Integrity checking in RDF databases using SPARQL constraints

A brief introduction to the subject of my training period

Leandro Lovisolo

INRA SupAgro and INRIA GraphiK
Montpellier, France

September 11, 2015

Motivation

Problem statement

Motivation

Problem statement

- We're trying to answer questions that require consulting heterogeneous data sources.

Motivation

Problem statement

- We're trying to answer questions that require consulting heterogeneous data sources.
    - Literature with inconsistent, semi-structured data.

Motivation

Problem statement

- We're trying to answer questions that require consulting heterogeneous data sources.
  - Literature with inconsistent, semi-structured data.
  - No standard naming convention.

Motivation

Problem statement

- We're trying to answer questions that require consulting heterogeneous data sources.
  - Literature with inconsistent, semi-structured data.
  - No standard naming convention.
  - No information about the reliability of the data sources.

Motivation

Problem statement

- We're trying to answer questions that require consulting heterogeneous data sources.
  - Literature with inconsistent, semi-structured data.
  - No standard naming convention.
  - No information about the reliability of the data sources.
  - Each data source has its specific browsing/querying mechanism (no common interface.)

Motivation

Sample problem domain: **biorefinery**

- Ligno-cellulosic biomass pre-treatment before enzymatic hydrolysis is an essential step to obtain good yields.

Motivation

Sample problem domain: **biorefinery**

- Ligno-cellulosic biomass pre-treatment before enzymatic hydrolysis is an essential step to obtain good yields.
- Several pre-treatment principles available, but **no clear criteria on how to choose the best one** taking into account environmental sustainability for a given biomass and biorefinery product (e.g. glucose.)

Proposed solution

- Represent scientific knowledge with ontologies using recommended standardized tools and languages for such purposes (semantic web technologies, RDF(S), OWL, etc.)

Proposed solution

- Represent scientific knowledge with ontologies using recommended standardized tools and languages for such purposes (semantic web technologies, RDF(S), OWL, etc.)
- Develop an ontology and data management web application (e.g. the **@Web platform**) that makes it easy for scientists to introduce data from scientific publications into an ontology, execute queries against an ontology, etc.

Proposed solution

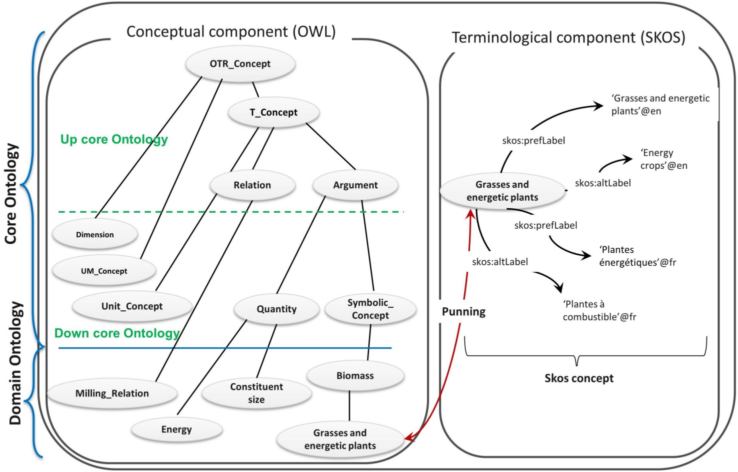
- Represent scientific knowledge with ontologies using recommended standardized tools and languages for such purposes (semantic web technologies, RDF(S), OWL, etc.)
- Develop an ontology and data management web application (e.g. the **@Web platform**) that makes it easy for scientists to introduce data from scientific publications into an ontology, execute queries against an ontology, etc.
- Create integrity constraints to automatically detect inconsistencies and errors in scientific publications and to automatically classify publications according to their topics.

Proposed solution

- Represent scientific knowledge with ontologies using recommended standardized tools and languages for such purposes (semantic web technologies, RDF(S), OWL, etc.)
- Develop an ontology and data management web application (e.g. the **@Web platform**) that makes it easy for scientists to introduce data from scientific publications into an ontology, execute queries against an ontology, etc.
- Create integrity constraints to automatically detect inconsistencies and errors in scientific publications and to automatically classify publications according to their topics.
    - *The focus of my internship!*

An example of a termino-ontological resource

Taken from the biorefinery application



Design goals for the core ontology

- **Simple** so as to make the annotator's task easier.

