Cah. Techn. Inra, 2010 69, 51-56

Utilisation d'ontologies pour la validation de mesures appliquée à la fermentation alcoolique

Virginie Rossard[1], Pascal Neveu[2], Evelyne. Aguera [3], Abdel-Ilah Dkhissi [2], Éric Latrille [1], MarcPerez [4], Christian Picou[4], N. Rozas[2], Jean-Marie Sablayrolles [4], Rallou Thomopoulos[5]

Résumé : *L'application présentée concerne l'utilisation de méthodes et d'outils issus du Web sémantique pour la validation de mesures effectuées sur des bioprocédés instrumentés. Elle est opérationnelle sur les cuves de fermentation de l'unité mixte de recherche des Sciences pour l'œnologie (UMR SPO) du centre Inra de Montpellier et de l'unité expérimentale de l'Inra de Pech-Rouge.*
L'objectif est d'invalider des séries de mesures de façon automatique en fonction des connaissances relatives aux événements temporels qui se produisent au cours du déroulement d'un procédé ou d'un groupe de procédé. En effet, ces événements (opérations, interventions, pannes, etc.) ont un impact sur les mesures effectuées. Des interfaces permettent de recueillir des informations durant l'exploitation du bioprocédé. Ces informations sont représentées dans un formalisme XML constituant une base de connaissances. Des catégories d'événements ont été identifiées en fonction de leurs signatures sur les données. L'ensemble de ces connaissances est formalisé sous forme d'ontologie à l'aide des langages OWL et RDF. L'exploitation et la manipulation de ces ontologies se font avec des requêtes SPARQL et l'API Jena. Les traitements numériques associés ont été réalisés par des fonctions du logiciel statistique R. Cette application donne un taux d'invalidations de données qui avoisine les 90 %.
L'utilisation d'outils standards comme ceux du Web sémantique nous apparaît comme efficace et intéressante pour la validation de données. Elle peut facilement s'étendre à d'autres usages plus larges dans le domaine des sciences du vivant.

Mots clés : bioprocédé, ontologie, Web sémantique, logiciel R, gestion de connaissances, mesures temporelles, validation.

Introduction

Pour les utilisateurs tels que microbiologistes, œnologues, … une des grandes préoccupations est la confiance qu'ils peuvent accorder aux mesures surtout si elles sont exploitables pour des travaux ultérieurs d'analyse et de raisonnement. C'est dans ce but que nous avons développé cette application d'invalidation de données.
L'application présentée concerne la gestion des fermentations alcooliques dans un contexte de recherche. Les cuves de fermentations alcooliques sont instrumentées pour mesurer différentes variables en continu, durant toute la fermentation (au minimum 6 variables toutes

[1] UR050 LBE (Laboratoire de Biotechnologie de l'Environnement) – INRA - F-11100 Narbonne
☏ +33 (0)4 68 42 51 58 ✉ virginie.rossard@supagro.inra.fr
[2] UMR079 ASB (Analyses des systèmes et biométrie) – INRA – F-34060 Montpellier
[3] UE0999 de Pech-Rouge – INRA – F-11430 Gruissan
[4] UMR1083 SPO (Sciences pour l'œnologie) – INRA – F-34060 Montpellier
[5] UMR1208 IATE (Ingénierie des agropolymères et technologies émergentes) – INRA – F-34060 Montpellier

les 10 secondes). Les variables les plus importantes sont le débit de CO2 qui caractérise l'activité fermentaire et la température qui est contrôlée, essentielle aussi bien pour le déroulement de la fermentation que les caractéristiques organoleptiques du produit (Sablayrolles, 2008). Les moyennes des dernières valeurs des variables, mesurées en ligne, sont stockées dans une base de données avec une fréquence de 20 minutes. Cette fréquence a été déterminée pour appréhender suffisamment finement la dynamique d'une fermentation alcoolique. Les installations différent, évoluent et sont réparties géographiquement. Actuellement, les deux sites Inra de Pech Rouge (Aude) et de Montpellier (Hérault) disposent respectivement de 16 cuves de 100 litres et de 30 fermenteurs de 1 litre. Ainsi 46 fermentations peuvent se dérouler simultanément sur une durée comprise entre 5 et 30 jours. Ces fermentations sont intégrées en temps réel. L'unité mixte de recherche des Sciences pour l'œnologie (UMR SPO) du centre Inra de Montpellier dispose dans la base de plus de 1 700 fermentations archivées. Ce nombre va augmenter rapidement car le nombre de cuves instrumentées ainsi que le nombre de paramètres mesurés en ligne sont sans cesse en augmentation.

