A decision support system for eco-efficient biorefinery process comparison using a semantic approach

Charlotte Lousteau-Cazalet, Abdellatif Barakat, Jean-Pierre Belaud, Patrice Buche, Guillaume Busset, Brigitte Charnomordic, Stéphane Dervaux, Sébastien Destercke, Juliette Dibie-Barthelemy, Caroline Sablayrolles, et al.

A decision support system for eco-efficient biorefinery process comparison using a semantic approach

Charlotte Lousteau-Cazalet [f], Abdellatif Barakat [f], Jean-Pierre Belaud [c,d], Patrice Buche [e,f,*], Guillaume Busset [b], Brigitte Charnomordic [g], Stéphane Dervaux [h], Sébastien Destercke [i], Juliette Dibie [g], Caroline Sablayrolles [a,b], Claire Vialle [a,b]

[a] Université de Toulouse, INP-ENSIACET, LCA (Laboratoire de Chimie Agro industrielle), Toulouse, France
[b] INRA, UMR 1010 CAI, Toulouse, France
[c] Université de Toulouse, INP-ENSIACET, LGC (Laboratoire de Génie Chimique), France
[d] CNRS UMR 5503, Toulouse, France
[e] LIRMM GraphIK, Montpellier, France
[f] INRA, UMR Ingn Agropolymeres & Technol Emergentes 1208, 2 Pl Pierre Viala, F-34060 Montpellier, France
[g] INRA UMR MISTEA, 2 Pl Pierre Viala, F-34060 Montpellier 1, France
[h] INRA – AgroParisTech, UMR 518 MIA-Paris, F-75231 Paris Cedex 05, France
[i] CNRS UMR Heudyasic, rue Personne de Roberval, F-60200 Compiègne, France

ARTICLE INFO

Keywords:
Decision support system
Biorefinery
Uncertainty management
Knowledge engineering
Ontology
Bioprocess eco-design

ABSTRACT

Enzymatic hydrolysis of the main components of lignocellulosic biomass is one of the promising methods to further upgrading it into biofuels. Biomass pre treatment is an essential step in order to reduce cellu lose crystallinity, increase surface and porosity and separate the major constituents of biomass. Scientific literature in this domain is increasing fast and could be a valuable source of data. As these abundant sci entific data are mostly in textual format and heterogeneously structured, using them to compute biomass pre treatment efficiency is not straightforward. This paper presents the implementation of a Decision Support System (DSS) based on an original pipeline coupling knowledge engineering (KE) based on semantic web technologies, soft computing techniques and environmental factor computation. The DSS allows using data found in the literature to assess environmental sustainability of biorefinery sys tems. The pipeline permits to: (1) structure and integrate relevant experimental data, (2) assess data source reliability, (3) compute and visualize green indicators taking into account data imprecision and source reliability. This pipeline has been made possible thanks to innovative researches in the coupling of ontologies, uncertainty management and propagation. In this first version, data acquisition is done by experts and facilitated by a termino ontological resource. Data source reliability assessment is based on domain knowledge and done by experts. The operational prototype has been used by field experts on a realistic use case (rice straw). The obtained results have validated the usefulness of the system. Further work will address the question of a higher automation level for data acquisition and data source reliabil ity assessment.

1. Introduction

The bioconversion of lignocellulosic biomass has been exten sively studied in the past 30 years. Enzymatic hydrolysis of the main components of the biomass is one of the promising methods to further upgrading it into biofuels (Fig. 1). The structural hetero geneity and complexity of cell wall constituents such as crys tallinity of cellulose microfibrils, specific surface area and porosity of matrix polymers are responsible for the recalcitrance of cellulosic materials. Biomass pre treatment is consequently an essential step in order to reduce cellulose crystallinity, increase surface and porosity and separate the major constituents of bio mass (e.g. cellulose, hemicellulose, lignin, phenolic acids). The objective of such pre treatments depends on the process type and biomass structure. For instance, pre treatment methods can be divided into different categories: mechanical, physical, chemi cal, physicochemical and biological or various combinations of these (Fig. 2). Each method has its drawbacks such as energy

* Corresponding author at: INRA, UMR Ingn Agropolymeres & Technol Emergentes 1208, 2 Pl Pierre Viala, F-34060 Montpellier, France.
E-mail address: patrice.buche@supagro.inra.fr (P. Buche).

http://dx.doi.org/10.1016/j.compag.2016.06.020

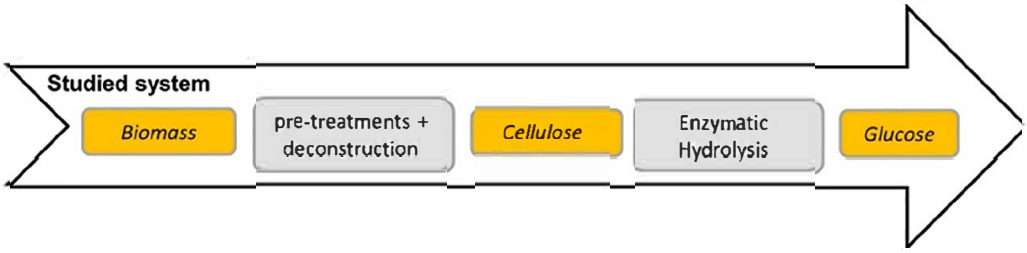

Fig. 1. Biorefinery pre-treatment process.

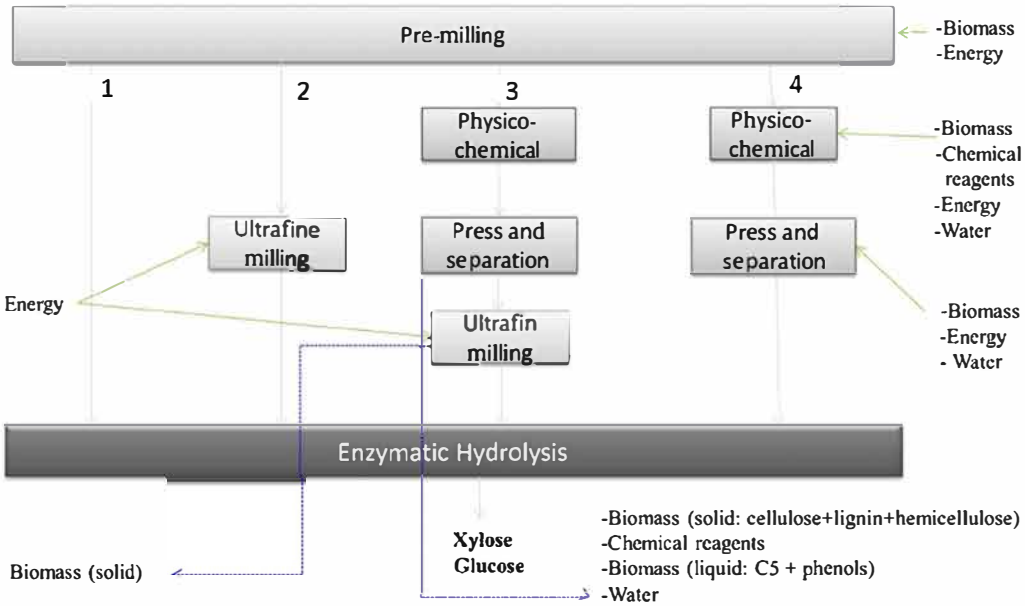

Fig. 2. Four different biorefinery processes to transform biomass into glucose.

consumption, corrosion of processing tools, water consumption, introduction of inhibiting effects, or the high number of separation and purification steps. Low or no water consumption during ligno cellulosic pre treatment can decrease the generated effluents, and also reduce the energy input for the biomass pre treatment (Zhu and Pan, 2010; Barakat et al., 2014).

In recent years, environmentally friendly pre treatments such as milling or ultrasonic, plasma and wet explosions have been studied for biomasses such as woods, bagasse, rice and wheat straw (Kumar et al., 2009; Zhu and Pan, 2010; Adapa et al., 2011; Schultz Jensen et al., 2011; Sheikh et al., 2013). Currently, these processes are not cost effective, not only because of high invest ment costs but also because they can be very heavy on energy. For example, the total energy requirement of milling processes depends on the physicochemical properties of biomass and on the ratio of particle size distribution of materials before and after milling, this ratio being strongly dependent on the equipment or machine used. The environmental factor, energy consumption and energy efficiency are classically used to compare the perfor mances, efficiencies and environmental impacts of different pre treatment processes (Zhu and Pan, 2010; Barakat et al., 2014; Chuetor et al., 2015). However, survey articles concerning these three criteria for chemical, physicochemical and mechanical treat ment of lignocellulosic biomass remain scarce. Moreover, the rapidly increasing scientific literature in this domain would make such surveys quickly obsolete. To take advantage of this huge and potentially valuable source of information, innovative tools able to integrate continuously new information are required.

The main obstacle holding back the use of those scientific data is their textual format and heterogeneous structure. Our first aim in this paper is to show the relevance of semantic web based KE methods to structure the experimental information and express it in a standardized vocabulary. Such structuring can be done using an ontology (the semantic part of our model) to represent the experimental data of interest (see step 1 in Fig. 3). Ontologies are knowledge representation models that facilitate linkage of open data and offer automated reasoning tools. Once structured in ontologies, collected information and data are made homogeneous and can be processed to compute criteria allowing the comparisons of processes.

Our second aim in this paper is to demonstrate the feasibility of a pipeline (see Fig. 3), taking as inputs process data found in scien tific documents, and whose final output is a ranking of those pro cesses integrating data source reliability. Note that our system is partially inspired from previous semantic approaches used to facil itate "a priori" calculation of environmental indicators in industrial symbiosis (Trokanas et al., 2015; Raafat et al., 2013).

To illustrate our proposal, we present a first attempt to compare different pre treatment processes (Fig. 2) in terms of sugar yield after enzymatic hydrolysis and of environmental factor, by reusing data already published in the scientific literature. Energy efficiency is out of the scope of this paper as there is a lack of data about energy consumption in the current literature. The illustrating example concerns glucose extraction in rice straw and compares the four processes presented in Fig. 2. These processes may include a sequence of unit operations, as shown in Table 1.

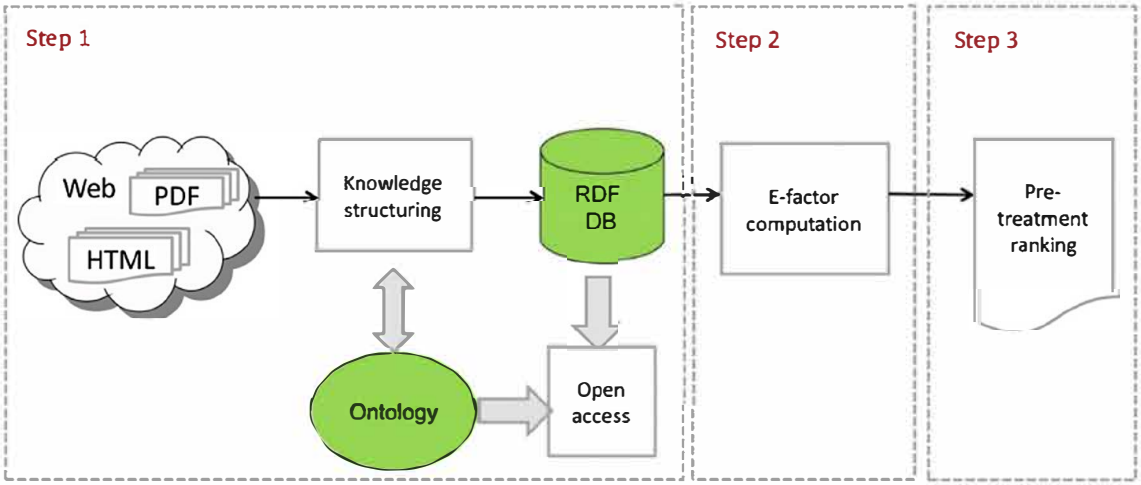

Fig. 3. Data treatment pipeline to compare biomass pre-treatment processes.

Table 1
Definition of process types in terms of unit operations.

