Software for Data and Knowledge Management in Winemaking Fermentations

Pascal Neveu[1], Virginie Rossard[2], Anne Tireau[1], Evelyne Aguera[3], Marc Perez[4], Christian Picou[4] and Jean-Marie Sablayrolles[4]

[1]*INRA/SupAgro, UMR 729 MISTEA, F-34060 Montpellier, France*

[2]*INRA/SupAgro, UR 50 LBE, F-11100 Narbonne, France*

[3]*INRA/SupAgro, UE 999 Pech-Rouge, F-11430 Gruissan, France*

[4]*INRA/SupAgro, UMR 1083 SPO, F-34060 Montpellier, France*

Keywords: Knowledge Management, Ontology, Fermentation, Winemaking.

Abstract: An increasing amount of data is generated by the on-line monitoring of biotechnological processes. Classical data management solutions, which have proved effective in many application domains, are not efficient at dealing with scientific data in life science. We describe a management software of data from wine fermentations and associated knowledge. The information dealt with in this framework relates to the knowledge of real-time events occurring during fermentations. The data have been entered into a database and we propose an organisation of this knowledge to improve efficiency, based on the use of methods and tools from the Semantic Web. A specific ontology of events (faults or enological operations) is used to automatically identify wrong on-line measurements which clearly improved data quality and understanding.

1 INTRODUCTION

Alcoholic fermentation, a key stage in winemaking, is likely to change considerably in the future. Indeed winemakers are increasingly of the opinion that tools for controlling fermentation according to the type of wine desired is becoming essential and the use of sensors, for real time monitoring, is becoming increasingly feasible.

Instrumentation can be used to automate the monitoring and allows the development of Information Systems suitable for the storage of data, kinetics and all types of information and facilitating the use of this knowledge to develop more ambitious treatments. Progressive semantic annotation of data (Hignette et al., 2007) and source descriptions are of particular interest for food processes (Sall et al., 2009). Such tools should prove very powerful in the long term, not only for reconstructing a history of events (traceability) but also providing extensive information to improve management practices (decision support).

This paper deals with a first step in this direction, focusing in particular on the validation of measurements, taking into account knowledge about faults and operations of particular interest in winemaking. It is original in that (i) it makes use of a database of results from the on-line monitoring around 4500 fermentations and (ii) it implements techniques (based on ontologies and reasoning) novel for this type of application. This study was carried out in the context of a European project, CAFE (Computer-aided food processes for control engineering). All of the developments are implemented in an Information System called *Alfis*[1].

2 MATERIALS AND METHODS

2.1 Fermentations

Very few industrial-sized tanks are currently equipped with an on-line fermentation monitoring system and there is therefore no corresponding database. However, we have many data acquired from pilot studies, or laboratory-scale studies. Indeed, fermentations are monitored by automatic measurements of CO_2 release and the temperature of the fermenters is controlled. In pilot scale, CO_2 output was measured with a mass flowmeter and the temperature of the tanks is controlled by opening or closing solenoid valves in either

[1]http://alfis.supagro.inra.fr/

Neveu P., Rossard V., Tireau A., Aguera E., Perez M., Picou C. and Sablayrolles J..
Software for Data and Knowledge Management in Winemaking Fermentations.
DOI: 10.5220/0004138803690372
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 369-372
ISBN: 978-989-8565-30-3

a hot water circuit or a circuit connected to a refrigeration unit.

CO_2 production and temperature were monitored on-line, with control and supervision software written in Labview language. This software can be used to add symbolic data and metadata describing enological operations such as the addition of nutrients or cap punching, observations and faults. Factors other than CO_2 and temperature were monitored manually, off-line and the data obtained were also input into the database. Complex data were recorded such as the principal by-products of fermentation, as determined by chromatographies (liquid or gas), histograms of the distribution of yeast populations into size classes.

2.2 The Information System

Each site is characterised in terms of its methods of measurement, equipment, organisation, etc. The Information System combines the acquisition of (i) online data from sensors controlled by supervision software, (ii) off-line data from biological analyses of fermentation by-products (see 2.1) and (iii) symbolic data (expert opinions, operations, faults, etc.). These data and metadata are formalised in XML and organised into a generic form. It provide a flexible, generic method for managing heterogeneous, multisource data.

Data from the various sites are available from the Web via a HTTP server (Apache) and a DBMS (MySQL). The database contains the measurements and various data. However, knowledge, such as professional know-how, cannot easily be used with a relational DBMS. Indeed we needed (i) searches requiring reasoning such as subsumption (ii) providing active hypertext links and semantic (e.g. to identify the specific fermentations used in an article and vice versa), and (iii) dynamic management of the data and knowledge relating to professional know-how (consistency sequence of events).

We used ontologies to formalise, make use of professional know how and for the "intelligent" management of all the faults, operations and decisions coming into play during fermentation. These ontologies are written in Ontology Web Language (OWL) and semantic annotations with the Resource Description Framework (RDF) syntax. These ontologies were designed with the "Ontology Development 101" (Noy and McGuinness, 2001) method. In our software applications, the ontologies and the RDF annotations are stored in the database then exploited by the semantic engine CORESE (Corby et al., 2004) which enables to infer and process SPARQL queries on RDF annotations and OWL ontologies. It also provides the property paths functionality. The Jena API is used for the RDF annotation management.