1 Description des données et technologies utilisées

1.1 Métadonnées

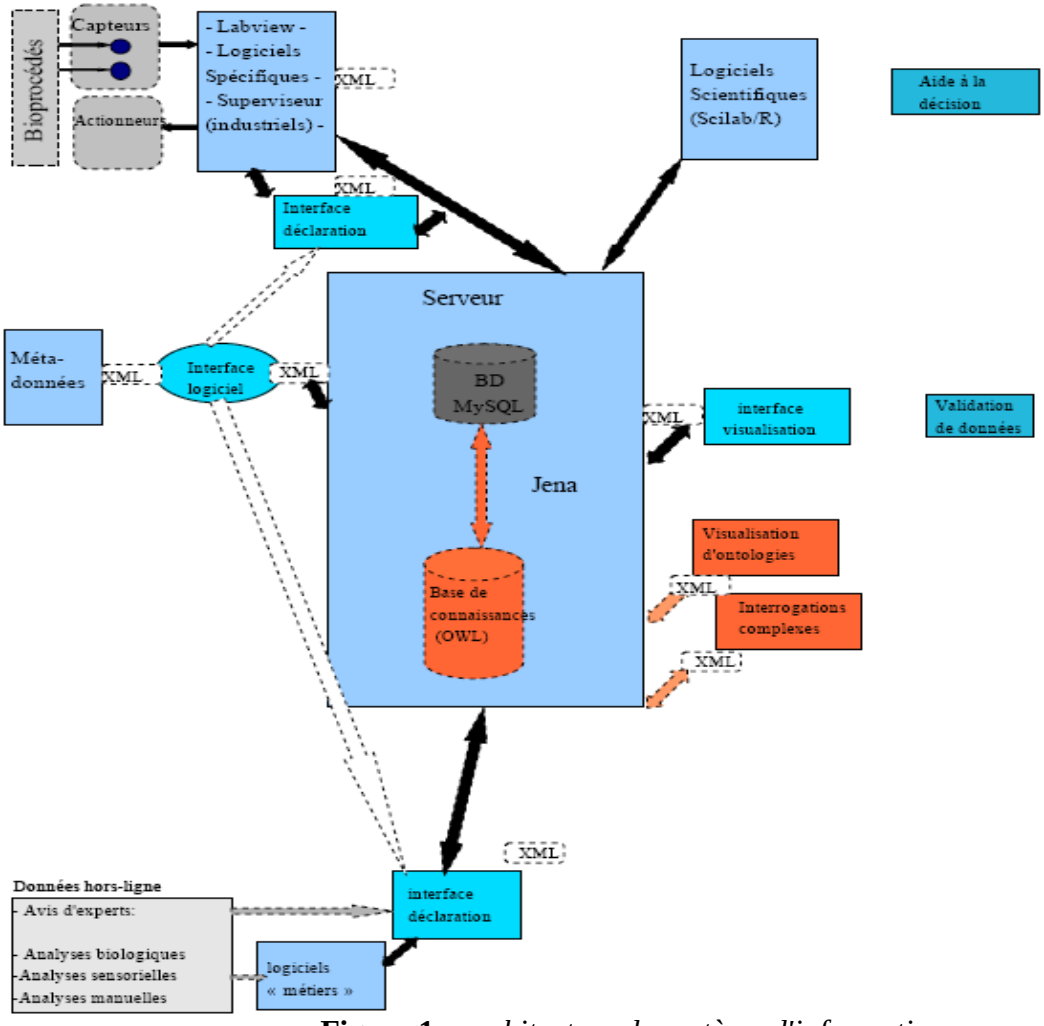

Figure 1 : *architecture du système d'information*

La **figure 1** montre l'architecture du système d'information en place, les sources d'information et leur hétérogénéité. Un moyen générique et souple est donc nécessaire pour gérer l'ensemble de ces données temporelles multi-sources. Dans un premier temps, les données sont décrites en **XML** (eXtensible Markup Language) afin d'annoter les données (acquisitions, calculs de confiance, règles métiers, etc.). C'est un moyen efficace pour collecter, gérer et distribuer des données (Neveu P. *et al.,* 2003). Un **schéma XML** (Van der Vlist E., 2002) a aussi été défini pour valider en ligne la structure des données. Pour cela, l'utilisation de métadonnées est essentielle pour comprendre et pour exploiter des données de façon pertinente. Ces données sont ensuite stockées dans un système de gestion de base de données : MySQL 4.1.
Pour cette application, nous nous intéressons aux mesures du débit de CO2 et de la température pour déterminer une confiance en exploitant des connaissances issues de commentaires.

1.2 Représentation des connaissances par les ontologies

D'après le W3C (World Wide Web Consortium), l'ontologie est une science décrivant les types d'entités (pannes, bioréacteurs, etc.) dans le monde et la façon dont elles sont reliées. C'est ainsi que des ontologies concernant les événements temporels (pannes et opérations) ont été mises en place pour cette