Process type	Sequence of unit operations
PM	Milling
PM-UFM	Pre-milling + Ultrafine milling
PM-PC-PS-UFM	Pre-milling + Physicochemical treatment + Press and separation + Ultrafin milling
PM-PC-PS	Pre-milling + Physicochemical treatment + Press and separation

The scope of application, as well as its relevance within the field of Biorefineries, have been defined in close collaboration with the 3BCAR network. The 3BCAR French Carnot network (http://www.3bcar.fr) brings together researchers from seven laboratories in France (500 researchers) involved in the design of biomass transformation processes into bioenergy, bio based materials and molecules. In the framework of the IC2ACV project financed by the 3BCAR network, the need arose for a Decision Support System (DSS) able to help researchers involved in the Biorefinery design. The purpose of this DSS is to assist decision makers in making rational choices based on data and knowledge expressed by domain experts in the scientific literature. It is a real boon to collaborate with the 3BCAR network as researchers act as a stakeholder advisory board and help to specify the project scope, by defining the key parameters that must be in the generic tool. Moreover, a preliminary communication (Busset et al., 2015) indicates that this tool is of great interest for the international research community in the field of Biorefinery and industrial companies. The DSS developed in this context aims at achieving progress in the complex issue of assessing the environmental impact of alternative Biorefinery systems. The presented DSS has the following functionalities.

(i) It is able to annotate, store and maintain potentially incomplete or imprecise data extracted from the scientific literature, in dedicated databases containing biorefinery system characteristics (e.g. glucose yield, mass balance in terms of water, chemical reagents) and process parameters (e.g. milling rotation speed, treatment duration, temperature, pressure);
(ii) It computes environmental impact indicators (called Efactor in the following) integrating data reliability aspects;
(iii) It ranks the different Biorefinery processes according to their environmental impact.

The underlying pipeline, shown in Fig. 3, is based on an original coupling of KE methods, soft computing techniques and environmental factor computation to assess environmental sustainability of Biorefinery systems (see steps 2 and 3 of Fig. 3). To our knowledge, there is currently no data treatment pipeline similar to the one designed and implemented in this study. This pipeline takes as input a set of scientific papers extracted from bibliographical resources on the Web and generates as output a ranking of biorefinery processes based on environmental impact assessment.

The paper is organized as follows. Our approach is compared to the state of the art in Section 2. Functional specifications of the DSS are introduced in Section 3. The corresponding software architecture is detailed in Section 4. Implementation details of the graphical user interface in Java[1], the RDF[2] database management system with Jena[3], the environmental impact calculations, as well as the DSS assessment are presented in Section 5. Section 6 concludes the paper.

2. Comparison with the state of the art

Our DSS, which will be described in more details in Section 4, is a pipeline composed of three steps: (1) annotation guided by an ontology of experimental data published in scientific papers, (2) annotated data extraction and Efactor indicator computation, (3) process ranking including yield and Efactor indicator visualization in graphical maps. To our knowledge, no similar methodology exists in the literature, in particular for steps 2 and 3. We can however provide some elements of comparison for the first step of the pipeline: experimental data structuring using the semantic tool (including data annotation and querying guided by an ontology) proposed in this paper, called @Web (for Annotated Tables from the Web).

In general, relevant experimental data are scattered in different parts of the document and expressed in different formats. For example, in Biorefinery related papers, unit operation controlled parameters are often described in sentences within the Material and Method section, while experimental results are presented in tables located in the Results and discussion section. In this context, the automation of information extraction and annotation from text and tables is a major issue that should be discussed. Let us first

[1] www.java.com.
[2] Resource Description Framework is a graph model dedicated to formal description of Web resources.
[3] https://jena.apache.org/.

discuss the automatic extraction in text of relevant and related pieces of information. The literature on this topic is twofold.

On one hand, a substantial amount of work on binary relation extraction has been done. The first approaches to discover relations between entities focused on a limited linguistic context and relied on discovering co occurrences and manually designed pattern matching (Huang et al., 2004). Rule based techniques defined as regular expressions over words or part of speech (POS) tags have been used to construct linguistic or syntactic patterns (Proux, 2000, Hao et al., 2005; Hawizy et al., 2011; Raja et al., 2013). However, manually defined rules require heavy human effort. Later on, machine learning based approaches, e.g. Support Vector Machines (SVMs) (Minard et al., 2011), were widely employed (Rosario, 2005, Miwa et al., 2009; Van Landeghem et al., 2009, Zhang, 2011) to solve classification tasks (Rosario and Hearst, 2004). Those methods have shown their usefulness but require a large amount of annotated data for training, which usually takes tremendous human efforts to achieve. Being based on numerical models, they may not be directly understandable by the final user.

On the other hand, the extraction task of n ary relations (i.e. relations having more than 2 arguments) is a more complex issue, though it is needed in our application context. Work was conducted dividing n ary relation extraction into three main steps. The first step consists in identifying entities (or arguments) using resources such as ontologies or dictionaries. The second one consists in identifying the trigger word of the relation using dictionary based methods or rule based approaches to construct patterns from dependency parse results (Le Minh et al., 2011), or using machine learning methods (Bjorne et al., 2009, Buyko et al., 2009, Bui et al., 2011, Zhou et al., 2014) in order to predict the trigger word of the relation. Finally in the third step, binary relations are constructed using the trigger word and machine learning methods are used to classify whether or not binary relations belong to the searched n ary relation, but with a substantial loss of accuracy. Relation extraction methods are in general based on those three independent steps. In our context, arguments of the n ary relation can be implicitly expressed in the text and usually appear in several sentences. Therefore, state of the art methods, which make the hypothesis of the presence of a trigger word, are not directly usable.

Let us now discuss automatic extraction of relevant information in tables. State of the art methods and tools (Knoblock et al., 2012; Buche et al., 2013; Tian et al., 2013; Zhang, 2014) make the assumption that data tables are organized in the same way than in relational databases: a data table is composed of vertical columns (each column corresponding to a single feature, for example Biomass, temperature, etc.), themselves composed of cells. Unfortunately, this assumption is not always valid for data tables published in scientific articles. Various features may be present in the same column (temperature and treatment duration may be given in the header of a column corresponding to the process yield) or tables may have two entries (vertical and horizontal). Therefore robust automatic data table pre treatment must be designed and validated in order to apply state of the art tools to data tables extracted from scientific documents.

Automatic extraction of relevant information from text and tables of scientific articles being an active research topic, it is not ready yet for use in an operational system. Hence in our pipeline (see Fig. 3) annotation is performed manually, the ontology being used to guide the annotator. In this way, we may consider the annotation process as semi automated.

The only tool comparable with @Web to implement the first step of the DSS is, to the best of our knowledge, Rosanne (Rijgersberg et al., 2011), an Excel "add in" application built upon the OM ontology, an ontology of quantities and units of measure. Rosanne allows quantities and units of measures associated with columns of an Excel table to be annotated using concepts from OM. As @Web does, Rosanne manages the notion of phenomenon, very similar to the notion of symbolic concept in @Web, which represents non numerical data, for example process name or type of material. The main difference is that @Web defines the notion of relation, which links together data (studied object with controlled parameters and results) in order to represent a whole experiment. This notion is important in the DSS, being used to extract annotated data in order to compute Efactor indicators. There is no such notion, nor an equivalent one, available in Rosanne. Authors of Rosanne made the choice to develop an Excel "plug in" that brings semantics to data under Excel, a widely used tool in the scientific community. Being a Web application, @Web was naturally designed as a collaborative platform to share documents (for example scientific articles) associated with annotated tables. Moreover, @Web proposes an end user graphical interface to query annotated tables (see Section 4.1) which is not available in the current version of Rosanne. For its part, Rosanne proposes an interesting functionality to merge several annotated tables sharing a column annotated with the same concept. In conclusion, @Web and Rosanne tools are complementary and are based on a partly common ontological representation, the quantity units component of @Web being very close to the one used by OM. Some perspectives are given in the conclusion regarding that complementarity.

3. Functional specifications of the system

Since the DSS functional specifications depend on the users, the first step was therefore to identify the potential users. They were identified in the 3BCAR network of researchers. Then, the functional specifications were determined during the IC2ACV project's meetings, gathering several experts of the 3BCAR network. Finally, the functional specifications were refined during annual meetings of 3BCAR about IC2ACV results.

The following functional specifications are implemented in the IC2ACV DSS:

1. gathering, integrating and structuring heterogeneous data available in the scientific literature about biomass transformation processes;
2. defining a simple and generic ontological model of the information which must be identified and annotated in the scientific papers. The model must be simple because it should be easily updated by biorefinery experts (who are not computer scientists) and generic in order to be useful for other kinds of data managed by 3BCAR researchers (for instance packaging characteristics);
3. allowing an open data access to original data and associated units of measure. This can be done thanks to permalinks[4] and dedicated querying system managing unit conversion to facilitate data reusability;
4. assessing the reliability of data (sources) and taking it into account in the environmental impact assessment;
5. managing imprecise data since experimental data associated with biomass (eg. glucose rate) and biomass process (eg. glucose yield) are subject to uncertainty;
6. taking into account the biological variability associated with biomass processes and the subsequent uncertainty propagation during the environmental impact indicator computation;
7. computing environmental factor indicators (mass balance indicators called Efactors);
8. visualizing the ranking of biomass processes according to process yield and Efactors.

[4] Permanent links are URLs designed to reference a piece of information in a permanent way or for a given period of time.

4. Architecture of the decision support system

This section details the three steps of the data treatment pipe line. In the first step, experimental data published in scientific papers are annotated thanks to an ontology implemented in OWL[5] 2/DL and assessed in terms of their source reliability. Annotated data are stored in a RDF database and available in open access via permalinks, a SPARQL[6] end point and a dedicated query ing system guided by the ontology. The second step consists in extracting annotated data from the RDF database to compute Efactor indicators and data reliability scores. This extraction is done using SPARQL queries generated by the dedicated querying system guided by the ontology. Process yields, Efactor indicators and data reliability scores can be visualized in the third step as graphical maps. This last step provides a synthetic and global overview of biomass pre treatment process ranking.

4.1. Heterogeneous experimental data integration (step 1)

To facilitate integration of scientific data coming from heteroge neous sources, one of the relevant solutions is to use ontologies (Noy, 2004; Doan et al., 2012). An ontology defines a set of primi tives to model a domain of interest: classes, attributes (or proper ties) and relations between members of the classes (Guarino et al., 2009). The ontology is used to create and/or reuse standardized vocabularies and to index data sources with those vocabularies in order to allow data source interoperability. Our system uses @Web to capitalize experimental data extracted from scientific documents found on the Web. Here are its main components. @Web implements a complete workflow (see Fig. 4) to manage experimental data: extraction and semantic annotation of data from scientific documents, data source reliability assessment and uniform querying of the collected data stored in a database opened on the Web. @Web relies on an Ontological and Terminological Resource (OTR) which guides scientific data semantic annotation and querying. An OTR associates a terminological component to an ontology in order to establish a clear distinction between the linguistic expressions in different languages (i.e. the term) and the notion it denotes (i.e. the concept) (Roche et al., 2009; McCrae et al., 2011; Cimiano et al., 2011). For instance, English terms "*Grasses and energetic plants*" and "*Energy crops*" and the French term "Plantes énergétiques" denote the same symbolic con cept *Grasses and energetic plants*. The OTR is designed to model scientific experiments. It is composed of two layers: a generic one and a specific one dedicated to a given application domain. Since the OTR is at the heart of the scientific data capitalization workflow, @Web can be reused for different application domains: only the specific part of the OTR must be redefined to re use @Web for a new domain (see Touhami et al., 2011 for a reuse in food packaging). Let us point out that the OTR satisfies functional specifications 2 of the IC2ACV DSS presented in Section 2.

@Web is composed of two sub systems (see Fig. 4). The first one is an annotation sub system for the acquisition and annotation, with concepts of the OTR, of experimental data extracted from sci entific documents; those annotated data are being stored into a database. This sub system also allows the reliability of data sources to be assessed using the approach of (Destercke et al., 2013). The second sub system is a flexible querying system based on the approach presented in (Destercke et al., 2011). @Web is implemented using the semantic web standards (XML[7], RDF, OWL, SPARQL): the OTR is defined in OWL2 DL, annotated tables in XML/RDF and the querying in SPARQL. We present in Section 4.1.1 the Biorefinery OTR used in @Web. Section 4.1.2 details the model used to assess data source reliability.

