Représentation et extraction de données de la littérature dans le domaine de la transformation des agro-ressources

Patrice Buche[1], Juliette Dibie[2], Martin Lentschat[3] et Mathieu Roche[4]

- [1]IATE, Montpellier Univ
- [2]MIA, Paris Saclay
- [3]LERASS, Toulouse Univ 1
- [4]TETIS, Montpellier Univ

Context and issues

Decision support systems based on the **scientific literature** (e.g., bioprocess efficiency / packaging solutions comparison) where following issues must be taken into account:

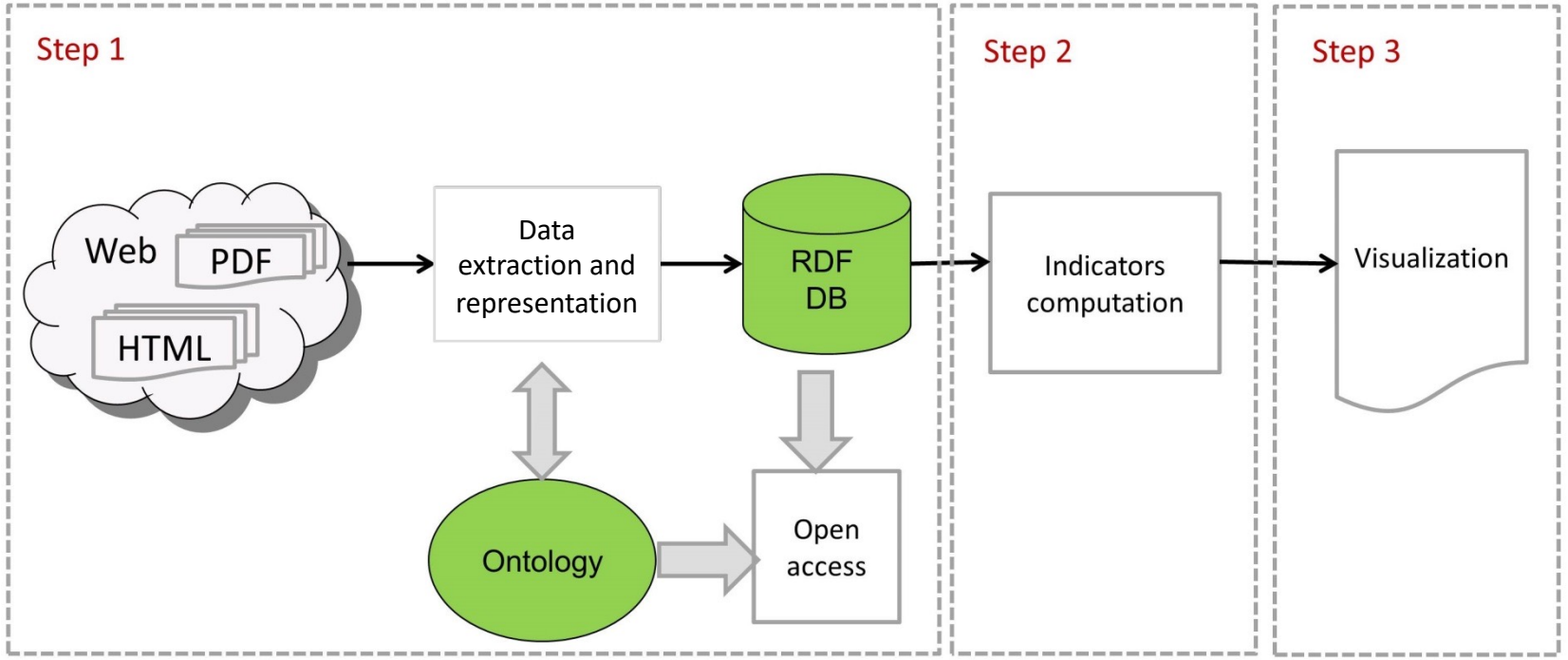
- Scientific data **heterogeneously** in terms of vocabulary and structuration (textual, tabular and image format);
- Data **reliability** due to the way experiments have been conducted.

Proposal

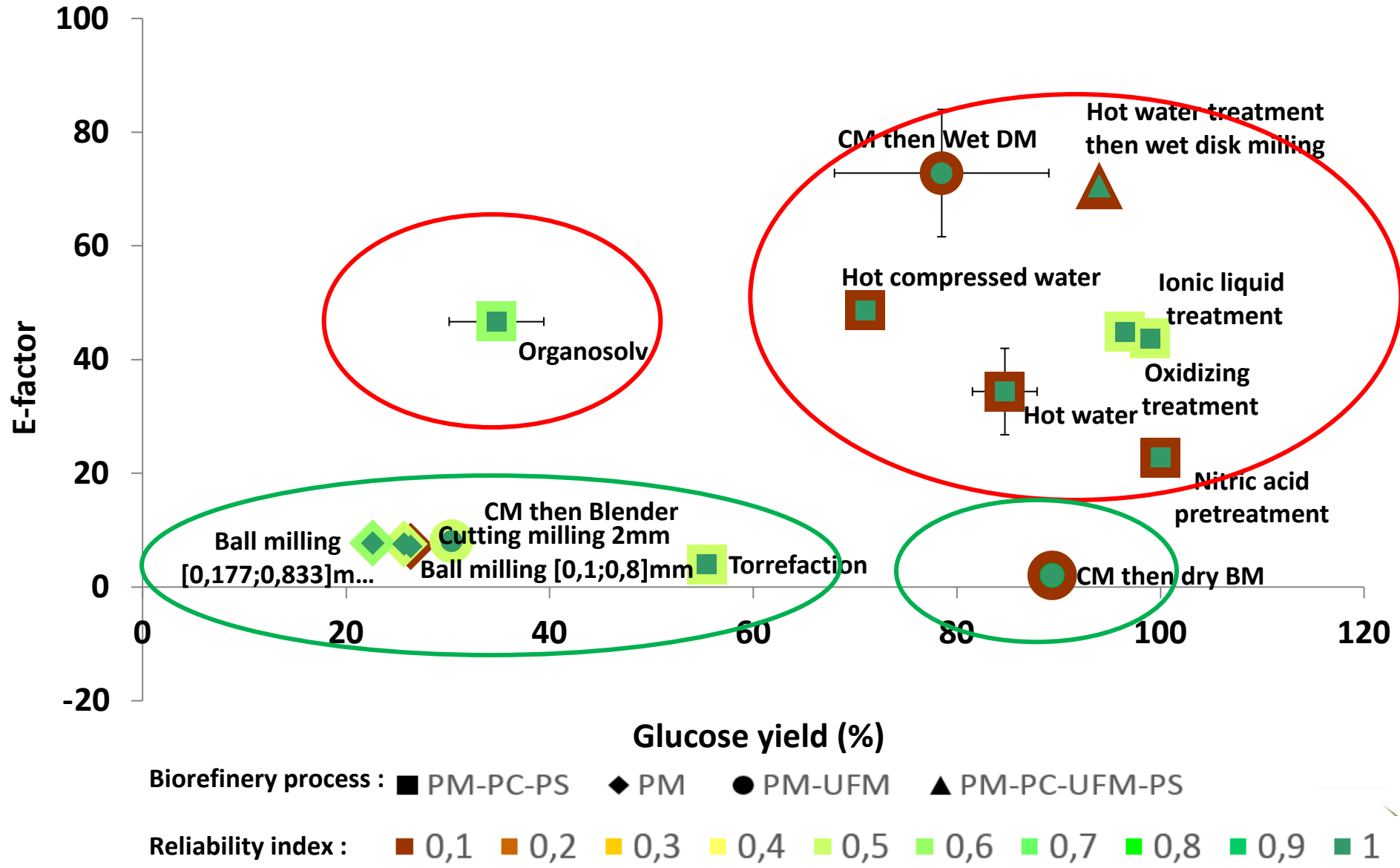
Decision Support Systems (DSS) workflows which deals with the preceding issues in order to:

1. **Extract, represent** and **integrate** in standardized database table format **qualitative** and **quantitative** experimental data by using **ontologies and table & text-mining**
2. **Assess** data source reliability
3. Depending on targeted decision support
   - **Compute** and **visualize** indicators which take into account data imprecision and reliability
   - **Rank solutions** based on mandatory and optional criteria