Design goals for the core ontology

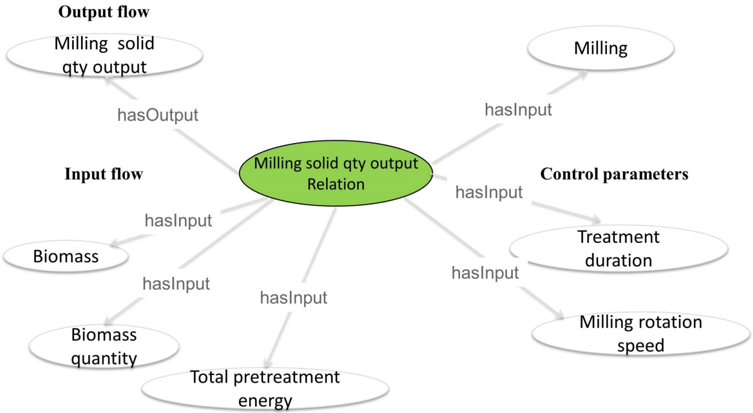
- **Simple** so as to make the annotator's task easier.
- **Generic** enough so that the approach can be applied to different, unrelated domains.

Design goals for the core ontology

- **Simple** so as to make the annotator's task easier.
- **Generic** enough so that the approach can be applied to different, unrelated domains.
    - Proven in the domains of biorefinery and packaging selection.

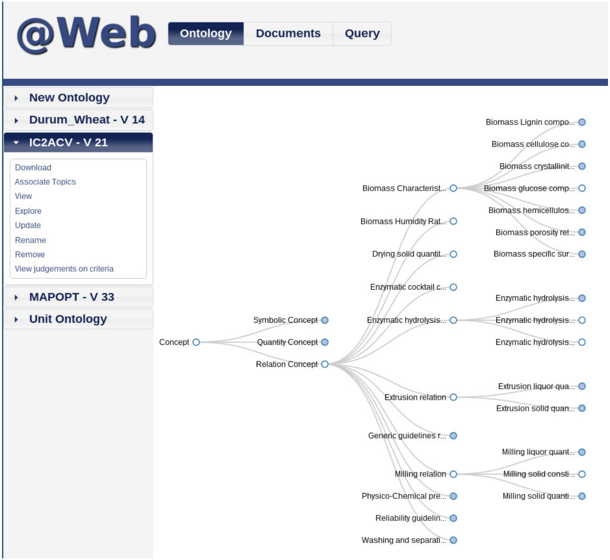
A sample relation

Also from the biorefinery domain



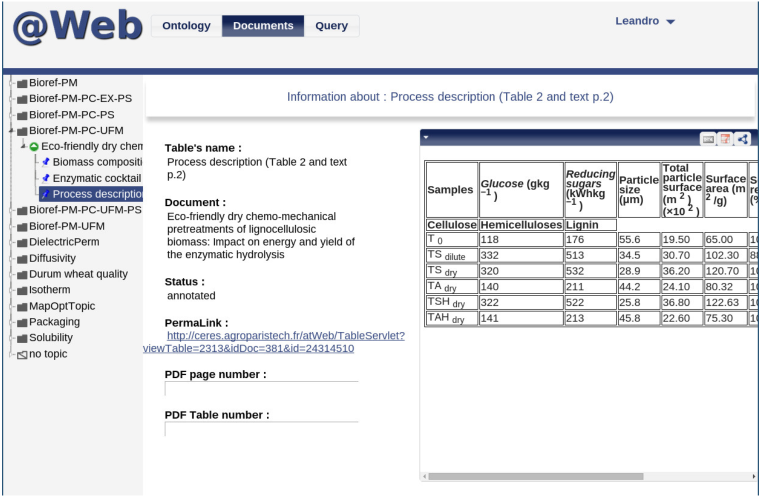
The **@Web** platform

Exploring an ontology



The **@Web** platform

Browsing documents



Information about : Process description (Table 2 and text p.2)

Table's name :
Process description (Table 2 and text p.2)