4.1.1. Biorefinery OTR model

The OTR is designed to represent scientific experiments in order to annotate data tables in a given domain (see Touhami et al., 2011 for more details). We made the choice to represent an experiment by using n ary relations between several experimental parameters and a given result. This structures information in a simple way as requested by functional specification 2 (see Section 3). As recom mended by W3C (Noy et al., 2006), we used the design pattern which represents a n ary relation thanks to a concept associated with its arguments via properties. Let us illustrate this notion by using the example of n ary relation *Biomass Glucose Composi tion Relation* (see Fig. 5). This relation is characterized by 4 argu ments: (1) the glucose rate, which is the experimental result, (2) the biomass, which is the studied object, and associated experi mental parameters being (3) the biomass state (untreated or trea ted) and (4) the experiment number reported in the document. This relation is used to create annotated tables, as shown in Table 2. It presents an example of annotated table extracted from the scien tific document (Hideno et al., 2009), which determines the glucose rate of rice straw, in two different experiments. The columns of the annotated table correspond to the arguments of the relation *Bio mass Glucose Composition Relation*.

An excerpt of Biorefinery OTR global structure is presented in Fig. 6. The conceptual component of Biorefinery OTR is composed of a core ontology to represent n ary relations between experimen tal data and a domain ontology to represent specific concepts of a given application domain. In the Up core ontology, generic con cepts Relation and Argument represent respectively n ary relations and arguments. The representation of n ary relations between experimental data requires a particular focus on the management of quantities and their associated units of measure. In the Down core ontology, generic concepts Dimension, UM_Concept, Unit_Concept and Quantity allow the management of quantities and their associated units of measure. The sub concepts of the gen eric concept Symbolic_Concept represent the non numerical argu ments of n ary relations between experimental data. The domain ontology contains specific concepts of a given application domain, in this paper the Biorefinery domain. They appear as sub concepts of the generic concepts of the core ontology. In the Biorefinery OTR, relations represent either experiments which characterize biomass (see Fig. 5) or experiments involving unit operations performed on biomass. For instance, the milling operation is represented by the n ary relation *Milling Solid Quantity Output Relation* (see Fig. 7).

It is characterized by 7 arguments and represents the milling solid quantity output, which is the milling experimental result for a given biomass associated with a set of experimental parame ters being the biomass input quantity, the total pre treatment energy used for the milling, the treatment duration, the milling rotation speed and the type of milling.

In the Biorefinery OTR, all concepts are represented as OWL classes, hierarchically organized by the subsumption relation sub ClassOf and pairwise disjoints.

The terminological component of the Biorefinery OTR contains the domain related set of terms used to annotate data tables. Sub concepts of the generic concepts Relation, Symbolic_Concept and Quantity, as well as instances of the generic concept Unit_Concept, are all denoted by at least one term of the terminological compo nent. Each of these sub concepts or instances are, in a given lan guage, denoted by a preferred label and optionally by a set of alternative labels, which correspond to synonyms or abbreviations.

[5] Web Ontology Language is a knowledge representation model built upon RDF.
[6] SPARQL (SPARQL Protocol and RDF Query Language) is the protocol and the query language which permits to search, add, edit or delete RDF graphs.
[7] Extensible Markup Language is a markup language.

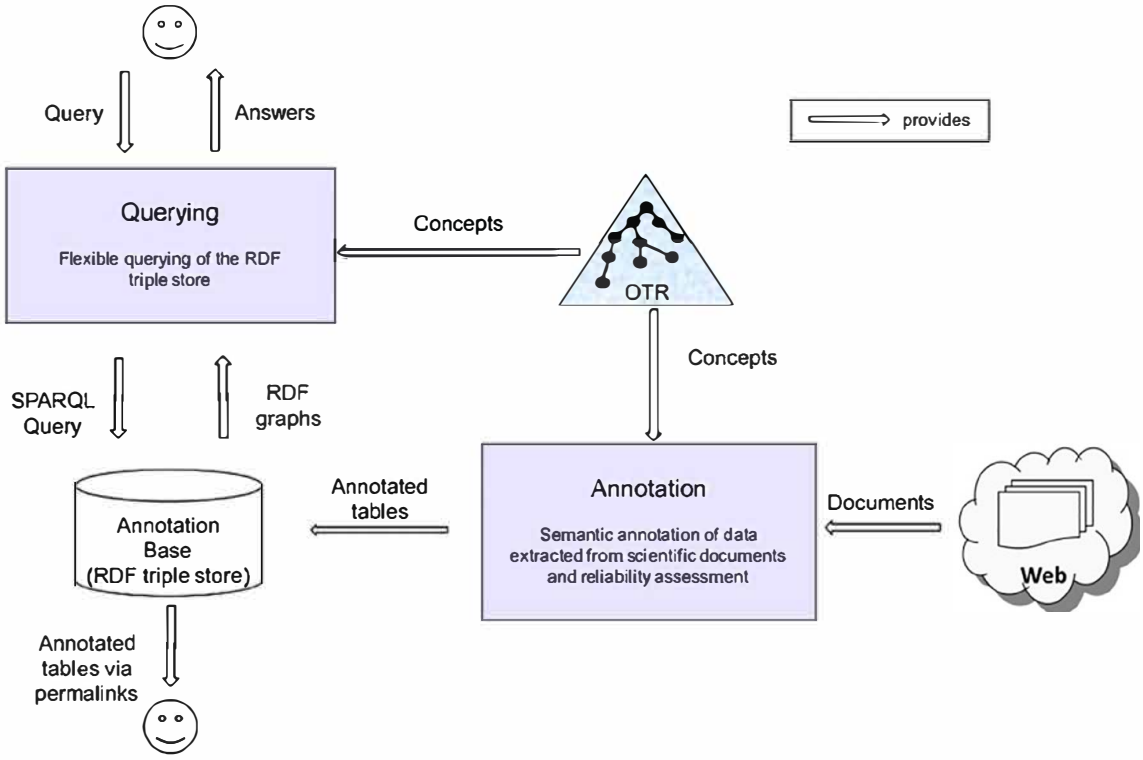

Fig. 4. Knowledge annotation and querying in @Web.

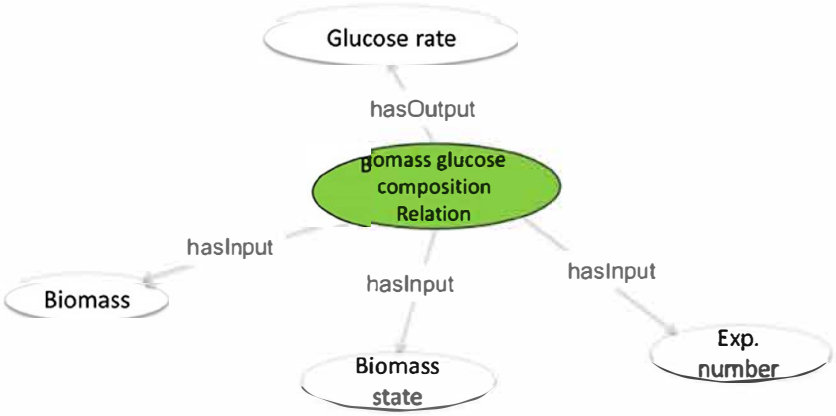

Fig. 5. A relation concept used to characterize a given biomass.

Table 2
Excerpt of the annotated table biomass composition.

No	Biomass	Biomass state	Experience number Unit: 1	Glucose rate Unit: %
1	Rice straw	Untreated biomass	1.000e+00	3.700e+01
2	Rice straw	Untreated biomass	2.000e+00	3.700e+01

Labels are associated with a concept or an instance thanks to SKOS[8] labelling properties, recommended by W3C to represent controlled vocabularies associated with concepts (see the "*Grasses and Energetic plants*" example given in the introduction of Section 4.1).

[8] Simple Knowledge Organization System.

4.1.2. Reliability assessment scores for Biorefinery

When gathering data from various documents, the question rapidly arises as to how reliable these data or these documents are. @Web proposes a reliability estimation tool, presented in details in (Destercke et al., 2013) and whose basis we recall here. This tool aims at providing an automatic, *a priori* (that is, avoiding a specific examination) estimation of the document and data reliability from a set of meta information related to the data and the document.

To this effect, S groups $A_1, \ldots, A_s$ of *meta information* important to assess reliability are first defined in accordance with decision makers and domain experts, a group A_i taking C_i values $a_{i1}, \ldots, a_{iC_i}$. For instance, the C_i values for the group "Sugar analysis method" would be the different available methods. Various types of meta information, summarized in Table 3, have been considered for the Biorefinery data sources:

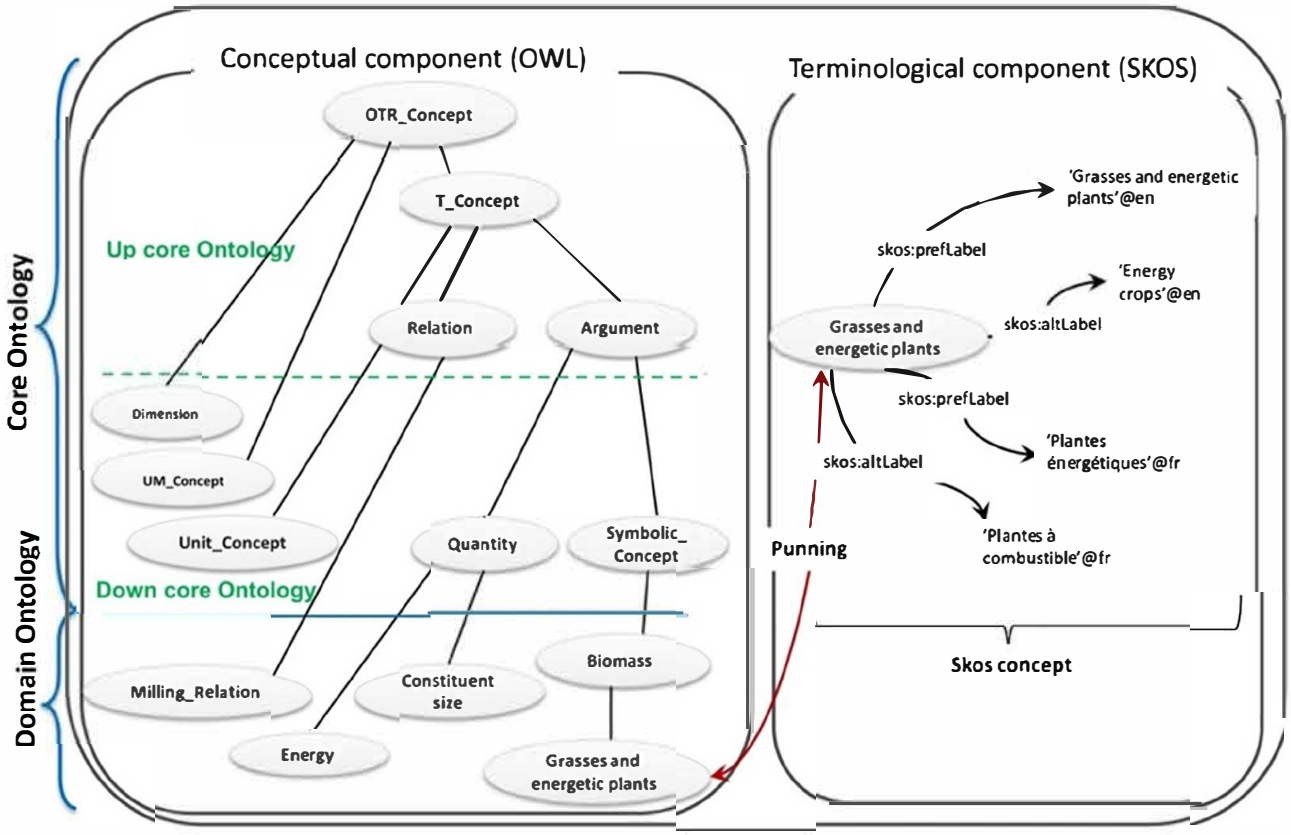

Fig. 6. Onto-terminological resource specialized for biorefinery.

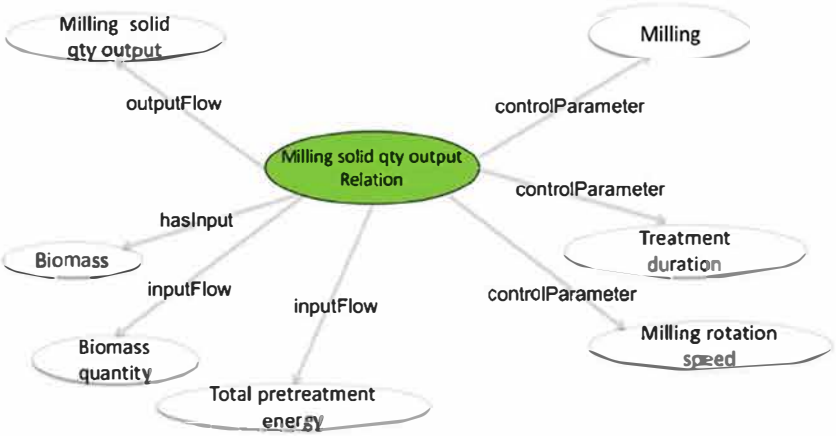

Fig. 7. A relation concept to model the milling unit operation.

