A decision support system using multi-source scientific data, an ontological approach and soft computing - application to eco-efficient biorefinery

Charlotte Lousteau-Cazalet, Abdellatif Barakat, Jean-Pierre Belaud, Patrice Buche, Brigitte Charnomordic, Stéphane Dervaux, Sébastien Destercke, Juliette Dibie-Barthelemy, Caroline C. Sablayrolles, Claire Vialle

A decision support system using multi-source scientific data, an ontological approach and soft computing – Application to eco-efficient biorefinery

Lousteau-Cazalet Charlotte
INRA IATE
Montpellier, France
charlotte.loustau-cazalet@ifp.fr

Barakat Abdellatif
INRA IATE
Montpellier, France
Abdellatif.Barakat@supagro.inra.fr

Belaud Jean-Pierre
INP-ENSIACET, LGC
Toulouse, France
JeanPierre.Belaud@ensiacet.fr

Buche Patrice
INRA IATE/LIRMM
Montpellier, France
buche@supagro.inra.fr

Busset Guillaume
INP-ENSIACET, CAI
Toulouse, France
guillaume.busset@ensiacet.fr

Charnomordic Brigitte
INRA MISTEA
Montpellier, France
bch@supagro.inra.fr

Dervaux Stéphane
INRA/AgroParisTech MIA-
F-75231 Paris
stephane.dervaux@versailles.inra.fr

Destercke Sébastien
CNRS Heudiasyc
F-60200 Compiègne
destercke@hds.utc.fr

Dibie Juliette
INRA/AgroParisTech MIA
F-75231 Paris
dibie@agroparistech.fr

Sablayrolles Caroline
INP-ENSIACET, CAI
Toulouse, France
caroline.sablayrolles@ensiacet.fr

Vialle Claire
INP-ENSIACET, CAI
Toulouse, France
claire.vialle@ensiacet.fr

***Abstract*— In decision tasks such as bioprocess efficiency comparison, scientific literature is a valuable source of data. This large number of scientific data is heterogeneously structured, mainly in textual format. Innovative tools able to integrate and treat constantly new information are required. In this context, the use of semantic web methods such as ontologies seems relevant to structure the experimental information. Imprecision and uncertainty can arise from data incompleteness and variability. This is particularly true for processes involving biological materials. Document reliability should also be considered. Soft computing methods have the potential to be the kingpin of specialized software that can be integrated in decision support systems (DSS) intended to solve these issues. This paper presents the implementation of a pipeline which permits to: (1) structure and integrate the experimental data of interest by using ontologies, (2) assess data source reliability, (3) compute and visualize indicators taking into account data imprecision.**

Keywords: Decision support system, uncertainty management, belief theory, fuzzy numbers, knowledge engineering, ontology, bioprocess eco-design

I. INTRODUCTION

Environmental sustainability assessment of processes is being increasingly viewed as an important tool to aid in the shift towards sustainability. It is a complex task that requires several steps and the examination of numerous factors: energy consumption, energy efficiency and environmental factor of chemical, physicochemical and mechanical treatments. If we consider the bioconversion of lignocellulosic biomass processes [3,31], comparative studies remain scarce, even if the topic of lignocellulosic biomass has been extensively studied in the past thirty years, yielding a great number of scientific papers focused on a specific study. Building DSS able to include scientific data extracted from the literature opens the way to a whole series of new (Meta)-analyses, making it possible to widen the scope of work [5], in order to build more realistic DSS and to help researchers involved in the process design to make rational decisions based on data and knowledge expressed by domain experts in the scientific literature.

However this topic is challenging in many ways. The first obstacle holding back the use of those scientific data is their textual format and heterogeneous structure. In this context, the use of ontologies is relevant [30,24] to structure the experimental information and express it in a standardized vocabulary. This permits to organize knowledge in order to perform automatic reasoning and to facilitate linked open data. The second challenge is to take into account data imprecision and incompleteness. Indeed, a scientific publication often presents data summaries with various formats, for instance intervals or [mean, standard deviation] pairs. These summaries are issued from sets of experiments, which are not available in the paper. The use of intervals or fuzzy numbers is well suited to deal with such imprecisions and uncertainties. The third difficulty consists in taking into account source (document) reliability when using these data in calculations. Belief theory provides elegant solutions to handle this point [8].

The DSS architecture proposed in this paper aims at coupling generic methods and reusable software modules to meet these challenges, while being instantiated to meet the specific needs of a particular application. More precisely, this DSS relies on the development of a system:

(i) able to annotate, store and maintain potentially incomplete or imprecise data extracted from the scientific literature in dedicated databases,
(ii) allowing the computation of indicators taking into account data imprecision,
(iii) evaluating document reliability.

Our approach is compared to the state of the art in Section 2, and the software architecture is detailed in Section 3. To illustrate our proposal, we present in Section 4 an example about glucose extraction in rice straw comparing four processes that may include a sequence of unit operations. Section 5 concludes the paper.