Data and Treatment. The incorporation of data and knowledge into databases makes it possible to standardise access to this information. We have developed applications in PHP, JAVA and R languages, through the R-ODBC package. Web applications generate a lot of scientific graphs (Highcharts library or by R) that often need be enriched. Statistical treatments, written in R, can access ad hoc data via a simple SQL request and are controlled by JAVA modules using the ontologies. The use of Semantic Web tools and methods makes it possible to separate knowledge from the control mechanisms of the computer languages routinely used. Classical approaches could be used to write an application, but it would be difficult to control the complexity of the effects of successive changes.

3 FERMENTATION KINETICS

In the database, fermentation kinetics is expressed as the rate of CO_2 production (dCO_2/dt). Several factors have a significant effect on fermentation kinetics and hence on the quality of wines. This knowledge had been formalized in OWL Lite and this makes possible data cleaning and analysis (curve clustering, data mining, etc) on a large number of kinetics.

Effect of Enological Operations. Various operations may significantly modify the shape of the curves of CO_2 production rate. Nitrogen supplements diammonium phosphate (DAP) in most cases are routinely added to cause a rapid and sustained increase in dCO_2/dt. These additions are usually combined with oxygenation, which entails opening of the tank for a few minutes. During red wine making, the solid matter rises to the surface and forms a cap. This cap is immersed several times during the course of fermentation. It increases dCO_2/dt, following the resuspension of yeasts trapped in the solid matter. Other less common processes are carried out during the production of particular types of wine. One such process included in this database is the partial reduction of alcohol content (Aguera et al., 2010) by vacuum distillation or CO_2 stripping.

Anomalous Points and Malfunctions. The graphs of the CO_2 production rate may include anomalous points that do not correspond to the true rate of

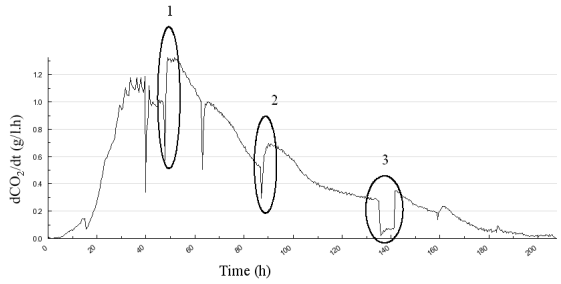

Figure 1: Effect of enological operations or faults. 1: combined addition of oxygen and di-ammonium phosphate, 2: cap punching, 3: breakdown.

CO_2 production (Figure 1). Such points may result from the momentary opening of the tank during an enological operation, decreasing dCO_2/dt or completely stopping CO_2 release. Other operations, such as sampling, may also disrupt dCO_2/dt measurements. Anomalous measurements may also be obtained for temperature, generally coinciding with anomalous dCO_2/dt readings.

The occasional anomalous results must be distinguished from disruptions caused by faults and failures, which may have several causes sensor or temperature control system for example and which may affect only one tank or the whole installation. These faults or failures may significantly affect the monitoring of fermentation (e.g. flowmeter fault), its progress (e.g. bad temperature control) or both. Detection is essential. Certain malfunctions can be detected only with numerical data, whereas the detection of others requires symbolic data or metadata.

4 RESULTS AND DISCUSSION

Database and Ontologies. A database was created and was improved by the design of two highly targeted ontologies. Then, applications were developed or defined for data validation and construction of complex queries over winemaking fermentations. The database provides access to measurements taken on-line, uni- and multidimensional measurements taken off-line, symbolic data and metadata. These data are organised according to ontologies of operations and faults (Figure 2). Since its initial establishment in 2004, a complete inventory of this database has been carried out. It contains about 4500 fermentations, 52 of which were found to have been stored incorrectly. Minor manual corrections were made to render the data for these fermentations accessible. No information was lost.

The number, type and nature of events occurring during the course of a fermentation generate a combined analysis that is difficult for applications to manage. The ontologies concern the rationale underlying the hierarchy used for complex queries and the validation of on-line measurements. Here ontologies supply real benefists. Indeed, the integration of new knowledge or knowledge management is easier than in classical applications and software modules are less sensitive to technical evolutions and modifications of experiments. In this case, ontologies provide an efficient access and understanding of data for different end-users and to formalize technical itinerary. For example, by using subsumption, it is possible to ask for all general faults (Figure 2) without the user knowing all the different types. This set-up makes it possible to reply to queries such as "which fermentations have one alcohol content reduction process and no control system faults?", without knowing all the types of control system fault or the methods of alcohol content reduction. Using transitive property such as "mustBePreceed" has improved consistency and completeness of the information about enological operations (Figure 3). The application thus has the know-how required to deal with complex queries of this type. The perfection of this knowledge was an important outcome of our approach.