Table 3
Metadata considered in the reliability assessment for Biorefinery.

Source	Production	Statistics
Type of source	Sugar analysis method	Energy measure repetitions
Citation count		Enzymatic hydrolysis repetitions
Publication date		Biochemical and physicochemical treatment repetitions

- meta information on the data source itself, for instance the source type (e.g. scientific publication, technical report), the source reputation, citation data;
- meta information related to means used to produce data. In papers based on experiments in Life Science, such information is typically included in a section called *Material and method*, which thoroughly describes the experimental protocol and material. Some methods may be known to be less accurate than others, but are still chosen for practical considerations;
- meta information related to statistical procedures: presence of repetitions, uncertainty quantification (i.e. variance, confidence interval), elaboration of an experimental design.

In practice, the groups are made so that their impact on reliability can be estimated independently, while a group A_i may contain multiple criteria (e.g. number of citations and publication date).

For each possible value of each group, the method then consists in assessing the reliability of a document or data having this particular value. After the groups have been formed, for each value $a_{ij}, i \quad 1,\ldots,S.j \quad 1,\ldots,C_i$, an expert of the field from which data are collected gives his/her opinion on about how reliable are the data whose meta information is a_{ij}. This opinion is expressed linguistically, chosen from a set of limited modalities (or combinations of them), e.g. *very unreliable, slightly unreliable, neutral, slightly reliable, very reliable* and *unknown*. The number of modalities is limited (usually 5 or 7), accounting for known limitations of human cognitive abilities (Miller, 1956). In practice, several experts of the field belonging to the 3BCAR network, specialists of different unit operations (respectively mechanical, physico chemical and biological) have been interviewed resulting in a consensual opinion.

To each document o, are then associated S linguistic assessments (according to the value taken by the corresponding meta information). A missing value in the meta information is simply treated as the linguistic assessment *unknown* in terms of reliability. In order to apply refined fusion techniques able to deal with potentially conflicting information (as some meta information may indicate an unreliable document, while others may designate a rather reliable one), the linguistic assessments are translated into a numerical format, using the notion of fuzzy sets, that are adequate numerical models of linguistic values. In order to have enough flexibility, these fuzzy sets are defined on an ordered finite reliability space $\Theta \quad \{\theta_1,\ldots,\theta_5\}$ of 5 elements, θ_1 being the lowest reliability value, θ_5 the highest. The number of elements could be higher than 5, but this is a reasonable choice. Indeed that number should remain odd in order to have a neutral element, not too low so that fuzzy sets corresponding to different terms can be numerically quite distinguishable and not too high so that computational problems do not arise. Each modality is then transformed into a fuzzy set on Θ (see Fig. 8 for an illustration).

The S fuzzy sets $\mu_{a_1^o},\ldots,\mu_{a_S^o}$ corresponding to document o's group reliability are then merged together using evidential theory and a maximal coherent subset approach which allows conflicting evidences to be taken into account. Such an approach aims at reconciling all sources while keeping as much information as possible, meaning that missing information (i.e. presence of the *unknown* modality) does not impact the result. The potential conflict in meta information (i.e. assessment of high reliability for one aspect but of low reliability for another one) is reflected in the imprecision of the final model: presence of conflict will end up in a quite imprecise estimation of the reliability, while its absence will result in a quite precise estimation. The result of this merging is a mass distribution $m_o: 2^\Theta \rightarrow [0,1]$ which reflects the global reliability of o (see Destercke et al., 2013) for more details).

The mass m_o provides an accurate synthesis of the different meta information contributions, and its analysis could allow one to automatically identify subsets of conflicting and of coherent meta information. Yet, it is still too complex to be analysed at a glance. For this reason, a further summarizing is provided, in which

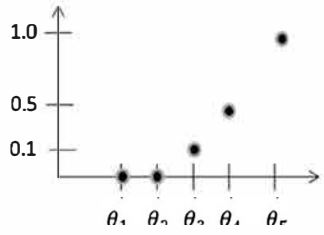

Fig. 8. Fuzzy set corresponding to the term *very reliable* defined on Θ with 5 elements.

reliability estimation in the form of an interval $[\underline{\mathbb{E}_o}, \overline{\mathbb{E}_o}]$ is proposed. The final score $\underline{\mathbb{E}_o}$ is computed using the following formula:

$$\underline{\mathbb{E}_o} \quad \sum_{E \subset \Theta} m_o(E) inf_{\theta_i \in E} f(\theta_i)$$

where $f(\theta_i) \quad i$, that is each θ_i is replaced by its corresponding rank (a natural choice, even if other functions could be chosen). $\overline{\mathbb{E}_o}$ is obtained with the same formula, replacing *inf* by *sup*. The length or imprecision of $[\underline{\mathbb{E}_o}, \overline{\mathbb{E}_o}]$ reflects to which extent the various pieces of meta information are consistent: $\underline{\mathbb{E}_o}$ provides in some way a "worst" possible reliability, while $\overline{\mathbb{E}_o}$ correspond to a "best" possible reliability, the two being close to each other when information is consistent. These scores can then be used in the querying system to rank annotated data associated with documents according to their reliability, where one can adopt different strategies: e.g., optimistic (pessimistic) by ranking according to $\overline{\mathbb{E}_o}$ ($\underline{\mathbb{E}_o}$).

These values will be also used in order to compute a global reliability score R within a collection of documents which includes several experiments of the same biorefinery process (see Section 4.2). For a given set of documents $o_1,\ldots,o_n$, the global reliability score R is a value in the interval $[0,1]$ with $R = 0$ for a set of unreliable documents and $R = 1$ for a set of very reliable documents. As reliability scores $[\underline{\mathbb{E}_{o_i}}, \overline{\mathbb{E}_{o_i}}]$ associated with documents o_i are imprecise, we compute standardized bounds $[\underline{R}, \overline{R}]$, given $\mathbb{E}_{min}$ and $\mathbb{E}_{max}$ the minimal and maximal reliability scores ($\mathbb{E}_{min} \quad 1, \mathbb{E}_{max} \quad 5$ in this paper), as follows:

$$\begin{cases} \underline{R} & \frac{\sum_{i=1}^{n} (\underline{\mathbb{E}_{o_i}} - \mathbb{E}_{min})}{n(\mathbb{E}_{max} - \mathbb{E}_{min})} \\ \overline{R} & \frac{\sum_{i=1}^{n} (\overline{\mathbb{E}_{o_i}} - \mathbb{E}_{min})}{n(\mathbb{E}_{max} - \mathbb{E}_{min})} \end{cases}$$

More details are presented in (Destercke et al., 2013); in particular various means to analyse the reliability results, such as the benefits one can retrieve from imprecise assessments or the way to detect subgroups of agreeing/disagreeing meta information.

4.2. Eco design indicators (steps 2 and 3)

As seen in the previous section, the first step of the pipeline presented in Fig. 3 allows knowledge extracted from heterogeneous data sources to be annotated and its reliability to be assessed. The second step consists in extracting annotated data from the RDF database to compute environmental impact indicators. Following functional specification 7 listed in Section 2, we now present the computation of mass balance indicators denoted Efactors. The Efactor indicator can be seen as the total input quantity of matter not valorised into glucose but required to produce 1 kg of glucose. This indicator is often used in survey papers dedicated to biorefinery processes comparison (Zhu and Pan, 2010; Barakat et al., 2014; Chuetor et al., 2015). All kinds of matters which are inputs of the process are taken into account (by example, the biomass, water, chemical reagents ...). For a given set of n documents $o_1,\ldots,o_n$, we consider for each document o_i the m experimental settings which are described in o_i denoted $e_{i1},\ldots,e_{im}$. In the following, we call experimental settings, the set of controlled parameter adjustments for the given process (by example in a milling, different durations are tested resulting in different experimental settings). Each experimental setting is associated with a given biomass, denoted biomass(e_{ij}), which belongs to the set of l studied biomasses $b_1,\ldots,b_l$. This biomass(e_{ij}) has been assigned (during the first step) to a given Biorefinery process, denoted process(e_{ij}), which belongs to the set of k alternative processes $p_1,\ldots,p_k$, (see Fig. 2), which will be compared in Section 5.2. Following the functional specification 7 expressed by 3BCAR

researchers, a matter balance indicator, denoted $Efactor(o_i, p, b)$ can be computed for experimental setting $\{e_{ij}\}$ belonging to a given document o_i.

Remark. As biomass quantities differ in the considered experiments, all values are normalized for 1 kg of initial biomass in order to compute comparable *Efactor* indicators. Efactor is defined as in (Chuetor et al., 2015):

Efactor definition.

$$Efactor = \frac{BiomassQty + ChemicalReagentQty + SolventQyt \quad GlucoseReleasedQty}{GlucoseReleasedQty} \tag{1}$$

GlucoseReleasedQty definition.

$$GlucoseReleasedQty \quad BiomassQty * GlucoseRate * GlucoseYield \tag{2}$$

where

- *BiomassQty* is the initial biomass quantity (kg).
- *ChemicalReagentQty* is the chemical reagent product quantity used in the process (kg).
- *SolventQty* is the quantity of solvent (water and/or solution) used in the process (kg).
- *GlucoseReleasedQty* (kg) is a quantity defined as the biomass quantity (input of the enzymatic hydrolysis unit operation) multiplied by the glucose rate (available in the raw biomass) and the glucose yield which depends on the considered experimental setting.

Functional specifications 5 and 6 expressed by 3BCAR researchers consist in (i) taking into account the uncertainty recorded in experimental data results and (ii) propagating uncertainties in the Efactor computation. The experimental results considered in this study are *GlucoseRate* and *GlucoseYield*. For each experiment, the available results may be given as a scalar value, as an interval, or as a tuple of the mean value and the standard deviation over the experimental repetitions. Consequently, experimental results *GlucoseRate* (resp. *GlucoseYield*) can be considered as a sample drawn from a random variable. We have noticed that, in all documents, the *GlucoseYield* random variable depends on experimental settings, which is not the case for the *GlucoseRate* random variable whose sampling shows no variation.

In the following, we propose for a given document o_i, a given biomass $b \in \{b_1, \ldots, b_l\}$ and a given process $p \in \{p_1, \ldots, p_k\}$, two ways to compute $Efactor(o_i, p, b)$. The first one consists in selecting the best experimental setting presented in document o_i and computing $Efactor^{best}(o_i, p, b)$. As the uncertainty level is not always available in experimental data results, we propose a second way to compute *Efactor* which consists in taking into account the information provided by the entire set of settings and computing $Efactor^{all}(o_i, p, b)$. It is an indirect way to provide information about the uncertainty associated with experimental data results of a document o_i. We also define, for a given biomass b and a given process p, an *Efactor* indicator calculated for the entire set of settings of the entire set of n documents $o_1, \ldots, o_n$.

Computing Efactor for the best experimental setting in document o_i: Having in mind the imprecision expressed for random variable *GlucoseYield*, a pessimistic point of view will prefer to guarantee the highest minimal *GlucoseYield*, while an optimistic one will prefer to guarantee the highest maximal *GlucoseYield*. In this paper, we have chosen the pessimistic point of view to select the best experimental setting. Let us consider $\overline{GY}_{e_{ij}}$ (resp. $\sigma_{GY_{e_{ij}}}$) the mean value (resp. the standard deviation) associated with the *GlucoseYield* random variable of experimental setting j described in document o_i.