II. COMPARISON WITH THE STATE OF THE ART

To the best of our knowledge, there is no comparable DSS implementing a full pipeline such as the one presented in Fig. 1, which allows to represent imprecise data extracted from heterogeneous textual documents in order to compare indicators (for example bioprocess efficiency indicators). This DSS, which will be described in more detail in Section 3, is a pipeline composed of three steps: (1) annotation guided by the ontology of experimental data published in scientific papers, (2) annotated data extraction and indicators computation, (3) indicators visualization in graphical maps. In general, relevant experimental data published in textual documents are scattered in different parts of the document and expressed in different formats. For example, in Process Engineering papers, operation control 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. Automatic extraction of scattered information from text and tables of scientific articles is an open research topic [15,17,13,4,28]. It is out of the scope of this paper dedicated to the implementation of a first version of an operational pipeline presented in Fig.1, in which annotation is a manual operation guided by the ontology. However, comparison could be done concerning the first step of the pipeline: experimental imprecise data representation 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).

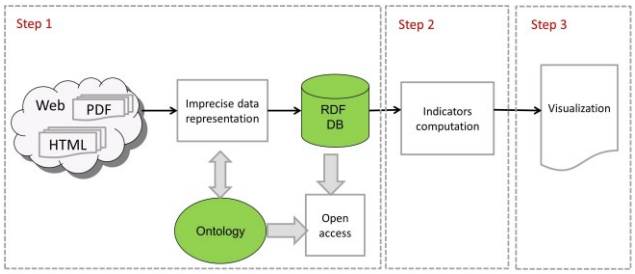

Fig. 1. Data treatment pipeline combining ontologies and soft computing tools.

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 build on 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. Moreover, as @Web, Rosanne manages the notion of phenomenon, very similar to the notion of symbolic concept in @Web, which represents non numerical data, as for instance studied objects. The main difference is that @web defines the notion of relation, which links data (studied object with controlled parameters and results) in order to represent a whole experiment. It is important in the DSS as this notion is used to extract annotated data in order to compute indicators. Moreover, @Web proposes an end-user graphical interface to query annotated tables using soft computing tools, in particular a bipolar fuzzy pattern matching algorithm [7] which takes into account the fact that data stored in @Web may be imprecise and of diverse reliability [8]. This is not available in the current version of Rosanne. From its side, Rosanne proposes an interesting functionality to merge several annotated tables sharing a column annotated with the same concept. As a 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.

III. ARCHITECTURE OF THE DECISION SUPPORT SYSTEM

Fig.1 details the three steps of the data treatment pipeline, which combines ontologies and soft computing tools. In the first step, experimental data published in scientific papers are annotated thanks to an ontology and assessed in terms of their source reliability. Annotated data (that may be imprecise) are stored in a RDF database and available in open access via permalinks, a SPARQL[1] end-point and a dedicated querying system guided by the ontology. The second step consists in extracting annotated data from the RDF database to compute indicators and data reliability scores. Indicators and data reliability scores can be visualized in the third step as graphical maps.

3.1 Heterogeneous experimental data integration (step 1)

To facilitate integration of scientific data coming from heterogeneous sources, one of the relevant solutions is to use ontologies [20,10,9]. @Web implements the first step of Fig. 1 as a complete workflow (see Fig. 2) to manage experimental data: extraction and semantic annotation of data from scientific documents, data source reliability assessment and bipolar flexible querying of the collected imprecise data stored in a database opened on the Web. @Web relies on an Ontological and Terminological Resource (OTR) which guides the scientific data semantic annotation and the querying. OTR 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 therefore be reused for different application domains: only the specific part of the OTR must

[1] SPARQL (**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage) is the protocol and the query language which permits to search, add, modify or suppress RDF graphs

be redefined to re-use @Web for a new domain (see [29] for a reuse in food packaging domain).

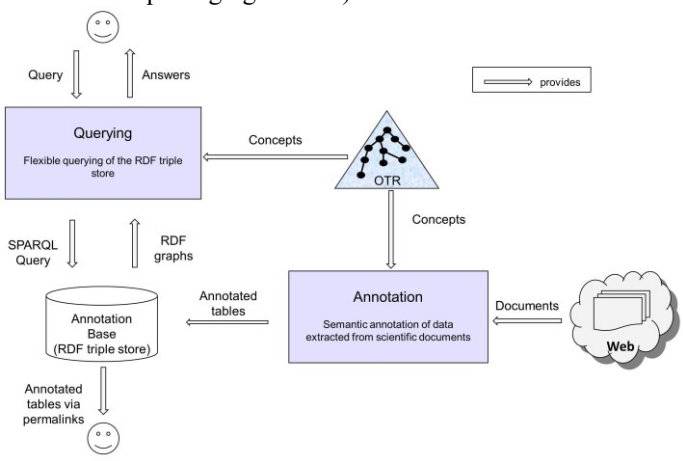

Fig.2. Knowledge annotation and querying in @Web system.