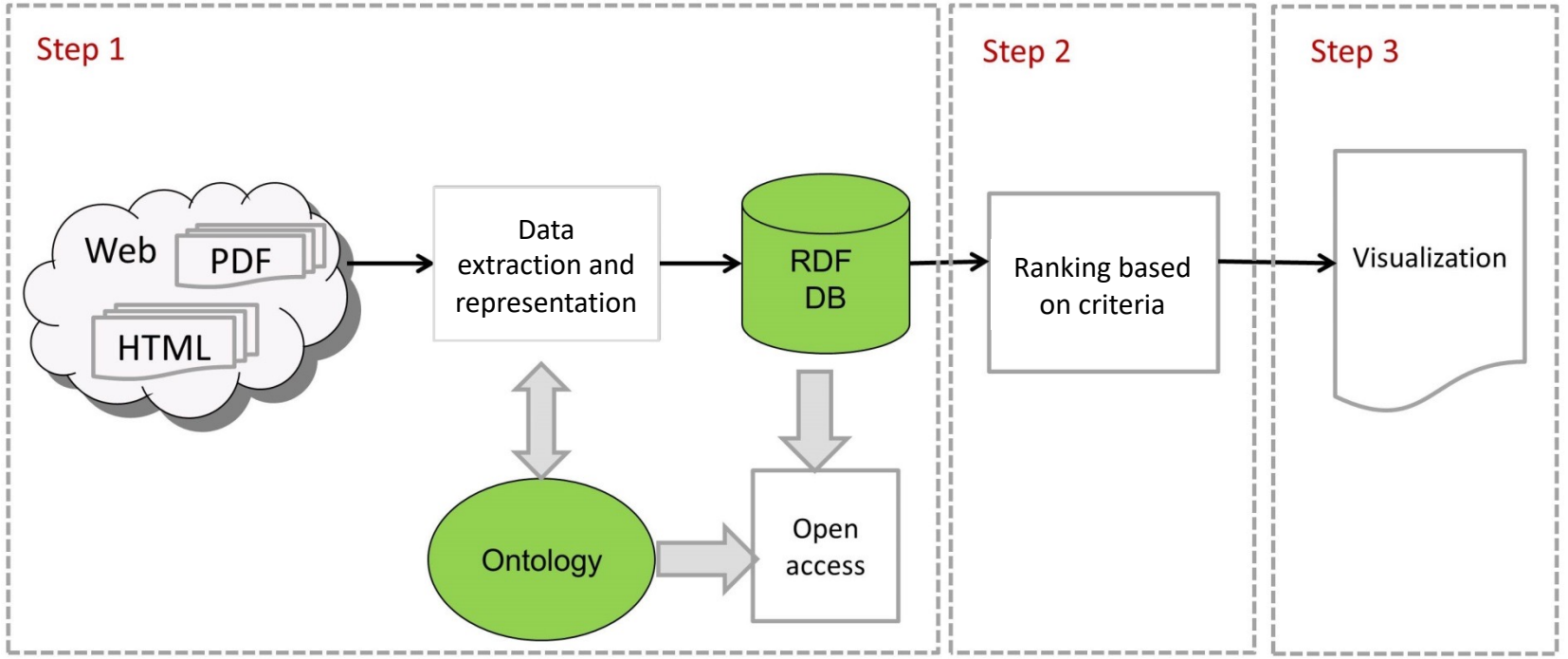
Example 1: Architecture of the Decision Support System (DSS) for bioprocess efficiency comparison



Multicriteria Decision Support System for Bioprocess Efficiency Comparison (Loustau-Cazalet et al. 2016)



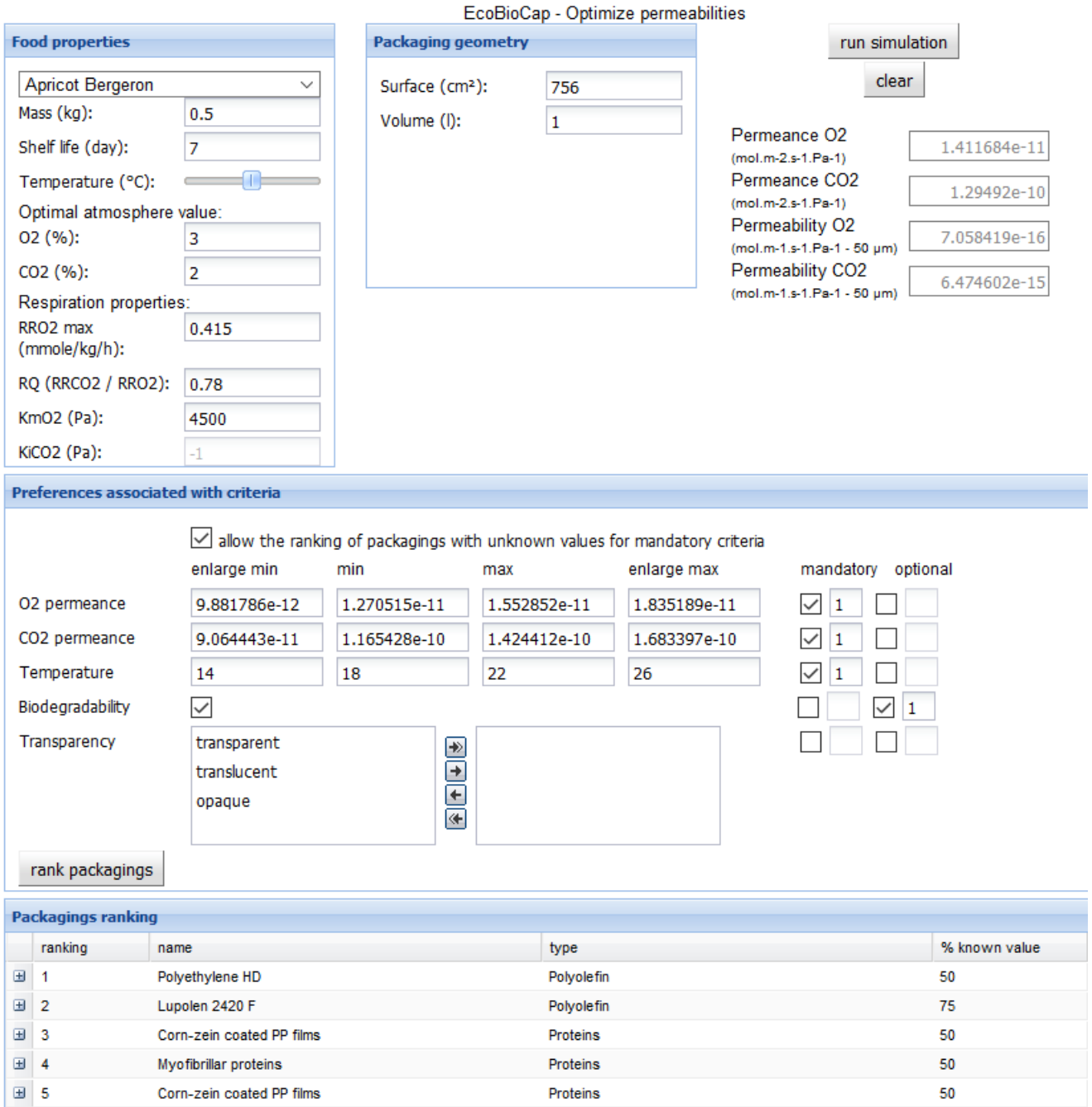
Example 2: Architecture of the Decision Support System (DSS) for food packaging selection



Food packaging decision support system architecture (Guillard et al. 2015, Pignères et al. 2025)

Current database content : 785 packaging solutions (3518 packaging characteristics)