We assume that the sample is drawn from a normal distribution, the sample size being unknown (this is a reasonable assumption in such experiments). We recall that under this assumption, the 95% confidence interval of the *GlucoseYield* random variable Is defined by $[\overline{GY}_{e_{ij}} \quad 2\sigma_{GY_{e_{ij}}}, \overline{GY}_{e_{ij}} + 2\sigma_{GY_{e_{ij}}}]$. We consider for each document o_i the m experimental settings which are described in o_i denoted $e_{i1}, \ldots, e_{im}$. Then, the best experimental setting with a confidence degree of 95%, denoted e_{ij^*}, is the one having the maximal lower bound of a 95% confidence interval:

Best experimental setting definition.

$$\left(\exists e_{ij^*} \in (e_{i1}, \ldots, e_{im}) \middle| \overline{GY}_{e_{ij^*}} \quad 2\sigma_{GY_{e_{ij^*}}} \quad \max_{j \in (1,\ldots,m)} (\overline{GY}_{e_{ij}} \quad 2\sigma_{GY_{e_{ij}}})\right) \tag{3}$$

For the four experimental settings described in (Amiri et al., 2014), the results are presented in Fig. 9 for the following rice straw pre treatment process type "Pre Milling then Physicochemical treatment then Press and Separation" (called PM PC PS). The best experimental setting corresponds to the pessimistic choice discussed above, i.e. the one having the maximal lower bound of the 95% confidence interval associated with the *GlucoseYield* random variable ([33.07 36.47]). Let us consider that *BiomassQty* = 1 kg, *SolventQty* = 8 kg, *ChemicalReagentQty* = 0.0005 kg. With the 95% confidence interval associated with the *GlucoseRate* random variable = [0.51995,0.57335] in (Amiri et al., 2014), we compute, following Eq. (1), $Efactor^{best}(o_i, p, b)$ = [42.04,51.34].

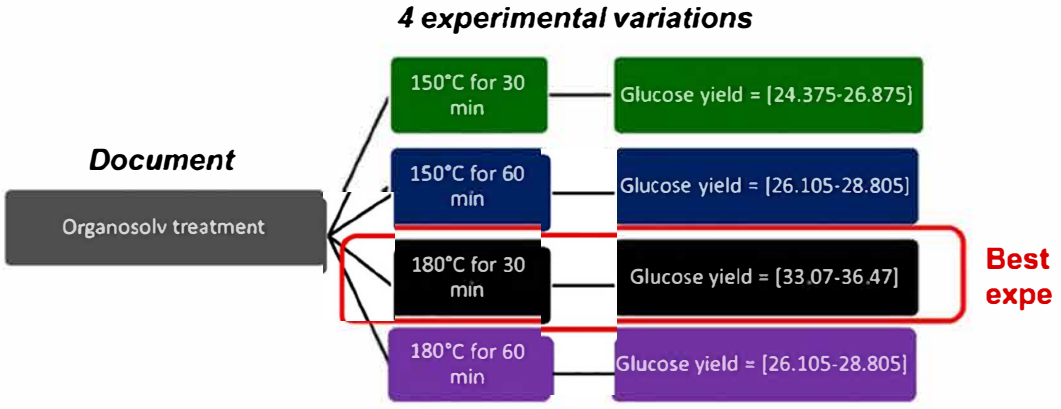

Fig. 9. 95% confidence intervals associated with the *GlucoseYield* random variable for experimental settings presented in Amiri et al. (2014).

Table 4
Seven examples of Efactor^{all} indicators for processes of type PM-PC-PS.

Biblio ref	Physicochemical pre-treatment	$Efactor^{all}_{min}$	$Efactor^{all}_{max}$
Hideno et al. (2012)	Hot compressed water	48.66	61.96
Amiri et al. (2014)	Organosolv	42.04	70.02
Sheikh et al. (2013)	Torrefaction	3.97	4.78
Hideno et al. (2009)	Hot water	31.21	45.67
Ilgook et al. (2014)	Nitric acid pre-treatment	22.76	24.01
Poornejad et al. (2013)	Oxidizing treatment	44.90	59.02
Poornejad et al. (2013)	Ionic liquid treatment	43.74	45.38

Computing Efactor for all of the settings in document o_i: In this case, we want to take into account *Glucose Yield* values obtained for all of the settings. As settings are inter dependent, we define the global *Glucose Yield* as the interval including the 95% confi dence intervals associated with the *GlucoseYield* values obtained in all of the settings. For instance, in Fig. 9, the global *Glucose Yield* = [24.375,36.47]. Following Eq. (1), $Efactor^{all}(o_i, p, \mathrm{b})$ = [42.04,70.01]. Unsurprisingly, $Efactor^{best}(o_i, p, \mathrm{b}) \subseteq Efactor^{all}(o_i, p, \mathrm{b})$ as $Efactor^{all}(o_i, p, \mathrm{b})$ takes into account all the experimental settings of the document o_i.

Computing Efactor for the entire set of settings of the entire set of documents: An aggregated mass balance indicator for a set of n doc uments $o_1, \ldots, o_n$ associated with a given biomass $b \in \{b_1, \ldots, b_l\}$ and a given process $p \in \{p_1, \ldots, p_k\}$ must also be computed, denoted $Efactor(p, \mathrm{b})$. In this case, we consider that experimental settings described in different documents are independent. Conse quently, we define $Efactor(p, \mathrm{b})$ as the interval $[Efactor_{min}(p, \mathrm{b}), Efactor_{max}(p, \mathrm{b})]$ where:

$$Efactor_{min}(p, \mathrm{b}) \quad \frac{\sum_{i\ 1}^{n} Efactor^{all}_{min}(o_i, p, \mathrm{b})}{n}$$

$$Efactor_{max}(p, \mathrm{b}) \quad \frac{\sum_{i\ 1}^{n} Efactor^{all}_{max}(o_i, p, \mathrm{b})}{n}$$

Table 4 presents seven examples of $Efactor^{all}$ indicators, com puted for a set of 6 documents including 23 settings of rice straw pre treatment processes of type PM PC PS, the type of the specific physicochemical pre treatment being given in the table. Based on this table, $Efactor(p, \mathrm{b})$ = [33.90,44.41] for p = PM PC PS and b = rice straw.

5. Implementation

In this section, we detail the implementation of the data treat ment pipeline presented in Fig. 3. In Section 5.1, we describe the @Web software. Section 5.2 deals with the implementation of Efac tor indicator computation and visualization.

5.1. @Web

We have presented in Section 4.1.1 the Biorefinery OTR which has been defined to model experimental data in the domain of biorefinery pre treatment processes. Biorefinery OTR is used in @Web for the task of experimental data source annotation and querying, using n ary relation concepts. @Web relies on the gen eric part of the OTR model (see the core ontology in Fig. 7) and allows the management of the domain ontology of Biorefinery OTR with its associated terminology. As @Web relies on the generic part of the OTR model, several OTR dedicated to different applica tion domains can be managed simultaneously in @Web. For instance, in our current implementation, an OTR dedicated to gas transfer in packaging materials has also been defined and is avail able at http://www6.inra.fr/cati icat atweb. Let us notice that the core ontology, which has been designed to be non modifiable, is not accessible to ontology managers. Moreover, we made the choice to manage units in a transversal way defining only one OTR of units of measure, because some units of measure may be used in different OTR. Units can therefore be used by all the OTR defined in @Web. The current version of the OTR of units of mea sure is available at http://www6.inra.fr/cati icat atweb (section @Web platform, thumbnail Ontology, option Unit Ontology).

Recorded tutorials of the current @Web version are available on line (http://www6.inra.fr/cati icat atweb/Tutorials). Here we focus our presentation on the annotation sub system of @Web, which implements the five sub steps presented in Fig. 10. The annotation sub system of @Web implements a complete workflow to extract experimental data from scientific documents and semantically annotate them with n ary relation concepts defined in the Biorefinery OTR.

In the first sub step, called *Document selection*, relevant docu ments according to the OTR are retrieved from the Web and man ually selected by a domain expert. This selection may be done using classical bibliographical tools (for instance Web of Science[9]). Documents can be uploaded in @Web from a desktop or from a col laborative repository management using Mendeley[10]. After docu ment loading, @Web manages their bibliographical references as well as their entire text both in HTML and PDF formats. Documents are grouped in topics. Four topics have been defined for biorefinery and correspond to the four pre treatment processes (see Fig. 2 and Table 1) which are compared in Section 5.2. For instance, (Amiri et al., 2014), whose experimental settings have been presented in Fig. 9, has been stored in PM PC PS topic which corresponds to the PM PC PS pre treatment process presented in Table 1.

The second sub step is dedicated to document reliability assess ment using the model presented in Section 4.1.2. In the current version, meta information associated with each document is man ually entered in order to compute reliability score. In Fig. 11, the reliability score reflects an imprecise assessment $[\underline{\mathbb{E}_o}, \overline{\mathbb{E}_o}]$ [1.5, 4.98], due to a conflict between expert opinions asso ciated with meta information:

- "citation age and citation number" and "source type" are con sidered as *very reliable*.
- "Enzymatic hydrolysis reproducibility" and "Biochemical and physicochemical analysis reproducibility" are considered as *hardly reliable* because only the average value of experimental results associated with those unit operations is given in the document.

All operations involving belief functions needed to compute reliability scores have been implemented in an R package. The package is called belief (Maillet et al., 2010), and it includes basic functions to manipulate belief functions and associated mass assignments (currently on finite spaces only).

In the third sub step shown in Fig. 10, called *Table extraction*, data tables are automatically extracted from HTML versions of doc uments using tag analysis. The discovered tables are then pre sented to the domain expert for validation as they represent a synthesis of some experimental data published in the document and may be used to facilitate the manual entering. The fourth sub step, called *Table annotation*, corresponds to the manual semantic annotation of the selected data tables using the concepts of the Biorefinery OTR. Taking into account the actual content of

[9] webofscience.com.
[10] http://www.mendeley.com/features/.

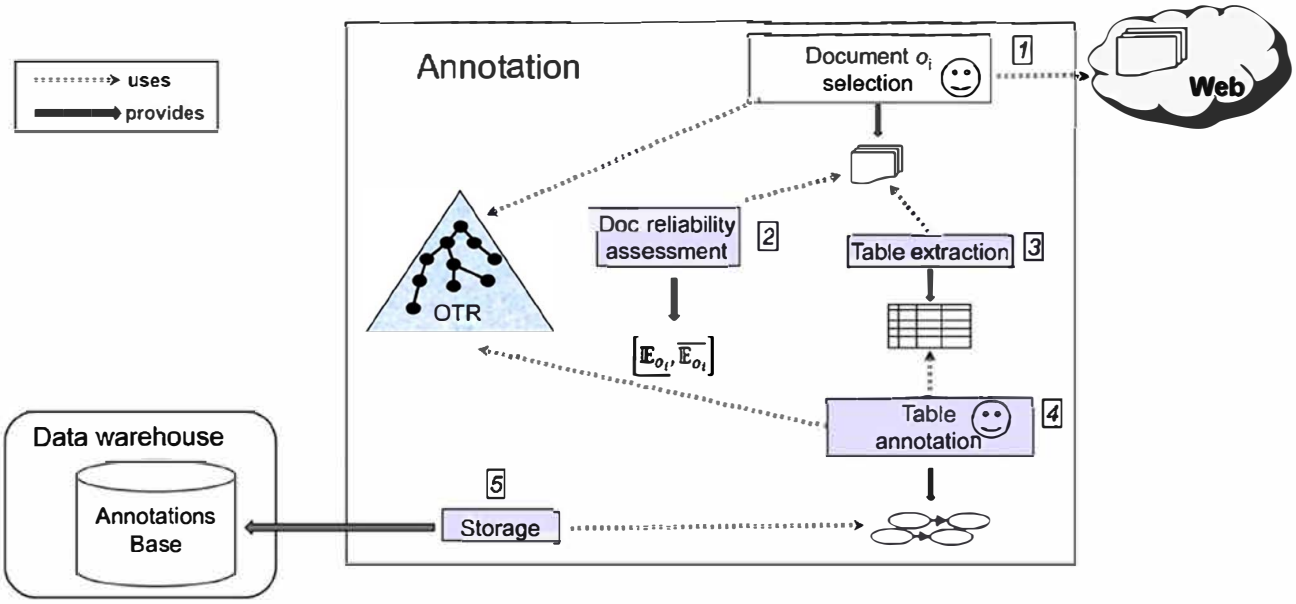

Fig. 10. The five sub-steps of the annotation sub-system in @Web.