@Web is composed of two sub-systems (see Fig.2) combining knowledge engineering and soft computing tools. The first one is an annotation sub-system for the acquisition and annotation, with concepts of the OTR, of experimental imprecise data extracted from data found in scientific documents; those annotated data being stored into a database. This sub-system also allows the reliability of data sources to be assessed using the approach of [8]. The second sub-system is a bipolar flexible querying system based on the approach presented in [7], which allows data stored in the database to be queried. @Web is implemented using the semantic web standards (XML[2], RDF, OWL[3], SPARQL): the OTR is defined in OWL2-DL, annotated tables in XML/RDF and the querying in SPARQL. We present in Section 3.1.1 the way OTR has been modeled to be used in @Web. Section 3.1.2 details the model used to assess data source reliability.

3.1.1 OTR model

The OTR is designed to annotate data tables representing scientific experiments results in a given domain (see [29] for more details). We made the choice to represent an experiment which involves a studied object, several experimental parameters and a result using n-ary relations in order to structure information in a simple way which can be easily understood by annotators. As recommended by W3C [20], we used the design pattern which represents a n-ary relation thanks to a concept associated with its arguments via properties. Fig. 3 illustrates this concept (Milling is a unit operation performed on a given biomass).

[2] Extensible Markup Language is a markup language.

[3] Web Ontology Language is a knowledge representation model built upon RDF.

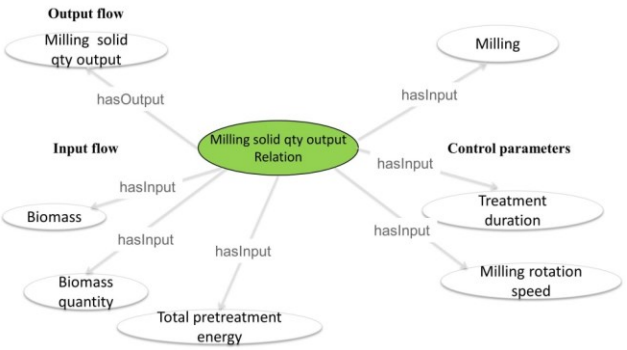

Fig.3. A Relation concept to model the milling unit operation.

An excerpt of the OTR global structure is presented in Fig.4. The conceptual component of OTR is composed of a **core ontology** to represent n-ary relations between experimental data and a **domain ontology** to represent specific concepts of a given application domain (Biorefinery in this example). In the up core ontology, generic concepts **Relation** and **Argument** represent respectively n-ary relations and arguments. 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. The terminological component of OTR contains the set of terms describing the studied domain and used to annotate data.

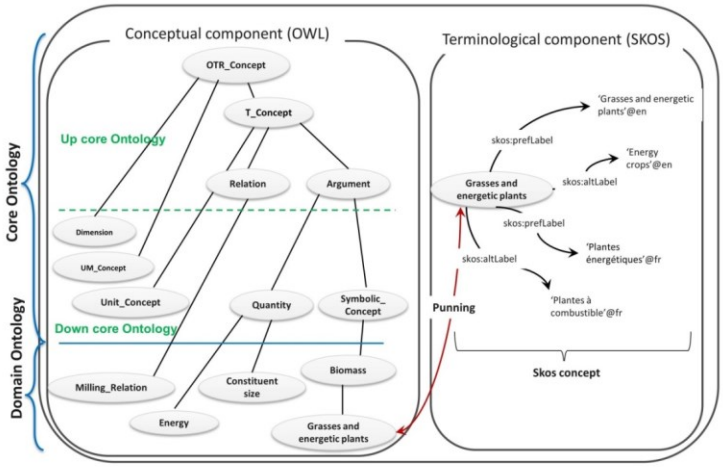

Fig.4.OTR model specialized for biorefinery.

3.1.2 Reliability assessment scores

We recall in this section the approach presented in [8] in order to compute reliability assessment scores associated with data sources extracted from the Web. Given a document *o* collected from some bibliographical resources on the web, the role of the reliability is to affect an interval-valued score $\left[\underline{\mathbb{E}_o}, \overline{\mathbb{E}_o}\right]$ that reflects the *a priori* reliability of information *o*. The interval is obtained through an expert system using meta-information, and the length or imprecision of $\left[\underline{\mathbb{E}_o}, \overline{\mathbb{E}_o}\right]$ reflects to which extent the various pieces of meta-information are consistent.