https://ico.iate.inrae.fr/EcoBioCap Querying/



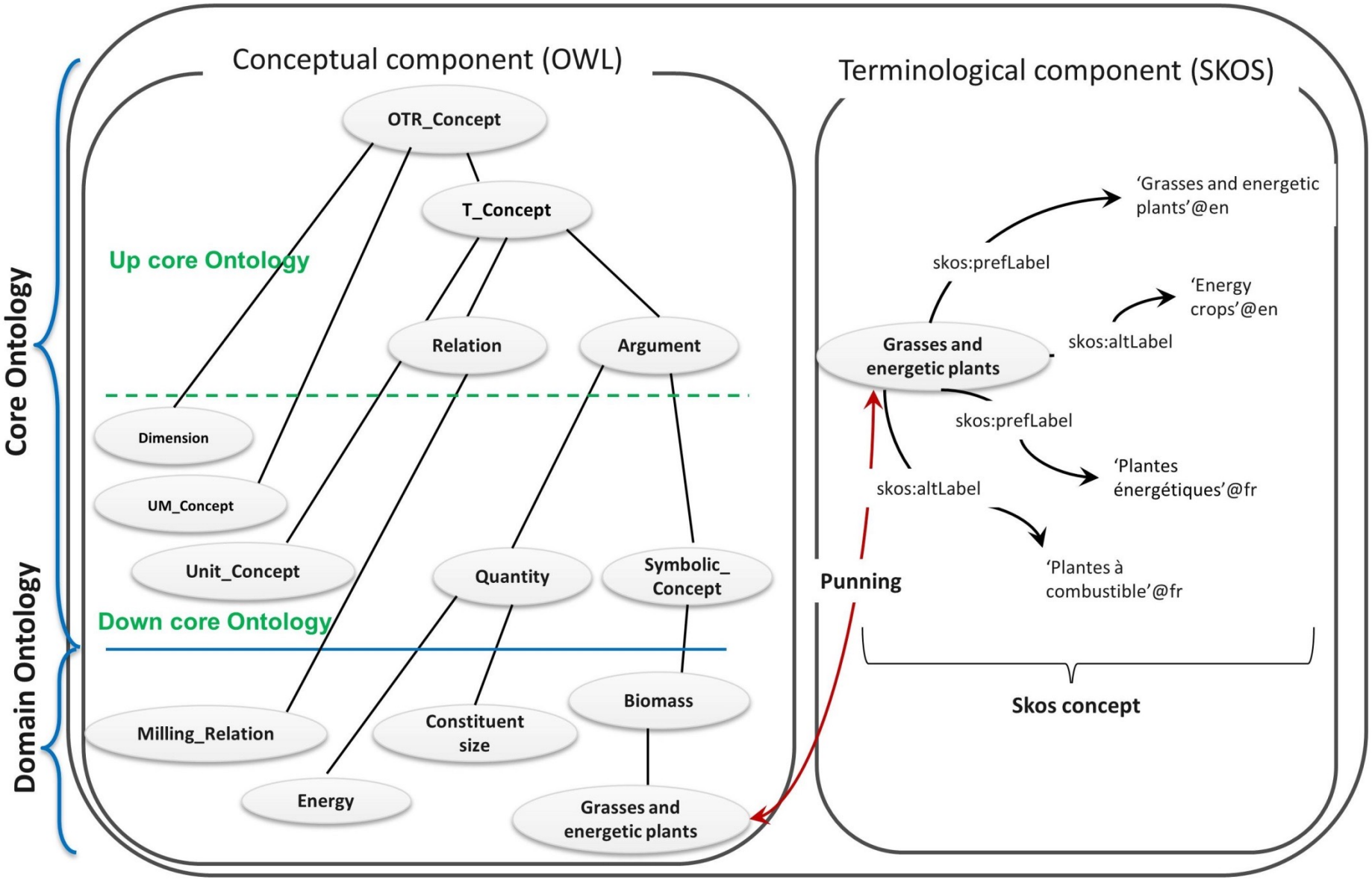
Key ideas to integrate data

- Representing knowledge in ontologies using international standards to guide extraction of data (semantic web languages as OWL/RDF)
- Assessing data source reliability
- Developing end-user oriented Web applications for ontology and data management based on semantic web languages
- Designing table & text-mining workflow to extract data from scientific litterature

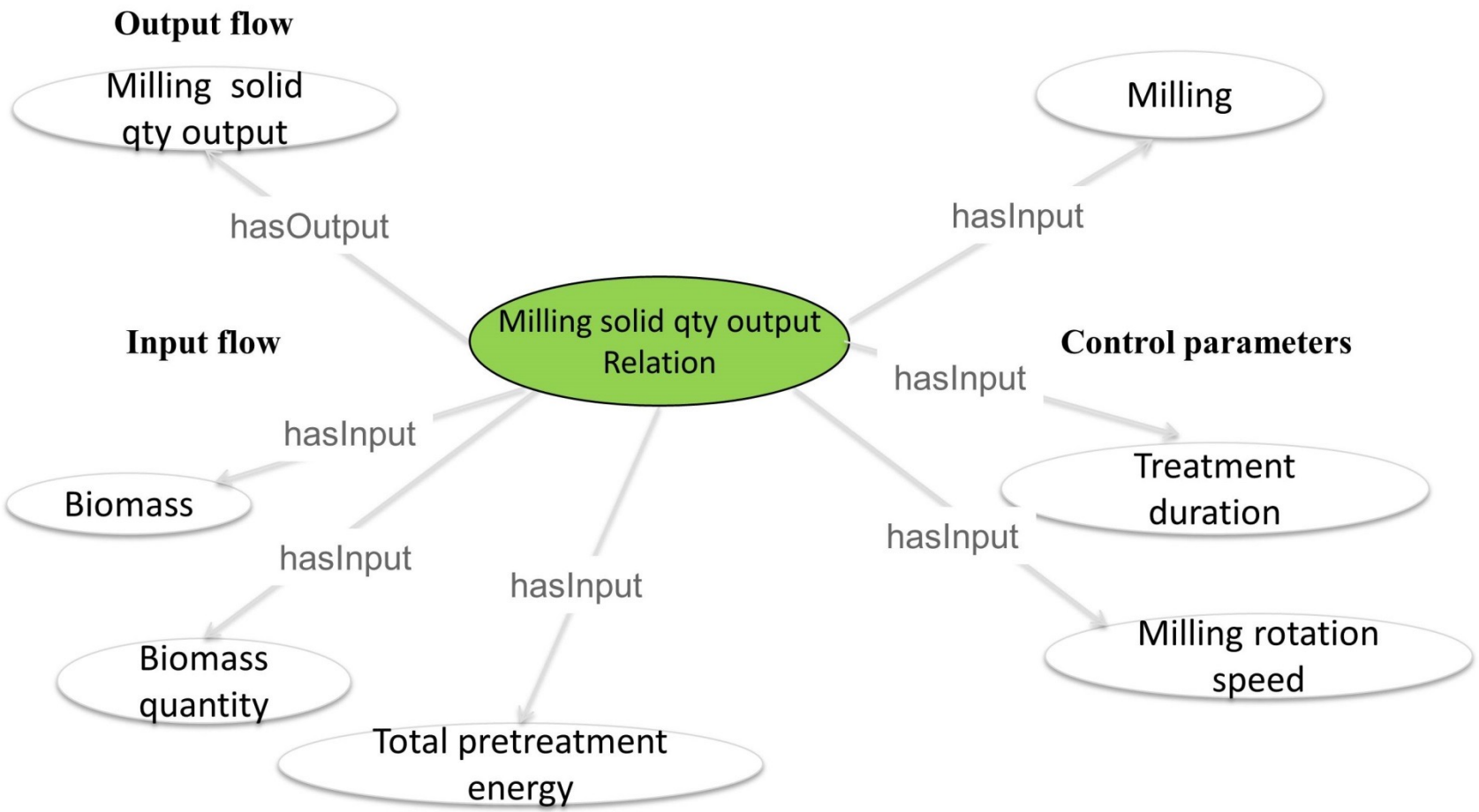
Key ideas to integrate data

- **Representing knowledge in ontologies using international standards to guide extraction of data (semantic web languages as OWL/RDF)**
- Assessing data source reliability
- Developing end-user oriented Web applications for ontology and data management based on semantic web languages
- Designing table & text-mining workflow to extract data from scientific litterature

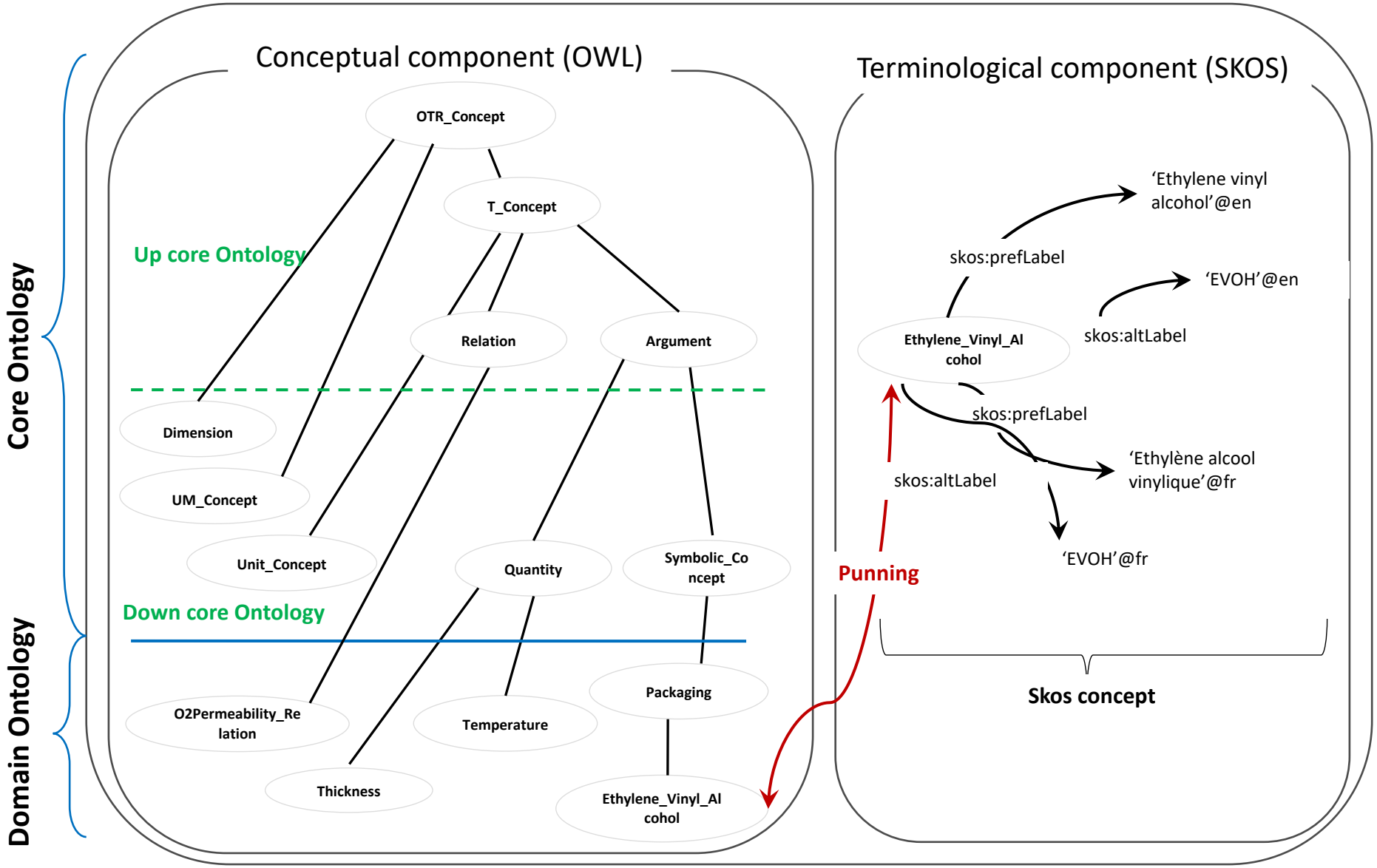
naryQ Onto Terminological Resource (OTR) with BIOREFINERY specialisation