application. Il s'agit d'abord d'une hiérarchie de spécialisation qui a été formalisée en OWL (Ontology Web Language). Cela concerne aussi la gestion des paramètres, associés aux différents cas, effectuée sous forme de triplet RDF, Ressource Description Framework, Ressource-Propriété-Valeur. OWL est un dialecte XML basé sur une syntaxe RDF (W3C, Rdf vocabulary language 1.0 2004). Le langage OWL permet une interprétation du contenu Web par les machines supérieures à celles offertes par les langages XML, RDF et le schéma RDFS, en fournissant un vocabulaire supplémentaire avec une sémantique formelle (W3C, OWL Web ontology language, 2004). Parmi les trois niveaux, nous avons choisi OWL-DL qui permet de raisonner (union, complément, intersection). WonderWeb OWL Ontology Validator est un "validateur" de structure d'OWL. Pour éditer ces ontologies, nous avons utilisé le logiciel Protégé©. Créé par l'université en informatique médicale de Stanford (*http://protege.stanford.edu/*), Protégé© est un logiciel libre qui a plusieurs avantages dont la flexibilité des interfaces, la compatibilité aux différentes plateformes, l'extensibilité avec des plugins. Cependant il n'intègre pas de validateur.
Protégé© permet au développeur, de créer des ontologies (méronymie, synonymie, taxonomie) et d'éditer le code OWL. L'interface WEB (**figure 3**) des commentaires permet de visualiser et de saisir à partir des feuilles de l'arbre le sujet souhaité. Nous utilisons conjointement les ontologies et le système de gestion de base de données « MySQL ».