The system is built as follows. First, an ordered finite reliability space $\Theta = \{\theta_1, \cdots, \theta_R\}$ is built, θ_1 being the lowest reliability value, θ_R the highest. Usually, $R = 5$ (as in this paper) or $R = 7$ to ensure a good compromise between complexity and expressiveness. A non-decreasing score function f on Θ is then defined, in our case $f(\theta_i) = i$.
Second, S groups $A_1, \cdots, A_s$ of *meta-information* that will be used to assess reliability are defined, a group A_i taking C_i values $a_{i1}, \cdots, a_{iC_i}$. Various types of meta-information have been considered for the data sources:

1. meta-information on the data source itself: for instance the source type (e.g. scientific publication, technical report), the source reputation, citation data;
2. meta-information related to means used to produce data. Such information is typically included in a section called *Material and method* in papers based on experiments in Life Science, 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;
3. 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, which can lead to make groups A_i containing multiple criteria (e.g. number of citation and publication date).
After the groups have been formed, for each value $a_{ij}, i = 1, \cdots, S, j = 1, \cdots, C_i$, an expert of the field from which data are collected gives his/her opinion about how reliable is 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. Each modality is then transformed into a fuzzy set. Fig. 5 illustrates such a fuzzy set, defined on Θ with R=5.

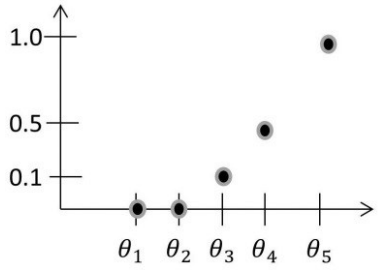

Fig.5. Fuzzy set corresponding to the term *very reliable.*

To each document *o* are then associated *S* fuzzy sets $\mu_{a_1^o}, \cdots, \mu_{a_S^o}$ defined on Θ corresponding to its *meta-information*. Those fuzzy sets are then merged together using evidential theory and a maximal coherent subset approach which allows conflicting evidences to be taken into account (i.e. assessment of high reliability for an aspect but of low reliability for another). The result of this merging is a mass distribution $m_o: 2^\Theta \rightarrow [0,1]$ which reflects the global reliability of *o* (see [8] for more details). Final score $\underline{\mathbb{E}_o}$ is then computed using the following formula:

Eq.1. Final score definition.

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

$\overline{\mathbb{E}_o}$ is obtained with the same formula, replacing *inf* by *sup*. These scores are then used in the querying system to order annotated data associated with documents thanks to their reliability. [8] presents various means to analyze the result of the reliability, such as the reasons that have led to an imprecise assessments and the detection of subgroups of agreeing/disagreeing meta-information.

3.2 Software workflow

In this section, due to the lack of space, we only present the step 1 of the workflow which implements the data treatment pipeline presented in Fig. 1. @Web relies on the generic part of the OTR model (see the core ontology in Fig. 4) and allows the management of the domain ontology (by example Biorefinery OTR). As @Web relies on the generic part of the OTR model, several OTR dedicated to different application 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 available at http://www6.inra.fr/cati-icat-atweb. Current version of the OTR of units of measure is also available at http://www6.inra.fr/cati-icat-atweb (section @Web platform, thumbnail Ontology, option Unit Ontology).

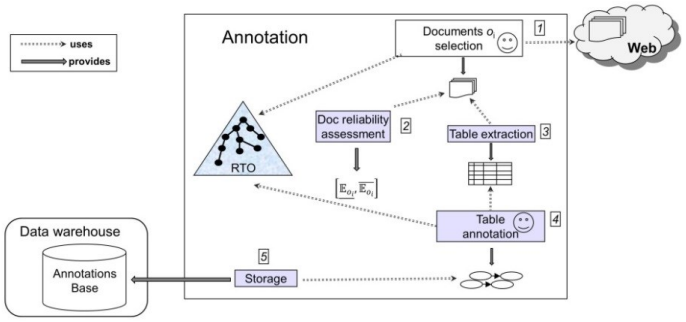

Fig.6. Five steps of the annotation sub-system in @web

The annotation workflow of @Web is implemented in five steps presented in Fig.6. Recorded tutorials of current @Web version are available on-line (http://www6.inra.fr/cati-icat-atweb/Tutorials) for readers interested by the complete workflow. In this papier, we focus on two steps. Firstly, we present the second step which is dedicated to document reliability assessment using the model presented in Section 3.1.2. In the current version, meta-information associated with each document is manually entered in order to compute

reliability score. In Fig.7, , reliability score has led to an imprecise assessment $\left[\underline{\mathbb{E}_o}, \overline{\mathbb{E}_o}\right] = [1.5, 4.98]$ due to conflict between expert opinions associated with meta-information: "citation age and citation number" and "source type" are considered *very reliable*, "Enzymatic hydrolysis reproducibility" and "Biochemical and physic-chemical analysis reproducibility" are considered *hardly reliable* because only the average value of experimental results associated with those unit operations are given in the document. All operations involving belief functions-needed to compute reliability score have been implemented in a R package. The package is called belief [22], and it includes basic functions to manipulate belief functions and associated mass assignments (currently on finite spaces only). Secondly, we focus on the fourth step, called *Table annotation*, which corresponds to the manual semantic annotation of the selected data tables using the concepts of Biorefinery OTR. Taking into account the actual content of the original table, the annotator selects from the n-ary relation concepts defined in the OTR those relevant to annotate the table. Several n-ary relations may be used in a given annotated table in order to annotate experimental data associated with a complete pretreatment process.