A relation concept representing milling unit operation

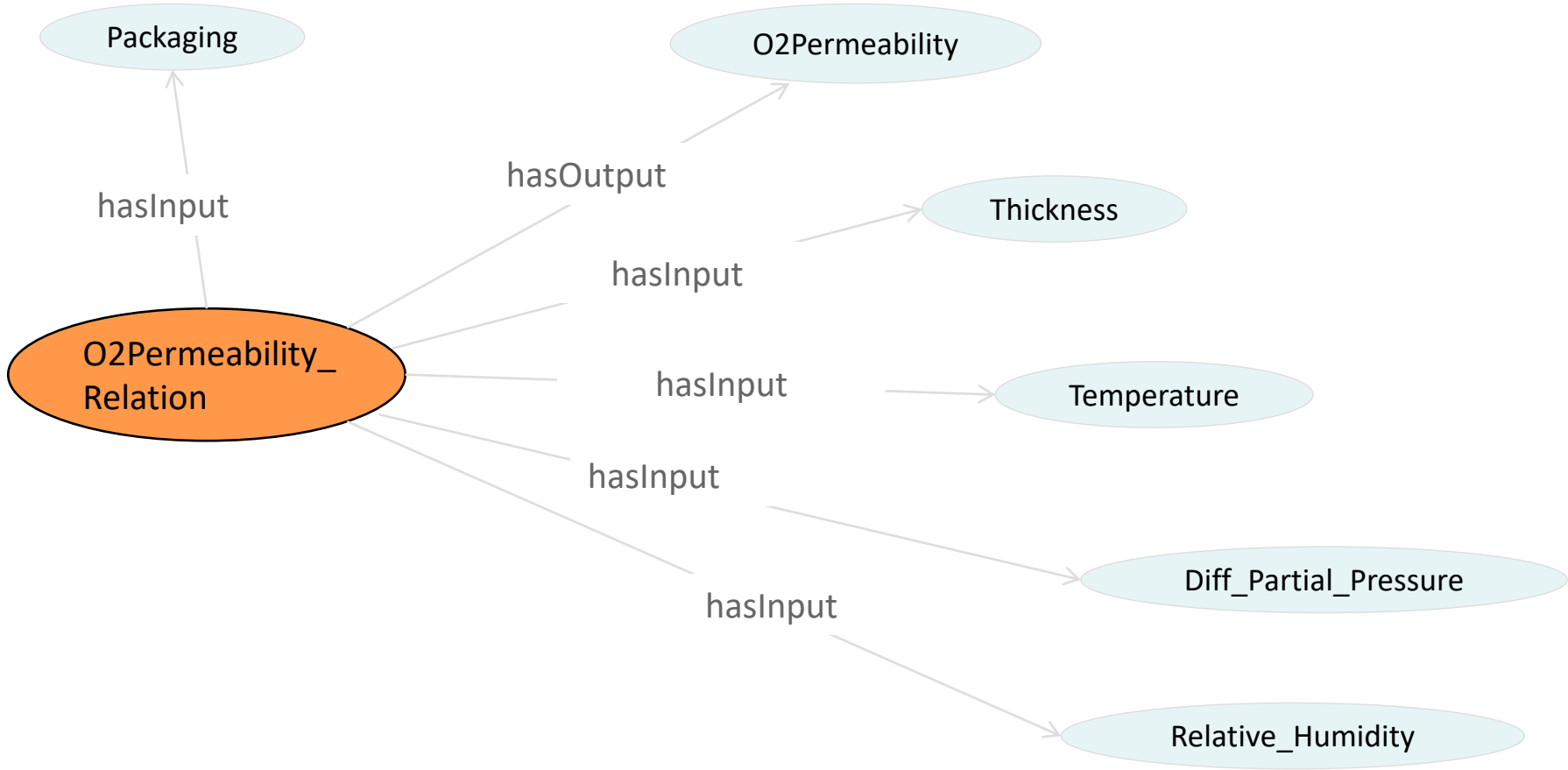


naryQ Onto Terminological Resource (OTR) with TRANSMAT specialisation



(Touhami et al. 2011, Buche et al. 2013c)

A relation concept representing packaging 02 permeability measurement



Key ideas to integrate data

- Representing knowledge in ontologies using international standards to guide extraction of data (semantic web languages as OWL/RDF)
- Assessing data source reliability
- **Developing end-user oriented Web applications for ontology and data management based on semantic web languages**
- Designing table & text-mining workflow to extract data from scientific litterature

@Web platform: Data source reliability assessment

Document's general information

Document's name :
Wet disk milling pretreatment without sulfuric acid for enzymatic hydrolysis of rice straw

Associate topic : Bioref-PM-PC-PS

Associate ontology : IC2ACV

Accepted Tables : 3

Rejected Tables : 1

Authors :
Akihiro Hideno, Hiroyuki Inoue, Kenichiro Tsukahara, Shinji Fujimoto, Tomoaki Minowa, Seiichi Inoue, Takashi Endo, Shigeki Sawayama

Journal : Bioresource Technology

Year : 2009

Volume : 100

Document's criteria values

Criterion age and citation number
Citation Number : more than 40
Age : 3 to 8 years old

Criterion age and top citation
Age : 3 to 8 years old
Top Citation : top 10%

Criterion source type
Source Type : journal article

Criterion Sugar analysis method
Sugar analysis method : HPLC

Criterion Energy measure reproducibility
Energy measure reproducibility : unknown

Criterion Enzymatic hydrolysis reproducibility
Enzymatic hydrolysis reproducibility : average

Criterion Biochemical and physico-chemical analysis reproducibility
biochemical and physico-chemical analysis repetition : average

Reliability model

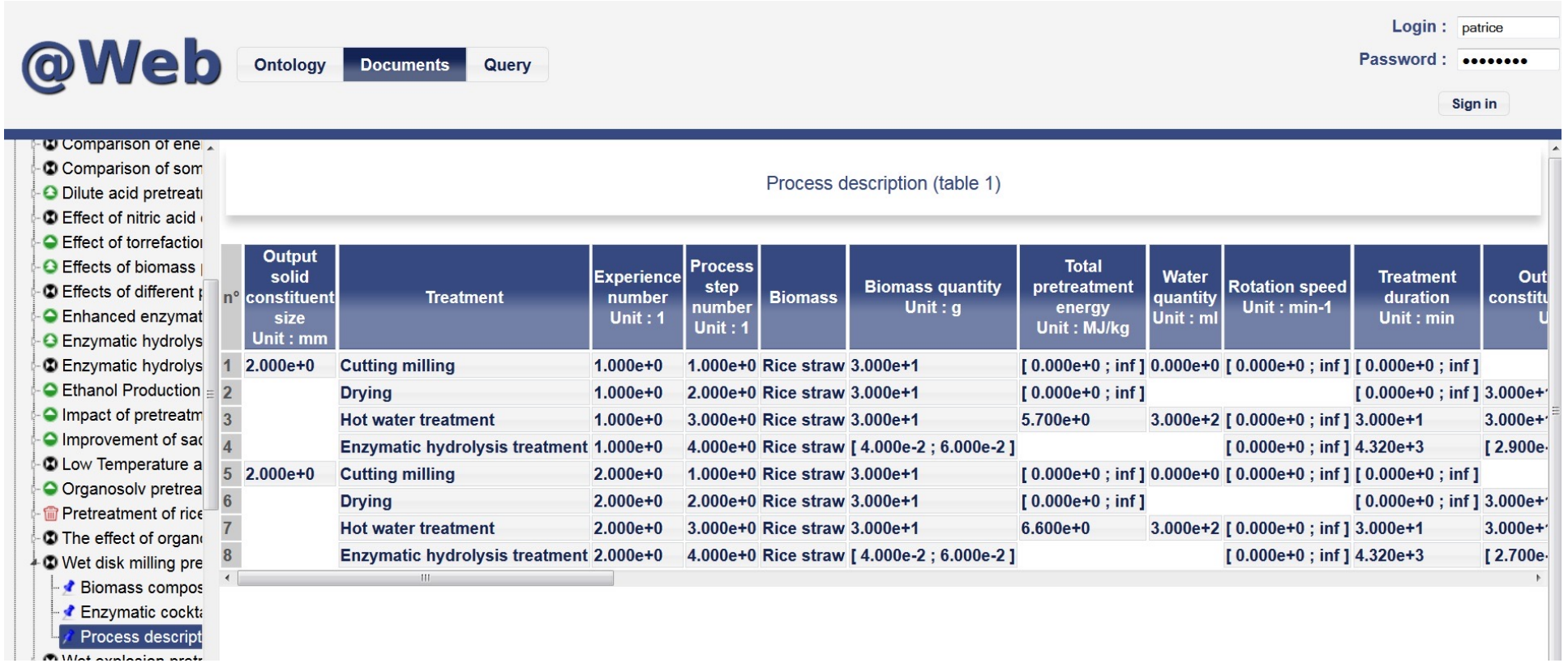
Reliability assessment's document information