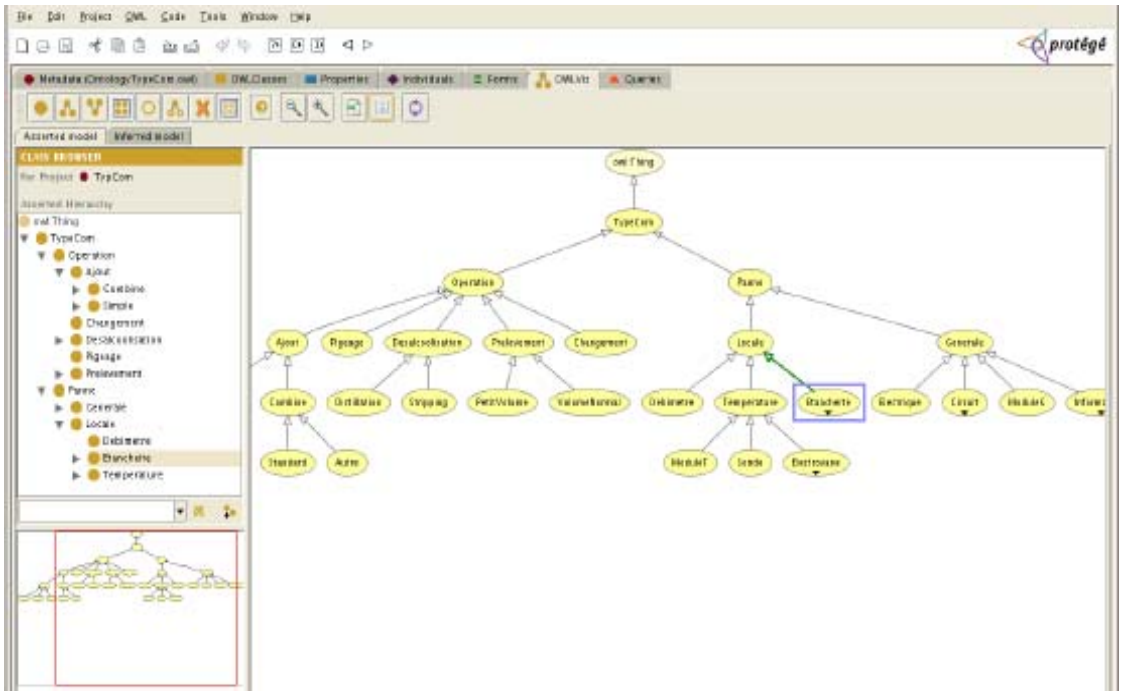

Figure 2 : Ontologie des pannes et des opérations

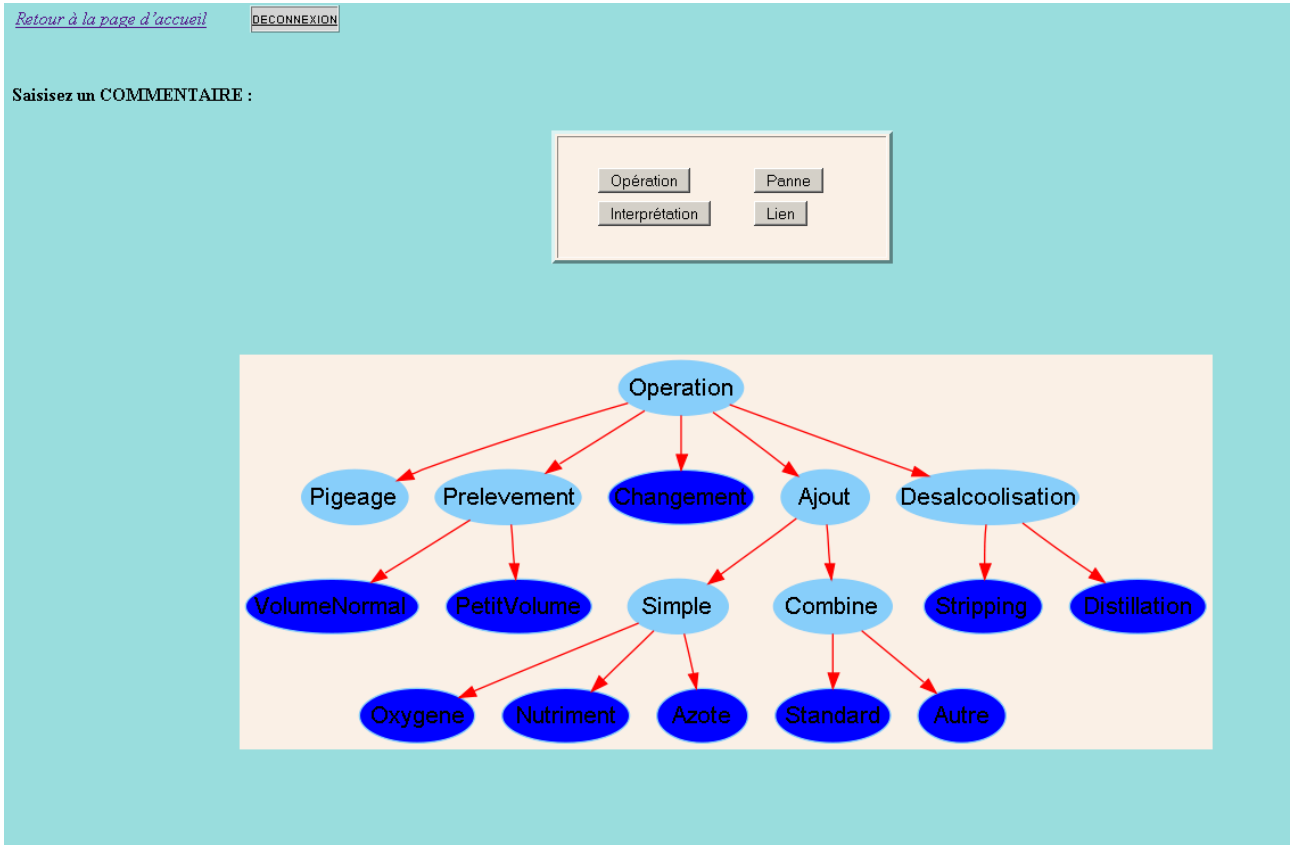

Figure 3 : *déclaration d'un événement temporel*

1.3 Le logiciel R

C'est un langage statistique et de graphiques scientifiques. Il est interactif et interprété. Il dispose de structures de programmation (branchements, boucles, ...). C'est une version libre de S-Plus. R est ouvert, portable et dispose de fonctionnalités puissantes : il est devenu un outil statistique très connu.
Il a été utilisé d'une part pour détecter les mesures atypiques à l'aide d'une fonction de lissage (ksmooth), des commentaires associés et de la note de confiance située dans la base de données : package RODBC ou RMySQL. D'une autre part il permet la visualisation des courbes, des données invalidées et les pannes ou les opérations survenues.

2. Fonctionnement de l'application

Toutes ces technologies combinées nous permettent d'invalider des données. L'application développée avec le logiciel R permet d'analyser des séries de mesures de façon quasi-automatique en fonction des connaissances relatives aux événements temporels (opérations, pannes, etc.). L'écriture d'une application pour invalider les données doit prendre en compte beaucoup de cas différents. L'utilisation d'ontologies nous a permis de gérer l'ensemble de ces cas et de séparer les données, les connaissances et les traitements. Cela donne une forme plus générique et évolutive à l'application.
La mise en œuvre se fait par l'intermédiaire d'une fonction R valideC qui affiche le graphique de chaque fermentation. Il repère les données invalides associées aux commentaires des experts (**figure 4**). Par exemple valideC("PR07-06") affiche le graphique de chaque fermentation effectuée en juin 2007 à Pech Rouge.