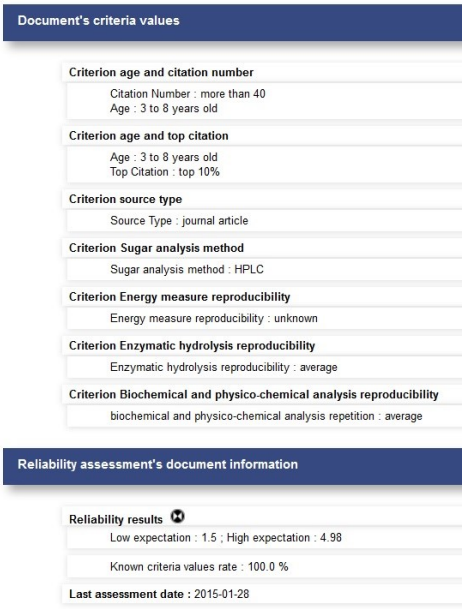

Fig.7. @Web reliability assessment associated with [Hideno et al. 2009].

n°	Output solid constituent quantity Unit : g	Treatment	Experience number Unit : 1	Process step number Unit : 1	Biomass	Biomass quantity Unit : g	Treatment duration Unit : min	Temperature Unit : oC	Total pretreatment energy Unit : MJ/kg	Output liquor quantity Unit : l	Water quantity Unit : l	Rotation speed Unit : min-1	Output solid constituent size Unit : mm
1	5.000e+1	Cutting milling	1.000e+0	1.000e+0	Rice straw	5.000e+1	[-inf ; inf]		[-inf ; inf]		0.000e+0	[-inf ; inf]	2.000e+0
2	5.000e+1	Drying	1.000e+0	2.000e+0	Rice straw	5.000e+1	[-inf ; inf]	6.000e+1	[-inf ; inf]				
3	5.000e+1	Dry ball milling	1.000e+0	3.000e+0	Rice straw	5.000e+1	5.000e+0		9.000e+0		0.000e+0	1.700e+3	
4	[3.100e-2 ; 4.900e-2]	Enzymatic hydrolysis treatment	1.000e+0	4.000e+0	Rice straw	[4.000e-2 ; 6.000e-2]	4.320e+3	4.500e+1				[-inf ; inf]	
5	5.000e+1	Cutting milling	2.000e+0	1.000e+0	Rice straw	5.000e+1	[-inf ; inf]		[-inf ; inf]		0.000e+0	[-inf ; inf]	2.000e+0
6	5.000e+1	Drying	2.000e+0	2.000e+0	Rice straw	5.000e+1	[-inf ; inf]	6.000e+1	[-inf ; inf]				
7	5.000e+1	Dry ball milling	2.000e+0	3.000e+0	Rice straw	5.000e+1	1.500e+1		2.700e+1		0.000e+0	1.700e+3	
8	[3.000e-1 ; 4.500e-2]	Enzymatic hydrolysis treatment	2.000e+0	4.000e+0	Rice straw	[4.000e-2 ; 6.000e-2]	4.320e+3	4.500e+1				[-inf ; inf]	

Table 1 Excerpt of the annotated table Process Description

Table 1 presents an example of an annotated table in @Web extracted from the scientific document [11], which describes a biorefinery pretreatment process composed of a sequence of four unit operations realized in experiments 1 and 2. The columns of the annotated table correspond to arguments of the relation *Milling_Solid_Quantity_Output_Relation* (see Fig. 3). For instance, we can see on the row n° 1 that the first unit operation is a cutting milling, instance of the relation *Milling_Solid_Quantity_Output_Relation*. The row n° 3 shows that the third unit operation of this process is a second 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 to several tasks. For example, it is possible to enter an imprecise quantitative value as an interval of values or a pair mean/standard deviation. In Table 2, the quantity *Output solid constituent quantity* is defined as the precise value 5g for *Cutting milling treatment* in row n°1 and as the interval [3.1e-2,4.9e-2] g. for *Enzymatic hydrolysis treatment* in row n°4. Missing data are denoted by the interval [-inf; inf]. In the fifth and last step of annotation, called *Storage*, the annotated data tables are stored in a RDF triple store which could be queried through either an end-user querying interface or a SPARQL endpoint for open data access. The data annotated with @Web may be queried through an end-user interface, which implements a flexible bipolar querying method described in [7,8]. It must be noticed that this querying method performs simultaneously three kinds of reasoning: (1) inference using specialization relation defined in the 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. Selection criteria can be expressed on relation arguments. They may be mandatory or desirable.