Reliability results
Low expectation : 1.5 ; High expectation : 4.98

Known criteria values rate : 100.0 %

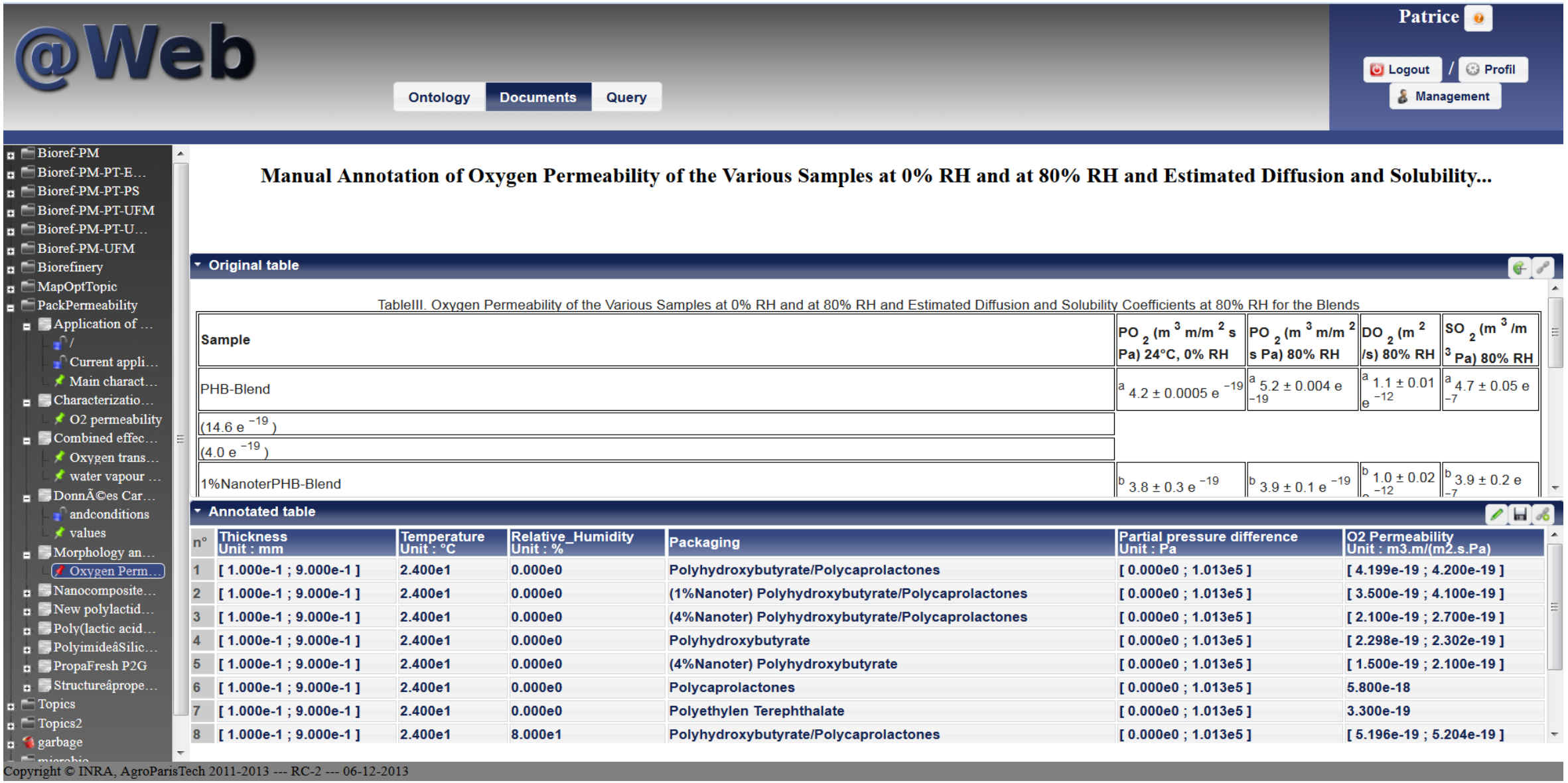
Last assessment date : 2015-01-28

@Web platform: Annotated table using several relation concepts



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 : MJ/kg	Water quantity Unit : ml	Rotation speed Unit : min-1	Treatment duration Unit : min	Out constitu U
1	2.000e+0	Cutting milling	1.000e+0	1.000e+0	Rice straw	3.000e+1	[0.000e+0 ; inf]	0.000e+0	[0.000e+0 ; inf]	[0.000e+0 ; inf]	
2		Drying	1.000e+0	2.000e+0	Rice straw	3.000e+1	[0.000e+0 ; inf]			[0.000e+0 ; inf]	3.000e+
3		Hot water treatment	1.000e+0	3.000e+0	Rice straw	3.000e+1	5.700e+0	3.000e+2	[0.000e+0 ; inf]	3.000e+1	3.000e+
4		Enzymatic hydrolysis treatment	1.000e+0	4.000e+0	Rice straw	[4.000e-2 ; 6.000e-2]			[0.000e+0 ; inf]	4.320e+3	[2.900e-
5	2.000e+0	Cutting milling	2.000e+0	1.000e+0	Rice straw	3.000e+1	[0.000e+0 ; inf]	0.000e+0	[0.000e+0 ; inf]	[0.000e+0 ; inf]	
6		Drying	2.000e+0	2.000e+0	Rice straw	3.000e+1	[0.000e+0 ; inf]			[0.000e+0 ; inf]	3.000e+
7		Hot water treatment	2.000e+0	3.000e+0	Rice straw	3.000e+1	6.600e+0	3.000e+2	[0.000e+0 ; inf]	3.000e+1	3.000e+
8		Enzymatic hydrolysis treatment	2.000e+0	4.000e+0	Rice straw	[4.000e-2 ; 6.000e-2]			[0.000e+0 ; inf]	4.320e+3	[2.700e-

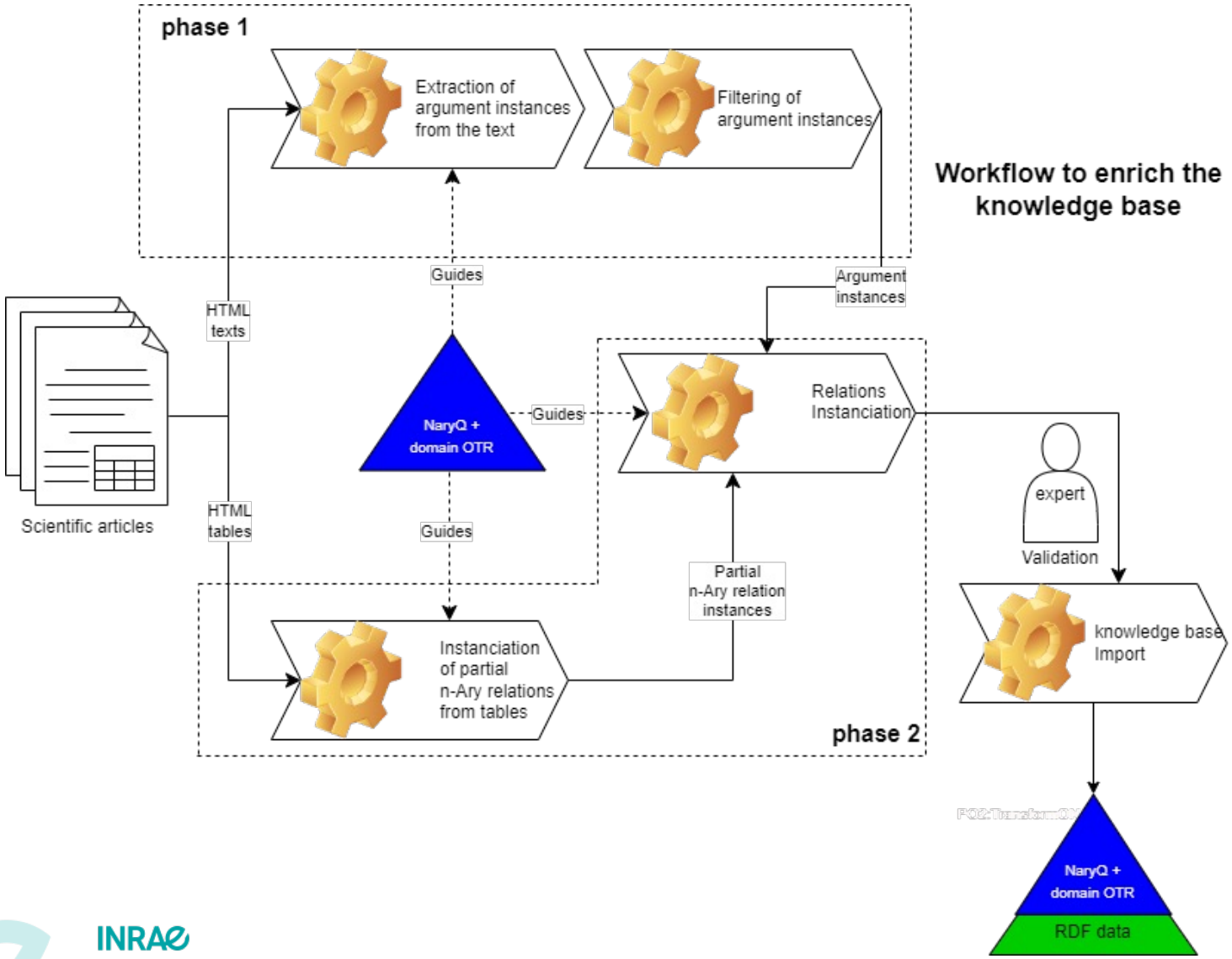
@Web platform: Editing an annotated table using a relation concept



