+3	[1.800e+1 ; 2.400e+1]	0.000e+0	Salt	0.000e+0
16		Washing and centrifugation	2.000e+0	5.000e+0	Rice straw	1.000e+3	[-inf ; inf]	0.000e+0	9.000e+3	1.000e+1	1.000e+3	[1.800e+1 ; 2.400e+1]	2.000e+1	Salt	0.000e+0
17		Enzymatic hydrolysis treatment	2.000e+0	6.000e+0	Rice straw	[4.000e-2 ; 6.000e-2]			[-inf ; inf]	4.320e+3	[2.800e-2 ; 4.200e-2]	4.500e+1			
18	3.000e+0	Cutting milling	3.000e+0	1.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]	0.000e+0	[-inf ; inf]	[-inf ; inf]					
19		Drying	3.000e+0	2.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]			[-inf ; inf]	[-inf ; inf]	6.000e+1			
20		Hot water treatment	3.000e+0	3.000e+0	Rice straw	1.000e+3	[-inf ; inf]	1.000e+1	0.000e+0	6.000e+1	1.000e+3	1.500e+2	0.000e+0	Salt	0.000e+0
21		Wet disk milling	3.000e+0	4.000e+0	Rice straw	1.000e+3	[-inf ; inf]	1.000e+1	[-inf ; inf]	[-inf ; inf]	1.000e+3	[1.800e+1 ; 2.400e+1]	0.000e+0	Salt	0.000e+0
22		Washing and centrifugation	3.000e+0	5.000e+0	Rice straw	1.000e+3	[-inf ; inf]	0.000e+0	9.000e+3	1.000e+1	1.000e+3	[1.800e+1 ; 2.400e+1]	2.000e+1	Salt	0.000e+0

Guidelines

Screenshot



PrefLabel

Milling solid quantity output relation (en)
Quantité de constituant solide issue du broyage (fr)

Hierarchy

Milling solid quantity output relation

AltLabel

ScopeNote

- When the output of a step is a slurry, you need to pick only one output type between « output liquor quantity » and « output solid quantity » depending on which phase is considered to be dominant between solid and liquid. If no indication is given about which phase is major in the slurry, the output will be set as solid by default and described as such in the sequel of the experiment unless other precisions are given. (en)
- The output quantity of a step is equal to the sum of the quantity of water used and the quantity of biomass present in the step. (en)
- la quantité en sortie d'une étape est calculée comme étant la somme de la quantité d'eau utilisée et de la quantité de biomasse présente à l'étape. (fr)
- Lorsque la sortie d'une étape se présente sous forme d'un mélange solide-liquide indissociable, le mélange sera considéré liquide (« quantité de liquide en sortie ») ou solide (« quantité de solide en sortie ») en fonction de la phase prédominante dans le mélange. Si les proportions liquide/solide du mélange ne sont pas connue, on choisira par défaut une « quantité de solide en sortie », que l'on conservera par la suite dans la description de l'expérience, sauf indication contraire donnée par la suite. (fr)

Relation

Result :

- Output solid constituent quantity

Access :

- **Treatment duration**
- **Biomass quantity**
- **Treatment**
- **Rotation speed**
- **Biomass**
- **Total pretreatment energy**
- **Experience number**
- **Water quantity**
- **Process step number**

Example guideline

Guideline

"The output quantity of a step is equal to the sum of the quantity of water used and the quantity of biomass present in the step."

Example guideline

Guideline

"The output quantity of a step is equal to the sum of the quantity of water used and the quantity of biomass present in the step."

$$output = waterInput + biomassInput$$

Example guideline

An annotated row that doesn't fulfill the guideline

n°	Output solid constituent size Unit : mm	Treatment	Experience number Unit : 1	Process step number Unit : 1	Biomass	Biomass quantity Unit : g	Total pretreatment energy Unit : kW.h.kg-1	Water quantity Unit : l	Rotation speed Unit : min-1	Treatment duration Unit : min	Output solid constituent quantity Unit : g
1	3.000e+0	Cutting milling	0.000e+0	1.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]	0.000e+0	[-inf ; inf]	[-inf ; inf]	
2		Drying	0.000e+0	2.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]			[-inf ; inf]	[-inf ; inf]
3		Wet disk milling	0.000e+0	3.000e+0	Rice straw	1.000e+3	[-inf ; inf]	2.000e+1	[-inf ; inf]	[-inf ; inf]	1.000e+3

Example guideline

An annotated row that doesn't fulfill the guideline

n°	Output solid constituent size Unit : mm	Treatment	Experience number Unit : 1	Process step number Unit : 1	Biomass	Biomass quantity Unit : g	Total pretreatment energy Unit : kW.h.kg-1	Water quantity Unit : l	Rotation speed Unit : min-1	Treatment duration Unit : min	Output solid constituent quantity Unit : g
1	3.000e+0	Cutting milling	0.000e+0	1.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]	0.000e+0	[-inf ; inf]	[-inf ; inf]	
2		Drying	0.000e+0	2.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]			[-inf ; inf]	[-inf ; inf]
3		Wet disk milling	0.000e+0	3.000e+0	Rice straw	1.000e+3	[-inf ; inf]	2.000e+1	[-inf ; inf]	[-inf ; inf]	1.000e+3

$$output = waterInput + biomassInput$$

Example guideline

An annotated row that doesn't fulfill the guideline

n°	Output solid constituent size Unit : mm	Treatment	Experience number Unit : 1	Process step number Unit : 1	Biomass	Biomass quantity Unit : g	Total pretreatment energy Unit : kW.h.kg-1	Water quantity Unit : l	Rotation speed Unit : min-1	Treatment duration Unit : min	Output solid constituent quantity Unit : g
1	3.000e+0	Cutting milling	0.000e+0	1.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]	0.000e+0	[-inf ; inf]	[-inf ; inf]	
2		Drying	0.000e+0	2.000e+0	Rice straw	[-inf ; inf]	[-inf ; inf]			[-inf ; inf]	[-inf ; inf]
3		Wet disk milling	0.000e+0	3.000e+0	Rice straw	1.000e+3	[-inf ; inf]	2.000e+1	[-inf ; inf]	[-inf ; inf]	1.000e+3

$$output = waterInput + biomassInput$$

$$1000 = 20 + 1000$$

Part II

RDF data validation: survey of the state of the art

Shape Expressions

Pending.

SHACL

Pending.

Plain SPARQL

Pending.

Part III

Implementation

Examples of real constraints

Pending.

Demo

Part IV

Conclusions

Conclusions

Pending.

Future work

Pending.

Thanks!