Document :
Eco-friendly dry chemo-mechanical pretreatments of lignocellulosic biomass: Impact on energy and yield of the enzymatic hydrolysis

Status :
annotated

PermaLink :
http://ceres.agroparistech.fr/atWeb/TableServlet?viewTable=2313&idDoc=381&id=24314510

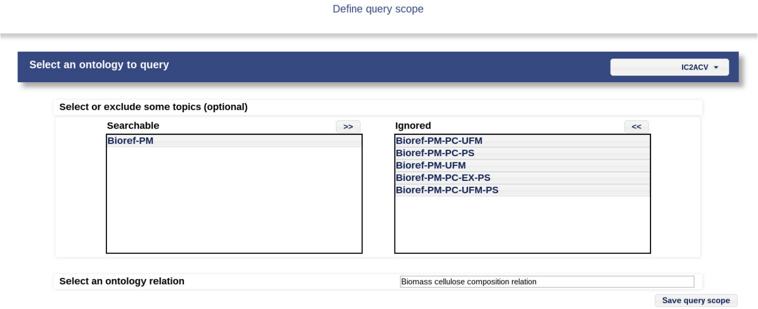
PDF page number :

PDF Table number :

Samples	*Glucose* (gkg $^{-1}$)	*Reducing sugars* (kWhkg $^{-1}$)	Particle size (µm)	Total particle surface (m ^{2}) (×10 ^{2})	Surface area (m ^{2} /g)	S re (%
Cellulose	Hemicelluloses	Lignin				
T $_{0}$	118	176	55.6	19.50	65.00	10
TS $_{dilute}$	332	513	34.5	30.70	102.30	88
TS $_{dry}$	320	532	28.9	36.20	120.70	10
TA $_{dry}$	140	211	44.2	24.10	80.32	10
TSH $_{dry}$	322	522	25.8	36.80	122.63	10
TAH $_{dry}$	141	213	45.8	22.60	75.30	10

The **@Web** platform

Querying an ontology: defining the search scope



Define query scope

Select an ontology to query

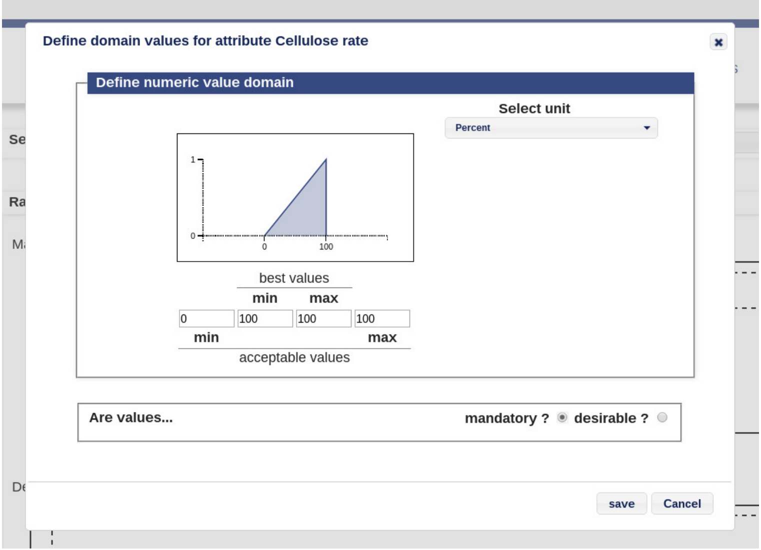
Select or exclude some topics (optional)

Searchable	Ignored
Bioref-PM	Bioref-PM-PC-UFM
	Bioref-PM-PC-PS
	Bioref-PM-UFM
	Bioref-PM-PC-EX-PS
	Bioref-PM-PC-UFM-PS