Key ideas to integrate data

- Representing knowledge in ontologies using international standards to guide extraction of data (semantic web languages as OWL/RDF)
- Assessing data source reliability
- Developing end-user oriented Web applications for ontology and data management based on semantic web languages
- **Designing semi-automatic table & text-mining workflow to extract data from scientific litterature**

N-Ary relation extraction workflow (PhD thesis, december 2021)



- Corpora:
  - .html and open access articles
  - PHASE 1:
    - 15 articles
    - 3 annotators
    - 1772 annotated argument instances
  - PHASE 2:
    - 10 articles
    - 1 annotator, 3 validators
    - 31 tables, 331 partial relations made of 779 argument instances
  - FINAL:
    - 10 articles
    - 1 annotator using previous two corpora
    - 331 reconstituted relations containing 1547 argument instances

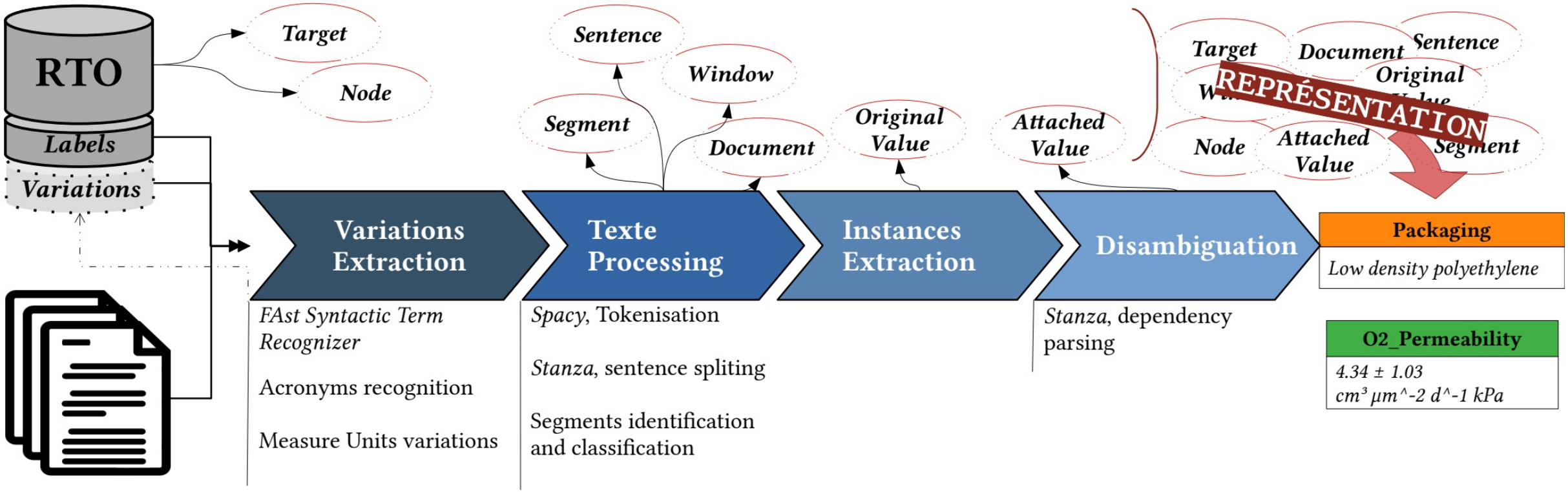
Main challenges

« The oxygen permeability of sorbitol–starch–water films (SSW) were studied. Sorbitol had an antiplasticisation effect for contents below 21% (wt), as indicated by a drop in oxygen permeability $1.43 * 10^{-16}$ cm^3 cm/cm^2 s Pa for the sample not containing sorbitol and $0.15 * 10^{-16}$ cm^3 cm/cm^2 s Pa for the sample containing sorbitol at a relative humidity value of 30%. »

- Format heterogeneity
- Data heterogeneity (e.g., symbolic, quantitative)
- Terminological variations (e.g., *'oxygen permeability' / 'permeability to oxygen'*)
- Argument instances scattering
- Multiplicity of argument and relations instances
- *"misleading"* argument instances

Argument instances recognition from the text

- PHASE I : Ontology driven process

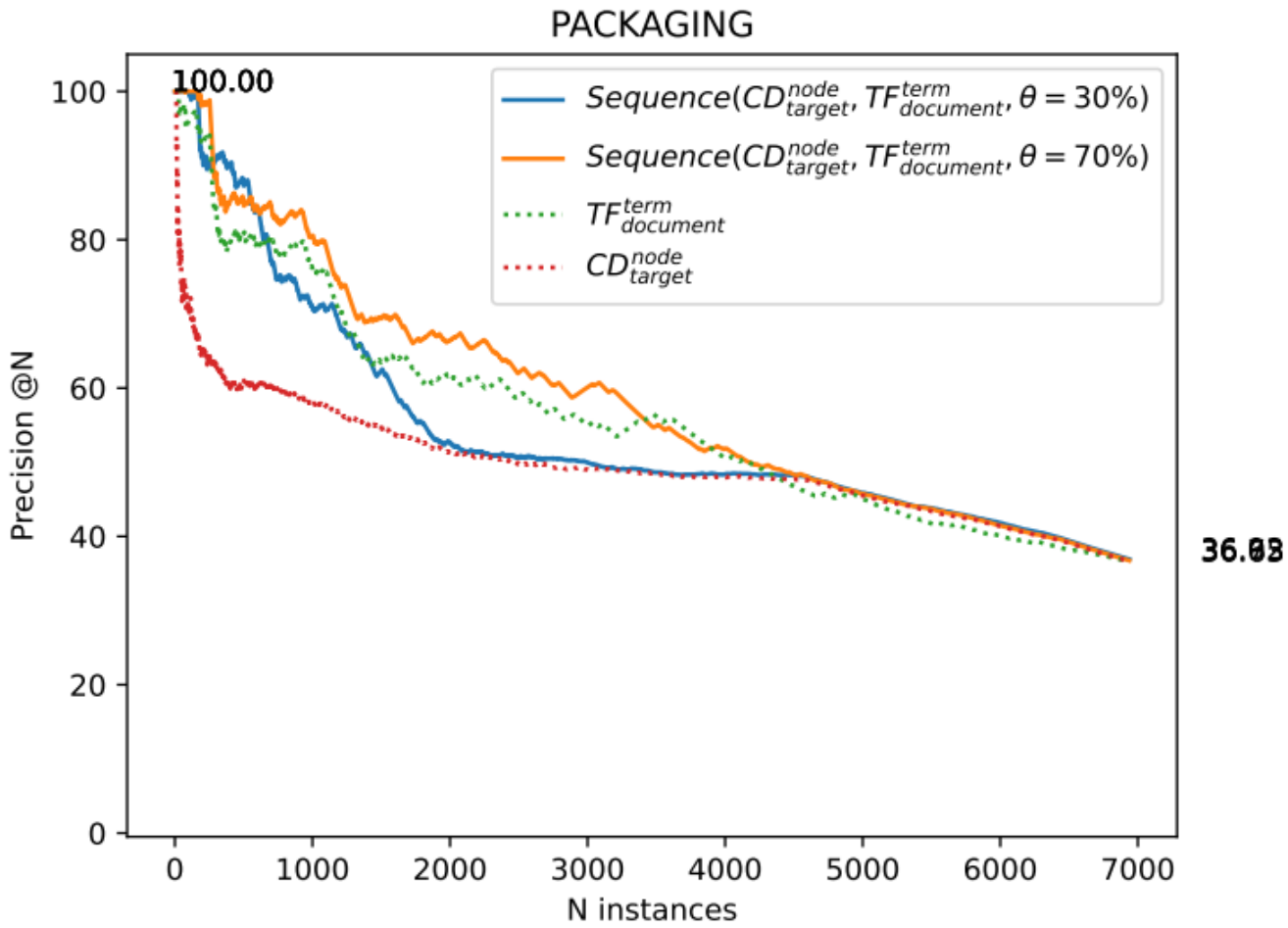


- Problem : high recall (.85) but low precision (.41)

2
2

Relevance filtering

- Multi criteria approach relying on SciPuRe features:
  - Ontological (Target, Node)
  - Lexical (Original and Attached Values)
  - Structural (Sentence, Segment, ...)
- Three scores:
  - Term Frequency
  - Inverse Category Frequency
  - Conceptual Distance
- Different strategies for different arguments