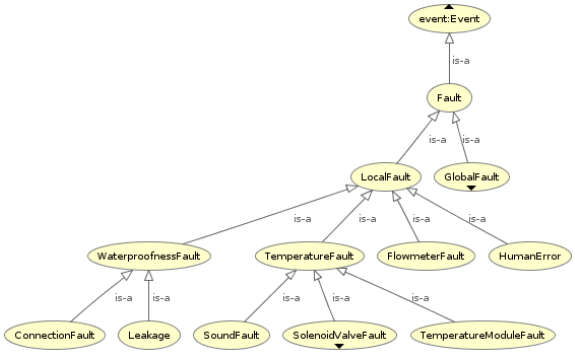

Figure 2: Extract of the ontology of local faults.

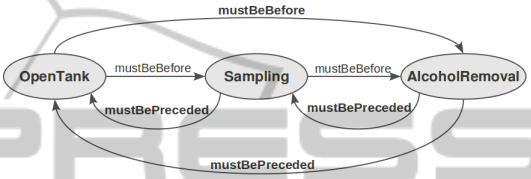

Figure 3: Example of semantic inference to infer new relations (in bold) between concepts from primer acquisitions.

Software Application for Measurement Validation. One of the major concerns of users is that only data providing an accurate picture of the state of the system should be stored. In our case, measurements for temperature and gas release may be disrupted by enological processes as well as faults. Moreover, the flow

rate is affected differently by opening the tank or a flowmeter fault. Using ontologies and the R language, we have been able to go beyond the limits of the systems generally used.

The component developed uses an ontology for symbolic data and the R language to fit the curves on measurements using a non parametric method (spline estimator) under constraints. The curve shape depend on the event sequence. To ensure a good fit, the number of nodes and the constraints are generated by using the event ontology and a new dedicated one ontology for impacts. This ontology of application was designed to classify impacts of events on kinetics and to be able to automatically generate a specific R treatment. So, this component uses the fitted curve, the type of event (operation or fault) and comments from experts (Neveu et al., 2008) to validate or invalidate the measurements. To quantify the efficiency of the application, we analysed 20 fermentations with many operations and malfunctions. A detailed analysis of the dCO_2/dt curves by experts pointed out 242 wrong or litigious measurements (among more than 20 000). Using the application, we found very similar results, with only 21 misdiagnoses, i.e. less than 9% difference. This result indicates a very good rate of data invalidation and should prove a good behavior of our ontologies.

Perspectives: Determination of the Usefulness of Fermentations. Another concern of users is the confidence they can have in the fermentation and whether the data can be used in other contexts for analytical and diagnostic purposes. Certain faults (default of temperature regulation, for example) happen at key moments and are therefore disastrous, preventing the subsequent use of the fermentation data. A description of this knowledge as a function of the type and time of the fault is essential, to enable the information system to determine whether a data set is valid. This requires taking the investigation further and seeing whether the whole data set can be interpreted.

5 CONCLUSIONS

Advances in alcoholic fermentation monitoring and control necessitate new tools for data management. Databases need to be associated with tools for the representation and management of knowledge. This study shows an example of a reliable and user-friendly application based on ontologies, with the aid of Semantic Web tools and technologies (JENA, CORESE, PROTEGE, etc.). These languages and softwares, such as RDF(S) and OWL provide reasoning and they are of great interest for the management of complex data.

The applications developed automatically detect the inconsistencies caused by the principal malfunctions during the fermentations and distinguishes between those of little importance and those that might disrupt the process. Disruptions related to enological operations (such as cap punching, pumping over) are also detected and analysed. The application was validated using a database containing about 4500 fermentations. It might represent a powerful tool to improve fermentation management and control in the near future.

The method of formalization that separates the expert knowledge allows to take into account easily changing business practices. It allows to easily adapted the developed software for other cases in food process (cheese making, brewinb, etc).

REFERENCES

Aguera, E., Bes, M., Roy, A., Camarasa, C., and Sablayrolles, J.-M. (2010). Partial removal of ethanol during fermentation to obtain reduced-alcohol wines. *American Journal of Enology and Viticulture*, 61 (1):53–60.

Corby, O., Dieng-Kuntz, R., and Faron-Zucker, C. (2004). Querying the semantic web with corese search engine. *Proceedings of European Conference on Artificial Intelligence (ECAI) Valencia, Spain*, pages 705–709.

Hignette, G., Buche, P., Dibie-Barthélemy, J., and Haemmerlé, O. (2007). Semantic annotation of data tables using a domain ontology. In *Discovery Science*, pages 253–258. Springer.

Neveu, P., Rossard, V., Aguera, E., Perez, M., Picou, C., and Sablayrolles, J.-M. (2008). Gestion de données et de connaissances pour les bioprocédés. *Extraction et Gestion des Connaissances*, Atelier : fouille de données temporelles:21–28.

Noy, N. and McGuinness, D. (2001). Ontology development 101. *Knowledge Systems Laboratory, Stanford University*.

Sall, O., Lo, M., Gandon, F., Niang, C., and Diop, I. (2009). Using xml data integration and ontology reuse to share agricultural data. *International Journal of Metadata, Semantics and Ontologies*, 4:93–105.