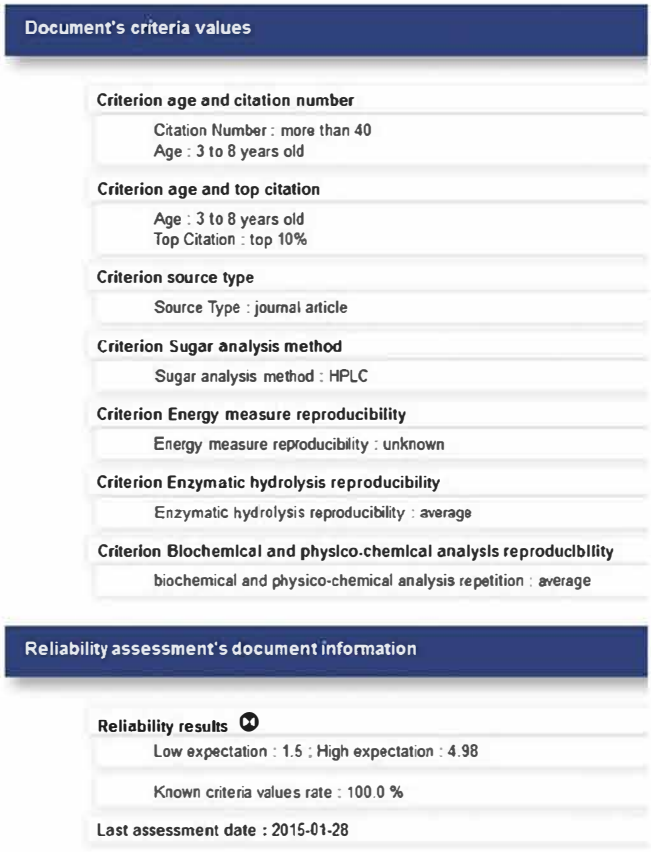

Fig. 11. @Web reliability assessment associated with Hideno et al. (2009).

the original table, the annotator selects from the n ary relation concepts defined in the Biorefinery OTR those relevant to annotate the table.

For instance, in Fig. 12, the expert selects several n ary relations including, for instance, *Enzymatic hydrolysis output solid constituent quantity relation* and *Milling Solid Quantity Output Relation* from the list of n ary relations concepts defined in the Biorefinery OTR. The signatures of both n ary relations concepts are visualized in a table, one signature per row. This will guide the expert in his/her entering task, allowing him/her not to forget to fulfil arguments of the selected n ary relation concepts. This is important for data reusability. This example shows that several relations may be used in a given annotated table in order to annotate experimental data associated with a complete pre treatment process as the one presented in Table 5.

This table presents an example of annotated table in @Web extracted from the scientific document (Hideno et al., 2009), which describes a biorefinery pre treatment process composed of a sequence of four unit operations occuring in experiments 1 and 2. The columns of the annotated table correspond to arguments of the relation *Milling Solid Quantity Output Relation* (see Fig. 6). For instance, we can see on the first row that the first unit operation is a cutting milling, instance of the relation *Milling Solid Quantity Output Relation*. The third row shows that the third unit operation of this process is another milling, dry ball milling, another instance of the relation *Milling Solid Quantity Output Relation*.

During the manual data entering guided by the OTR, @Web proposes assistance in several tasks. For instance, when entering quantity values and their associated units of measure, the expert may select a unit in the list of units associated with the quantity in the Biorefinery OTR. The expert may also drag and drop the quantitative values from the original table to the annotated table, which makes the data entry easier and reduces the risk of errors. As requested in functional specification 5 listed in Section 3, it is possible to enter a quantitative value as an interval or a mean/standard deviation pair. For instance, in Table 5, the quantity *Output solid constituent quantity* is defined as the precise value 5 g for *Cutting milling treatment* in row no. 1 and as the interval]3.1e 2, 4.9e 2] g. for *Enzymatic hydrolysis treatment* in row no. 4. Missing data are denoted by the interval [inf; inf]. @Web may also assist the expert in the symbolic concepts entry task, by allowing him/her to navigate in the hierarchy of symbolic concepts belonging to the Biorefinery OTR. For instance, in Fig. 13, the expert may navigate into the Biomass concept hierarchy from the Biorefinery OTR and see the labels of the selected concept in the upper right corner of the snapshot.

In the fifth and last step of annotation, called *Storage*, the annotated data tables are stored in a RDF triple store which can be quer

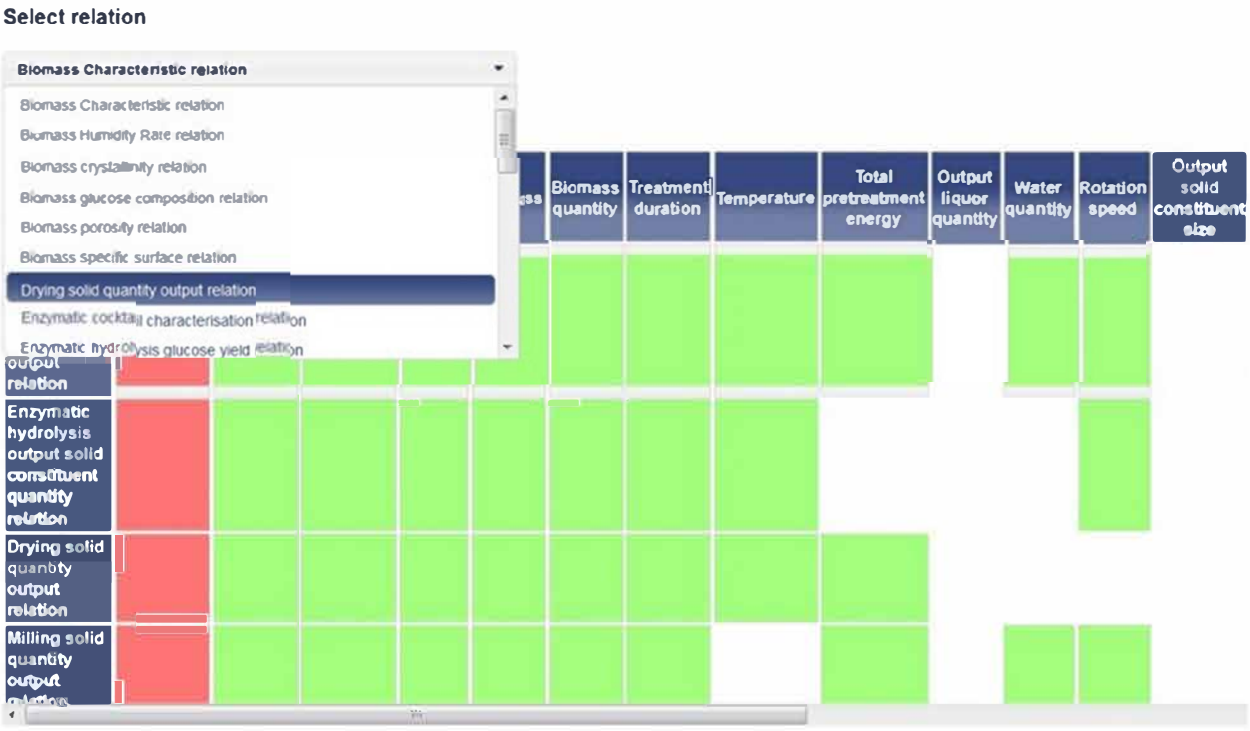

Fig. 12. Selection of several relevant n-ary relation concepts from the Biorefinery OTR in order to annotate experimental data associated with a complete pre-treatment process.

ied through either an end user querying interface or a SPARQL end point for open data access.

The data annotated by the annotation sub system of @Web may be queried through an end user interface, which implements func tional specification 3 presented in Section 3. A detailed presenta tion of the flexible bipolar querying method which has been used to implement the querying sub system of @Web is given in (Destercke et al., 2011, 2013). It should be noticed that this query ing method simultaneously performs three kinds of reasoning: (1) inference using specialization relation defined in the Biorefinery OTR, (2) ranking according to fuzzy pattern matching between preferences expressed in the query and imprecise data, (3) ranking according to preferences expressed about data source reliability. In this paper, we present the implementation of the querying sub system through an example which illustrates the way data needed for Efactor indicator computation are extracted from the RDF triple store. For instance, the query presented in Fig. 14 has been built in order to compute the Efactor indicator associated with pre treatment processes of type Organosolv pre treatment for rice straw. First, the user selects the ontology IC2ACV which is the name associated with the Biorefinery OTR in @Web. Secondly, the user selects one of the concept relations defined in IC2ACV to build the query. In the example of Fig. 14, the *Physicochemical pre treatment solid quantity output relation* has been selected because its arguments allow the elaboration of the matter balance required to compute the Efactor. Selection criteria can be expressed on relation arguments. They may be mandatory or desir able. Mandatory means that only instances of relation *Physico chemical pre treatment solid quantity output relation* which fulfil the selection criterion will be retrieved. In the example of Fig. 14, only results associated with Rice straw will be retrieved. Desirable criteria allow the ranking of results to be refined.

In Fig. 14, instances of *Physicochemical pre treatment solid quan tity output relation* which correspond to Organosolv treatment will be ranked first. Fig. 15 presents the results associated with the query expressed in Fig. 14. We can see that the first four results correspond to the 4 experimental settings presented in Fig. 9 for Amiri et al. (2014). Biomass (resp. solvent chemical reagents) quantity which is required to compute Efactor is presented in the Biomass quantity column (resp. Acid quantity). Results may be downloaded in a CSV file for Efactor computation.

5.2. Eco design indicator computation and visualization

In this section, we present the implementation of Efactor com putation and visualization which corresponds to the second step of the pipeline presented in Fig. 3. In Section 4.2, we have defined three kinds of Efactor indicators for a given set of n documents $o_1, \ldots, o_n$, and for each document o_i, m experimental settings which are described in o_i denoted $e_{i1}, \ldots, e_{im}$:

- $Efactor^{best}(o_i, p, b)$ *computes Efactor for the best experimental set ting in document* o_i.
- $Efactor^{all}(o_i, p, b)$ *computes Efactor for all settings in document* o_i.
- $Efactor(p, b)$ *computes Efactor for the entire set of settings of the entire set of documents.*

In the implementation, we consider that the set of documents on which Efactor has been computed corresponds to a topic in @Web. In this article, we have considered four topics, each of them associated with one of the four pre treatment processes presented in Fig. 2. We have seen in the previous section that we use @Web queries to extract csv data files in order to compute the Efactor associated with a given topic. We have implemented the computa tion of the three Efactor indicators in an Excel file in which have been previously stored data extracted from @Web. Graphical rep resentations are generated in VBA programming language executed in an Excel file to display an X Y plot for a given topic and a given biomass where X corresponds to Efactor and Y to glu cose yield. For instance, in Fig. 16, we show a ranking of pre treatments based on Efactor computation for the best experiment

Table 5
Excerpt of the annotated table process description.

No	Biomass	Output solid constituent size Unit: mm	Treatment	Experience number Unit: 1	Process step number Unit: 1	Biomass quantity Unit: g	Total pretreatment energy Unit: MJ/kg	Water quantity Unit: ml	Rotation speed Unit: min-1	Treatment duration Unit: min	Output solid constituent quantity Unit: g	Temperature Unit: °C
1	Rice straw	2.000e+00	Cutting milling	1.000e+00	1.000e+00	3.000e+01	[0.000e+00; inf]	0.000e+00	[0.000e +00; inf]	[0.000e+00; inf]		
2	Rice straw		Drying	1.000e+00	2.000e+00	3.000e+01	[0.000e+00; inf]			[0.000e+00; inf]	3.000e+01	6.000e+01
3	Rice straw		Hot water treatment	1.000e+00	3.000e+00	3.000e+01	5.700e+00	3.000e+02	[0.000e +00; inf]	3.000e+01	3.000e+01	1.600e+02
4	Rice straw		Enzymatic hydrolysis treatment	1.000e+00	4.000e+00	[4.000e−02; 6.000e−02]			[0.000e +00; inf]	4.320e+03	[2.900e−02; 4.500e−02]	4.500e+01
5	Rice straw	2.000e+00	Cutting milling	2.000e+00	1.000e+00	3.000e+01	[0.000e+00; inf]	0.000e+00	[0.000e +00; inf]	[0.000e+00; inf]		
6	Rice straw		Drying	2.000e+00	2.000e+00	3.000e+01	[0.000e+00; inf]			[0.000e+00; inf]	3.000e+01	6.000e+01
7	Rice straw		Hot water treatment	2.000e+00	3.000e+00	3.000e+01	6.600e+00	3.000e+02	[0.000e +00; inf]	3.000e+01	3.000e+01	1.800e+02
8	Rice straw		Enzymatic hydrolysis treatment	2.000e+00	4.000e+00	[4.000e−02; 6.000e−02]			[0.000e +00; inf]	4.320e+03	[2.700e−02; 4.200e−02]	4.500e+01

of considered documents. Each point corresponds to a given pre treatment of rice straw presented in a given document. For each point, the category of pre treatment is represented by geometric symbol (for instance • for PM UFM, see the legend of Fig. 16).