IV. CASE STUDY: BIOPROCESS EFFICIENCY

We now present the application of the pipeline presented in Fig. 1 to a case study of bioprocess efficiency [1,2]. The DSS aims at solving the dilemma of assessing the environmental impact of alternative biorefinery systems, namely glucose extraction in rice straw. Several processes are being compared on the basis of scientific data extracted from bibliographical resources on the Web. Fig. 8 displays the studied system.

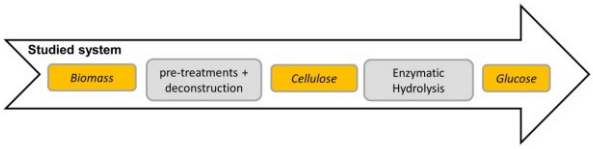

Fig. 8 - Biorefinery pre-treatment process.

Efactor is a classical indicator used to compare bioprocess efficiency. For a given set of n documents $o_1,\cdots,o_n$, we consider for each document o_i the m experimental settings which are described in o_i denoted $e_{i1},\cdots,e_{im}$. Each experimental setting is associated with a given biomass, denoted $\text{biomass}(e_{ij})$, which belongs to the set of l studied biomasses $b_1,\cdots,b_l$. For a given biomass b and a given process p, 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,. *Efactor* can be seen as the total input quantity of matter not valorized into glucose but required to produce 1 kg of glucose. *Efactor* is presented as follows in [6]:

Eq.2. *Efactor* definition.

$$Efactor = \frac{B + C + S - GRQ}{GRQ}$$

where

- B is the initial constant biomass quantity (kg),
- C is the chemical reagent product constant quantity used in the process (kg),
- S is the constant quantity of solvent (water and/or solution) used in the process (kg),
- GRQ (kg) is a quantity defined as the biomass quantity (input of the enzymatic hydrolysis unit operation) multiplied by the *glucose rate* (glucose available in the raw biomass, denoted *GR*) and the *glucose yield* (glucose extracted from the biomass, denoted *GY*) which depends on the considered experimental result.

The experimental results considered in this case study are *GR* and *GY*. Consequently, *GR* (resp. *GY)* can be considered as a sample drawn from a random variable. We have noticed that, in a given document o_i, the *GY* random variable depends on experimental settings, which is not the case for the *GR* random variable whose sampling shows no variation. In the following, we propose for a given document o_i, a given biomass $b \in \{b_1,\cdots,b_l\}$ and a given process $p \in \{p_1,\cdots,p_k\}$, to compute *Efactor*(o_i,p,b) in selecting the best experimental setting presented in document o_i and computing $Efactor^{best}(o_i,p,b)$. Having in mind the imprecision expressed for random variable *GY*, a pessimistic point of view will prefer to guarantee the best minimal *GY*, while an optimistic one will prefer to guarantee the best maximal *GY*. 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 *GY* 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 is unknown; this is usually a reasonable assumption in such experiments). 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:

$$\overline{GY}_{e_{ij^*}} - 2\sigma_{GY_{e_{ij^*}}} = \max_j\left(\overline{GY}_{e_{ij}} - 2\sigma_{GY_{e_{ij}}}\right)$$

The procedure is illustrated using rice straw biorefinery treatment process data from [2]. The best experimental setting corresponds to the one having the maximal lower bound of the 95% confidence interval associated with the *GY* random variable. Let us consider that $B = 1$ *kg*, $S = 8$ *kg*, $C = 0.0005$*kg* and the 95% confidence interval associated with the *GR* random variable = [0.51995, 0.57335] in [2], we compute, following Eq. 2, $Efactor^{best}(o_i,p,b)$ = [42.04, 51.34].

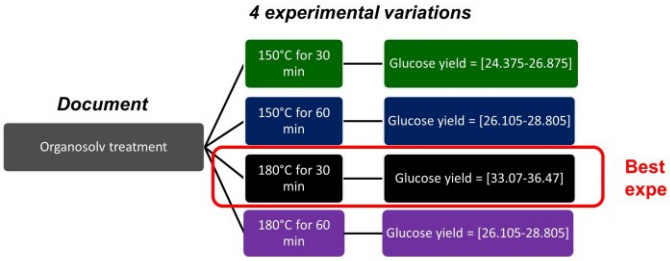

Fig. 9: 95% confidence intervals associated with the *GY* random variable for experimental settings presented in [2].

We have seen in the previous section that we use @Web queries to extract in csv files data in order to compute *Efactor* associated with a given topic. We have implemented the computation of the *Efactor* indicator in a R program. Graphical representations are generated to display an X-Y plot for a given topic and a given biomass where X corresponds to *Efactor* and Y to glucose yield. For instance, in Fig.10. , we show a ranking of biorefinery treatments based on *Efactor* computation for best experiment of considered documents. Each point corresponds to a given biorefinery treatment of rice straw presented in a given document. For each point, the

category of treatment is represented by a geometric symbol (see the legend of Fig.10.).