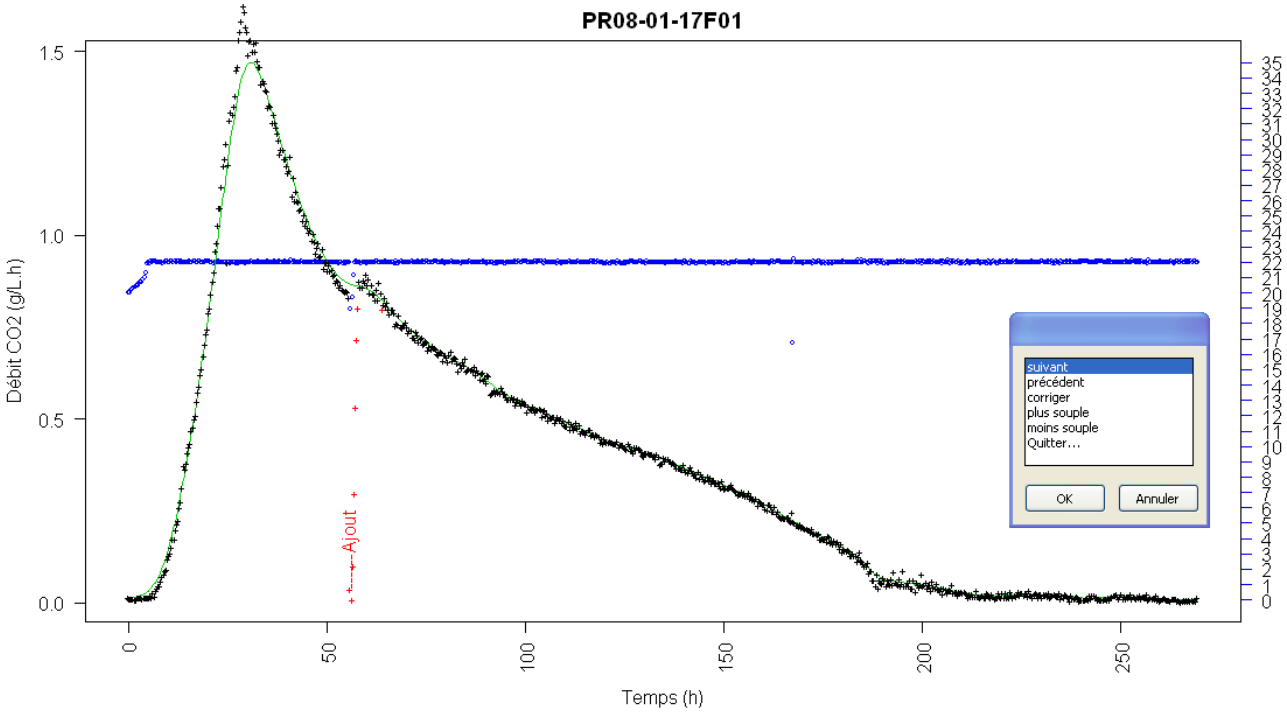

Figure 4 : déclaration d'un événement temporel

Les mesures, en dessous de la courbe après une opération ou une panne, sont invalidées (repérées en rouge). Les utilisateurs contrôlent l'application grâce à un menu qui permet de

faire défiler les fermentations, de corriger les mesures ainsi que d'assouplir ou de durcir manuellement le lissage de la courbe. La représentation graphique affiche :

- en noir les données validées dont l'indice de confiance est bon (1) ;
- en rouge les mesures invalidées dont l'indice de confiance est faible (0) ;
- en bleu les mesures de température ;
- des repères rouges qui indiquent une opération ou une panne d'après les connaissances figurant dans le Système d'Information ;
- une courbe verte qui est le lissage de la courbe.

On peut changer manuellement la validité des points en cliquant dessus et valider ces données en passant à la courbe suivante ou précédente. Si l'utilisateur clique sur un point invalide (rouge), il revalide la donnée et le point devient valide (noir). À l'inverse un clic sur un point valide (noir) la donnée devient invalide (rouge). L'ajustement de la courbe de fermentation (en vert) est tracé grâce à la fonction ksmooth dont on règle la fenêtre de sensibilité. Ici, la zone d'influence est de 8h et peut être modifiée grâce à un clic sur "plus souple" ou "moins souple" du menu.
Les résultats de cette application ont donné un taux d'invalidations de données de 90 %.

Conclusion et Perspectives

Pour répondre à une demande explicite des biologistes concernant la validation de données, nous avons créé d'une part une ontologie et d'autre part un programme R, pour permettre, automatiquement, l'invalidation d'un grand nombre de mesures. Ces mesures sont celles de la température ou du débit de gaz qui peuvent être perturbées par des opérations ainsi que par des pannes. Ce travail est basé sur la cohérence entre les mesures et les événements temporels. Avec l'ontologie nous avons construit une application générique et évolutive. et nos premiers résultats montrent que ces approches méthodologiques et technologiques sont prometteuses dans ce cadre applicatif. Le développement de telles applications est crucial par rapport aux coûts humains et financiers engendrés autour de ces expérimentations.
Nous avons la volonté également de mettre en œuvre cette approche sur des bioprocédés continus comme ceux utilisés en dépollution. La durée d'exploitation de ce type de procédés est de plusieurs années. Ces systèmes nous permettront d'éprouver nos réalisations face à une masse d'événements temporels beaucoup plus grande et un contexte d'exploitation différent.

Bibliographie

Neveu P., Lardon L., Hacquart C., Simon B. and Steyer J.-P. (2003) PlantML, un langage pour la gestion répartie de bioprocédés, 3ème colloque STIC et Environnement, Rouen, France, 19-20 juin 2003, pp. 197-200

Sablayrolles J.-M. (2008) Fermented beverages: the example of winemaking, Advances in fermentation technology, vol. Pandey, Larroche, Soccol and Dussap (ed), Asiatech Publishers, New Delhi, pp. 321-347

Van der Vlist E. (2002) « XML Schema ». O'Reilly,

W3C (2004) «Rdf vocabulary description language 1.0: Rdf schema », recommendation, http://www.w3c.org/TR/rdf-schema

W3C (2004) «OWL Web ontology language », recommendation, http://www.w3c.org/TR/owl-features