Flexible querying of Web data for predictive modelling of risk in food

P. Buche[1], O. Couvert[3], J. Dibie-Barthélemy[1, 2], D. Doussot[1, 2], E. Mettler[4], L. Soler[1]

[1]INRA MIA, Unité Mét@risk UR 1204
16, rue Claude Bernard - 75 231 PARIS Cédex 05 - FRANCE

[2]AGROPARISTECH, UFR Informatique
16, rue Claude Bernard - 75 231 PARIS Cédex 05 - FRANCE

[3]ADRIA Développement
Creac'h Gwen – 29196 QUIMPER Cedex

[4]Soredab (Groupe SOPARIND BONGRAIN)
La Tremblaye
78125 La Boissière-Ecole - FRANCE

Email of corresponding author: patrice.buche@paris.inra.fr

JUSTIFICATION: One of the main stakes of many predictive modelling in foods and in particular in risk in food assessment is to gather experimental data and to maintain and update these data. In the framework of the Sym'Previus platform (http://www.symprevius.org), we have designed a complete data integration system opened on the Web which allows a local database to be complemented by data extracted from the Web. The local data was classified by means of a predefined vocabulary organized in taxonomy, called ontology, which is used to extract pertinent data from the Web.

OBJECTIVE : Our aim is to integrate the data found on the Web with the local data by means of a flexible querying system which allows the end-user to retrieve the nearest local and Web data corresponding to his/her selection criteria. Our solution allows the end-user to query simultaneously and uniformly local and Web data in order to feed the predictive modelling tools available on the Sym'Previus platform.

METHODOLOGY : The original flexible querying system presented in this paper is able to query Web data previously annotated thanks to the ontology of the domain. We first remind the semi-automatic annotation method (implemented in the @WEB tool) which retrieves data from data tables found in scientific documents extracted from the Web and annotates them. Second, we present the original contribution of the paper, which consists in the design of the flexible querying system, called MIEL++, which permits to query simultaneously the local data and the semantic annotated Web data, in a transparent way to the end-user, thanks to the ontology. This system is flexible because (i) it allows the end-user to express preferences in his/her selection criteria and (ii) it takes into account, in the answers building, the different kinds of fuzziness of the semantic annotated Web data. This second point is essential to deal with the uncertainty of the Web data and with the imperfection of their annotations.

RESULTS : An experimentation of the @WEB and the MIEL++ tools are currently conducted on two distinct corpora: food predictive microbiology and chemical risk in foods. Over the 123 (resp. 34) relations in the manual annotation of predictive microbiology (resp. chemical risk) corpus, 119 (resp. 27) were correctly recognized in our automatic process, 4 (resp. 7) were not recognized and there were 30 (resp. 2) relations that were kept for the tables while they should not have been recognized.

IMPLICATION : Probabilistic simulations of Sym'Previus software needs a lot of data in food products to take the food matrix into account, and to assess food variability in bacterial growth simulations. A prototype of the @WEB and the MIEL++ tools will be soon integrated

with the predictive modelling tools of the Sym'Previus project. These automatic links between web data and simulation tools allows a new step in risk assessment to be performed.