Select an ontology relation: Biomass cellulose composition relation

The **@Web** platform

Querying an ontology: search parameters



Define domain values for attribute Cellulose rate

Define numeric value domain

Select unit

Percent

best values

min **max**

0 | 100 | 100 | 100

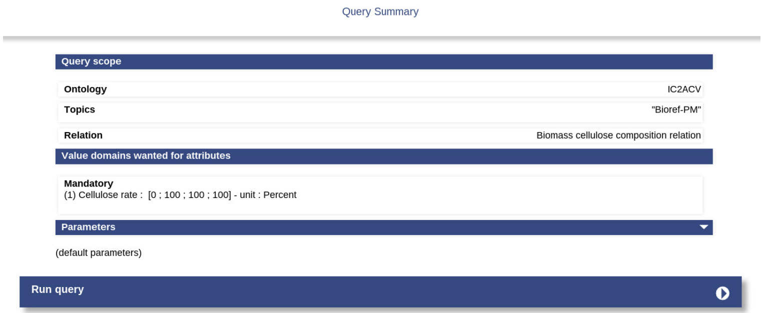
min **max**

acceptable values

Are values... **mandatory ?** ● **desirable ?** ○

The **@Web** platform

Querying an ontology: executing a query



Query Summary

Query scope

Ontology	IC2ACV
Topics	"Bioref-PM"
Relation	Biomass cellulose composition relation

Value domains wanted for attributes

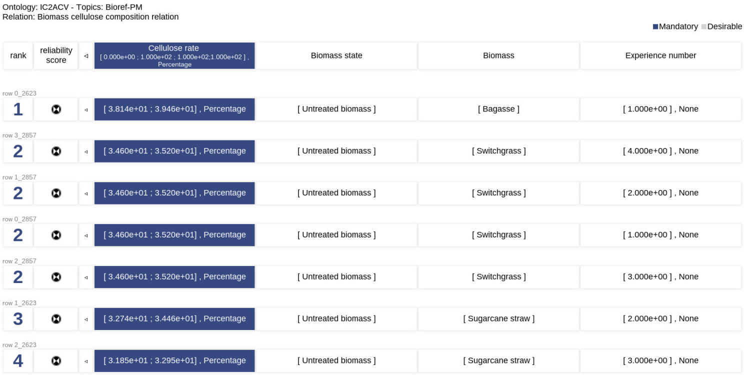
Mandatory
(1) Cellulose rate : [0 ; 100 ; 100 ; 100] - unit : Percent

Parameters

(default parameters)

The **@Web** platform

Querying an ontology: results



Ontology: IC2ACV - Topics: Bioref-PM
Relation: Biomass cellulose composition relation

■ Mandatory □ Desirable

	rank	reliability score	◃	Cellulose rate [0.000e+00 ; 1.000e+02 ; 1.000e+02;1.000e+02] , Percentage	Biomass state	Biomass	Experience number
row 0_2623	1	⊗	◃	[3.814e+01 ; 3.946e+01] , Percentage	[Untreated biomass]	[Bagasse]	[1.000e+00] , None
row 3_2857	2	⊗	◃	[3.460e+01 ; 3.520e+01] , Percentage	[Untreated biomass]	[Switchgrass]	[4.000e+00] , None
row 1_2857	2	⊗	◃	[3.460e+01 ; 3.520e+01] , Percentage	[Untreated biomass]	[Switchgrass]	[2.000e+00] , None
row 0_2857	2	⊗	◃	[3.460e+01 ; 3.520e+01] , Percentage	[Untreated biomass]	[Switchgrass]	[1.000e+00] , None
row 2_2857	2	⊗	◃	[3.460e+01 ; 3.520e+01] , Percentage	[Untreated biomass]	[Switchgrass]	[3.000e+00] , None
row 1_2623	3	⊗	◃	[3.274e+01 ; 3.446e+01] , Percentage	[Untreated biomass]	[Sugarcane straw]	[2.000e+00] , None
row 2_2623	4	⊗	◃	[3.185e+01 ; 3.295e+01] , Percentage	[Untreated biomass]	[Sugarcane straw]	[3.000e+00] , None

The annotator's task

- Given a scientific publication and a desired ontology, capture data from the publication using the appropriate concepts in the ontology.