Reliability scores associated with each document, whose com putation has been presented in Section 4.1.2, have been repre sented in two colors for each point. The surrounding (resp. inner) color corresponds to the upper bound (resp. lower bound). For instance, the point "CM then dry BM"[11] (corresponding to pre treatment category PM UFM) has a glucose yield around 90% and a low Efactor. It is associated with reliability scores which correspond to an imprecise assessment due to disagreeing meta information represented by an external circle painted in red and an internal one in green (see Reliability index in Fig. 16).

Fig. 17 presents a ranking of pre treatments realized on rice straw based on Efactor computation for all experimental settings of four topics. Each point corresponds to a given rice straw pre treatment studied in the entire set of documents. For instance, the point PM PC PS corresponds to the Efactor associated with topic PM PC PS computed using $Efactor^{all}$ indicators presented in Table 4. It integrates 23 experimental settings extracted from 6 documents. For each topic, reliability scores associated with each document have been merged into a global reliability score, as defined in Section 4.1.2.

5.3. DSS assessment and discussion

Results obtained on rice straw with the DSS have been pre sented to 3BCAR experts in biorefinery. Those results have been positively assessed by experts who used tables and graphics asso ciated with Efactor indicators produced by the DSS to perform the following analysis. Fig. 17 shows that the highest glucose yield (86% ± 2%) from rice straw was obtained after wet disk milling (PM PC UFM PS). Nitric acid, oxidizing and ionic liquid pre treatment (PM PC PS) also achieves a good glucose yield (72.03% ± 3.32%). But these experimental conditions result in a high $Efactor_{min}$ estimated to about 70.6 (resp. 33.90) for PM PC UFM PS (resp. PM PC PS). In Fig. 16, it must be noticed that a low *Efactor* (2.03 ± 0.14) was estimated for Cutting Milling (CM) coupling to Ball Milling (BM) with about 90% of glucose yield (89.4% ± 2%) even if data source reliability is not fully established (see reliability indicator in Fig. 16 and associated metadata in Fig. 11). In general, water or chemical pre treatments of rice straw produced more glucose compared to mechanical or dry pre treatment (mechanical, torrefaction ...), but also generated more effluents with a high *Efactor*. Results presented in Fig. 16 clearly show that dry pre treatments (milling, torrefaction ...) are simpler technologies which are in general less effective in the production of glucose, but without the need of any chemical or water inputs. They have a low environmental impact (low *E*factor), thus mini mizing waste generation while maximizing value of the lignocellu losic biomass.

The results obtained on the Rice straw use case demonstrate the usefulness of the DSS data treatment pipeline feasibility. In this experimentation, users have particularly appreciated the following functionalities:

- The DSS permits to enrich continuously the RDF database with new scientific data and gives the possibility to compare them with already stored scientific data.
- The OTR provides a simple reading grid to homogenize hetero geneous textual data, even if the annotation remains manual.

[11] which means Cutting Milling then dry Ball Milling.

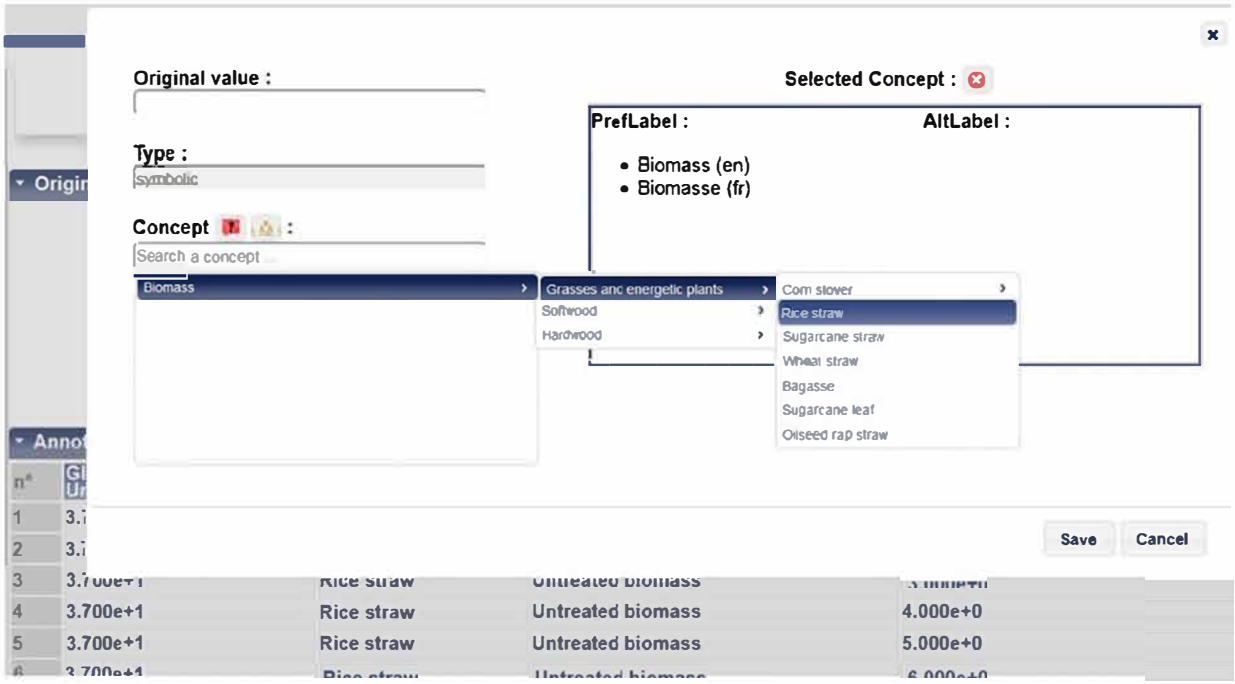

Fig. 13. Selection of a symbolic concept when navigating in the hierarchy of Biomass concepts during table annotation.

Fig. 14. Example of query expressed through @Web user interface.

- The DSS permits to easily navigate from a graphical representation of indicators (see Section 5.2) to detailed annotated data stored in the RDF database. It has been recognized to be useful to design new experimental study protocols. For instance, Fig. 16 made biorefinery experts think of designing a new experimental protocol to study more precisely the impact of torrefaction and particle size on the glucose yield. The range of particle sizes is easily available by consulting the annotated tables (see Table 5, Output solid constituent size column) or by running queries on the concept relations which provide this information.
- The DSS is a collaborative platform which can be easily shared by a community of researchers. Ontology and annotated tables are available in open access mode. Researchers who want to enrich the RDF database just need to obtain a login to the DSS.

rank	Biomass [Rice straw]	Treatment [Organosolv treatment]	Acid	Acid quantity	Biomass quantity
Click on a hidden column Alkali quantity2 Total pretreatment energy Alcohol Salt Salt quantity Gaz quantity Process step number Alcohol quantity Alkyluent quantity					
row 20_2471					
1	Rice straw	Organosolv treatment	Sulfuric acid	[5.000e-01],mg	[5.000e+01],g
row 14_2471					
1	Rice straw	Organosolv treatment	Sulfuric acid	[5.000e-01],mg	[5.000e+01],g
row 8_2471					
1	Rice straw	Organosolv treatment	Sulfuric acid	[5.000e-01],mg	[5.000e+01],g
row 2_2471					
1	Rice straw	Organosolv treatment	Sulfuric acid	[5.000e-01],mg	[5.000e+01],g
row 23_2495					
2	Rice straw	Chemical and physico-chemical pretreatment	Acid	[0.000e+00],g	[1.000e+00],g
row 7_2495					
2	Rice straw	Chemical and physico-chemical pretreatment	Acid	[0.000e+00],g	[1.000e+00],g

Fig. 15. Excerpt of the results associated with the query presented in Fig. 14.

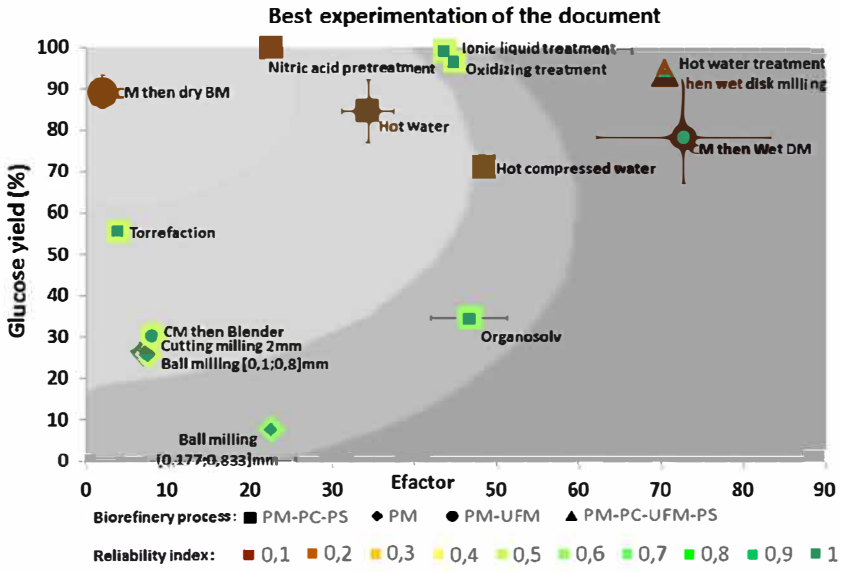

Fig. 16. Efactor associated with rice straw for best experiment of documents.

The current version of Efactor computation step remains partly manual as data extracted from the database using @Web queries (see Section 5.1) must be put in an Excel file to compute the indicator. We are currently working on an advanced version which will extract the data directly from the RDF database.

As discussed in Section 2, the complete automation of experimental data extraction from textual documents is still a challenge to be met. We believe that the combination of KE and text mining methods will permit to make essential advances. In the short term, our approach will focus on adding assistants in the @web software, in order to speed up manual annotation guided by the OTR. In the very near future, we will implement an assistant based on the coupling of OTR and text mining approaches, to complete n ary relation annotation by suggesting the more relevant sentences in which a given argument of the n ary relation appears. Moreover, time cost necessary to annotate experimental data should be compared to the one required to produce similar data in the laboratory.

In our experimentation, it took the annotators about 80 days to design the ontology, select and read scientific papers and to manually enter more than 400 experimental results concerning 4 biomasses and 6 pre treatment processes described in 32 publications. This comes to an average time of 0.2 day per experimental result. It must be put in perspective with the time spent to produce experimental results in the laboratory. In our experiment,

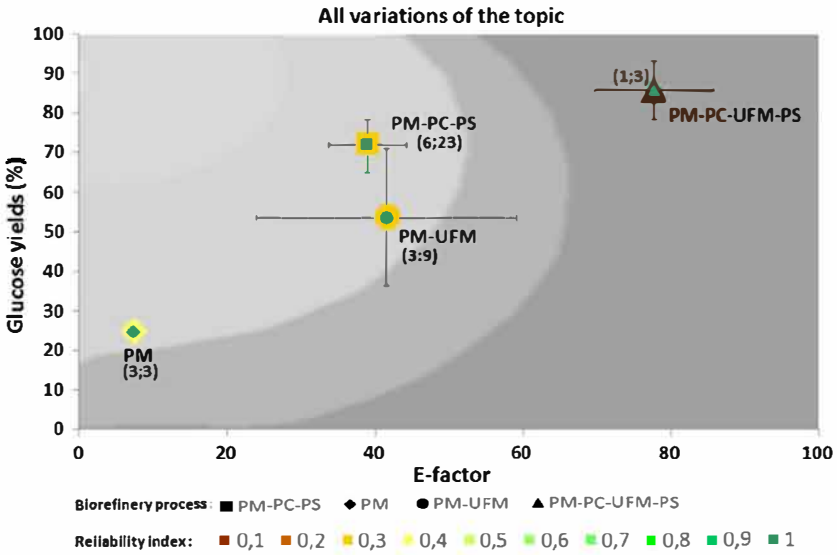

Fig. 17. Efactor associated with rice straw for all experimental settings of each topic.

the annotators who also made experimentations in the laboratory made 36 laboratory experiments in 160 days. This represents an average of 4.4 days per experimental result. The time spent for manual entering of an experimental result from the literature was considered by end users small enough compared to the time spent to produce a similar experimental result at lab scale (ratio 1/22). Moreover, the DSS design is envisioned in an iterative approach in which annotated data will be reused to develop new functionalities as the one presented in Section 6.

6. Conclusion and prospect

In this paper we have proposed a decision support system for eco efficient biorefinery process selection based on an ontology based semantic approach. The ontology is used to guide the anno tation of potentially incomplete or imprecise experimental data retrieved from the bibliography in order to store them in a struc tured database. Moreover a model has been used to assess the reli ability of data sources. Finally, a ranking of biorefinery processes has been computed in terms of glucose yield and Efactor indicators taking into account data imprecision and reliability.