2
3

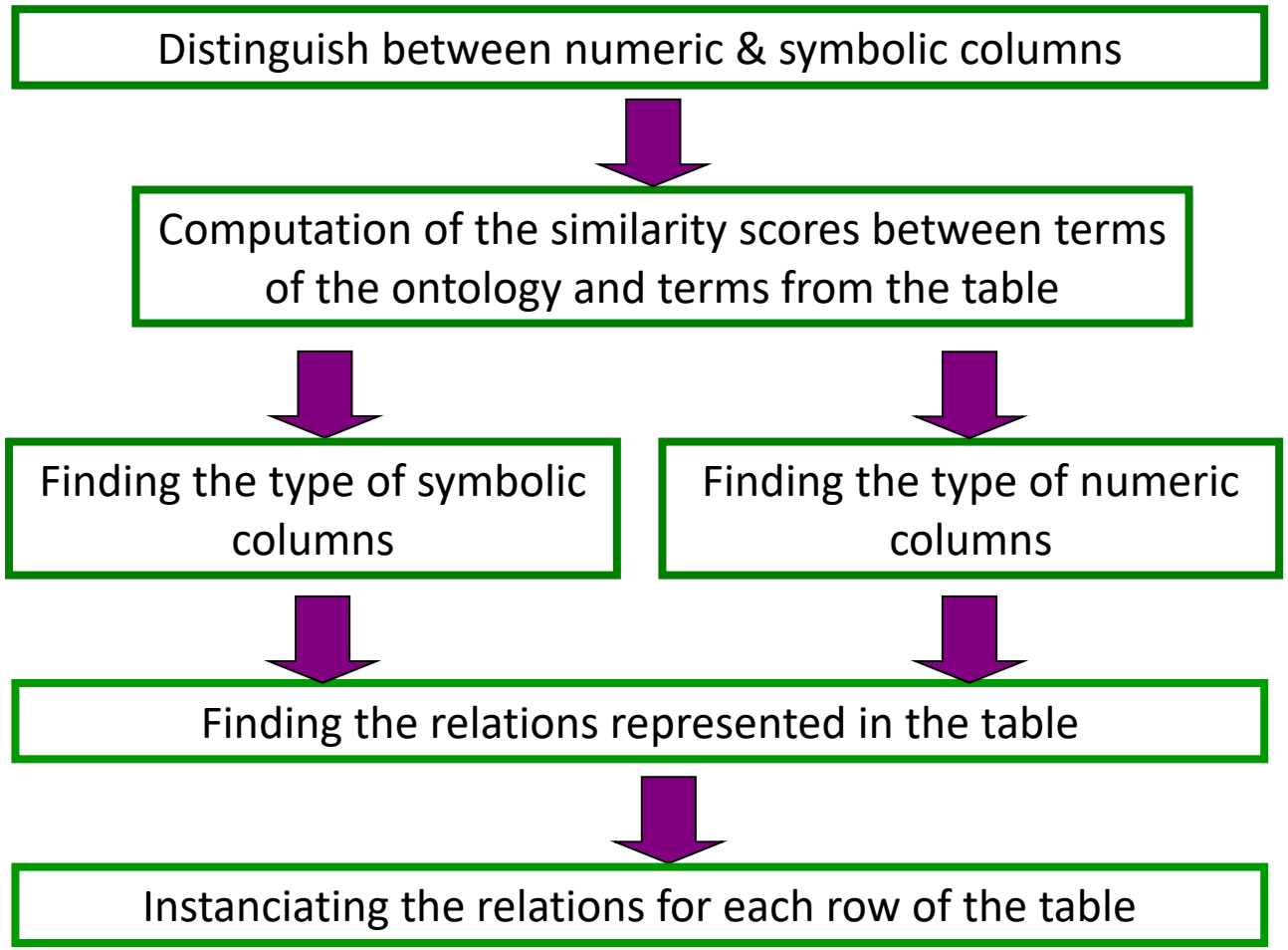
Data table extraction and annotation guided by the OTR (Buche et al. 2013)

Table 1. Water vapour permeability (WVP) at 25 °C of PE films coated with chitosan compared to chitosan self standing films prepared with different casting solvents and plasticizers.

Sample	Δ RH 70%	Δ RH 45%	Δ RH 33%
	WVP×10^{-13} (g/msPa)	WVP×10^{-13} (g/msPa)	WVP×10^{-13} (g/msPa)
PE	4.62±0.73f	5.55±0.23f	7.72±2.58f
CS coated PE	12.37±1.14f	6.67±0.23f	7.88±2.39f
PECSEinv	9.14±1.09f	6.41±3.28f	2.85±0.34f
CSA	4161.31±656.17a,b	2199.80±1048.33d,e	25.71±2.20f
CSE	4100.77±588.88a,b	2884.37±346.43b,c,d,e	38.71±2.61f
CSAGLY	5410.08±1543.67a	1905.39±149.64e	26.14±1.24f
CSEGLY	3481.46±343.88b,c,d	2635.38±414.28c,d,e	105.17±6.57f

PE, polyethylene; CS coated PE, chitosan (CSE) coated polyethylene; PESCEinv, coating exposed to dry compartment; CSA, chitosan film prepared with aqueous acid solvent; CSE, chitosan film prepared with hydroalcoholic acid solvent; CSAGLY and CSEGLY, glycerol plasticized samples.

Different letters (a–f) indicate significant differences between formulations ($p<0.05$).



Data table extraction and annotation (cont.)

- Automatic extraction of partial relation instances
  - Based on existing approach (Buche & al., *"Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource"*)
  - Modified to handle different table formats
  - Produce a relation representation factorizing argument representations

StaRe (Scientific Table Representation)

	Descripteur	Valeur
ONTOLOGIQUE	Relation	H2O_Permeability_Relation
	Result_Argument	**SciPure**

Target	Node	Original_Value	Attached_Value
H2O_Perm.	H2O_Perm.	$1.27 * 10^2$ $cm^3 mm^{-2} s^{-1} bar$	*Water Perm.*

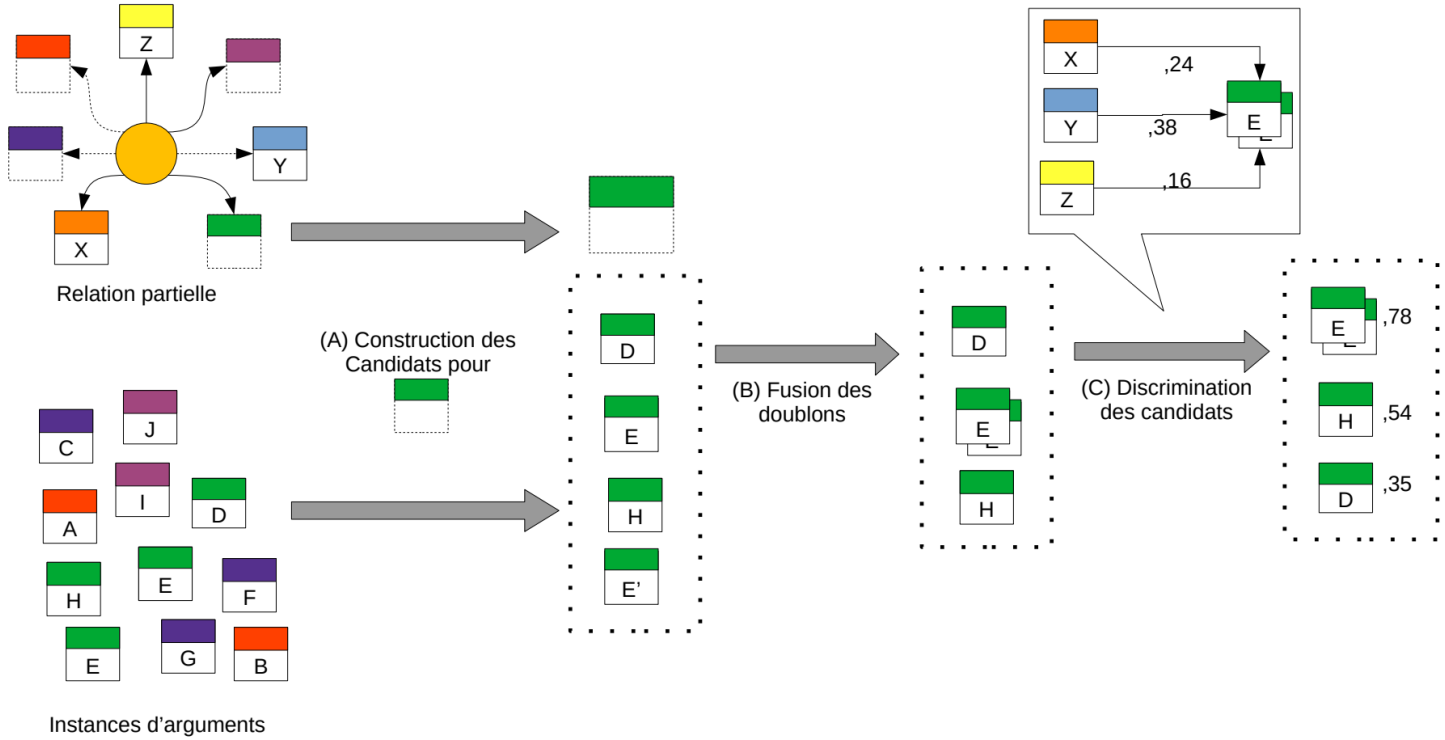
	Target	Node	Original_Value	Attached_Value
ONTOLOGIQUE	Arguments	Packaging	Chitosan	*Chitosan films*	*Chitosan films*
		Method	∅	∅	∅
		R_H	∅	∅	∅
		Temperature	Temperature	25° C	*Temp. (° C)*
		Thickness	∅	∅	∅

STRUCT.	Table	*Table 3*
	Caption	*Water permeability of tested packaging at 25° C*
	Segment	*Results and Discussion*
	Document	*Barrier properties of chitosan coated polyethylene*
	DOI	*10.1016/j.memsci.2012.02.037*