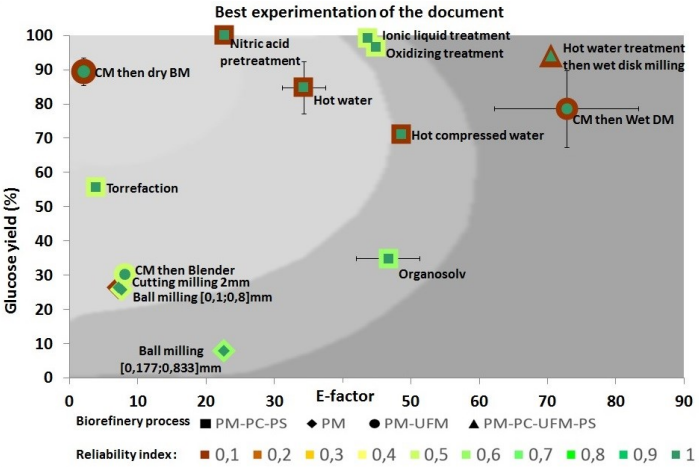

Fig.10. *Efactor* associated with rice straw for best experiment of documents

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

Table 2. Metadata considered in the reliability assessment.

Reliability scores associated with each document, whose computation has been presented in section 3.1.2, are based on metadata given in Table 2. They have been represented 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" [4] (corresponding to biorefinery process PM-UFM in Fig. 10) 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.10.).

Results obtained on rice straw with the DSS have been presented to experts in biorefinery. Those results have been positively assessed by experts who used tables and graphics associated with *Efactor* indicators produced by the DSS to perform the following analysis. In Fig.10, 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 and associated metadata in Fig.7). In general water or chemical pretreatments of rice straw produced more glucose compared to mechanical or dry pretreatment (mechanical, torrefaction…), but produced more effluents with a high *Efactor*. Results presented in Fig.10 clearly demonstrate that dry pretreatments (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 with an low environmental impact (low *Efactor*), thus minimizing waste generation while maximizing value of the lignocellulosic feedstock.

[4] which means Cutting Milling then dry Ball Milling

V. CONCLUSION AND PROSPECT

In this paper we have proposed a decision support system based on the integration of multisource scientific data and on the calculation of overall indicators. The DSS combines an ontology-based semantic approach and soft computing tools (fuzzy logic and belief theory) to handle data imprecision and reliability. The ontology is used to guide the annotation of potentially incomplete or imprecise experimental data retrieved from the bibliography in order to store them in a structured database. Moreover a model has been used to assess the reliability of data source, and a ranking of results is done taking into account data imprecision and reliability. The potential of the approach has been illustrated with a study of environmental impact factors of biomass conversion processes. Used with the reliability indicators, the DSS gives interesting information for an early stage of decision making at research or laboratory scale. The current development of data warehouses makes it possible for such approaches to gain in efficiency and to give more and more realistic results.

VI. ACKNOWLEDGEMENTS

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

REFERENCES

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

[2] H. Amiri, K. Karimi and H. Zilouei, Organosolv pretreatment of rice straw for efficient acetone, butanol, and ethanol production. Bioresource technology : 152: 450-456, 2014.

[3] A. Barakat, S. Chuetor, F. Monlau, A. Solhy, X. Rouau, Eco-friendly dry chemo-mechanical pretreatments of lignocellulosic biomass: Impact on energy and yield of the enzymatic hydrolysis. Applied Energy, 113: 97-105, 2014.

[4] P. Buche, J. Dibie-Barthélemy, L. Ibanescu and L.Soler, Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource. IEEE Trans. Knowl. Data Eng. 25(4): 805–819, 2013.

[5] G. Busset, J.P. Belaud, P. Buche, A. Barakat, C. Lousteau-Cazalet, C. Vialle and C. Sablayrolles, Environmental Life Cycle Analysis using knowledge engineering based approach for assessing sustainability of biorefinery systems. Proceedings of BFFM'2015 (Biorefinery for Food, Fuels and Materials 2015 symposium).

[6] S. Chuetor, R. Luque, C. Barron, A. Solhy, X. Rouau and A. Barakhat, Innovative combined dry fractionation technologies for rice straw valorization to biofuels. Green Chemistry 17: 926-936, 2015.

[7] S. Destercke, P. Buche and V. Guillard, A flexible bipolar querying approach with imprecise data and guaranteed results. Fuzzy Sets and Systems 169(1): 51-64, 2011.

[8] S. Destercke, P. Buche and B. Charnomordic, Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse. IEEE Trans. Knowl. Data Eng. 25(1): 92-105, 2013.

[9] A. Doan, A.Y. Halevy and Z.G. Ives, Principles of Data Integration. Morgan Kaufmann, 2012.

[18] N. Fridman, Semantic integration: A survey of ontology-based approaches. SIGMOD Record 33(4): 65–70, 2004.