The interest of Efactors is to give an overview of the process mass balance. It may be considered as a "local" indicator (because it is based on data at process scale) and "inventory level" indicator (because it uses physical flows). It completes the classical process yield in order to rank different options. The less input consuming and waste generating process is generally preferred. Used with the reliability indicators, Efactor gives interesting information at an early decision stage at research or laboratory scale. Other indi cators could be considered regarding the environmental balance. Therefore, a perspective of our work will be to compute life cycle indicators to complete the list of assessment criteria for decision makers. The application of life cycle assessment (LCA) according to the ISO 14040 standards (ISO, 2006) would deepen the analysis through the system boundaries extension (inclusion of inputs life cycle with the use of life cycle inventory databases) and through the potential environmental impacts calculation. Impact assess ment aims at transforming inventory results into environmental indicators (also called impacts categories). To compute those envi ronmental indicators, we will have to deal with the lack of data in the literature concerning the energy consumption and energy effi ciency of chemical, physicochemical and mechanical treatment of rice straw for example. Ideally, these indicators will complete the Efactor and glucose yield. More generally, this approach may be applied to any kind of biomass (food or no food) transformation process. Consequently, the number of publications which could be valorized using this approach is potentially very high. Moreover, it must be noticed that the first step of the treatment pipeline (data integration) may be applied to a lot of kinds of scientific data in order to perform numerical treatments (meta analysis, decision support tools ...). For example, we already use this approach to create decision support systems which determine optimal selec tion and dimensioning of food packagings (Guillard et al., 2015), by reusing literature data about matter transfer. Another exciting perspective would be to develop interoperability between @Web and Rosanne, to take the best of both tools in order to improve automatic annotation of relevant information from scientific documents.

Acknowledgement

This work has been realized in the framework of the IC2ACV Carnot 3BCAR project.

References

Adapa, P., Tabil, L., Schoenau, G., 2011. Grinding performance and physical properties of non-treated and steam exploded barley, canola, oat and wheat straw. Biomass Bioenergy 35 (1), 549–561.

Amiri, H., Karimi, K., Zilouei, H., 2014. Organosolv pre-treatment of rice straw for efficient acetone, butanol, and ethanol production. Bioresour. Technol. 152, 450–456.

Barakat, A., Chuetor, S., Monlau, F., Solhy, A., Rouau, X., 2014. Eco-friendly dry chemo-mechanical pre-treatments of lignocellulosic biomass: impact on energy and yield of the enzymatic hydrolysis. Appl. Energy 113, 97–105.

Bjorne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., Salakoski, T., 2009. Extracting complex biological events with rich graph-based feature sets. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. BioNLP '09. PA, USA, Association for Computational Linguistics, Stroudsburg, pp. 10–18.

Buche, P., Dibie-Barthélemy, J., Ibanescu, L., Soler, L., 2013. Fuzzy web data tables integration guided by an ontological and terminological resource. IEEE Trans. Knowl. Data Eng. 25 (4), 805–819.

Bui, Q.-C., Sloot, P.M.A., 2011. Extracting biological events from text using simple syntactic patterns. In: Proceedings of the BioNLP Shared Task 2011 Workshop, BioNLP Shared Task '11. PA, USA, Association for Computational Linguistics, Stroudsburg, pp. 143–146.

Busset, G., Belaud, J.-P., Buche, P., Barakat, A., Lousteau-Cazalet, C., Vialle, C., Sablayrolles, C., 2015. Environmental Life Cycle Analysis using knowledge engineering based approach for assessing sustainability of biorefinery systems.

In: Proceedings of BFFM'2015 (Biorefinery for Food, Fuels and Materials 2015 Symposium).
Buyko, E., Faessler, E., Wermter, J., Hahn, U., 2009. Event extraction from trimmed dependency graphs. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, BioNLP '09. PA, USA, Association for Computational Linguistics, Stroudsburg, pp. 19–27.
Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M., 2011. LexInfo: a declarative model for the lexicon-ontology interface. J. Web Sem. 9 (1), 29–51.
Chuetor, S., Luque, R., Barron, C., Solhy, A., Rouau, X., Barakhat, A., 2015. Innovative combined dry fractionation technologies for rice straw valorization to biofuels. Green Chem. 17, 926–936.
Destercke, S., Buche, P., Guillard, V., 2011. A flexible bipolar querying approach with imprecise data and guaranteed results. Fuzzy Sets Syst. 169 (1), 51–64.
Destercke, S., Buche, P., Charnomordic, B., 2013. Evaluating data reliability: an evidential answer with application to a web-enabled data warehouse. IEEE Trans. Knowl. Data Eng. 25 (1), 92–105.
Doan, A., Halevy, A.Y., Ives, Z.G., 2012. Principles of Data Integration. Morgan Kaufmann Publishers Inc.
Guarino, N., Oberle, D., Staab, S., 2009. What is an ontology? In: Staab, S., Studer, R. (Eds.), Handbook on Ontologies, International Handbooks on Information Systems. Springer Berlin Heidelberg, pp. 1–17.
Guillard, V., Buche, P., Destercke, S., Menut, L., Guillaume, C., Gontard, N., 2015. A Decision Support System for designing biodegradable packaging for fresh produce. Comput. Electron. Agric. 111, 131–139.
Hao, Y., Zhu, X., Huang, M., Li, M., 2005. Discovering patterns to extract protein-protein interactions from the literature: part ii. Bioinformatics 21, 3294–3300.
Hawizy, L., Jessop, D., Adams, N., Murray-Rust, P., 2011. ChemicalTagger: a tool for semantic text-mining in chemistry. J. Cheminformatics 3 (1), 17.
Hideno, A., Inoue, H., Tsukahara, K., Fujimoto, S., Minowa, T., Inoue, S., Endo, T., Sawayama, S., 2009. Wet disk milling pre-treatment without sulfuric acid for enzymatic hydrolysis of rice straw. Bioresour. Technol. 100, 2706–2711.
Hideno, A., Inoue, H., Tsukahara, K., Fujimoto, S., Minowa, T., Inoue, S., Endo, T., Sawayama, S., 2012. Combination of hot compressed water treatment and wet disk milling for high sugar recovery yield in enzymatic hydrolysis of rice straw. Bioresour. Technol. 104, 743–748.
Huang, M., Zhu, X., Payan, D.G., Qu, K., Li, M., 2004. Discovering patterns to extract protein-protein interactions from full biomedical texts. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, JNLPBA '04. PA, USA, Association for Computational Linguistics, Stroudsburg, pp. 22–28.
Ilgook, K., Bomi, L., Ji-Yeon, P., Sun-A, C., Jong-In, H., 2014. Effect of nitric acid on pre-treatment and fermentation for enhancing ethanol production of rice straw. Carbohydr. Polym. 99, 563–567.
Knoblock, C.A., Szekely, P.A., Ambite, J.L., Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyan, M., Mallick, P., 2012. Semi-automatically mapping structured sources into the semantic web. In: ESWC, vol. 2012, pp. 375–390.
Kumar, P., Barrett, D.M., Delwiche, M.J., Stroeve, P., 2009. Methods for pre-treatment of lignocellulosic biomass for efficient hydrolysis and biofuel production. Ind. Eng. Chem. Res. 48 (8), 3713–3729.
Le Minh, Q., Truong, S.N., Bao, Q.H., 2011. A pattern approach for biomedical event annotation. In: Proceedings of the BioNLP Shared Task 2011 Workshop, BioNLP Shared Task '11. PA, USA, Association for Computational Linguistics, Stroudsburg, pp. 149–150.
Maillet, N., Charnomordic, B., Destercke, S., 2010. Belief: Contains basic functions to manipulate belief functions and associated mass assignments. R Package Version 1.0. [Online]. Available: <http://CRAN.R-project.org/package=belief>.
McCrae, J., Spohr, D., Cimiano, P., 2011. Linking lexical resources and ontologies on the semantic web with lemon. In: ESWC, vol. 1, pp. 245–259.
Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63 (2), 81–97.
Noy, N., 2004. Semantic integration: a survey of ontology-based approaches. SIGMOD Rec. 33 (4), 65–70.
Noy, N., Rector, A., Hayes, P., Welty, C., 2006 Defining n-ary relations on the semantic web. W3C Working Group Note <http://www.w3.org/TR/swbp-n-aryRelations>.
Poornejad, N., Karimi, K., Behzad, T., 2013. Improvement of saccharification and ethanol production from rice straw by NMMO and [BMIM][OAc] pre-treatments. Ind. Crops Prod. 41, 408–413.
Minard, A.-L., Ligozat, A.-L., Brigitte Grau, B., 2011. Multi-class SVM for Relation Extraction from Clinical Reports. RANLP 59, 604–609.
Miwa, M., Saetre, R., Miyao, Y., Tsujii, J., 2009. A rich feature vector for protein-protein interaction extraction from multiple corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, EMNLP '09. PA, USA, Association for Computational Linguistics, Stroudsburg, pp. 121–130.
Raafat, T., Trokanas, N., Cecelja, F., Bimi, X., 2013. An ontological approach towards enabling processing technologies participation in industrial symbiosis. Comput. Chem. Eng. 59, 33–46.
Raja, K., Subramani, S., Natarajan, J., 2013. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database: The Journal of Biological Databases and Curation*, *2013*, bas052. http://dx.doi.org/10.1093/database/bas052.
Rijgersberg, H., Wigham, M., Top, J.L., 2011. How semantics can improve engineering processes: a case of units of measure and quantities. Adv. Eng. Inform. 25 (2), 276–287.
Roche, C., Calberg-Challot, M., Damas, L., Rouard, P., 2009. Ontoterminology – a new paradigm for terminology. In: KEOD, pp. 321–326.
Rosario, B., Hearst, M.A., 2004. Classifying semantic relations in bioscience texts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04. PA, USA, Association for Computational Linguistics, Stroudsburg.
Rosario, B., Hearst, M.A., 2005. Multi-way relation classification: Application to protein-protein interactions. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05PA, USA, Association for Computational Linguistics, Stroudsburg. 732–739.
Schultz-Jensen, N., Kadar, Z., Thomsen, A.B., Bindslev, H., Leipold, F., 2011. Plasma-assisted pre-treatment of wheat straw for ethanol production. Appl. Biochem. Biotechnol. 165 (3–4), 1010–1023.
Sheikh, M.M., Kim, C.H., Park, H.J., Kim, S.H., Kim, G.C., Lee, J.Y., Sim, S.W., Kim, J.W., 2013. Effect of torrefaction for the pre-treatment of rice straw for ethanol production. J. Sci. Food Agric. 93 (13), 3198–3204.
Tian, A., Sequeda, J., Miranker, D.P., 2013. QODI: Query as context in automatic data integration. In: International Semantic Web Conference, vol. 1, pp. 624–639.
Touhami, R., Buche, P., Dibie-Barthélemy, J., Ibanescu, L., 2011. An ontological and terminological resource for n-ary relation annotation in web data tables. In: OTM 2011 (2). LNCS, vol. 7045. Springer, pp. 662–679.
Trokanas, N., Cecelja, F., Raafat, T., 2015. Semantic approach for pre-assessment of environmental indicators in Industrial Symbiosis. J. Clean. Prod. 96, 349–361.
Van Landeghem, S., Saeys, Y., De Baets, B., Van de Peer, Y., 2009. Analyzing text in search of bio-molecular events: A high-precision machine learning framework. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, BioNLP '09. PA, USA, Association for Computational Linguistics, Stroudsburg, pp. 128–136.
Zhang, H., Huang, M., Zhu, X., 2011. Protein-protein interaction extraction from bio-literature with compact features and data sampling strategy. In 4th International Conference on Biomedical Engineering and Informatics, BMEI 2011, Shanghai, China, October 15-17, 1767–1771.
Zhang, Z., 2014. Towards efficient and effective semantic table interpretation. In: Proceedings of Semantic Web Conference, vol. 1, pp. 487–502.
Zhou, D., Zhong, D., He, Y., 2014. Biomedical relation extraction: From binary to complex. Comp. Math. Methods in Medicine vol. 2014, 18. http://dx.doi.org/10.1155/2014/298473 298473.
Zhu, J.Y., Pan, X.J., 2010. Woody biomass pre-treatment for cellulosic ethanol production: technology and energy consumption evaluation. Bioresour. Technol. 101 (13), 4992–5002.