Relation reconstitution

- Three steps to add an argument instance:
  A) Select candidates instances
  B) Detect duplicates
  C) Rank candidates

  Repeat ;)



- The ranking of a candidate depends of its association with every existing argument instances in the partial relation
- Different ranking strategies:
  - Structural → distance from original table and expert guide
  - Frequency → associate instances that appear together
  - Word embedding similarity → associate instances from similar context

Relation reconstitution

- Different approaches for different strategies

Method	Criteria	Number of candidates			
		1	3	5	10
Structural	proximity	.35	.56	.58	.65
Structural	+expert guide	.45	.56	.61	**.74**
Frequency	Jaccard	**.48**	.54	.61	.66
Frequency	PMI	.44	.53	.60	.68
Word embeddings	roberta	.40	**.59**	.64	.70
Word embeddings	scibert	.39	.57	**.65**	.70

Best F-scores obtained (with 20% relevance filtering)

- Web app: https://gitlab.softwareheritage.org/eskode/webapp_ARTEXT4LOD

Conclusion

Conclusion

- Design and implementation of a generic workflow for manual or semi-automatic annotation of scientific data from the literature
- The method takes into account 1) textual data, 2) data in tables and figures, 3) qualitative and quantitative data
- Design of 2 DSS and a CO2 solubility in food prediction model to demonstrate the reuse of annotated scientific data
- Data source reliability assessment used for visualization and ML training set modification (tested on microfiltration unit operation data)
- Resources: 3 OTR: TRANSMAT, BIOREFINERY, MICROFILTRATION and 6 data papers including 2 papers dedicated to text-mining approach

Conclusion (cont.)

Limitations, Hypothesis :

- Data table annotation heuristic is based on the assumption that arguments of n-ary relations are expressed in columns
- Data annotation is guided by OTR which requires to define in terms of n-ary relations which data will be extracted and annotated
- Table & text mining performance depends on OTR domain content

What’s next ?

3
1

What next?

- Publication in 2026 of a book dedicated to knowledge engineering approaches applied to agro-resource transformation (QUAE edition) including a summarization of this work

- Revisiting the approach taking into account LLMs

Thank you for your attention :-))

Any questions ?

patrice.buche@inrae.fr, martin.lentschat@univ-tlse2.fr

Bibliographic references: methodological papers

- P. Buche, J. Dibie-Barthelemy, L. Ibanescu and L. Soler, "Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource," *in IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 805-819, April 2013, doi: 10.1109/TKDE.2011.245.

- Sébastien Destercke, Patrice Buche, Brigitte Charnomordic. Evaluating data reliability : an evidential answer with application to a web-enabled data warehouse. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25 (1), pp.92-105. ⟨10.1109/tkde.2011.179⟩. ⟨hal-01267940⟩

- Valérie Guillard, Patrice Buche, Sébastien Destercke, Nouredine Tamani, Madalina Croitoru, et al.. A Decision Support System to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. *Computers and Electronics in Agriculture*, 2015, 111, pp.131-139. ⟨10.1016/j.compag.2014.12.010⟩. ⟨hal-01104835⟩

- Charlotte Lousteau-Cazalet, Abdellatif Barakat, Jean-Pierre Belaud, Patrice Buche, Guillaume Busset, et al.. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture*, 2016, 127, pp.351-367. ⟨10.1016/j.compag.2016.06.020⟩. ⟨lirmm-01346685⟩

- Martin Lentschat, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche. A new method to extract n-Ary relation instances from scientific documents. *Expert Systems with Applications*, 2022, 209, pp.118332. ⟨10.1016/j.eswa.2022.118332⟩. ⟨hal-03767632⟩

- Martin Lentschat, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche. Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications. *International Journal of Intelligent Information and Database Systems*, 2022, 15 (1), pp.78. ⟨10.1504/IJIIDS.2022.120146⟩. ⟨hal-03616243⟩

Bibliographic references: data papers

- Mélanie Munch, Patrice Buche, Luc Menut, Julien Cufi, Valérie Guillard. CO2 solubility and composition data of food products stored in data warehouse structured by an ontology. *Data in Brief*, 2023, 47, pp.108950. ⟨10.1016/j.dib.2023.108950⟩. ⟨hal-03998593⟩
- Martin Lentschat, Patrice Buche, Luc Menut, Romane Guari, Mathieu Roche. Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables. *Data in Brief*, 2022, 41, pp.108000. ⟨10.1016/j.dib.2022.108000⟩. ⟨hal-03610433⟩
- Patrice Buche, Stéphane Dervaux, Nadine Leconte, Maellis Belna, Manon Granger-Delacroix, et al.. Milk microfiltration process dataset annotated from a collection of scientific papers. *Data in Brief*, 2021, 36, pp.107063. ⟨10.1016/j.dib.2021.107063⟩. ⟨hal-03227408⟩
- Martin Lentschat, Patrice Buche, Juliette Dibie-Barthelemy, Luc Menut, Mathieu Roche. Food packaging permeability and composition dataset dedicated to text-mining. *Data in Brief*, 2021, 36, pp.107135. ⟨10.1016/j.dib.2021.107135⟩. ⟨hal-03250848⟩
- Charlène Fabre, Patrice Buche, Xavier Rouau, Claire Mayer-Laigle. Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. *Data in Brief*, 2020, 33, pp.106430. ⟨10.1016/j.dib.2020.106430⟩. ⟨hal-03004903⟩
- Valérie V. Guillard, Patrice Buche, Juliette Dibie-Barthelemy, Stéphane Dervaux, Filippo Acerbi, et al.. CO2 and O2 solubility and diffusivity data in food products stored in data warehouse structured by ontology. *Data in Brief*, 2016, 7, pp.1556-1559. ⟨10.1016/j.dib.2016.04.044⟩. ⟨hal-01357730⟩

35

Key ideas to integrate data

- Representing knowledge in ontologies using international standards to guide extraction of data (semantic web languages as OWL/RDF)
- **Assessing data source reliability**
- Developing end-user oriented Web applications for ontology and data management based on semantic web languages
- Designing table & text-mining workflow to extract data from scientific litterature

Revisiting the approach using LLMs

- Necessity to divide the extraction in sub-tasks
- Models deployed locally :
  - Are not able to process full documents
  - Need to be ‘put on a leash’ (no-chat and greedy mode with a determined max number of new tokens)
- The argument instance extraction is approached as a text annotation task with few-shot examples
  - Difficulties with distant or multiple value-units pairs
- Tables need restructuring / standardization before annotation
- The argument instances relevance task is hard to design (might be ignored ?)
- Relation reconstitution present argument instances in (various) context and ask for a related/not-related output

Evaluating data source reliability (Destercke et al. 2013)

IC2ACV 's domain values' reliability judgements

source type	unknown	not at all reliable	not at all or hardly	hardly reliable	hardly or average	average reliable	average or reliable	reliable	reliable or very	very reliable
> Book						-+				
> Conference proceeding				-						
> Encyclopedia article						-+				
> Journal article										++
> Patent	?									
> Report				-						
> Web page				-						
> Thesis						-+				
> Technical sheet									[+,++]	

@Web platform: Data source reliability assessment

Document's general information

Document's name :
Wet disk milling pretreatment without sulfuric acid for enzymatic hydrolysis of rice straw

Associate topic : Bioref-PM-PC-PS

Associate ontology : IC2ACV

Accepted Tables : 3

Rejected Tables : 1

Authors :
Akihiro Hideno, Hiroyuki Inoue, Kenichiro Tsukahara, Shinji Fujimoto, Tomoaki Minowa, Seiichi Inoue, Takashi Endo, Shigeki Sawayama

Journal : Bioresource Technology

Year : 2009

Volume : 100

Document's criteria values

Criterion age and citation number
Citation Number : more than 40
Age : 3 to 8 years old

Criterion age and top citation
Age : 3 to 8 years old
Top Citation : top 10%

Criterion source type
Source Type : journal article

Criterion Sugar analysis method
Sugar analysis method : HPLC

Criterion Energy measure reproducibility
Energy measure reproducibility : unknown

Criterion Enzymatic hydrolysis reproducibility
Enzymatic hydrolysis reproducibility : average

Criterion Biochemical and physico-chemical analysis reproducibility
biochemical and physico-chemical analysis repetition : average

Reliability model

Reliability assessment's document information

Reliability results
Low expectation : 1.5 ; High expectation : 4.98

Known criteria values rate : 100.0 %

Last assessment date : 2015-01-28

Evaluating data source reliability (Destercke et al. 2013)

- Use meta data to evaluate (biorefinery) data source reliability
- Model with belief functions (flexibility, subset combination, ability to model ignorance)
- Output=estimated reliability $\left[\underline{\mathbb{E}_d}, \overline{\mathbb{E}_d}\right]$(Lower, upper expectations)
- Usefulness: interval comparison and ranking on reliability

d_i	(#cit vs age)	repetitions	source type	$\left[\underline{\mathbb{E}_{d_i}}, \overline{\mathbb{E}_{d_i}}\right]$
d_1	-	No	Journal paper	[1.40,4.66]
d_2	(79,3)	Yes	Journal paper	[4.67,4.97]
d_3	-	No	Technical sheet	[1.05,1.40]