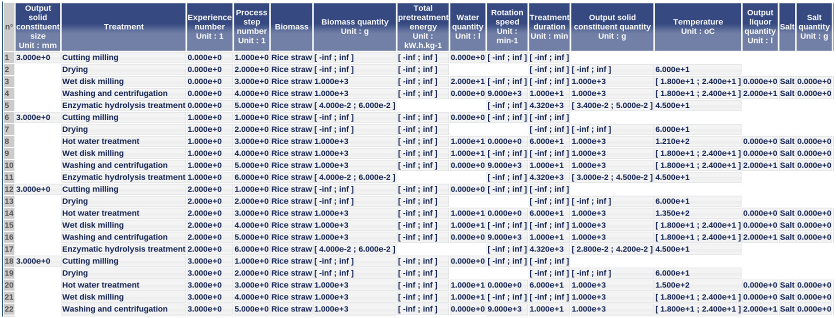
The annotator's task

- Given a scientific publication and a desired ontology, capture data from the publication using the appropriate concepts in the ontology.
- Create and update concepts in the ontology as they're discovered during the annotation process (i.e. in an iterative fashion.)

The annotator’s task

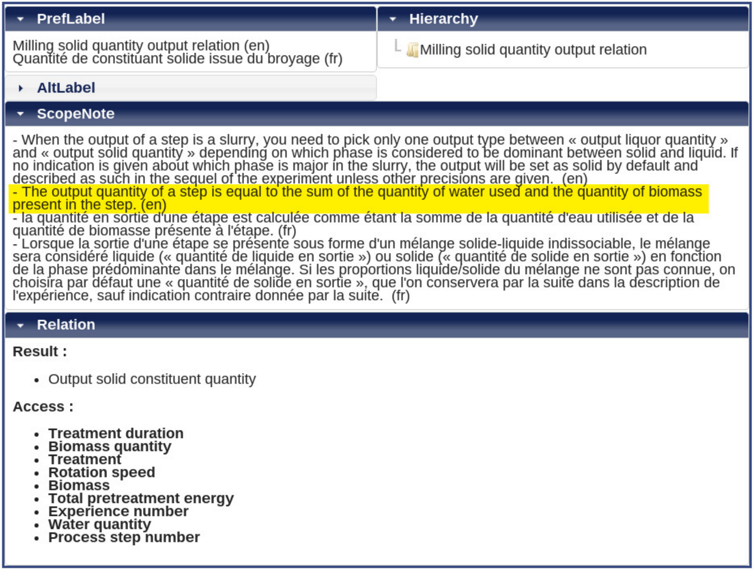
- Given a scientific publication and a desired ontology, capture data from the publication using the appropriate concepts in the ontology.
- Create and update concepts in the ontology as they’re discovered during the annotation process (i.e. in an iterative fashion.)
- Write and edit **guidelines** associated to each concept explaining when and how a concept should be used.

An example of data captured from a scientific publication



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

A sample guideline



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**

Some sample guidelines that can be easily translated into SPARQL constraints

Integrity constraints

- *“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.”*

Some sample guidelines that can be easily translated into SPARQL constraints

Integrity constraints

- *"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."*
- *"The second milling step must give an "Output solid constituent size" smaller than 0,5-1 mm."*

Some sample guidelines that can be easily translated into SPARQL constraints

Classification constraints

- *“Topic Bioref-PM-PC-UFM-PS : included experiments are composed of a pre-milling step, followed by a physico-chemical treatment, then by an ultrafine milling step (ball milling, wet disk milling, etc.), a press and separation step (washing and filtration), and finally the enzymatic hydrolysis step. This topic requires a press and separation step because there are a lot of effluents in the physico-chemical step or because the milling is made with effluent. The second milling step must give an “Output solid constituent size” smaller than 0,5-1 mm. (en)”*

Examples of guidelines that **cannot** be easily translated into SPARQL constraints

- *“In all treatments, when the authors indicate “overnight”, we considered a duration treatment between 10 and 15 hours”*

Examples of guidelines that **cannot** be easily translated into SPARQL constraints

- *"In all treatments, when the authors indicate "overnight", we considered a duration treatment between 10 and 15 hours"*
- *"Furthermore, we consider that the glucose rate equals to glucan rate divided by 0.9."*

Statistics

A promising approach

In the biorefinery ontology alone we have:

- 11 occurrences of the phrase *"equal to"*
- 5 occurrences of the phrase *"equals to"*
- 11 occurrences of the phrase *"sum of"*
- 3 occurrences of the phrase *"divided by"*
- 2 occurrences of the phrase *"multiplied by"*

spread across guidelines associated with 30 relation concepts.

At least 10 of them can be easily translated into SPARQL constraints.

Thanks!