[19] N. Fridman and A. Rector, Defining N-ary Relations on the Semantic Web, 2006. URL http://www.w3.org/TR/swbp-n-aryRelations/

[10] N. Guarino, D. Oberle and S. Staab, What is an ontology? In: S. Staab, R. Studer (eds.) Handbook on Ontologies, International Handbooks on Information Systems, 1–17. Springer Berlin Heidelberg, 2009.

[15] L. Hawizy, D. Jessop, N. Adams, P. Murray-Rust, ChemicalTagger: a tool for semantic text-mining in chemistry. Journal of cheminformatics 3(1):17, 2011. URL http://www.biomedcentral.com/1758-2946/3/17

[11] A. Hideno, H. Inoue, K. Tsukahara, S. Fujimoto, T. Minowa, S. Inoue, T. Endo and S. Sawayama, Wet disk milling pretreatment without sulfuric acid for enzymatic hydrolysis of rice straw. Bioresource Technology 100: 2706-2711, 2009.

[12] A. Hideno, H. Inoue, K. Tsukahara, S. Fujimoto, T. Minowa, S. Inoue, T. Endo and S. Sawayama, Combination of hot compressed water treatment and wet disk milling for high sugar recovery yield in enzymatic hydrolysis of rice straw. Bioresource Technology 104:743-748, 2012.

[17] D.M. Jessop, S.E. Adams and P. Murray-Rust, Mining chemical information from open patents. Journal of cheminformatics 3(1): 40, 2011. URL http://www.biomedcentral.com/1758-2946/3/40

[16] I. Kima, B. Leea, J-Y.Parkb, S-A. Choib and J-I.Han, Effect of nitric acid on pretreatment and fermentation for enhancing ethanol production of rice straw. Carbohydrate Polymers 99:563-567, 2014.

[13] C.A. Knoblock, P.A. Szekely, J.L. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan and P. Mallick, Semi-automatically Mapping Structured Sources into the Semantic Web. ESWC 2012: 375-390.

[14] P. Kumar, D.M. Barrett, M. J. Delwiche and P. Stroeve, Methods for Pretreatment of Lignocellulosic Biomass for Efficient Hydrolysis and Biofuel Production. *Industrial & Engineering Chemistry Research*, *48*, (8), 3713-3729, 2009.

[22] N. Maillet, B. Charnomordic and S. Destercke, R-belief: Contains basic functions to manipulate belief functions and associated mass assignments, 2010, rRpackage version 1.0. [Online]. Available: http://CRAN.R-project.org/package=belief

[20] N.F. Noy, A. Rector, P. Hayes and C. Welty, Defining N-ary relations on the semantic web. W3C working group note http://www.w3.org/TR/swbp-n-aryRelations, 2006.

[23] L. Olsson, Pretreatment of lignocellulosic materials for efficient bioethanol production. In *Biofuels*, Ed. Springer-Verlag Berlin: Berlin, 2007; (108): 41-65, 2007.

[21] N. Poornejad, K. Karimi and T. Behzad, Improvement of saccharification and ethanol production from rice straw by NMMO and [BMIM][OAc] pretreatments. Industrial Crops and Products 41: 408-413. 2012.

[24] T. Raafat, N. Trokanas, F. Cecelja, X. Bimi, An ontological approach towards enabling processing technologies participation in industrial symbiosis. Computers and Chemical Engineering (59): 33– 46, 2013.

[25] H. Rijgersberg, M. Wigham and J.L. Top, How semantics can improve engineering processes: A case of units of measure and quantities. Advanced Engineering Informatics 25(2): 276–287, 2011.

[26] M.I. Sheikh, C-H. Kim, H-J. Park, S-H. Kim, G-C. Kim, J-Y. Lee, S-W. Sima and J.W. Kimb: Effect of torrefaction for the pretreatment of rice straw for ethanol production. J Sci Food Agric 93, (13): 3198-204, 2013.

[27] N. Schultz-Jensen, Z. Kádár, A.B. Thomsen, H. Bindslev and F. Leipold, Plasma-assisted pretreatment of wheat straw for ethanol production. Appl Biochem Biotechnol 165, (3-4): 1010-23, 2011.

[28] A. Tian, J. Sequeda and D.P. Miranker, QODI: Query as Context in Automatic Data Integration. International Semantic Web Conference (1) 2013: 624-639.

[29] R. Touhami, P. Buche, J. Dibie-Barthélemy and L. Ibanescu, An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. In: OTM 2011 (2). Volume 7045 of LNCS., Springer, 662-679, 2011.

[30] N. Trokanas, F. Cecelja and T. Raafat, Semantic approach for pre-assessment of environmental indicators in Industrial Symbiosis. Journal of Cleaner Production (96): 349-361, 2015.

[31] J.Y. Zhu and X.J. Pan, Woody biomass pretreatment for cellulosic ethanol production: Technology and energy consumption evaluation. Bioresource Technology 101, (13): 4992-5002, 2010.